

Are bioassessments based on the reference condition approach affected by rapid approaches to sample collection and processing?

Amanda VALOIS,^{1*} Keith SOMERS,² Chantal SARRAZIN-DELAY,³ Wendel (Bill) KELLER³

¹Department of Zoology, University of Otago, PO Box 56, Dunedin 9054, New Zealand; ²Dorset Environmental Science Centre, Ontario Ministry of the Environment and Climate Change, Dorset, Ontario, Canada, P0A 1E0; ³Living with Lakes Center, Laurentian University, Sudbury, Ontario, Canada, P3E 2C6

*Corresponding author: amanda.valois@otago.ac.nz

ABSTRACT

Benthic invertebrates are used by a number of agencies worldwide as indicators for assessing stream health, which has resulted in the development of a variety of protocols for collecting and processing benthic samples. The large number of methods used means that calibration of data collection is not always possible, but if different methods produce similar estimates of community composition and metric values, then sharing of data can make bioassessments more efficient. This study explored the effect of two approaches to subsampling and sorting of benthic invertebrates on community composition, calculation of metrics, and assessment of stream health. We compared two commonly used sampling methods: a rapid approach, employing live, unaided sorting and a standard approach using microscope sorting of preserved samples, through a comparison of replicate samples collected from 61 streams. This study found that both methods resulted in similar estimates of community composition at a site, as determined by the Bray-Curtis similarity index. However, the live sorting methodology resulted in greater family richness and higher estimates of metrics that reflect large taxa (i.e., %EPT). Despite differences in a number of metrics, both methods performed equally well at identifying impairment in the test sites, with live-sorting samples slightly more sensitive.

Key words: Benthic invertebrates; rapid bioassessment; stream health.

Received: November 2015. Accepted: March 2016.

INTRODUCTION

Assessing stream health by monitoring benthic invertebrate communities is an integral part of water quality management worldwide. Benthic biomonitoring is included in a number of national and international programmes, e.g., Canada's Environmental Effects Monitoring (EEM) program, New Zealand's National River Water Quality Network (NRWQN), and the European Union's (EU) Water Framework Directive (WFD). Benthic species are favoured in bioassessment as they integrate local conditions throughout the aquatic phase of their life-span, complementing water chemistry grab samples and toxicity testing (Karr, 1993). However, monitoring these communities often requires a large field sampling effort, which, coupled with the time and cost associated with sorting and identifying samples, can result in expensive studies and large time lags in data return (Barbour *et al.*, 1999). Concerns associated with expense and time requirements have resulted in the development of a number of rapid bioassessment protocols (RBPs) in an effort to reduce costs and provide quicker turn-around of data (Wright *et al.*, 1984; Lenat, 1988; Barbour *et al.*, 1999, Dickens and Graham, 2002; Flotemersch *et al.*, 2006).

RBP's can be simplified at a number of different

stages. The processing of benthic invertebrate samples is time consuming largely because of the considerable effort required to separate all of the organisms from the debris that is collected with the samples (Rosillon, 1987; Ciborowski, 1991) and so, changes to subsampling strategies are the most common area for simplification (Barbour *et al.*, 1999). In the United States, most state agencies that collect macroinvertebrates for stream biomonitoring do not process samples in their entirety but subdivide them into a more manageable size, either processing a fixed proportion of the sample or removing a fixed number of organisms (Carter and Resh, 2001). Unbiased subsampling can be highly technical, involving specially designed equipment such as a Marchant box (Marchant, 1989) or involve relatively quick and inexpensive methods such as grid trays (Sovell and Vondracek, 1999) or sieves (Cuffey *et al.*, 1993). Subsampling can be focused on a fixed count or time strategy, counting taxa found in subsamples until a given taxa count (Lenat, 1988; Vinson and Hawkins, 1996) or time is reached (Chessman and Robinson, 1987). Subsampling in the field is often accompanied by removing organisms from the debris live, with no visual aids, while lab subsampling is followed by sorting of preserved samples under the microscope.

Although there are many subsampling and processing options available when selecting a bioassessment protocol, the underlying concept is that the resulting data must reflect the community composition in the whole sample to provide a meaningful assessment of impairment (Growth *et al.*, 1997). Additional considerations include the comparability of the data collected to other protocols. This becomes important when datasets need to be combined, for example, when evaluating temporal trends using historical data collected by different methods or combining data for regional or large scale analyses (Rehn *et al.*, 2007). For example, Wilson *et al.*, (2015) showed how standardizing biomonitoring data collected by different agencies could allow for increased spatial and temporal resolution in the bioassessment of large rivers in the Northeastern USA. Different data sets are also combined when local agencies cooperate to contribute to national or international water quality objectives. For example, the EU's WFD initiated a number of studies to determine the comparability of different benthic sampling protocols in contributing to their aims of classifying Europe's freshwater resources (Birk and Hering, 2006). In New Zealand, the Ministry for the Environment established a *New Zealand Macroinvertebrate Working Group* to develop standard protocols based on procedures currently in use by Regional Councils so as to maximise the value of existing data sets (Stark *et al.*, 2001).

Here, we examine two common RBP's used in Canada, a *standard* approach (Reynoldson *et al.*, 2003), adopted by the federal CABIN (Canadian Aquatic Biomonitoring Program) and EEM (Environmental Effects Monitoring) programs and a *live-sort* approach (David *et al.*, 1998) used by watershed and provincial organizations. These two assessment approaches both provide a consistent method for assessing biological health and allow for data sharing among users. Both methods use a kick and sweep collection in the riffle run habitat and employ fixed count subsampling of 300 organisms, however, the standard method standardizes sampling effort to time rather than area (live-sort). The main differences between the two methods are in the subsampling methodology (Marchant box *vs* teaspoon method), and the sample sorting methodology (microscopic sorting of preserved samples versus visual sorting of live samples). The use of differing methods creates the potential for inconsistent results when data sets are combined and may affect the assessment outcomes (*e.g.*, determination of impairment; Diamond *et al.*, 1996).

The purpose of this comparison is therefore to evaluate whether differences in sample processing result in differences in community composition among replicate samples and if those differences are reflected in biological metrics and the assessment of biological impairment. By combining a comparison of differences in benthic invertebrate metrics with a direct assessment of biological im-

pairment, we can determine whether data sets generated using the two methods can be combined and whether assessment results can be used interchangeably.

