Hydrobiologia (2006) 566:401–415
M.T. Furse, D. Hering, K. Brabec, A. Buffagni, L. Sandin & P.F.M. Verdonschot (eds), The Ecological Status of European Rivers: Evaluation and Intercalibration of Assessment Methods
DOI 10.1007/s10750-006-0081-8

Direct comparison of assessment methods using benthic macroinvertebrates: a contribution to the EU Water Framework Directive intercalibration exercise

Sebastian Birk* & Daniel Hering
Department of Hydrobiology, University of Duisburg-Essen, Universitätsstr. 5, D-45117, Essen, Germany
*(*Author for correspondence: E-mail: sebastian.birk@uni-essen.de)*

Key words: ecological quality classification, biological assessment index, macroinvertebrates, common intercalibration type, STAR project, EU Water Framework Directive

Abstract

The aim of the intercalibration exercise presently performed by the EU is to identify and resolve significant inconsistencies between the ecological quality classifications of EU Member States and the normative definitions of the EU Water Framework Directive. Based on benthic macroinvertebrate data of two European stream types (small siliceous mountain streams and medium-sized lowland streams in Central and Western Europe) we correlated the indices of 10 river quality assessment methods (ASPT, BMWP, DSFI, German Multimetric Index, Saprobic Indices) applied in Austria, Czech Republic, Denmark, Germany, Poland, Slovak Republic, Sweden and United Kingdom. National class boundaries were compared via regression analysis. Assessment methods of the same type (Saprobic Indices, BMWP/ASPT scores) showed best correlation results ($R^2 > 0.7$). The good quality status boundaries of the national methods deviated up to 25%; thus indicating the necessity to harmonize the national classification schemes. Prerequisites of the presented intercalibration approach are (1) a sufficiently large and consistent dataset representative of the respective common intercalibration types and (2) agreement on common type specific reference conditions.

Introduction

In the individual European countries the practice of evaluating ecological river quality is very different (Metcalfe-Smith, 1994; Knoben et al., 1995; Birk & Hering, 2002). Although river monitoring programmes in most countries are based on the benthic macroinvertebrate community, design and performance of individual methods to assess rivers with this organism group vary significantly. On the one hand this is due to different traditions in stream assessment. While in many Central and Eastern European countries modifications of the Saprobic System have been applied for decades as standard methods (Birk & Schmedtje, 2005), other countries rely on the Biological Monitoring Working Party score (BMWP, 1978), which has been adjusted for the use in various countries (Armitage et al., 1983; Just et al., 1998; Alba-Tercedor & Pujante, 2000; Kownacki et al., 2004). On the other hand the EU Water Framework Directive had a great effect on European freshwater management, since it outlines an innovative concept of bioassessment: not the impact of single pressures on individual biotic groups but the deviation of the community from undisturbed conditions is decisive for ecological status classification. In many EU Member States efforts are being made to adapt the national programmes to these new requirements; however, different approaches are being used, since in some countries a single stressor (e.g. organic pollution) is overwhelming, while in other

regions different stressors are of equal importance and simultaneously affect river inhabiting communities.

To overcome the difficulties in comparing the various national assessment methods the Directive outlines an intercalibration procedure of the methods' outputs. Member States are enabled to establish or to maintain their own methods; a definition of high, good or moderate biological quality is provided centrally through the intercalibration exercise. The aim of the intercalibration exercise is to identify and to resolve significant inconsistencies between the quality class boundaries established by Member States and indicated by the normative definitions of the Directive (CIS WG 2.A Ecological Status, 2004).

The first efforts to compare different national assessment methods in Europe go back to 1975. Three intercalibration campaigns organized by the Commission of the European Communities included comparisons of field sampling, sample treatment and quality assessment applied in Germany, Italy and United Kingdom (Tittizer, 1976; Woodiwiss, 1978; Ghetti & Bonazzi, 1980). These early studies established strong correlations between the individual assessment methods and compared the methods directly. This approach towards intercalibration was then followed by various authors both to demonstrate the relationship of methods and to point out discrepancies between national quality classifications (Ghetti & Bonazzi, 1977; Rico et al., 1992; Friedrich et al., 1995; Biggs et al., 1996; Morpurgo, 1996; Stubauer & Moog, 2000). In their preparatory study for the Water Framework Directive Nixon et al. (1996) explicitly recommended direct comparison to be used for the intercalibration of assessment methods.

However, the official intercalibration exercise for the Water Framework Directive has adopted an alternative approach due to the lack of a sufficiently large and consistent international database covering all of Europe: indirect comparison via intercalibration common metrics, thus, generating a 'common' multimetric assessment procedure, which is more or less applicable in most of Europe and comparing national assessment methods against this common method (Buffagni et al., 2006).

In this paper we

(1) evaluated the principal suitability of directly comparing assessment methods for intercalibration procedures;
(2) tested a variety of different regression techniques to refine the practical application of direct comparison for intercalibration purposes;
(3) directly compared assessment methods frequently applied for two broadly defined European river types and suggest steps for harmonising class boundaries.

Methods

Overview

This study was based on a two-step analysis: first, different assessment methods, which are presently being used in national water management, were calculated with the same taxa lists. The results of the individual assessment methods were then directly compared by regression analysis.

All data used in this study resulted from the AQEM project (Hering et al., 2004) and the STAR project (Furse et al., 2006). Only data on invertebrate samples restricted to two broadly defined stream types were used. With the data from each stream type up to 10 national assessment systems were calculated, which were first normalized by calculating ecological quality ratios (EQR) (i.e., transferring the results into a common scale ranging from 0 to 1). These normalized assessment results were fed into a regression analysis, to translate the index results of country A into the index results of country B. Comparison of more than two methods was enabled by including the index of country C and translating these results into the index results of country B ('common scale'). In addition, the assessment results were correlated to environmental gradients. In a second step, the class boundaries between the individual quality classes, as applied by the national assessment systems, were compared.

To test the impact of different regression techniques on the results, linear and nonlinear techniques were compared.

Samples and sites

This study was based on benthic invertebrate data sampled in the EU projects AQEM and STAR with standardized field and laboratory protocols (Furse et al., 2006). The data were limited to two broadly defined stream type groups: small, siliceous mountain streams and medium-sized lowland streams in Central and Western Europe. In the official intercalibration exercise for the Water Framework Directive, these stream types were named 'small-sized, mid-altitude brooks of siliceous geology' (R-C3) and 'medium-sized, lowland streams of mixed geology' (R-C4) in Central Europe (Table 1).

