A case study of three techniques for assessing comparability of national methods of ecological classification

Nigel Willby, University of Stirling, UK

Sebastian Birk, University of Duisburg-Essen, DE

August 2010.

Summary

1. This report considers three possible approaches for assessing the comparability of national ecological classification methods within the intercalibration process.

- **Technique 1**: Indirect approach for boundary comparison via the regression of a **common metric** against each national EQR. This approach follows the original Phase 1 Option 2 but includes a refinement to convert results to class equivalents. This ensures that outputs can be inter-compared between all Techniques.
- **Technique 2**: Direct approach for boundary comparison via the regression of a **pseudo-common metric**, formed from the average of an independent set of countries, against the national EQR of the remaining country. This approach was presented in Annex V of the Phase 2 Guidance document.
- **Technique 3**: Direct approach for boundary comparison via **multiple pairwise comparisons** of EQRs across a population of commonly assessed sites. This is a refinement of the alternative approach offered by NL.

We apply these Techniques to a single common dataset from the Central Baltic GIG river macrophyte IC group, in order to test their assessment of the comparability of five national methods of ecological classification.

2. **Benchmark normalisation** is an essential precursor to the comparison of national classifications. It serves to homogenise common datasets, thus compensating for typological differences between MS which may otherwise influence the comparability of their classifications.

3. There are two components to assessing comparability; **bias** (the direction and magnitude of deviation of a national class boundary from the harmonisation guideline – the global mean position of all methods at a given boundary – expressed as class equivalents; this reflects the level of ambition of different methods), and **class agreement** (a measure of the confidence that two or more national methods will classify any given site the same). Several related metrics are used to measure class agreement, depending on the Technique employed for comparing classifications (the average absolute class difference, measured across all participating countries; the proportion of classifications differing by <0.5 classes; the multi-rater kappa coefficient).

4. Prior to comparing classifications it must be demonstrated that methods are significantly correlated with each other or with a common metric (**feasibility criteria**). To ensure a minimum acceptable level of class agreement the size of the correlation rather than its probability is critical. A minimum correlation coefficient of r = 0.5 - 0.6 will be statistically significant across compliant IC datasets regardless of their size (n > 14) and should ensure an adequate level of relatedness of methods to satisfy class agreement criteria. However, a

much higher correlation (r = 0.8 - 0.9) will be needed in order to state that methods have a high probability of classifying a site the same.

5. Technique 1 requires an ecologically meaningful common metric to act as a yardstick for the comparison of methods. If sites can be commonly assessed by all national methods it will always be preferable to apply Technique 2 or 3. A biological common metric may, however, be valuable for expressing results in a more easily understood format, or for demonstrating compliance with the normative definitions.

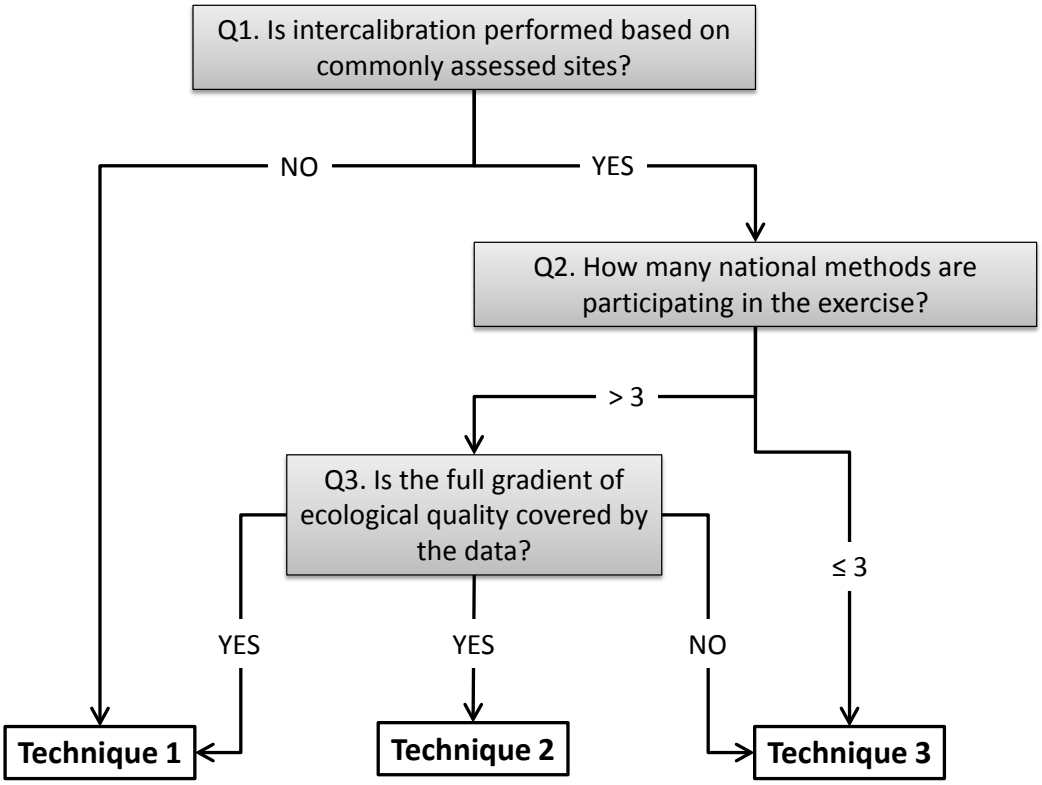
6. In regression-based approaches (Technique 1 and 2) it is necessary to convert the deviation in a national method, relative to the global mean of all methods, into **class equivalents**. Since the deviation is expressed in terms of a common metric it is necessary to know the turnover in the common metric between values equivalent to national EQRs of 1 and 0. This defines the length of the ecological quality gradient in common metric terms. Subdividing this gradient by five provides an estimate of average class width which can then be used to scale deviation as class equivalent units. This approach requires an assumption that average class width across the full quality gradient is a reasonable approximation of the width of High and Good classes in national methods.