METHODS

Study area and sampling protocols

Benthic invertebrates were collected from 61 streams in Northern Ontario, Canada - latitude 46°-51°W, longitude 79°-94° N (Fig. 1). The streams were selected to represent least impacted reference sites ($n=28$) and a variety of impacted (*i.e.*, test) sites ($n=33$) in the region. Over a century of industrial activities in the Sudbury area, in the eastern portion of our study region has led to widespread acidification and metal contamination, which is reflected in slightly elevated concentrations of nickel, zinc, aluminium and sulfate in some streams (Tab. 1). The same sites in a given stream were sampled on the same date in the fall (September/October) of 2006 using both the standard and live-sort protocols.

Invertebrates were collected from each stream by both methods using a kick-and-sweep with a standard D-frame net with a 500 μm mesh. The standard method used a 3-minute, bank-to-bank zigzag traveling kick-and-sweep primarily within riffle habitat (Tab. 2). Subsampling was done using a Marchant box whereby the contents of each randomly selected cell were processed until a minimum of 300 organisms was enumerated (Marchant, 1989). Organisms were extracted using a microscope (63X power), providing an estimate of abundance where the number of organisms extracted is extrapolated to the entire sample. The live-sort protocol involved sampling a defined area (three 1-m² quadrats) using a kick-and-sweep procedure rather than a timed collection (Tab. 2) (David *et al.*, 1998). The live-sort sample processing method involved subsampling from a bucket with a spoon and unaided sorting of live organisms in a white tray until 100 organisms were collected from each quadrat. The data were combined to create a 300 organism count for each stream. In contrast, with the standard method, no attempt was made to determine the abundance of taxa relative to the area sampled as quantitative subsampling was not used.

All benthic invertebrates were identified to family level, including Oligochaeta and Hydracarina, which are typically only identified to class or order. The following ten metrics were chosen to summarize the benthos count data: Family richness, Ephemeroptera, Plecoptera, Trichoptera (EPT) richness, %EPT, %Chironomidae, %Oligochaeta, %Mollusca, Simpson's diversity, Simpson's evenness, and the first two axes from a correspondence analysis (CA) ordination (rare removed - taxa occurring in 5% or fewer streams (≤ 3 streams) and with no more than 10 individuals per 300 count sample) of the family abundance data (*i.e.*, CA axis 1 and CA axis 2).

These metrics have been shown to be sensitive in differentiating impacted sites from reference sites under a variety of anthropogenic stressors (Bowman *et al.*, 2006; Carlson *et al.*, 2013; Narangarvuu *et al.*, 2014). The number of metrics was limited to ten to ensure sufficient degrees of freedom for further statistical analyses.

Habitat descriptors were measured at each site using protocols outlined in Reynoldson *et al.*, (2003) and David *et al.* (1998) and the resulting variables retained for analysis are summarized in Tab. 1. Chemical and physical water quality measurements (*e.g.*, pH, conductivity, alkalinity, nutrients, metals, ions, temperature, and dissolved oxygen) were collected and measured using standard protocols (Reynoldson *et al.*, 2003).

Community composition and metric similarity

We used several univariate and multivariate methods to compare the composition of the benthic invertebrate

community collected by the two sampling methods. The Bray-Curtis (BC) Similarity Index was used to summarize differences in community composition between two methods for each stream. The BC Index is a commonly used abundance-based index for comparing community similarity in benthic invertebrate studies, providing an easily interpretable distance metric (Hawkins and Norris, 2000; Lorenz and Clarke, 2006). Biological assessments often use summary metrics rather than community composition, and samples can give similar values for metrics even though they contain different taxa (Lorenz and Clarke, 2006). Therefore, we also examined differences among the metrics calculated from the two methods. Paired *t*-tests were used to assess whether the metrics agreed on average by testing the null hypothesis that the mean difference for each metric between the methods is zero. A rejection of the null hypothesis indicates that there is consistent tendency for a metric to be higher in one method than the other (called the bias). The variation about this mean is