Two hundred ninety four samples taken at 125 sites located in four different countries in spring and summer were analysed for the small mountain streams. The lowland stream type embraced a total of 217 samples taken at 71 sites in four different countries in spring, summer and autumn.

The ecological quality of each sampling site was pre-classified based on expert judgement of the field researchers having sampled the streams and, if available, additional knowledge derived from previous studies. Each site was assigned to one of five quality classes ('high', 'good', 'moderate', 'poor', 'bad') referring to the estimated main stressor's degree of impairment. For the AQEM sites, the pre-classification of most sites was replaced by the post-classification after sampling due to additional environmental parameters gained during the field work (physical–chemical and hydromorphological variables).

National assessment methods and quality classifications

Altogether ten biological assessment indices were compared in this analysis (Table 2), all of which are either in current usage in certain European countries or are about being implemented into water management as standard techniques. Most represented biotic index or score methods (Saprobic Index (SI), Biological Monitoring Working Party (BMWP) Score, Average Score Per Taxon (ASPT), Danish Stream Fauna Index (DSFI)). All indices were part of the respective national method

Table 1. Overview of samples included in the analysis

Stream type	Country	Stream type	Ecoregion no.	Number of samples
Small siliceous mountain streams	Austria	Small-sized shallow mountain streams	9	36
	Czech Republic	Small-sized shallow mountain streams	9, 10	40
		Small-sized streams in the Central Sub-alpine mountains	9	32
		Small-sized streams in the Carpathians	10	28
	Germany	Small streams in lower mountainous areas of Central Europe	9	86
		Small-sized Buntsandstein-streams	9	24
	Slovak Republic	Small-sizes siliceous mountains streams in the West Carpathians	10	48
Medium-sized lowland streams	Denmark	Medium-sized deeper lowland streams	14	46
	Germany	Mid-sized sand bottom streams in the German lowlands	14	86
	Sweden	Medium-sized deeper lowland streams	14	14
		Medium-sized streams on calcareous soils	14	35
	United Kingdom	Medium-sized deeper lowland streams	18	36

Table 2. Overview of national assessment methods

Stream type	Country	Assessment index	Category	Abundance	Reference
Small siliceous mountain streams	Austria	SI (AT) – Austrian Saprobic Index	BI	Y	Moog et al. (1999)
	Czech Republic	SI (CZ) – Czech Saprobic Index	BI	Y	CSN 757716 (1998)
	Germany	SI (DE) – German Saprobic Index	BI	Y	Friedrich & Herbst (2004)
	Poland	BMWP (PL) – Polish Biological Monitoring Working Party score	BI	N	Kownacki et al. (2004)
	Slovak Republic	SI (SK) – Slovak Saprobic Index	BI	Y	STN 83 0532-1 to 8, (1978/79)
	United Kingdom	ASPT (UK) – Average Score Per Taxon	BI	N	Armitage et al. (1983)
Medium-sized lowland streams	Denmark	DSFI (DK) – Danish Stream Fauna Index	BI	N	Skriver et al. (2000)
	Germany	GD (DE) – Module 'General Degradation' of the German Assessment System Macrozoobenthos	MI*	Y	Böhmer et al. (2004)
	Sweden	ASPT (SE) – Average Score Per Taxon applied in Sweden	BI	N	Swedish Environmental Protection Agency (2000)
		DSFI (SE) – Danish Stream Fauna Index applied in Sweden	BI	N	
	United Kingdom	ASPT (UK) – Average Score Per Taxon	BI	N	Armitage et al. (1983)

BI, biotic index; MI, multimetric index. *Includes the following single metrics: relative abundance of ETP taxa, German Fauna Index Type 15, number of Trichoptera taxa, Shannon–Wiener diversity, share of rheobiontic taxa, share of shredders (%).

planned for biological monitoring in the context of the Water Framework Directive. With the exception of DSFI and ASPT, applied in Sweden, calculation of index values was based on a nationally adjusted indicator species list.

For the indices applied in Austria, the Czech Republic, Germany and Denmark, stream type specific reference values existed; these described the value of an index to be expected under 'undisturbed conditions'. The system used in the United Kingdom predicted site specific reference values, Sweden defined reference conditions for broad-scale natural geographical regions but in Poland and the Slovak Republic reference values have not yet been established. All indices distinguished between five classes of biological quality. The British and Swedish methods and the German multimetric index defined class boundary values as EQR. The Polish BMWP and the Saprobic Systems used quality classes given as absolute index values. The Austrian, Czech and German quality bands were stream type specific. An overview of nationally defined reference conditions and class boundaries is given in Table 3.

Data preparation

National assessment methods were calculated to the taxa lists of each sample. Absolute index values were converted into EQR by dividing the calculated (observed) value by the index specific reference value. Since, for the Saprobic Indices, biological quality decreased with increasing index values these were converted by the following equation:

Table 3. Original reference and class boundary values of the national assessment methods

Index	SI (AT)	SI (CZ)	SI (DE)	BMWP (PL)	SI (SK)	ASPT (UK)
Small siliceous mountain streams						
Reference (abs)	≤1.50	≤1.20	≤1.25	n.a.	n.a.	≥6.62*
High-good	1.50	1.20	1.40	100	1.79	1.00
Good-moderate	2.10	1.50	1.95	70	2.30	0.89
Moderate-poor	2.60	2.00	2.65	40	2.70	0.77
Poor-bad	3.10	2.70	3.35	10	3.20	0.66
Lit. source	–	Brabec et al. (2004)	Rolauffs et al. (2003)	Kownacki et al. (2004)	STN 83 0532-1 to 8 (1978/79)	National Rivers Authority (1994)
Index	DSFI (DK)	GD (DE)	BMWP (PL)	ASPT (SE)	DSFI (SE)	ASPT (UK)
Medium-sized lowland streams						
Reference (abs)	7	1	n.a.	≥4.7	≥5	≥6.38*
High-good	7	0.80	100	0.90	0.90	1.00
Good-moderate	5	0.60	70	0.80	0.80	0.89
Moderate-poor	4	0.40	40	0.60	0.60	0.77
Poor-bad	3	0.20	10	0.30	0.30	0.66
Lit. source	–	Böhmer et al. (2004)	Kownacki et al. (2004)	Swedish Environmental Protection Agency (2000)	Swedish Environmental Protection Agency (2000)	National Rivers Authority (1994)

Abs, absolute value. *Values were derived by RIVPACS predictions for the corresponding stream type group based on averaged environmental parameter values and combined season information for the analysed samples.

$$EQR_{SI} = 1 - \frac{\text{observed SI value} - \text{reference SI value}}{\text{maximum SI value} - \text{reference SI value}}$$

To validate the national reference values, an index specific reference value was calculated as the 75th percentile of all samples taken at sites pre- or post-classified as high quality status (excluding outliers). For the small mountain streams, sampling sites located in Austria (6 samples), Czech Republic (14 samples), Germany (13 samples) and Slovak Republic (1 sample) were used. For the lowland type sites from Denmark (13 samples), Germany (26 samples), Sweden (2 samples) and United Kingdom (9 samples) were the basis of this calculation.