7. Regression-based approaches are not amenable to direct measures of class agreement, although the average of the absolute residual (absolute difference between observed and predicted values) appears to provide a reasonable estimate of average class difference.

8. Technique 3, (pairwise comparisons of national EQR values between each method and all other methods across all commonly assessed sites), provides a useful alternative to Technique 2 when the number of participating countries is small (<4), or the ecological gradient covered by a common dataset is constrained. Technique 3 allows an empirical measure of bias and class agreement between all methods. It seems to offer a universal solution to the intercalibration of those national methods that meet a minimum standard of relatedness. To implement Technique 3 it is necessary to first map all methods onto a class boundary system, using a **piecewise linear transformation**, where classes are standardised to a width of 0.2 EQR units.

9. Techniques 1 and 2 provide a simple means of establishing the bias of individual methods relative to the global mean of all methods. By inverting the regression model between the common metric and the national EQR it is straightforward to determine where national class boundaries should be positioned in order to secure an acceptable level of bias. In Technique 3 class boundaries are adjusted by an iterative process, changing the class boundary of the most biased country by 0.01 EQR units each step until all countries obtain an acceptable level of bias.

How to select the appropriate Technique to compare national methods



10. Investigation of these three Techniques supports the use of a harmonisation band defined by limits of **bias equivalent to ±0.25 classes**. National class boundaries with a bias inside this band can be considered to be harmonised, with all methods then displaying a comparable level of ambition. All Techniques investigated here produce similar results in terms of identifying those national methods displaying the greatest bias at HG or GM boundaries. The results do not support the use of an alternative approach to eliminate bias entirely through the synchronisation of class boundaries to the global mean. Fully synchronised boundaries differed between methods and resulted in negligible improvement in class agreement compared to that achieved using harmonised class boundaries. There is a risk that, unless common datasets are very large, synchronised class boundaries will be dataset-specific, as well as ignoring sources of uncertainty introduced earlier in the process of applying national methods to data of other nationalities.

11. Class agreement is primarily a product of the relatedness of national classifications. It will be relatively unresponsive to harmonisation of class boundaries unless these display extreme levels of bias (i.e. that is large relative to the random error in the relationship between methods or between methods and a common metric). Class agreement will be most sensitive to small changes in the position of class boundaries when methods are highly correlated from the outset.

12. The use of a **multi-rater kappa coefficient** in Techniques 2 and 3 provides a useful additional and well established measure of class agreement that can be employed to check agreement in relation to the classification of sites as <Good or ≥Good, and <High or High. However, it is preferable to reduce bias to acceptable levels by harmonising the position of class boundaries, rather than to manipulate class boundaries to optimise the value of kappa. In general the suitability of kappa for comparing class agreement between exercises will be influenced by dataset size and the relative proportions of sites above and below the boundary being tested.