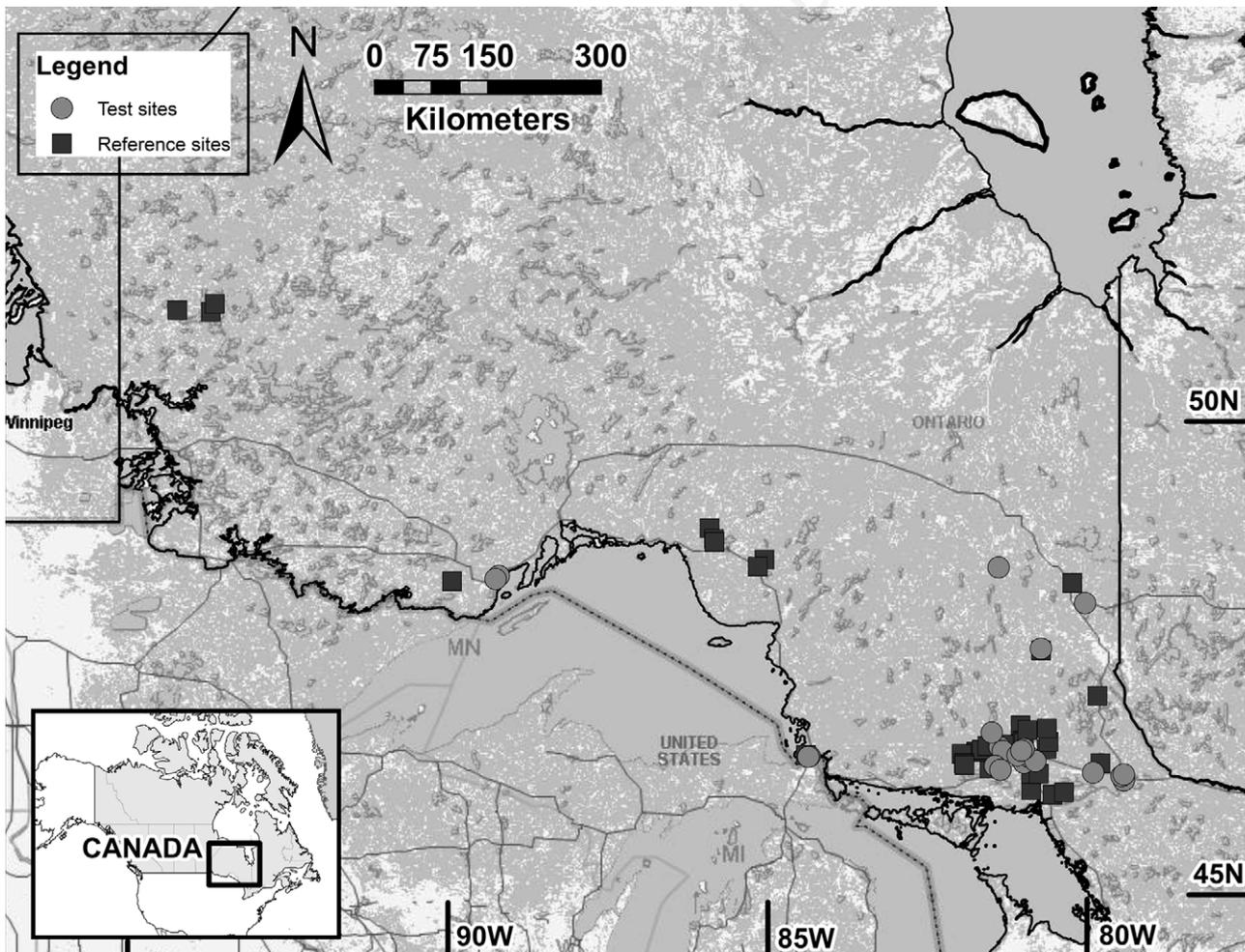


Fig. 1. Locations of sampling sites (n=61) in Northern Ontario, Canada.

estimated by the standard deviation of the differences.

To evaluate the degree of agreement in metrics between the two methods, simple linear regressions were performed to give a correlation coefficient, slope, and intercept for each metric. The Pearson correlation coefficient (r) was used as a measure of association. Perfect correlation ($r=1$) means that the values calculated by the two methods increase directly in proportion to one another, but does not imply they are identical. Systematic differences can be present, which are reflected in the slope and y-intercept of the regression equation. Two-tailed Student's t -test's with significance $\alpha=0.05$ were used to check if the slope differed from 1 (proportional systemic error) and if the intercept deviated from 0 (constant systematic error).

Metrics were \log_{10} square root or square root of the arcsine transformed where transformation helped meet assumptions of normality (%Chironomidae, %Oligochaeta only).

Assessment outcomes

We used a reference-condition approach to assess the ability of the two methods to distinguish impaired sites from reference condition. Appropriate reference sites for each stream were identified following Parsons *et al.* (2010). A Redundancy Analysis (RDA) with stepwise regression was performed on a correlation matrix of centered and standardized variables to identify those water chemistry variables (Tab. 1) that explained a significant amount of the variance in the abundance of the benthic

Tab. 1. Water chemistry and habitat characteristics of the study sites, with means (median for substrate and other variables estimated by % cover) and ranges given for the reference (n=28) and test (n=33) sites.

	Reference streams	Test streams
Water chemistry		
Aluminum (mg L ⁻¹)	71.07 (4-244)	126.1 (8.0-404.0)
DOC (mg L ⁻¹)	9.95 (3.08-27.74)	8.58 (2.24-29.01)
Calcium (mg L ⁻¹)	8.08 (1.64-20.61)	15.77 (0.70-119.28)
Iron (mg L ⁻¹)	0.27 (0.01-0.76)	0.34 (0.03-0.95)
Conductivity (uS cm ⁻¹)	66.5 (16.4-244.0)	246.3 (6.38-1184)
pH	7.14 (6.40-8.62)	6.97 (5.22-7.92)
Nickel (mg L ⁻¹)	2.07 (0.86-9.66)	121.6 (0.83-1581)
TP (mg L ⁻¹)	76.4 (6.6-217.0)	92.3 (24.7-229.6)
SO ₄ (mg L ⁻¹)	5.98 (1.17-10.47)	42.43 (2.43- 690.5)
TSS (mg L ⁻¹)	1.31 (0.2-4.6)	6.71 (0.2-52.8)
Zn (mg L ⁻¹)	2.22 (0.05-5.87)	20.7 (0.5-256.5)
Habitat variables		
Average depth (cm)	35.1 (8.7-83.8)	36.3 (12.0-84.7)
Average velocity (m ² s ⁻¹)	0.19 (0.02-0.34)	0.18 (0.02-0.47)
Wetted width (m)	10.48 (2.0-60.0)	7.02 (1.07-17.2)
Canopy cover (%)	5 (0-90)	0 (0-100)
Substrate-silt (%)	0 (0-90)	0 (0-70)
Substrate-sand (%)	10 (0-70)	20 (0-100)
Substrate-cobble (%)	30 (0-95)	10 (0-90)
Substrate-boulder (%)	0 (0-90)	0 (0-50)
Macrophytes (%)	0 (0-30)	5 (0-80)
Detritus (%)	0 (0-100)	0 (0-100)
Algae (%)	0 (0-90)	0 (0-95)
Organic matter (%)	0 (0-50)	0 (0-100)
Woody debris (%)	0 (0-10)	0 (0-10)

DOC, dissolved oxygen concentration; TP, total phosphorus; TSS, total suspended solids.

Tab. 2. Comparison of sampling and sorting methods between the rapid and standard approach.

Method	Rapid	Live-sort
Sampling area	1 m ² quadrat	Travelling: 3 minutes
Subsampling	Bucket and spoon method	Marchant box
Sorting	Unaided	Microscope
Number of samples	3	1
Number of organisms per samples	100+	300+
Density calculations	100 X 3	300 X % subsampled

invertebrate families in the reference-site communities. Rare families representing <5% of the total abundance were not used in the analysis. Separate RDAs were performed using i) rapid method; and ii) standard method benthic data to determine if the significant explanatory variables differed among the two sets of data. Patterns among the significant water chemistry variables identified in the RDA and 13 habitat descriptors (Tab. 1) were summarized using a Principal Component Analysis (PCA) of a correlation matrix PCA axes 1 through 4 were retained after evaluating the scree plot. The site scores for the 4 principle components were used to calculate Euclidean Distances between each test stream and the reference streams. This calculation allowed us to identify the nearest neighboring (NN) reference sites for each test stream to ensure that we compared test streams with chemically and physically similar reference streams. Although 30-50 reference sites are recommended (Bowman and Somers, 2006), we only had 28 reference sites available. As a result, we used the 15 most similar reference sites for the assessment of each test site.

Each test site was compared to its set of NN reference sites using a multivariate one sample *t*-test called Test Site Analysis (TSA; Bowman and Somers, 2006). TSA uses the Mahalanobis or generalized distance (D) to quantify the difference between a potentially impaired test site and the mean for a set of reference sites. The analysis involves a non-central (or equivalence) test that evaluates whether the test site falls outside of the normal range for the reference sites (Kilgour *et al.*, 1998). If the difference between the test site and the mean of the reference sites exceeds the normal range (*i.e.*, is significant at $P < 0.05$) the test site is deemed to be significantly impaired. If the *P*-value approaches 1 (*i.e.*, $P > 0.95$), the test site can be

considered in reference condition. If the *P*-value falls between these two values (*i.e.*, $0.05 < P < 0.95$) the test site is characterized as potentially impaired. TSA can include calculations of T^2 and partial T^2 statistics that identify the metrics that were most important in determining the difference between the test and reference sites. TSA was performed using the ten biological metrics for each individual test site compared to the mean values for the 15 NN reference sites. For each test site, separate TSA's were performed for the standard and live-sort method samples to determine if the two methods classified the test sites into different impairment categories. The statistical package CANOCO (ter Braak and Šmilauer, 2003) was used for all ordination analyses. All TSA calculations were done in Excel (ver. 14).