Conversion into the EQR scale resulted in values ranging from 0 to >1 since several samples revealed biological index values representing higher quality than the respective reference value. These values were not transformed into the value '1' in order to improve the correlation and regression analysis by enlarging the quality gradient.

Correlation and regression analysis

The magnitude of the relation between two assessment methods was specified by the 'coefficient of determination'. Beside linear regression, we applied nonlinear modelling via automatic curve-fitting using the software TableCurve 2D (SYSTAT Software Inc., 2002).

Comparison of quality class boundaries

In order to compare the national quality classes the boundary values of the different assessment methods were transformed into a 'common scale'. In this study two common scales were used: (1) The national method showing the highest mean correlation of all indices. (2) The 'integrative multimetric index for intercalibration' (IMI-IC), an artificial index designed here for the purpose of intercalibration. This index was defined as the mean of all index values calculated for a sample.

The transformation was done based on the results of linear regression analyses, in which the predictor variables were represented by the national indices and the response variables by the 'common scale'. Each boundary value transformed by regression was given including its 95% confidence interval. Class boundaries showing overlapping ranges (translated class boundary ± confidence interval) were considered as being equal.

Based on environmental variables, abiotic gradients were generated for each stream type and the pressure gradients best correlating to the methods analysed in this intercalibration approach were identified. Indirect gradient analysis was aimed at the identification and quantification of physical–chemical and hydromorphological gradients that can be assigned to human impairment. Therefore, Principle Component Analysis (PCA) was run separately on correlation matrices of physical–chemical, catchment land use, hydromorphological and microhabitat variables of the mountain and lowland dataset (see Feld et al., in prep.). A dimensionless value of abiotic pressure, including the 95% confidence interval, was assigned to each national class boundary via regression analysis. These pressure data were used to support class boundary comparisons.

Results

Definition of reference values

The 75th percentiles of reference values were specified in Table 4. Each reference was based on a slightly different number of samples due to the elimination of outliers. Except for the German indices and the assessment methods for which no reference was nationally defined (Polish BMWP and Slovak SI), the 75th percentile, as calculated in this study, generally represented higher biological quality than the minimum values of the national reference.

Descriptive statistics of national indices calculated from the AQEM–STAR datasets

The overall mean of normalized index values (0–1) for the small mountain streams amounted to 0.87, while the same statistic for medium-sized lowland streams was 0.77 (Table 5). The maximum values of all indices except DSFI exceeded 1.0. This was due to the selection of the 75th percentile of AQEM–STAR high status sites as the reference value. The values of the Polish BMWP and the German GD covered ranges of more than 1.0, while the Austrian and German SI, and the British and Swedish ASPT showed value ranges of less than 0.65.

Correlation and regression of national assessment methods

The correlation analysis revealed differences between assessment methods (Table 6). The linear equations of the regression analysis of national methods against methods representing a common scale (best correlating national index, IMI-IC) were displayed in Table 7.

For small mountain streams coefficients of determination ranged from 0.20 (Slovak SI and

Table 4. Reference values of national assessment methods derived by using the 75th percentile of index values calculated from samples taken at high status sites

Index	SI (AT)	SI (CZ)	SI (DE)	BMWP (PL)	SI (SK)	ASPT (UK)
Small siliceous mountain streams						
75th percentile	1.46 (32)	0.91 (34)	1.44 (33)	187 (33)	1.21 (30)	7.26 (33)
Index	DSFI (DK)	GD (DE)	BMWP (PL)	ASPT (SE)	DSFI (SE)	ASPT (UK)
Medium-sized lowland streams						
75th percentile	7	0.67	150	6.57	7	6.57

For small mountain streams the number of high status sites' samples is individually specified in brackets. Values of lowland streams are based on 50 samples.

Table 5. Descriptive statistics of national indices calculated from the AQEM–STAR datasets (normalized index values)

	Mean	Minimum	Maximum	25th percentile	75th percentile	Range	Quartile range
Small siliceous mountain streams ($n = 294$)							
SI (AT)	0.902	0.526	1.112	0.833	0.972	0.585	0.138
SI (CZ)	0.853	0.374	1.112	0.761	0.963	0.739	0.202
SI (DE)	0.920	0.444	1.055	0.895	0.984	0.611	0.088
BMWP (PL)	0.768	0.102	1.273	0.636	0.936	1.171	0.299
SI (SK)	0.890	0.444	1.281	0.798	0.984	0.837	0.186
ASPT (UK)	0.908	0.448	1.077	0.869	0.988	0.629	0.119
Medium-sized lowland streams ($n = 217$)							
DSFI (DK) and DSFI (SE)	0.767	0.286	1.000	0.571	1.000	0.714	0.429
GD (DE)	0.709	0.090	1.149	0.552	0.896	1.060	0.343
BMWP (PL)	0.741	0.173	1.480	0.580	0.900	1.307	0.320
ASPT (SE) and ASPT (UK)	0.869	0.457	1.091	0.797	0.956	0.634	0.159

Table 6. Coefficients of determination based on linear and nonlinear regression ($p < 0.05$)

Index	SI (AT)		SI (CZ)		SI (DE)		BMWP (PL)		SI (SK)		ASPT (UK)	
	Linear	Nonl.	Linear	Nonl.	Linear	Nonl.	Linear	nonl.	Linear	Nonl.	Linear	Nonl.
Small siliceous mountain streams ($n = 294$)												
SI (AT)	1.00	–	0.62	–	0.70	0.74	0.36	0.39	0.73	0.77	0.45	0.46
SI (CZ)	0.62	–	1.00	–	0.62	0.64	0.31	0.35	0.55	–	0.38	–
SI (DE)	0.70	0.73	0.62	0.70	1.00	–	0.53	0.63	0.48	0.56	0.69	0.73
BMWP (PL)	0.36	0.37	0.31	0.34	0.53	–	1.00	–	0.20	0.23	0.69	0.70
SI (SK)	0.73	–	0.55	–	0.48	0.51	0.20	0.21	1.00	–	0.24	0.26
ASPT (UK)	0.45	0.50	0.37	0.45	0.69	0.70	0.69	0.75	0.24	0.36	1.00	–
IMI-IC$_{R\text{-}C3}$	0.79	0.80	0.72	0.74	0.86	0.87	0.72	0.75	0.62	0.66	0.75	–
PE1	0.31	0.33	0.23	0.27	0.46	–	0.37	0.38	0.19	0.23	0.53	–