13. All the three Techniques investigated here produce similar results in terms of measures of class agreement, and in identifying those national methods that display the greatest bias. The presentation of bias and class agreement in terms of class equivalents ensures a standard framework for evaluating the results of comparing national methods.

Matrix for judging acceptability of comparisons based on levels of class bias and class agreement

		Class agreement		
mean av. abs. class diff*		<0.5	<1.0	>1.0
mean % ±0.5 classes$^{\phi}$		>50	30-50	<30
kappa coefficient†		>0.6	0.6-0.4	<0.4
Bias #	< ±0.25 classes			
	< ±0.5 classes			
	> ±0.5 classes			

* Measure applicable to Techniques 1, 2 and 3. $^{\bullet}$ Measure applicable only in Technique 3. † Measure applicable in Techniques 2 and 3. $^{\#}$ Refers to the national method exhibiting the greatest bias

14. A matrix is provided for judging whether national methods can be considered comparable. This relies on desirable, and minimum acceptable standards for bias and class agreement. Bias of <0.25 classes is desirable and ensures that methods show very similar levels of ambition. Bias of up to 0.5 classes may be acceptable under certain circumstances. Bias exceeding 0.5 classes must be unacceptable under all classes because it implies that a method is closer to a different boundary than the one being evaluated (or that two countries can differ by >1 class at the same boundary). Class agreement criteria of <0.5 classes average absolute class difference (or >50% of comparisons differing by <0.5 classes, or kappa value >0.6) are desirable since they indicate a good-high probability that different

methods will classify a site the same. Class agreement of <1.0 classes average absolute class difference (or >30% of comparisons differing by <0.5 classes, or kappa value >0.4) represent acceptable limits of class agreement where different methods have a moderate to good probability of classifying a site the same. Below these limits class agreement will differ by more than a class width and it can be inferred that classifications have a low probability of classifying a site the same.

15. Those exercises where national methods meet desirable standards for bias **and** class agreement can be considered to pass (i.e. successfully intercalibrate). When an exercise cannot meet desirable standards for both bias and class agreement but can achieve acceptable standards it could be considered to achieve a qualified pass provided the participating countries can provide an adequate explanation for the greater class bias or lower class agreement. Exercises in which at least one method exceeds acceptable levels of bias, or the average class agreement across participating countries is poor, would be considered to have failed. Boundary harmonisation may be sufficient to achieve a pass by ensuring an acceptable level of bias. Improvements to national methods themselves may be required if class agreement is persistently poor. However, the feasibility criteria concerning the relatedness of methods should ensure that failures based on class agreement are rare.

Table of Contents

1. Introduction

The WFD requires that national methods for ecological assessment are intercalibrated to ensure that class boundaries display consistency with normative definitions and are comparable between Member States. As part of the first phase of the intercalibration process (2004-2008) three Options were developed to assist in the comparison of national classifications. A number of exercises across a range of water body types and Biological Quality Elements were successfully completed based on these Options. However, due to their specific characteristics it has proved problematic to apply these Options successfully in some IC exercises. Moreover, the outputs of these different Options, in terms of measures of comparability and agreement between classifications, are difficult to compare directly.

Annex V of the Guidance to support the second phase of the IC process (2009-2012) contained details of a technique to allow direct comparison of national EQRs (Option 3) following a similar approach to that used in Option 2 where national methods are compared indirectly via regression against a common metric (Willby & Birk 2010a). Shortly before this the IC Fish Group published a proposal to allow optimisation of class agreement between methods via direct comparison of classifications (Pont & Delaigue, 2010). Subsequent to our proposals a number of MS provided specific comments, suggestions for refinements to the proposed approach (e.g. Horký, 2010, Pešta & Horký, 2010), and in the case of NL, an alternative approach (van den Berg, 2010). Annex V also emphasised the need for benchmark normalisation of national classifications before attempting tests of comparability, in order to ensure that the influence of typological differences between participating countries on classification results was minimised. The current document takes the relevant available techniques and applies them to a single common dataset from the Central Baltic GIG river macrophyte IC group, in order to test their assessment of the comparability of five national methods of ecological classification.

Specifically we focus on three basic techniques, viz

- Technique 1: Indirect approach for boundary comparison via the regression of a common metric against each national EQR. This approach follows the original Phase 1 Option 2 but includes a refinement to ensure that results are presented as class equivalents.