RESULTS

Community composition

The majority of families encountered were collected by both methods, with a few exceptions (Standard method only: Trhypachthoniidae, Aeolosomatidae, Halacaridae, Limncharidae, Malaconothridae and Onychiuridae; Live-sort method only: Molannidae, Baetiscidae, Belostomatidae, Leptohiphidae, Lestidae, Momoniidae, Neoacaridae, Noctuidae, Notonectidae, and Sericostomatidae). These families were often very rare (*i.e.*, found in <5% of the sites); however, in some cases, the families were found in >5% of sites (Standard Approach: Aeolosomatidae, Halacaridae, and Trhypachthoniidae). The BC similarity index was used to compare the communities collected by each method for each stream individually. BC similarities were generally high (44-99% - Tab. 3). Similarity appeared to be related to the degree of disturbance. Reference streams

Tab. 3. The mean difference and standard deviation between the two methods for each metric is given. Significance was tested using paired *t*-tests. A negative mean difference indicates that the values for that metric were more often higher in the standard over the rapid protocol. Simple linear regressions were calculated ($\text{Metric}_{(\text{Standard})} = m * \text{Metric}_{(\text{live-sort})} + b$) and the correlation coefficient (*r*), slope (*m*), and intercept (*y*) is illustrated below. Systematic error was tested using *t*-tests on both the slope ($b=1$) and intercept ($m=0$).

Metric	Paired t-test				Regression				Slope (=1)		Intercept (=0)	
	Mean	SD	t	P	r	m	b	P	t	P	t	P
Community descriptors												
Family richness	3.7	4.45	6.50	**	0.78	0.75	2.18	**	-3.19	*	1.13	0.260
EPT richness	2.04	2.21	7.22	**	0.91	0.83	-0.55	**	-3.28	*	-1.05	0.299
Simpson's diversity	0.09	0.17	4.28	**	0.59	0.57	0.21	**	-4.19	**	2.76	0.008
Simpson's evenness	0.03	0.10	2.73	*	0.05	0.05	0.16	0.65	-8.03	**	6.17	**
Relative abundance descriptors												
% EPT	18.40	16.47	8.72	**	0.83	0.53	-2.20	**	-10.47	**	-1.11	0.273
%Chironomidae	-14.68	20.16	-5.69	**	0.63	0.57	29.70	**	-4.69	**	7.61	**
%Oligochaeta	-7.38	11.64	-4.95	**	0.78	1.33	5.25	**	2.40	0.019	3.11	*
%Mollusca	-1.51	3.31	-3.57	**	0.78	0.82	2.09	**	-2.22	0.030	4.29	**
CA Axis 1	-0.11	0.23	-3.88	**	0.93	0.80	0.11	**	-4.58	**	4.55	**
CA Axis 2	0.01	0.18	0.56	0.57	0.92	0.84	-0.02	**	-3.56	*	-0.898	0.373

* $P < 0.01$; ** $P < 0.001$.

were very similar (77-99%) but impacted streams ranged more widely (44-94%). Three sites in particular, all test sites, had low similarity between the two methods.

Metric similarity

Mean differences between metric values for each method differed significantly from zero for all metrics except for CA axis 2 scores (paired *t*-test, $P < 0.001$, Tab. 3). The live-sort approach resulted in consistently higher estimates of family richness, EPT richness, and % EPT. In no stream was family richness equal between the two methods. On average, the difference in richness between the two methods was 3.7 families; however, the difference was quite large in some streams (> 10). For example, at one site more than twice as many families were collected using the live-sort approach (richness=30) in comparison to the standard method (richness=12). Those families found only in samples processed using the live-sort method were usually uncommon/rare (less than 10 individuals found across all sites). The live-sort approach also resulted in greater estimates of EPT richness (average difference=2 families). A greater proportion of those samples were composed of EPT, which was up to 57% higher in those samples from the rapid method (average difference=18.4%).

Proportionally, the standard method samples had a greater composition of families comprised of small organisms (e.g., Chironomidae and Oligochaeta). This was particularly evident at one site, where the lower family richness could be attributed to the almost complete dominance of Chironomidae in the standard sample (89%). Estimates of %Chironomidae were up to 44% higher in standard samples. Although in general, samples collected using the standard method had generally higher %Chironomidae, the variation in this metric was very large ($SD = 20.16$). At one site, the live-sort sample recorded twice as many chironomids than the matching standard sample. The %Oligochaetes in a sample was also higher in the standard samples (77% of the time). The difference in % Oligochaetes was due to much higher values in the standard sample, up to 44% higher. When live-sort samples were higher than standard samples, the difference was usually only a few percent ($< 12\%$). The average difference in proportion of Oligochaetes between standard and live-sort samples was 7.38%. CA axis 1 also reflected this gradient of more Chironomids and Oligochaetes in standard samples and more EPT in live-sort samples.

The equation for the best fit line relating standard and live-sort methods samples for each of the four metrics is given in Tab. 3. Four metrics exhibited high association: family richness ($r = 0.78$), EPT richness ($r = 0.91$), % EPT ($r = 0.83$), and % Mollusca ($r = 0.78$) (Fig. 2). For the two-tailed Student *t*-tests with the null hypothesis of the slope equalling one and the intercept equalling zero, *P* values were less than 0.05 for the majority metrics, indicating

both proportional (slope $\neq 1$) and constant (intercept $\neq 0$) errors in agreement between the two methods. Most errors were proportional, with differences between the two methods increasing at higher metric values.

Assessing impairment

An RDA of the standard data identified two water chemistry variables that significantly explained the variation in abundance of benthic invertebrate families in the reference streams; aluminium and total suspended solids (TSS). The live-sort RDA identified conductivity and TSS as significant predictors, but aluminium was not found to be significant. Typically, variables which are affected by anthropogenic impacts are not included when selecting reference streams and therefore aluminium was excluded from the analysis. Although it may be affected by acidification and urbanization, conductivity and TSS was retained to capture the large variation seen among reference streams, as has been shown in previous studies (Rosenberg *et al.*, 1999; Parsons *et al.*, 2010). A PCA using these two water chemistry variables (TSS and conductivity) and all 13 habitat descriptors (Tab. 1) was used to determine the nearest neighbours (NN).