Index	DSFI (DK) and DSFI (SE)		GD (DE)		BMWP (PL)		ASPT (SE) and ASPT (UK)	
	Linear	Nonl.	Linear	Nonl.	Linear	Nonl.	Linear	Nonl.
Medium-sized lowland streams ($n = 217$)								
DSFI (DK) and DSFI (SE)	1.00	–	0.61	–	0.53	0.54	0.65	–
GD (DE)	0.61	–	1.00	–	0.41	0.46	0.49	–
BMWP (PL)	0.53	0.54	0.41	–	1.00	–	0.51	–
ASPT (SE) and ASPT (UK)	0.65	0.67	0.49	0.50	0.51	0.57	1.00	–
IMI-IC$_{R\text{-}C4}$	0.90	–	0.76	–	0.73	0.75	0.80	–
HY1	0.23	–	0.35	–	0.12	0.13	0.24	0.26

IMI-IC, integrative multimetric index for intercalibration (see text for explanation); PE1, pollution/eutrophication gradient; HY1, hydromorphological gradient.

Polish BMWP) to 0.77 (Austrian SI and Slovak SI). Nonlinear regression gained higher R^2 values in 23 out of 36 relations. The mean difference in R^2 values between linear and nonlinear regressions was 0.04. The maximum difference in R^2 values of 0.12 was between linear and nonlinear equations for the relationship between SI (SK) and ASPT (UK). German SI had the highest average

Table 7. Coefficients of linear regression equations (*a* – slope, *b* – intercept) for the common scales and the abiotic gradients

Index	SI (AT)		SI (CZ)		SI (DE)		BMWP (PL)		SI (SK)		ASPT (UK)	
Parameter	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*	*a*	*b*
Small siliceous mountain streams												
SI (DE)	0.784	0.212	0.562	0.440	1.000	0	0.319	0.675	0.511	0.465	0.687	0.296
IMI-IC$_{R\text{-}C3}$	0.992	−0.021	0.717	0.261	1.100	−0.138	0.441	0.535	0.688	0.261	0.850	0.102
PE1	−0.845	1.000	−0.567	0.720	−1.089	1.236	−0.450	0.577	−0.542	0.721	−0.976	1.120

Index	DSFI (DK) and DSFI (SE)		GD (DE)		BMWP (PL)		ASPT (SE) and ASPT (UK)	
Parameter	*a*	*b*	*a*	*b*	*a*	*b*	*a*	b
Medium-sized lowland streams								
DSFI	1.000	0.000	0.579	0.356	0.344	0.570	1.349	−0.405
IMI-IC$_{R\text{-}C4}$	0.825	0.154	0.566	0.386	0.357	0.580	1.301	−0.343
HY1	−0.627	0.934	−0.583	0.857	−0.360	0.720	−1.078	1.396

IMI-IC, integrative multimetric index for intercalibration (see text for explanation); PE1, pollution/eutrophication gradient; HY1, hydromorphological gradient.

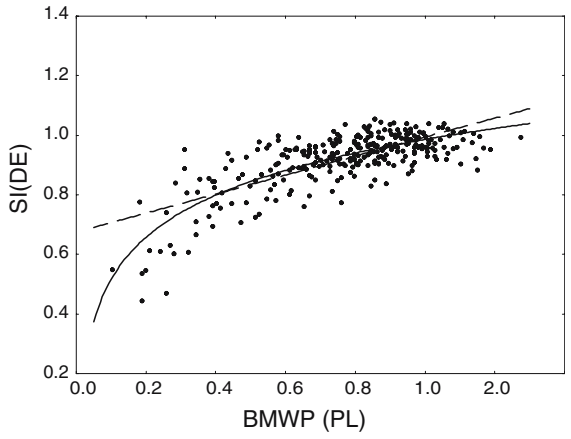

Figure 1. Regression of BMWP (PL) against SI (DE). Both linear ($R^2 = 0.53$, dashed) and nonlinear ($R^2 = 0.63$) regression lines are plotted.

correlation to the other assessment methods ($R^2 = 0.67$). The IMI-IC for this stream type was characterized by coefficients of determination ranging from 0.62 (Slovak SI) to 0.87 (German SI). In Figure 1 regression lines of BMWP (PL) against SI (DE) were exemplarily plotted for linear and nonlinear regression.

R^2 values for regressions of methods for the lowland streams varied between 0.41 (German GD and Polish BMWP) and 0.67 (British and Swedish ASPT, and Danish and Swedish DSFI). In 6 out of 16 correlations, nonlinear regression provided a higher proportion of the variance explained. Mean difference of the linear and nonlinear coefficients of determination was $R^2 = 0.02$ and the maximum difference was $R^2 = 0.06$ (Polish BMWP and British ASPT). DSFI showed the highest mean correlation for the lowland samples ($R^2 = 0.60$). The IMI-IC had coefficients of correlation ranging from 0.73 (Polish BMWP) to 0.90 (Danish and Swedish DSFI). All correlations were significant at $p < 0.05$. Since none of the differences between the linear and nonlinear coefficients of determination were significant, we assumed linear relationships between indices in the following analyses.