- Technique 2: Direct approach for boundary comparison via the regression of a pseudo-common metric, formed from the average of an independent set of countries, against the national EQR of the remaining country. This approach was presented in Annex V of the Phase 2 Guidance document.

- Technique 3: Direct approach for boundary comparison via multiple pairwise comparisons of EQRs across a population of commonly assessed sites. This is a refinement of the alternative approach offered by NL (van den Berg, 2010) and builds on our recent theoretical evaluation of this approach (Willby & Birk, 2010b).

2. Dataset

This document considers a dataset from the CB GIG river macrophyte intercalibration exercise as a template for comparing national classification methods using the three approaches outlined above. In this analysis we use a set of data from RC-3 rivers (small upland siliceous streams). A total of 215 sites surveyed in a more or less standard manner (100 m reach, visual assessment of cover of aquatic taxa present) were assessed by each of five different national methods, from one each of AT, DE, FR, UK and WL. The available data also included sites in two CB GIG countries, CZ and LU, who did not have completed methods available at the time of analysis. Prior to national classifications being applied the data was homogenised by converting different scales for assessment of plant cover to a common international system. Data on filamentous algae were not universally available in all national datasets, and were therefore removed. The composition of the commonly assessed sites is summarised in Table 1.

Table 1. Composition of commonly assessed RC-3 sites used in the analysis

MS	Benchmark sites	Non-benchmark sites	Total	Biotype
AT	2	18	20	Continental
CZ	5	8	13	Continental
DE	2	64	66	Continental
FR	26	24	50	Continental
LU	0	9	9	Continental
WL	5	30	35	Intermediate
UK	7	15	22	Atlantic

3. Benchmark normalisation

Within the benchmark sites in this dataset two clear biotypes are represented; an 'Atlantic type' confined to the UK and WL, and a 'Continental type', occurring in all countries excluding the UK (see Birk & Willby, 2010a). The differences between these biotypes can be evaluated using a biological common metric. In this case we refer to a biological common metric, mICM, that was developed using the correlations between individual macrophyte taxa and a synthetic pressure gradient based on the global average of national EQRs from a subset of sites in the common dataset (Birk & Willby, 2010b). The median common metric value differs significantly between benchmark sites belonging to these two biotypes (mICM of 0.55 and 0.24 for atlantic and continental subtypes respectively, Mann Whitney test, $U = 29.0$, $p = <0.001$). Due to this difference, and the fact that both subtypes are not distributed across all countries being intercalibrated, benchmark normalisation is necessary before methods can be compared. If there was no significant difference in mICM value between biotypes, or all biotypes occurred in all countries we would proceed without any further

adjustments. For the purposes of intercalibration it is convenient to express the benchmark normalised common metric in the form of an EQR, although it is important to recognise that this is not an EQR in the strict sense (i.e. a measure of deviation from reference condition) if a benchmark lower than reference condition is applied. In this example, the mICM was converted to an EQR by first subtracting the minimum observed mICM value (-0.72) from both observed and benchmark values. Thus:

mICM EQR = (Obs mICM - -0.72)/(mICM benchmark - -0.72)

Table 2 summarises the information on benchmark sites in relation to the national EQR values. Three countries are specific to the continental benchmark sites; DE, FR and AT. These countries tend to rate atlantic benchmark sites more highly than the continental sites typical of their own countries. The UK is specific to atlantic benchmark sites and thus has a much lower view of benchmark continental sites. Without making any adjustment for differences in benchmark type, the UK classification of continental sites would always appear unduly harsh whilst the classification of atlantic sites by continental countries would appear generous. Differences in classification evident from a comparison of results would then be due primarily to typological differences rather than to differences in the placement of class boundaries. Under these circumstances analysis of the comparability of methods could be misleading. Thus the purpose of benchmark normalisation is to homogenise a common dataset so that it can be uniformly assessed by all national methods regardless of the nationality of the sites assessed. Wallonia (WL) presents a special case since it harbours examples of both biotypes. One cannot assume that the relative proportion of examples of the two biotypes from WL is typical of WL as a whole. Similarly we cannot assume that the relative proportion of these two biotypes across all the contributing countries in the common dataset is typical of WL. Therefore the benchmark for WL is provided by a simple unweighted mean of the continental and atlantic benchmarks.