PCA axes 1 and 2 are presented in Fig. 3. Principal Component axis 1 represented a substrate/ flow gradient; velocity (-0.74) and % gravel/cobble (-0.68) had high negative loadings whereas % clay/silt (0.89) and % organic matter (0.64) had positive loadings. Principal Component axis 2 separated deeper (0.64) and wider sites (0.60) from small, shallow sites high in TSS (-0.54). Principal Component axis 3 had positive loadings for conductivity (0.50) and % algae (0.54), separating these sites from those high in % detritus (-0.56) and % woody debris (-0.55). Principal Component axis 4 separated two contrasting substrate types, those high in % sand (-0.68) from those high in % boulder (0.60).

We used TSA to evaluate the magnitude of difference between each test site and the appropriate NN reference sites for each of the standard and live-sort data. Agreement between assessment outcomes was high, with 88% (29 out of 33) of test sites assessed as impaired or potentially impaired. Similarly, there was moderate agreement between the generalized distance (D), which represents the distance between each site and the reference sites, generated by the two methods (Fig. 4; $r = 0.61$, $P = 0.002$). Generally, the live-sort method detected a greater difference (higher D values) from reference condition at most test sites. The higher D values in live-sort samples resulted in different assessment outcomes at 11 test sites. In five sites, differences in assessment outcomes were large (Tab. 4). At four of these sites, the live-sort method found them to be significantly impaired while the standard method indicated they were in reference condition. A comparison of both methods to identify each reference site as in reference condition (by

comparing it to the appropriate NN reference sites) revealed that neither method identified any of the reference sites as impaired. However, two of the reference sites were identified as potentially impaired by the live-sort method while only one of the reference sites was identified as potentially impaired by the standard method.

We re-ran the TSA's for each site, removing the metrics individually, to provide a partial T^2 values for each metric. This allowed us to determine whether the metrics that were important in separating the test community from reference conditions were the same in both samples (standard versus live-sort). For example, at one site, % EPT added the most unique information (partial $T^2=1140$, $F=9.3$, $P=0.03$) in the standard sample, and the removal of this metric found the site to be only potential impaired. In contrast, %Oligochaeta was the most important metric distinguishing this test site from reference when live-sort

sample data was used (partial $T^2=135$, $F=105.1$, $P=0.0001$). Removal of this metric from the analysis resulted in no significant difference from reference conditions for this test site. Among the nine sites found to be significantly impaired by both methods, the most important metric separating the test site from the reference sites was the same at five sites. Overall, %Oligochaeta was most often the most distinguishing metric among sites found to be significantly impaired using live-sort samples (9 out of 19 sites). In contrast, %Chironomidae was most often important in the standard samples (7 out of 14 sites).

DISCUSSION

In this study, we compared two common benthic bioassessment methods to determine the extent to which differences in subsampling and sorting methodology af-

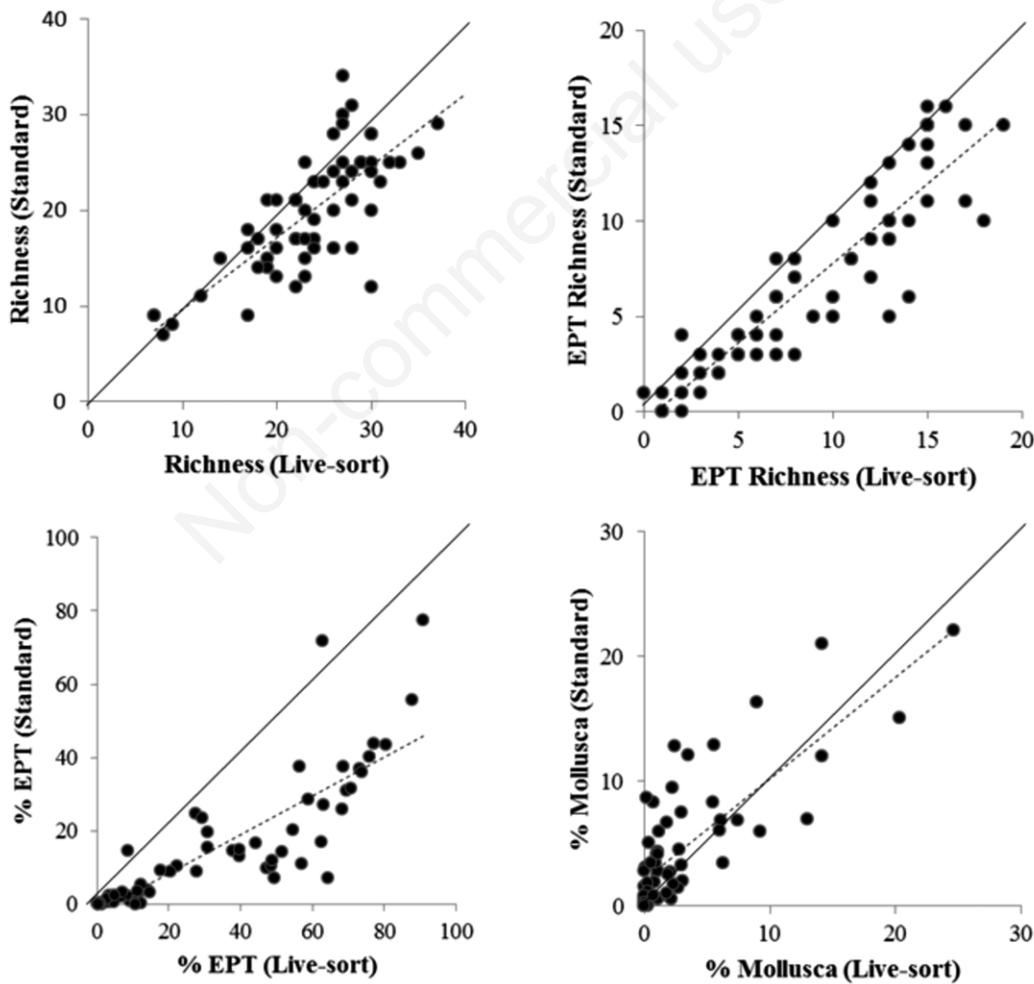


Fig. 2. Correlation between EPT richness, family richness, relative abundance of EPT, and relative abundance of Mollusca in the standard and live-sort method samples. The solid line shows the line of perfect agreement and the dashed line shows the least-squares regression relationship ($P<0.001$).

fect the assessment of biological impairment in streams. The two methods compared differed primarily in their approach to sorting: the standard method involved preserving samples for later subsampling using a Marchant box, with sorting and identification under a microscope, while the second approach used a live-sort methodology with no microscopic aid. Similarity in community composition between samples was high; however, significant differences between many indicator metrics were evident. The increased likelihood of finding rare taxa in the live-sort method resulted in differences in metrics which were affected by richness, including Simpson's diversity, Simpson's evenness, family richness, and EPT richness. Despite differences in the inclusion of rare taxa, both methods performed similarly in identifying biological impairment, with live sorting methods identifying 97% (32/33) of test sites as impaired or potentially impaired while the standard method identified 88% (29/33) as such.