Correlation to environmental gradients (PCA)

Index values of the small mountain streams showed the strongest relationship with the PCA gradient reflecting nutrient enrichment and organic pollution. Determination coefficients of this gradient and the assessment methods varied from 0.19 (Slovak SI) to 0.53 (British ASPT). Index values of the lowland streams showed highest correlations with the main hydromorphological gradient that comprised physical features of the river channel, its banks and immediate vicinity,

Table 8. EQR values of the high-good (H/G) and good-moderate (G/M) quality class boundaries transferred into 'common scale'

Class boundary	Common scale	SI (AT)		SI (CZ)		SI (DE)		BMWP (PL)		SI (SK)		ASPT (UK)	
		Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.
Small siliceous mountain streams													
H/G	SI (DE)	0.984	0.008	0.949	0.008	1.016	–	0.846	0.011	0.870	0.010	0.983	0.008
	IMI-$IC_{R\text{-}C3}$	0.955	0.008	0.911	0.008	0.979	0.008	0.771	0.010	0.806	0.011	0.952	0.009
	PE1	0.169	0.023	0.206	0.019	0.130	0.023	0.336	0.022	0.291	0.023	0.144	0.019
G/M	SI (DE)	0.799	0.012	0.895	0.007	0.801	–	0.794	0.016	0.776	0.020	0.907	0.006
	IMI-$IC_{R\text{-}C3}$	0.721	0.012	0.842	0.008	0.743	0.009	0.700	0.015	0.680	0.021	0.858	0.007
	PE1	0.368	0.032	0.262	0.019	0.364	0.025	0.409	0.032	0.391	0.045	0.251	0.014
Class boundary	Common scale	DSFI (DK)		GD (DE)		BMWP (PL)		ASPT (SE)		DSFI (SE)		ASPT (UK)	
		Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.	Boundary value	95% confid.
Medium-sized lowland streams													
H/G	DSFI	1.000	–	1.048	0.012	0.724	0.018	0.809	0.016	0.900	–	0.944	0.025
	IMI-$IC_{R\text{-}C4}$	0.979	0.012	1.061	0.008	0.744	0.012	0.827	0.011	0.897	0.009	0.958	0.017
	HY1	0.307	0.054	0.162	0.021	0.480	0.036	0.426	0.035	0.370	0.042	0.318	0.054
G/M	DSFI	0.714	–	0.875	0.016	0.610	0.016	0.674	0.019	0.800	–	0.795	0.016
	IMI-$IC_{R\text{-}C4}$	0.744	0.008	0.892	0.011	0.628	0.011	0.697	0.012	0.814	0.007	0.814	0.010
	HY1	0.486	0.035	0.335	0.030	0.552	0.034	0.534	0.041	0.432	0.035	0.437	0.034

In addition, the values of the abiotic gradients (PE1, HY1) corresponding to the national class boundaries are displayed. For each value derived by regression the 95% confidence interval is specified. IMI-IC, integrative multimetric index for intercalibration (see text for explanation); PE1, pollution/eutrophication gradient; HY1, hydromorphological gradient.

including information on the degree of impairment. The coefficients of determination ranged between 0.12 (Polish BMWP) and 0.35 (German GD).

Comparison of national quality classes

The comparison of biological quality classes was based on the transformation of boundary values of the assessment methods into a common scale. This allowed for a direct juxtaposition of class boundaries in Table 8.

Small-sized siliceous mountain streams

The common scales used in the comparison procedure for the mountain streams were SI (DE) and IMI-$IC_{R\text{-}C3}$ (multimetric index composed of all national assessment methods). In SI (DE) scale, the high-good boundaries of SI (AT) and ASPT (UK) were similar considering the 95% confidence interval. ASPT (UK) and SI (CZ) showed overlapping good-moderate boundary intervals and thus shared equal class boundaries. The same applied for the group of indices SI (AT), SI (DE), BMWP (PL) and SI (SK). Based on IMI-$IC_{R\text{-}C3}$ the high-good boundaries of SI (AT) and ASPT (UK) shared common intervals. For the good-moderate boundary the comparison showed similar values for SI (AT), BMWP (PL) and SI (SK).

The pollution/eutrophication gradient showed similar pressure between high-good boundaries of SI (AT), SI (CZ), SI (DE), ASPT (UK), and BMWP (PL) and SI (SK). For the good-moderate boundary corresponding levels of chemical impairment were between SI (AT) and SI (DE), SI (SK) and BMWP (PL), and SI (CZ) and ASPT (UK). The average confidence interval amounted to 0.025 units.

Medium-sized, lowland, mixed geology

The DSFI and IMI-$IC_{R\text{-}C4}$ (multimetric index composed of all national assessment methods) were used as common scales for the boundary comparisons of the lowland stream type. Using DSFI as the common scale, none of the national indices showed similar high-good class boundaries but the good-moderate boundaries of DSFI (SE) and ASPT (UK) were corresponding. The average confidence interval amounted to 0.017 DSFI units.

In the IMI-$IC_{R\text{-}C4}$ scale, the high-good boundaries of DSFI (DK) and ASPT (UK) had similar values and the good-moderate boundaries of DSFI (SE) and ASPT (UK) corresponded closely. Confidence intervals showed an average value of 0.011 units.

Boundary comparisons using the hydromorphological gradient were difficult because the large confidence intervals (0.038 units in average) resulted in overlapping boundary ranges. Both good quality boundaries of GD (DE) showed the lowest level of pressure. For the good-moderate boundary, levels of pressure were similar between DSFI (DK), DSFI (SE) and ASPT (UK), and between BMWP (PL) and ASPT (SE).

Discussion

Role of reference conditions in the intercalibration exercise

Within the intercalibration exercise, class boundaries of national assessment methods need to be defined as EQR. The position of each boundary on this relative scale is dependent on (1) the definition of reference conditions and (2) the procedure of setting class boundaries. If the former is not properly dealt with in the intercalibration process, the different nationally defined reference values may strongly impact upon comparability.

In this study we have defined a common reference, which is based on sites in several countries. As a result of this common reference, it was possible to include several methods in the comparison, even if countries have not yet defined reference values for a specific method. A further advantage of common references is that differences in national approaches to define references are avoided. On the other hand, common references are in danger of not adequately accounting for the differences between more specific streams types.

More importantly, countries have applied different procedures to define reference values and quality classification schemes. While this study is restricted to the analysis of national class boundary settings, it must be an objective of the official

intercalibration exercise to overcome differences in the references too.

Relations between assessment methods

In this study, the calculation of national assessment metric values is based on taxa lists derived by application of the standardized STAR–AQEM field and laboratory protocol. Thus, the correlation analyses of index values mainly reveal the numerical relation between these indices and is less biased by differences in field and laboratory procedures. The character of these relations depends on the architecture of the individual indices, e.g. number and indicative value of taxa included in the evaluation, type of abundance information used and the assessment formula. The effect of different national sampling methods on the comparability of taxa lists and metric results as a major constraint of intercalibration is investigated by Friberg et al. (2006). Buffagni et al. (2006) present a practical approach enabling the use, in intercalibration, of datasets derived by the national monitoring programmes.

An additional factor, impacting on the relationships, is the dataset itself, in particular the number of samples, the biogeographical gradient, the types of pressures influencing sampling sites and the range of degradation covered. The different ranges of index values (cf. Table 5) indicate a larger degradation gradient being covered by the lowland dataset. This is, in particular, obvious from the Polish BMWP and British ASPT values, which have been calculated for both datasets.