Table 2. Median mICM and national EQR values assigned to benchmark sites belonging to different biotypes

		national EQR				
	mICM	**DE**	**UK**	**FR**	**WL**	**AT**
Atlantic benchmark sites	0.55	1.00	1.05 *	1.28	1.25	0.95
Continental benchmark sites	0.24	0.83 *	0.80	1.03 *	1.00	0.93 *
Mean	0.4	0.92	0.93	1.16	1.12 *	0.94

The benchmark normalisation of national EQR values themselves is required for all three techniques discussed here. This is achieved by dividing each national EQR by its benchmark depending on the origin of the data being assessed. Thus, for DE, for example, the face value EQR assigned to a site is divided by 1 (i.e. is unchanged) when assessing Atlantic sites (in this case any site in the UK), is divided by 0.83 when assessing continental sites (those in DE, FR and AT), and is divided by 0.92

when assessing sites in WL. Similarly in the case of UK, the face value EQR is divided by 1.05 when assessing sites in the UK, is divided by 0.8 when assessing sites in those countries supporting only the continental biotype (DE, FR, AT), and is divided by 0.93 when assessing sites in WL which contains examples of both biotypes. In terms of classification of sites based on their benchmark normalised EQRs the class boundaries of each country, normalised by the benchmark group to which that country belongs, must be applied (Table 3). Thus, in the FR classification, on the benchmark normalised scale, the GM boundary is converted from 0.72 to 0.72/1.03 = 0.702 based on the continental benchmark for FR. All sites in the common dataset with a benchmark normalised EQR of <0.702 according to the FR method would then be classified as ≤Moderate, regardless of their national origin.

Table 3. Original class boundaries on national scale and equivalent values after benchmark normalisation.

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries in national method	Ref	0.875[1]	1.000	1.000	1.000	0.938[1]
	HG	0.750	0.800	0.855	0.925	0.875
	GM	0.550	0.600	0.720	0.610	0.625
benchmark normalised class boundaries	Ref	1.050	0.952	0.975	0.893	1.011
	HG	0.900	0.762	0.833	0.826	0.943
	GM	0.660	0.571	0.702	0.544	0.674

1. Based on median of EQR of sites in common dataset nationally assessed as being at High status

4. Technique 1: Indirect approach for boundary comparison

4.1 Mapping methods onto a common metric

A common dataset contains biological data from a given IC type to which countries are required to apply their classifications to the best of their ability. When data collected by one method cannot be assessed by other methods, e.g. due to the size of sampling unit or method of sampling, and vice versa, it is necessary to construct a common metric to act as a mutual yardstick for comparison.

The Option 2 method of intercalibration relies on establishing a regression model linking the common metric to each national EQR, and then comparing the values on the common metric scale, equivalent to each of the national class boundaries. If different national methods are mutually incompatible the regression analysis itself can be accomplished at the level of independent national datasets rather than through analysis of the potentially more limited amounts of national data in a common dataset. In the present dataset the volume of data provided by any one country is relatively small and it would clearly be desirable to base a regression model linking the national EQR to the common metric on a larger volume of data than was provided by each MS (Table 2). However, a common dataset, in which all countries have attempted to apply their methods, even if in a rather

compromised form, will prove invaluable for the initial process of developing a biological common metric that displays a good relationship with all national methods.

In the present example, to compare national class boundaries by regression against a common metric for independent sets of data we would select data from each country in turn and regress the benchmark normalised common metric, mICM EQR, against the national EQR. Thus, for DE, we would regress the mICM, normalised using a benchmark of 0.24 (Table 2), against the DE national EQR for data collected in DE and then assess where, on the mICM EQR scale, the HG and GM boundaries for DE lie. The process would then be repeated for UK, this time normalising the mICM value using a benchmark of 0.55 (Table 2), and regressing the benchmark normalised mICM against the UK national EQR for data collected in UK. If several countries with the same common biotype can satisfactorily assess each other's data this could be used to increase the volume of data considered by each national method. Thus, for example, if the three countries belonging to the continental benchmark group, DE, FR and AT, can all apply their methods to data from all three countries, it would be preferable to evaluate the DE method in relation to sites in DE, FR and AT, rather than against sites from DE alone, especially if the availability of data from any one country is limited.