Similarity in community composition between the two methods, as assessed using the BC similarity index, was high at the majority of sites, particularly among reference sites. This similarity was higher than between replicate samples in an evaluation of sample coherence by Lorenz and Clarke (2006), who found an average similarity of 70% between replicates. They also found low similarity between live-sorting and lab-sorting methods (e.g., between the STAR-AQEM and Italian IBE method), suggesting that these two protocols result in significantly different estimates of community composition. However, these differences may have also resulted from differences in stream microhabitats sampled, which also varied between the methods. Interestingly, we found that reference sites exhibited more within site similarity than test sites, suggesting method differences may be less important in reference streams.

The similarity observed in community composition be-

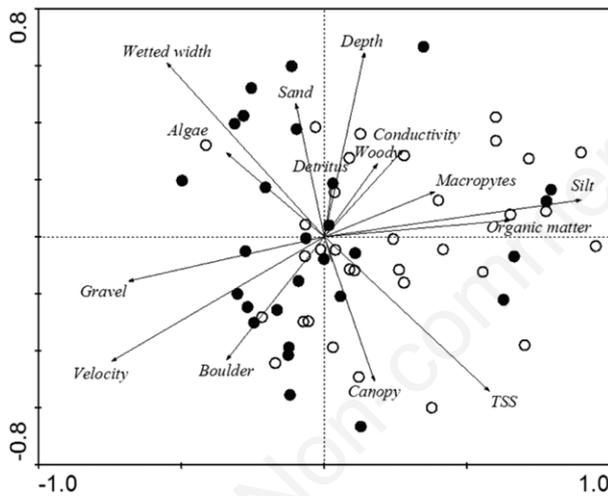


Fig. 3. Principal Component Analysis biplot of the 13 habitat and 2 water chemistry variables in the 28 reference sites (closed circles). Test sites (open circles) were passively projected onto the ordination diagram.

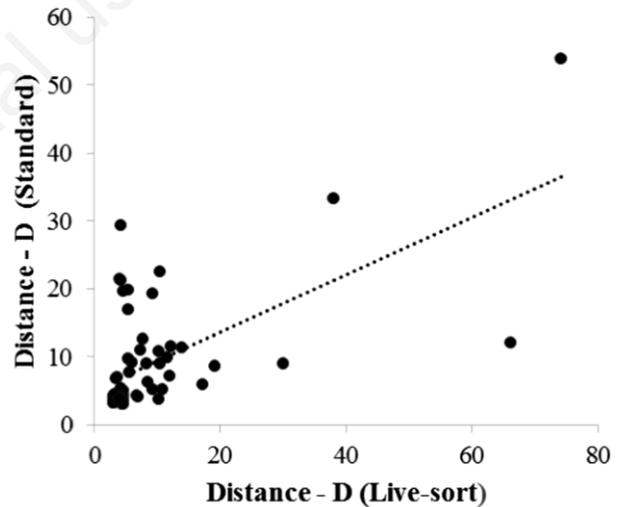


Fig. 4. Correlation between Mahalanobis (generalized distance - D) in the standard and live-sort method samples. The dashed line shows the least-squares regression relationship ($P < 0.001$).

Tab. 4. Results for test streams in which the two methods gave different assessments of impairment are given, including the generalized distance (D), effect size (F), and non-central P-value.

Site	D	F	P	Test site analysis result				
				Assessment	D	F	P	Assessment
Live-sort method more sensitive								
SUD202	4.00	16.8	0.98	Reference	21.47	485.0	<0.01	Impaired
SUD208	4.27	19.2	0.96	Reference	29.36	907.1	<0.01	Impaired
NMCN	4.27	19.2	0.96	Reference	21.21	473.5	<0.01	Impaired
SUD216	3.67	14.2	1.00	Reference	6.97	51.2	0.31	Potential
Standard method more sensitive								
SUD214	10.33	112.3	0.029	Impaired	3.77	15.0	0.99	Reference

tween methods is not necessarily reflected in indicator metrics, which may be more heavily influenced by the inclusion or absence of rare taxa. The under-representation of certain taxa in live-sorted samples has called into question the accuracy of this method and whether it can be compared with preserved and microscope-aided sorting (Smith *et al.*, 1999; Humphrey *et al.*, 2000; Nichols and Norris, 2006). However, the live-sort method has also been shown to collect more taxa than microscope sorting (Borisko *et al.*, 2007). Although we found that the standard approach identified some taxa not found in the live-sort samples in our study, live-sort samples resulted in more taxa and those taxa were generally from large, mobile groups. This resulted in higher estimates of family richness and EPT richness, sometimes up to twice as many taxa in these samples. In comparison, the standard methods results in samples often dominated by chironomids.

Variation in metrics exists for a number of reasons, including small-scale natural variability or patchiness in benthic invertebrates within a stream site (Downes *et al.*, 1993). In a hierarchical study of variation in biological metrics extending from differences between river basins to differences among replicate samples, Springe *et al.* (2006) found high variation in metrics among replicates and that metrics connected with taxonomic composition like EPT taxa, and especially taxonomic groups (%), were the most variable. Large samples, covering a sufficient area to adequately represent the benthic community at a site, are collected so as to reduce the chance of this small-scale spatial variation influencing between stream assessments. However, such large samples often require subsampling, another major source of metric variability. Clarke *et al.* (2006) found that sub-sampling variation contributed more than 50% of the overall variance between replicate samples for some metrics. They demonstrated that sub-sampling variance was greatest for those metrics which are based on the numbers of taxa present. Although replicate samples were not collected in this study, the consistent bias present in some metrics suggests that method differences in sorting were responsible for the within stream variation we have found in richness and diversity.