For the mountain stream data, relationships are strongest between the values of the different Saprobic Indices of Austria, Czech Republic, Germany and Slovak Republic and between the score methods applied in Poland and the United Kingdom. In general, the strength of correlations between the different Saprobic Indices results from similarities in indicator taxa and their indication values (Table 9). For instance, the Austrian and Slovak Saprobic Indices ($R^2 > 0.73$) share the largest number of indicator taxa and are most closely related concerning indicator taxa value and weight. Schmidt-Kloiber et al. (2006) provide a comprehensive analysis of saprobic indicator taxa applied in Europe.

For the lowland stream dataset, BMWP (PL) and ASPT (UK) correlate less strongly ($R^2 < 0.60$), which can be explained by the different taxonomic composition of the lowland dataset compared to that of the mountain streams. The two indices have 66 indicator taxa in common, amounting to a share of 73% (Polish BMWP) and 80% (British ASPT), respectively. BMWP indicator values of the common taxa in the Polish and UK systems are correlated with $R^2 = 0.73$.

Method comparisons of earlier studies show similar results. Based on 232 samples from various lowland and mountain stream types in Germany, Friedrich et al. (1995) found correlations of $R^2 = 0.71$ between ASPT (UK) and a previous version of the German Saprobic Index. The weak relation of ASPT and the Austrian Saprobic Index has already been demonstrated by Stubauer & Moog (2000), who used a large dataset covering all Austrian stream types ($n = 588$; $R^2 = 0.52$). Analyses of Birk & Rolauffs (2003) revealed strong correlations between the Austrian and German Saprobic Indices ($n = 262$; $R^2 = 0.75$).

Several indices revealed higher coefficients of determination when applying a nonlinear fit, in

Table 9. Comparison of the saprobic indicator taxa lists of Austria, Czech Republic, Germany and Slovak Republic: Share of common taxa and coefficients of determination derived from correlation analysis of indicator values and indicator weights

	SI (AT)			SI (CZ)			SI (DE)			SI (SK)		
	Share of common taxa (%)	Indicator value	Indicator weight	Share of common taxa (%)	Indicator value	Indicator weight	Share of common taxa (%)	Indicator value	Indicator weight	Share of common taxa (%)	Indicator value	Indicator weight
SI (AT)	–	1.00	1.00	56	0.64	0.14	72	0.74	0.04	77	0.88	0.53
SI (CZ)	36	0.64	0.14	–	1.00	1.00	54	0.74	0.14	53	0.73	0.31
SI (DE)	35	0.74	0.04	41	0.74	0.14	–	1.00	1.00	41	0.73	0.04
SI (SK)	45	0.88	0.53	48	0.73	0.31	49	0.73	0.04	–	1.00	1.00

particular if BMWP (PL) was involved. This index combines the parameters taxon richness and sensitivity into a single value which may cause the observed relationship. Also, due to the large range of values covered by the method, the nonlinearity of the relationships became evident (cf. Fig. 1). Nevertheless, these difference of the coefficients of determination are not significant. Therefore, the simple model of linear relationship between indices is most appropriate in this example of direct comparison.

Comparison of class boundary values

While earlier intercalibration studies focussed on the comparison of quality class bands (Ghetti & Bonazzi, 1977; Friedrich et al., 1995; Morpurgo, 1996), the Water Framework Directive specifically requires the comparability of the high-good and good-moderate quality class boundaries. Thus, the intercalibration exercise is focussed on the range medium to high biological quality. The original procedure outlined in the Directive is restricted to the use of just a few intercalibration sites, selected because they represent the boundary status between quality classes. However, this approach seems not to be feasible, since sites known to be on class boundaries cannot be selected prior to the intercalibration is completed and those boundaries are defined. Furthermore, the uncertainty of intercalibration results is high if the analysis is based on insufficient data.

Therefore, the primary step, in comparing national class boundary values and best identifying the type and magnitude of the relationship between national assessment methods, should be based on a large number of samples covering the entire quality gradient. In a further step, regression analysis should be used to transform boundary values into other assessment scales. By applying an acceptable level of uncertainty (e.g., confidence interval of 95% derived from regression analysis), ranges of index values can be compared.

The comparison of assessment methods has revealed discrepancies between national classification schemes of more than 25% in particular cases (e.g. high-good boundary of German SI and Polish BMWP translated in German SI scale). The extent of differences between class boundaries is largely dependent on the common scale used for comparison. While class boundaries clearly differ if compared through the German Saprobic Index scale, no differences occur between the same boundaries if compared through a multimetric index. Each method used as a common scale is somewhat related to other assessment methods as expressed by the correlation coefficient and the regression equation.

Based on these findings we recommend using the intercalibration approach described in this paper only for comparison of methods addressing similar components of the biocoenosis, e.g. for methods that are closely related such as ASPT, BMWP and the Saprobic Indices, or methods that are fully compliant with the requirements of the Water Framework Directive (i.e., methods evaluating taxonomic composition and abundance, ratio of disturbance sensitive to insensitive taxa and diversity of the macroinvertebrate community). This principle makes sure that 'like with like' comparisons are applied in intercalibration and minimizes errors in the comparison analysis due to the selection of inappropriate common scales. Furthermore, the relation between assessment methods needs to be carefully evaluated. Nonlinear correlations yielding significantly better fit and smaller confidence intervals are to be favoured over weaker linear relations.

When shall boundaries be considered as different?

Intercalibration encompasses two steps: Firstly, national quality boundaries are compared. If this analysis discovers major differences in classification schemes, they need to be harmonized in a second step. For the first step, we have described a possible procedure to translate boundary values into a common scale, which determines whether or not boundary values are corresponding. According to our results only a few class boundaries are similar, which thus requires the remaining boundaries to be harmonized.