In the present example, we are dealing with a situation where all countries can apply their methods reciprocally to data from other countries. This means that a direct comparison of EQR values of commonly assessed cases is possible. For the purposes of argument we will consider that these five countries have opted to compare their methods indirectly via a biological common metric. In principle, in this situation, it will always be preferable to use a direct method of comparison, since developing an appropriate common metric requires additional effort and it is unlikely that such a metric will be able to adequately capture the key aspects of all national methods. However, it may prove easier in some situations to visualise and communicate intercalibration results via a common metric, and where necessary, to convince an individual MS of the need to modify a method or its associated class boundaries.

In the present situation there are several different benchmarks *within* a common dataset since each MS can apply its method to sites of all other MSs. To allow boundaries to be compared indirectly it is necessary to benchmark both the common metric *and* the national EQRs. If we consider the UK method, in the absence of any benchmarking there is a clear differentiation between continental and atlantic benchmark sites, as perceived by the UK (Table 2) on both x and y axes (Figure 1a). Benchmarking only the common metric itself (partial benchmarking) will serve to increase the model error due to increased divergence on the x axis of the biotypes that are represented (Figure 1 b). In such cases a global regression model will not adequately represent the different biotypes present, as well as being highly sensitive to differences in the volume of data contributed by different countries, or belonging to different benchmark groups. Thus in the present case, 74% of the cases assessed originate from the continental group of countries (DE, FR and AT), while only 11% belong to the atlantic group. We could follow a partial benchmarking approach, as described above, to compare, for example, how UK assessed atlantic rivers with how DE assessed continental rivers, but one could not satisfactorily assess the comparative performance of two different methods across several different biotypes simultaneously. This requires benchmarking of both the common metric and the national EQRs (Figure 1 c).