The potential for differences associated with live-sorting has been predicted as a key factor that may affect the sensitivity of predictive models (Humphrey *et al.*, 2000). Large differences in assessment outcomes resulting from different sources of data calls into question the validity of assessments and their importance in directing changes in watershed management (Herbst and Silldorf, 2006). However, previous studies have shown that differences in community composition and metrics between samples processed under different sorting methodology, including live sorting, do not necessarily lead to differences in assessment (Nichols and Norris, 2005; Borisko *et al.*, 2007).

Growns *et al.*, (1997) found that, despite sample differences in some metrics, all six sample processing methods evaluated discriminated between reference and impacted sites. In our study, both methods performed similarly in identifying biological impairment. However, the live-sort method identified more sites as impaired in comparison to the standard methods. This suggests that caution should be used when assessments combine data from both methods.

The development of rapid bioassessment protocols has resulted in much debate on the most efficient methods, with respect to time, cost, *etc.*, for processing benthic invertebrate samples (Courtemanch, 1996; Growns *et al.*, 1997; King and Richardson, 2002, Nichols and Norris, 2006). Although we did not evaluate cost and processing times, we would expect that live-sorting would be significantly faster. Smith *et al.* (1999) found that the cost of laboratory sorted samples was about 20% greater than that of live picking, which was 40% faster than laboratory sorting. Other benefits to the live-sort approach include a reduction in the amount of material that is returned to the lab which is beneficial for hard to reach sites (Chessman, 1995), the need for less expensive equipment and less preservatives, all which benefit community groups (Campbell, 2007). While these are important features of sorting protocols, ultimately, they must result in data which is representative of the site and fits the objectives of the study (Barbour *et al.*, 1999). For example, the standard method evaluated here uses a Marchant box, which has been criticized for being expensive and time consuming (Perrin *et al.*, 2005), but is necessary for estimates of density. Additionally, in assessments in which genus or species identification of Chironomidae provide the most sensitive indication of impact (for example, in wetland biomonitoring; King and Richardson, 2002), the live-sort method may not be adequate for these objectives.

CONCLUSIONS

The goal of bioassessment should not be to completely describe the benthic community of a site, but rather characterize it in a way that allows it to be separated from other sites in an ecologically relevant manner (Baker and Huggins, 2005). As different agencies and programmes have different objectives when undertaking bioassessments, they may not require the same level of effort in sample processing but may yield comparable data for certain objectives (Barbour *et al.*, 1999). For some objectives, the extra time and expense required for microscope aided sorting may not necessarily be repaid with an increase in sensitivity. Here we showed that in our streams, live-sorting provides similar assessments of impairment to the more labour intensive laboratory sorting. This result is particularly beneficial to volunteer led community groups, which often do not have the resources that are available to government and regulatory agencies, but can

still generate data that provide meaningful assessment of impairment.

REFERENCES

- Baker D, Huggins D, 2005. Sub-sampling techniques for macroinvertebrates, fish, and benthic algae sampled in biological monitoring of streams and rivers. Kansas Biological Survey, Lawrence, KS Open-File Report No. 132: 26 pp.
- Barbour MT, Gerritsen J, Snyder BD, Stribling JB, 1999. Rapid bioassessment protocols for use in wadeable streams and rivers: periphyton, benthic invertebrates, and fish. Second edition. EPA 841-B-99-002, U.S. EPA, Office of Water, Washington: 344 pp.
- Birk S, Hering D, 2006. Direct comparison of assessment methods using benthic macroinvertebrates: a contribution to the EU Water Framework Directive intercalibration exercise. *Hydrobiologia* 566:401-415.
- Borisko JPB, Kilgour W, Stanfield LW, Jones FC, 2007. An evaluation of rapid bioassessment protocols for stream benthic invertebrates in Southern Ontario, Canada. *Water Qual. Res. J. Can.* 42:184-193.
- Bowman MF, Somers KM, 2006. Evaluating a novel Test Site Analysis (TSA) bioassessment approach. *J. N. Am. Benthol. Soc.* 25:712-727.
- Bowman MF, Somers KM, Reid RA, Scott LD, 2006. Temporal response of stream benthic macroinvertebrate communities to the synergistic effects of anthropogenic acidification and natural drought events. *Freshwater Biol.* 51:768-782.
- Campbell WB, 2007. Innovation in evaluating freshwater macroinvertebrates in Mexico: Community-based volunteers and water quality biomonitoring. *Mexican Aquatic Entomology Symposium*: 91-104.
- Carlson PE, Johnson RK, Mckie BG, 2013. Optimizing stream bioassessment: Habitat, season, and the impacts of land use on benthic macroinvertebrates. *Hydrobiologia* 704:363-373.
- Carter, JL, Resh VH, 2001. After site selection and before data analysis: Sampling, sorting, and laboratory procedures used in stream benthic macroinvertebrate monitoring programs by USA state agencies. *J. N. Am. Benthol. Soc.* 20:658-682.
- Chessman BC, 1995. Rapid assessment of rivers using macroinvertebrates: a procedure based on habitat-specific sampling, family level identification and a biotic index. *Aust. J. Ecol.* 20:122-129.
- Chessman BC, Robinson DP, 1987. Some effects of the 1982-83 drought on water-quality and macroinvertebrate fauna in the lower Latrobe river, Victoria. *Aust. J. Mar. Freshw. Res.* 38:289-299.
- Ciborowski JH, 1991. Estimating processing time of stream benthic samples. *Hydrobiologia* 222:101-107.
- Clarke RT, Lorenz A, Sandin L, Schmidt-Kloiber A, Strackbein, J, Kneebone NT, Haase P, 2006. Effects of sampling and sub-sampling variation using the STAR-AQEM sampling protocol on the precision of macroinvertebrate metrics. *Hydrobiologia* 566:441-459.
- Courtemanch DL, 1996. Commentary on the subsampling procedures used for rapid bioassessments. *J. N. Am. Benthol. Soc.* 15:381-385.
- Cuffney TF, Gurtz ME, Meador MR, 1993. Methods for collecting benthic invertebrate samples as part of the National Water-Quality Assessment Program, US Geological Survey. Available from: <http://water.usgs.gov/nawqa/protocols/OFR-93-406/>
- David SM., Somers KM, Reid RA, Hall RJ, Girard RE, 1998. Sampling protocols for the rapid bioassessment of streams and lakes using benthic macroinvertebrates, Second edition. Ontario Ministry of the Environment Report, Dorset: 29 pp.
- Diamond JM, Barbour MT, Stribling JB, 1996. Characterizing and comparing bioassessment methods and their results: A perspective. *J. N. Am. Benthol. Soc.* 15:713-727.
- Dickens CWS, Graham PM, 2002. The South African Scoring System (SASS) version 5 rapid bioassessment method for rivers. *Afr. J. Aquat. Sci.* 27:1-10.
- Downes BJ, Lake PS, Schreiber ESG, 1993. Spatial variation in the distribution of stream invertebrates: implications of patchiness for models of community organization. *Freshwater Biol.* 30:119-132.
- Flotemersch JE, Stribling, JB, Paul J, 2006. Concepts and approaches for the bioassessment of non-wadeable streams and rivers. EPA/600/R-06/127. Office of Research and Development, US Environmental Protection Agency, Cincinnati.
- Growns JE, Chessman BC, Jackson JE, Ross DG, 1997. Rapid assessment of Australian rivers using macroinvertebrates: cost and efficiency of 6 methods of sample processing. *J. N. Am. Benthol. Soc.* 16:682-693.
- Hawkins CP, Norris RH, 2000. Effects of taxonomic resolution and use of subsets of the fauna on the performance of RIVPACS-type models, p. 217-228. In: J.F. Wright, D.W. Sutcliffe and M.T. Furse (eds.), *Assessing biological quality of freshwaters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, Cumbria.
- Herbst DB, Silldorf EL, 2006. Comparison of the performance of different bioassessment methods: similar evaluations of biotic integrity from separate programs and procedures. *J. N. Am. Benthol. Soc.* 25:513-530.
- Humphrey CL, Storey AW, Thurtell L, 2000. AUSRIVAS: operator sample processing errors and temporal variability - implications for model sensitivity, p.143-163. In: J.F. Wright, D.W. Sutcliffe and M.T. Furse (eds.), *Assessing biological quality of freshwaters: RIVPACS and other techniques*. Freshwater Biological Association, Ambleside, Cumbria.
- Karr JR, 1993. Defining and assessing ecological integrity: Beyond water quality. *Environ. Toxicol. Chem.* 12:1521-1531.
- Kilgour BW, Somers KM, Matthews DE, 1998. Using the normal range as a criterion for ecological significance in environmental monitoring and assessment. *Ecoscience* 5:542-550.
- King RS, Richardson CJ, 2002. Evaluating subsampling approaches and macroinvertebrate taxonomic resolution for wetland bioassessment. *J. N. Am. Benthol. Soc.* 21:150-171.
- Lenat DR, 1988. Water quality assessment of streams using a qualitative collection method for benthic macroinvertebrates. *J. N. Am. Benthol. Soc.* 7:222-233.
- Lorenz A, Clarke RT, 2006. Sample coherence - a field study approach to assess similarity of macroinvertebrate samples. *Hydrobiologia* 566:461-476.
- Marchant R, 1989. A subsampler for samples of benthic invertebrates. *B. Aust. Soc. Limnol.* 12:49-52.
- Narangarvu D, Hsu C, Shieh S, Wu F, Yang P, 2014. Macroinvertebrate assemblage patterns as indicators of water quality