The use of abiotic pressure data in intercalibration allows for additional interpretation of results. Sandin & Hering (2004) applied organic pollution gradients to set intercalibration class boundaries defining a standard level of pollution. We particularly propose to use pressure information for the process of boundary comparison. Figure 2 displays the relative position of the national good-moderate

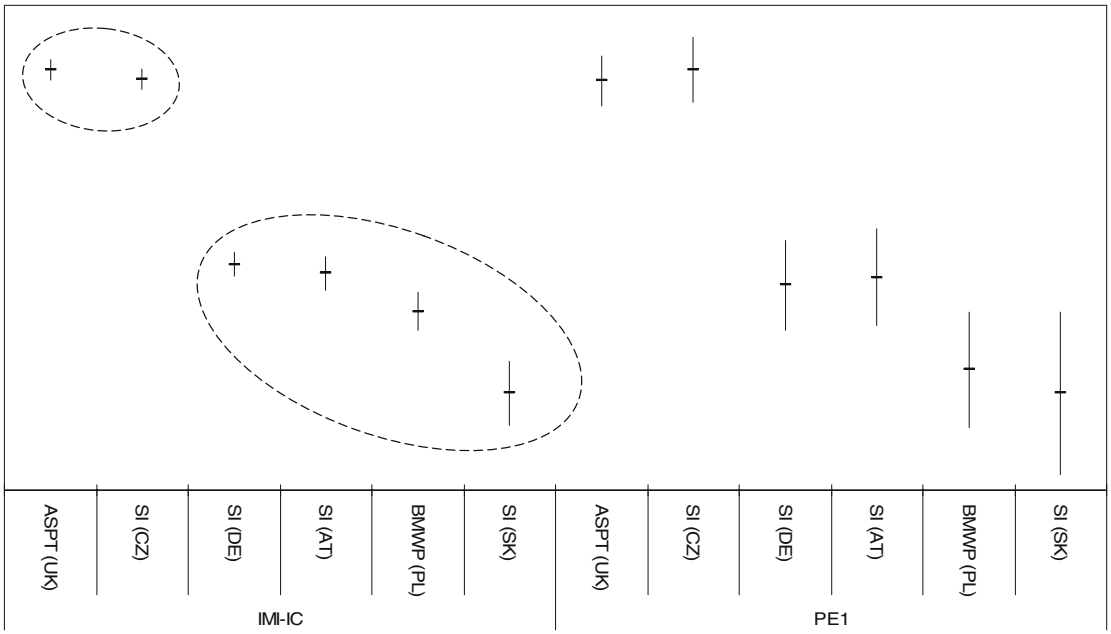

Figure 2. Relative comparison of good-moderate class boundary values (incl. 95% confidence intervals) using IMI-$IC_{R\text{-}C3}$ and corresponding chemical pressure values of the small siliceous mountain streams. Based on the results of the pressure data analysis two groups of similar boundaries are highlighted by dashed circles.

boundaries, including confidence intervals translated into a common biotic scale and an abiotic pressure scale (pollution/eutrophication gradient). Comparisons based on the interpretation of biotic data indicate that four out of six class boundaries are deviating (cf. Table 8), while the consideration of pressure data (Fig. 2) reveals only two groups of boundaries with overlapping pressure intervals. Thus, harmonization is only needed between the two groups of boundaries.

Conclusions

Intercalibration represents a crucial step towards the implementation of a pan-European water quality standard. Besides scientific issues, which we partly addressed in this paper, it holds a major social challenge. Although assessment methods are in general scientifically sound instruments, the element of quality classification is a concession to the practical requirements of decision making in water policy. According to the Water Framework Directive the quality assigned to a site can decide on restoration efforts to be spent or saved. Therefore, intercalibration is of political interest since the definition of quality boundaries sets the environmental standard to be achieved. Furthermore, intercalibration holds an ethical component: By selecting certain quality criteria we agree on a level of anthropogenic degradation acceptable for our freshwater systems. Although beyond its scope science needs to consider all these aspects in the preparation of reasonable and tenable results.

Acknowledgements

STAR was funded by the European Commission, 5th Framework Programme, Energy, Environment and Sustainable Development, Key Action Water, Contract no. EVK1-CT-2001-00089. The authors gratefully acknowledge the fruitful discussions held with Andrea Buffagni, Stefania Erba and Marcello Cazzola on intercalibration topics. Thanks are due to Peter Rolauffs and Christian Feld for their support in data preparation and analysis, and to Jean-Nicolas Beisel, Wouter van

de Bund and Mike Furse for their valuable comments on the manuscript.

References

Alba-Tercedor, J. & A. M. Pujante, 2000. Running-water biomonitoring in Spain: opportunities for a predictive approach. In Wright, J. F., D. W. Sutcliffe & M. T. Furse (eds), Assessing the Biological Quality of Fresh Waters - RIVPACS and Other Techniques. FBA, Ambleside, 207–216.

Armitage, P. D., D. Moss, J. F. Wright & M. T. Furse, 1983. The performance of a new biological water quality score system based on macroinvertebrates over a wide range of unpolluted running-water sites. Water Research 17: 333–347.

Biggs, J., A. Corfield, D. Walker, M. Whitfield & P. Williams, 1996. A preliminary comparison of European methods of biological river water quality assessment. NRA Thames Region Operational Investigation. Environment Agency Technical Report No. 0I/T/001. National Rivers Authority Thames Region, Reading.

Birk, S. & D. Hering, 2002. Waterview web-database: a comprehensive review of European assessment methods for rivers. FBA News 20: 4.

Birk, S. & P. Rolauffs, 2003. A preliminary study comparing the results between the Austrian, Czech and German saprobic systems for the intercalibration of cross-border river basin districts. In Deutsche Gesellschaft für Limnologie (DGL) – Tagungsbericht (Köln). DGL, Werder, 74–79.

Birk, S. & U. Schmedtje, 2005. Towards harmonization of water quality classification in the Danube River Basin: overview of biological assessment methods for running waters. Archiv für Hydrobiologie, Supplement Large Rivers 16: 171–196.

BMWP (Biological Monitoring Working Party)., 1978. Final Report of the Biological Monitoring Working Party: Assessment and presentation of the biological quality of rivers in Great Britain. Department of the Environmental Water Data Unit, London.

Böhmer, J., C. Rawer-Jost, A. Zenker, C. Meier, C. K. Feld, R. Biss & D. Hering, 2004. Assessing streams in Germany with benthic invertebrates: development of a multimetric invertebrate based assessment system. Limnologica 34: 416–432.

Brabec, K., S. Zahradkova, D. Nemejcova, P. Paril, J. Kokes & J. Jarkovsky, 2004. Assessment of organic pollution effect considering differences between lotic and lentic stream habitats. Hydrobiologia 516: 331–346.

Buffagni, A., S. Erba, M. Cazzola, J. Murray-Bligh, H. Soszka & P. Genoni, 2006. The STAR common metrics approach to the WFD intercalibration process: Full application for small, lowland rivers in three European countries. Hydrobiologia 566: 379–399.

CIS WG 2.A Ecological Status (ECOSTAT), 2004. Guidance on the intercalibration process. Agreed version of WG 2.A Ecological Status meeting held 7–8 October 2004 in Ispra. Version 4.1. 14. October 2004. ECOSTAT, Ispra.

CSN 757716., 1998. Water quality, biological analysis, determination of saprobic index. Czech Technical State Standard, Czech Standards Institute, Prague.

Feld, C. K., T. Ofenböck, O. Moog & D. Hering, in prep. Assessing hydromorphological degradation and organic pollution in European rivers – selection of suited metrics derived from benthic macroinvertebrates. Manuscript.

Friberg, N., L. Sandin, M. T. Furse, S. E. Larsen, R. T. Clark & P. Haase, 2006. Comparison of macroinvertebrate sampling methods in Europe. Hydrobiologia 566: 365–378.

Friedrich, G. & V. Herbst, 2004. Eine erneute Revision des Saprobiensystems – weshalb und wozu?. Acta Hydrochimica et Hydrobiologica 32: 61–74.

Friedrich, G., E. Coring & B. Küchenhoff, 1995. Vergleich verschiedener europäischer Untersuchungs- und Bewertungsmethoden für Fließgewässer. Landesumweltamt Nordrhein-Westfalen, Essen.

Furse, M., D. Hering, O. Moog, P. Verdonschot, R. K. Johnson, K. Brabec, K. Gritzalis, A. Buffagni, P. Pinto, N. Friberg, J. Murray-Bligh, J. Kokes, R. Alber, P. Usseglio-Polatera, P. Haase, R. Sweeting, B. Bis, K. Szoszkiewicz, H. Soszka, G. Springe, F. Sporka & I. Krno, 2006. The STAR project: context, objectives and approaches. Hydrobiologia 566: 3–29.

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

Ghetti, P. F. & G. Bonazzi, 1980. Biological water assessment methods: Torrente Parma, Torrente Stirone, Fiume Po. 3rd Technical Seminar. Final Report. Commission of the European Communities, Brussels.

Hering, D., O. Moog, L. Sandin & P. F. M. Verdonschot, 2004. Overview and application of the AQEM assessment system. Hydrobiologia 516: 1–20.

Just, I., F. Schöll & T. Tittitzer, 1998. Versuch einer Harmonisierung nationaler Methoden zur Bewertung der Gewässergüte im Donauraum am Beispiel der Abwässer der Stadt Budapest. Umweltbundesamt, Berlin.

Knoben, R. A. E., C. Roos & M. C. M. van Oirschot, 1995. Biological Assessment Methods for Watercourses. UN/ECE Task Force on Monitoring and Assessment, Lelystad.

Kownacki, A., H. Soszka, D. Kudelska & T. Fleituch, 2004. Bioassessment of Polish rivers based on macroinvertebrates. In Geller, W. et al. (eds), Proceedings of the International 11th Magdeburg Seminar on Waters in Central and Eastern Europe: Assessment, Protection, Management. 18–22 October 2004, UFZ Leipzig, 250–251.

Metcalfe-Smith, J. L., 1994. Biological water-quality assessment of rivers: Use of macroinvertebrate communities. In Calow, P. & G. E. Petts (eds), The Rivers Handbook – Hydrological and Ecological Principles. Blackwell Scientific Publications, Oxford, 144–170.

Moog, O., A. Chovanec, J. Hinteregger & A. Römer, 1999. Richtlinie zur Bestimmung der saprobiologischen Gewässergüte von Fliessgewässern. Bundesministerium für Land- und Forstwirtschaft, Wien.

Morpurgo, M., 1996. Confronto fra Indice Saprobico (Friedrich e DIN, 1990) e Indice Biotico Esteso (Ghetti e IRSA, 1995). Biologia Ambientale 14: 30–36.

National Rivers Authority, 1994. The Quality of Rivers and Canals in England and Wales (1990 to 1992) as Assessed by

a New General Quality Assessment Scheme. HMSO, London.

Nixon, S. C., C. P. Mainstone, T. Moth Iversen, P. Kristensen, E. Jeppesen, N. Friberg, E. Papathanassiou, A. Jensen & F. Pedersen, 1996. The harmonized monitoring and classification of ecological quality of surface waters in the European Union. Final Report. European Commission Directorate General XI, Brussels.

Rico, E., A. Rallo, M. A. Sevillano & M. L. Arretxe, 1992. Comparison of several biological indices based on river macroinvertebrate benthic community for assessment of running water quality. Annales de Limnologie 28: 147–156.

Rolauffs, P., D. Hering, M. Sommerhäuser, S. Rödiger & S. Jähnig, 2003. Entwicklung eines leitbildorientierten Saprobienindexes für die biologische Fließgewässerbewertung. Umweltbundesamt, Berlin.

Sandin, L. & D. Hering, 2004. Comparing macroinvertebrate indices to detect organic pollution across Europe: a contribution to the EC water framework directive intercalibration. Hydrobiologia 516: 55–68.

Schmidt-Kloiber, A., W. Graf, A. Lorenz & O. Moog, 2006. The AQEM/STAR taxalist – a pan-European macroinvertebrate ecological database and taxa inventory. Hydrobiologia 566: 325–342.

Skriver, J., N. Friberg & J. Kirkegaard, 2000. Biological assessment of running waters in Denmark: introduction of the Danish stream fauna index (DSFI). Verhandlungen der Internationalen Vereinigung für theoretische und angewandte Limnologie 27: 1822–1830.

STN (Slovenská Technická Norma) 83 0532-1 to 8, 1978/79. Biologický rozbor povrchovej vody. (Biological analysis of surface water quality.) Slovak Standardisation Institute, Bratislava.

Stubauer, I. & O. Moog, 2000. Taxonomic sufficiency versus need for information – comments based on Austrian experience in biological water quality monitoring. Internationale Vereinigung für theoretische und angewandte Limnologie: Verhandlungen 27: 1–5.

Swedish Environmental Protection Agency, 2000. Environmental quality criteria: lakes and watercourses. Swedish Environmental Protection Agency, Stockholm.

SYSTAT Software Inc., 2002. TableCurve 2D – Version 5.01. SSI, Richmond CA.

Tittizer, T., 1976. Comparative study of biological–ecological water assessment methods. Practical demonstration on the river Main. 2–6 June, 1975 (summary report). In Amavis, R.-J. (ed.) Principles and Methods for Determining Ecological Criteria on Hydrobiocoenosis: Proceedings of the European Scientific Colloquium Luxembourg, Nov. 1975. Pergamon Press, Oxford, 403–463.

Woodiwiss, F. S., 1978. Comparative study of biological–ecological water quality assessment methods. Second practical demonstration. Summary Report. Commission of the European Union, Brussels.