- in the Xindian watershed, Taiwan. *J. Asia Pac. Entomol.* 17:505-513.
- Nichols S, Norris RH, 2006. River condition assessment may depend on the sub-sampling method: field live-sort versus laboratory sub-sampling of invertebrates for bioassessment. *Hydrobiologia* 572:195-213.
- Parsons BP, Watmough SA, Dillon PJ, Somers KM, 2010. A bioassessment of lakes in the Athabasca Oil Sands Region, Alberta, using benthic macroinvertebrates. *J. Limnol.* 69(Suppl.1):105-117.
- Perrin CJ, Dolecki D, Salter S, 2005. A recommended approach for sorting and sub-sampling benthic invertebrate samples for use in the Skeena RCA analysis. Report prepared by Limnotek Research and Development Inc., for B.C. Ministry of Water Land and Air Protection, Smithers: 9 pp.
- Rehn AC, Ode PR, Hawkins CP, 2007. Comparisons of targeted-riffle and reach-wide benthic macroinvertebrate samples: implications for data sharing in stream-condition assessments. *J. N. Am. Benthol. Soc.* 26:332-348.
- Reynoldson TB, Logan C, Pascoe T, Thompson SP, 2003. CABIN (Canadian Aquatic Biomonitoring Network) invertebrate biomonitoring field and laboratory manual. Environment Canada, National Water Research Institute.
- Rosenberg, DM, Reynoldson TB, Resh VH, 1999. Establishing reference conditions for benthic invertebrate monitoring in the Fraser River catchment, British Columbia, Canada. DOE-FRAP 1998-32. Fraser River Action Plan, Environment Canada, Vancouver, British Columbia.
- Rosillon D, 1987. Seasonal variation in the benthos of a chalk trout stream, the River Sanson, Belgium. *Hydrobiologia* 126:253-262.
- Smith, MJ, Kay WR, Edward DHD, Papas PJ, Richardson KS, Simpson JC, Pinder AM, Cale DJ, Horwitz PHJ, Davis JA, Yung FH, Norris RH, Halse SA, 1999. AusRivAS: using macroinvertebrates to assess ecological condition of rivers in Western Australia. *Freshwater Biol.* 41:269-282.
- Springe G, Sandin L, Briede A, Skuja A, 2006. Biological quality metrics: their variability and appropriate scale for assessing streams. *Hydrobiologia* 566:153-172.
- Stark JD, Boothroyd IKG, Harding JS, Maxted JR, Scarsbrook MR, 2001. Protocols for sampling macroinvertebrates in wadeable streams. New Zealand Macroinvertebrate Working Group Report No. 1. Prepared for the Ministry for the Environment. Sustainable Management Fund Project No. 5103: 57 pp.
- Sovell LA, Vondracek B, 1999. Evaluation of the fixed-count method for Rapid Bioassessment Protocol III with benthic macroinvertebrate metrics. *J. N. Am. Benthol. Soc.* 18:420-426.
- ter Braak, CFJ, Šmilauer, P, 2003. CANOCO Reference Manual and User's Guide to Canoco for Windows: Software for Canonical Community Ordination (version 4.52). Micro-computer Power, New York.
- Vinson MR, Hawkins CP, 1996. Effects of sampling area and subsampling procedure on comparison of taxa richness among streams. *J. N. Am. Benthol. Soc.* 15:392-399.
- Wilson MJ, McTammany ME, Bilger MD, Reese SP, Hayes BR, 2015. Combining data from multiple agencies to assess benthic macroinvertebrate communities in a large gravel-bed river. *Freshwater Sci.* 34:593-605.
- Wright JF, Moss D, Armitage PD, Furse MT, 1984. A preliminary classification of running-water sites in Great Britain based on macroinvertebrate species and the prediction of community type using environmental data. *Freshwater Biol.* 14:221-256.