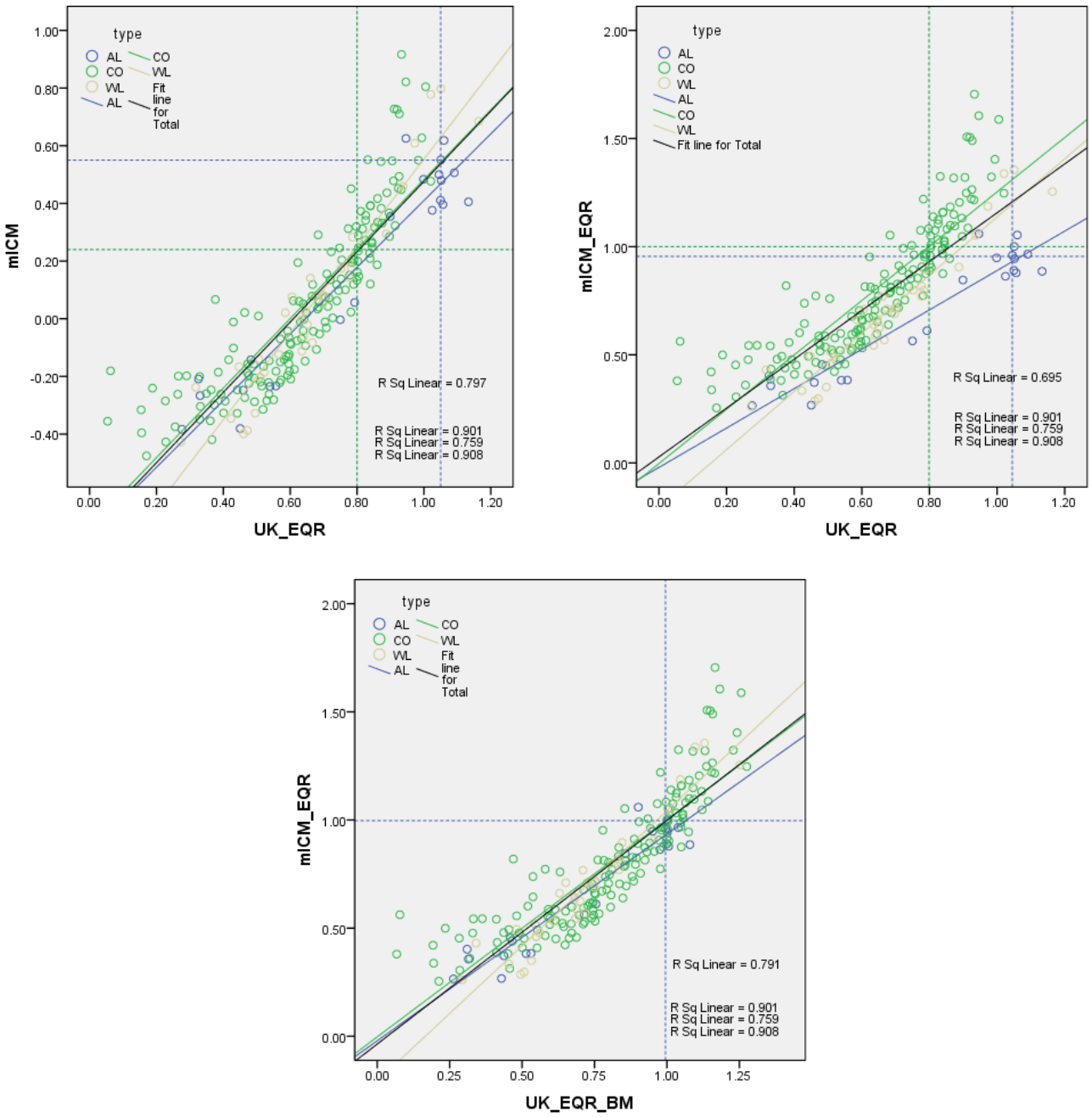


Figure 1. Effect of benchmarking common metric (mICM) and national metric (UK_EQR) using the UK method as an example. Top left (a) no benchmarking; Top right (b) benchmarking only of mICM increases global model error. Bottom (c) benchmarking of both mICM and national EQR eliminates biotype separation on x and y axes allowing classification of all data to be assessed via a single model. Green and blue dashed lines represent benchmarks of continental and atlantic biotypes respectively on mICM (y) or UK_EQR (x) scales.

The need to benchmark both axes in these circumstances was not reported in Annex V of the recent IC Guidance document, although it is probably a relatively special case to use Technique 1 in a

situation such as that presented here; direct comparison of classifications via Technique 2 or 3 would normally be used in preference when there is a dataset of commonly assessed sites.

In Technique 1 the mICM EQR serves as the yardstick against which national class boundaries are to be compared. Thus, we need to establish where on the benchmarked mICM (mICM EQR) scale the benchmarked national EQR, EQR_B, values lie which correspond to the HG and GM boundaries in Table 3. Figure 2 displays the best fit between each national EQR_B and the mICM EQR. In all cases an exponential model provided a significantly improved fit relative to a simple linear model indicating that most of the variation in the mICM EQR occurs over the upper three quality classes. In this example the model fit is very good in all cases. If it was relatively poor an alternative modelling technique such as mixed linear or non-linear regression could be trialled (Horký, 2010). The mICM EQR versus benchmarked national EQR models are summarised in Table 4.

Table 4. Summary of mICM EQR models for each of the five national methods being compared

	Model Summary				Parameter Estimates	
	R Square	F	Sig.	SE	Constant	b1
DE	.735	598	.000	0.165	.220	1.399
UK	.822	995	.000	0.114	.234	1.426
FR	.848	1205	.000	0.132	.191	1.575
WL	.849	1210	.000	0.132	.191	1.577
AT	.778	757	.000	0.158	.143	1.886

Note: the model form is exponential so mICM EQR = Constant * exp (b1 * x) where x = the benchmarked national EQR

A pre-requisite at this stage is that all national methods being compared should be significantly correlated with the common metric. Clearly this requirement is easily fulfilled in the present example. If it were not, uncorrelated methods would need to be removed before continuing with the comparison. In general it will not be desirable for the range of correlation with the common metric to differ strongly between methods. Any methods with a weak correlation with the common metric should be investigated, with the object of obtaining a satisfactory correlation by, for example, excluding or reweighting component metrics.

In reality it is difficult to set a suitable probability threshold for the significance of the correlation coefficient since IC datasets vary greatly in size depending, for example, on BQE and water body type, while the ease of obtaining a highly significant correlation ($p < 0.001$) increases with sample size. In fact the correlation coefficient itself, rather than its probability, is more critical as far as measuring agreement between methods is concerned, because the coefficient reflects the ratio of covariance between a method and the common metric to the product of their standard deviations. Thus setting a minimum value for r of 0.5-0.6 (which will be significant at $p = 0.05$ in even the very smallest IC datasets) is arguably more useful than specifying a fixed threshold of p, such as 0.001,

which would potentially require a very (unnecessarily) high value of r (>0.8) in very small datasets and allow an unduly relaxed value of r (<0.3) in very large datasets.

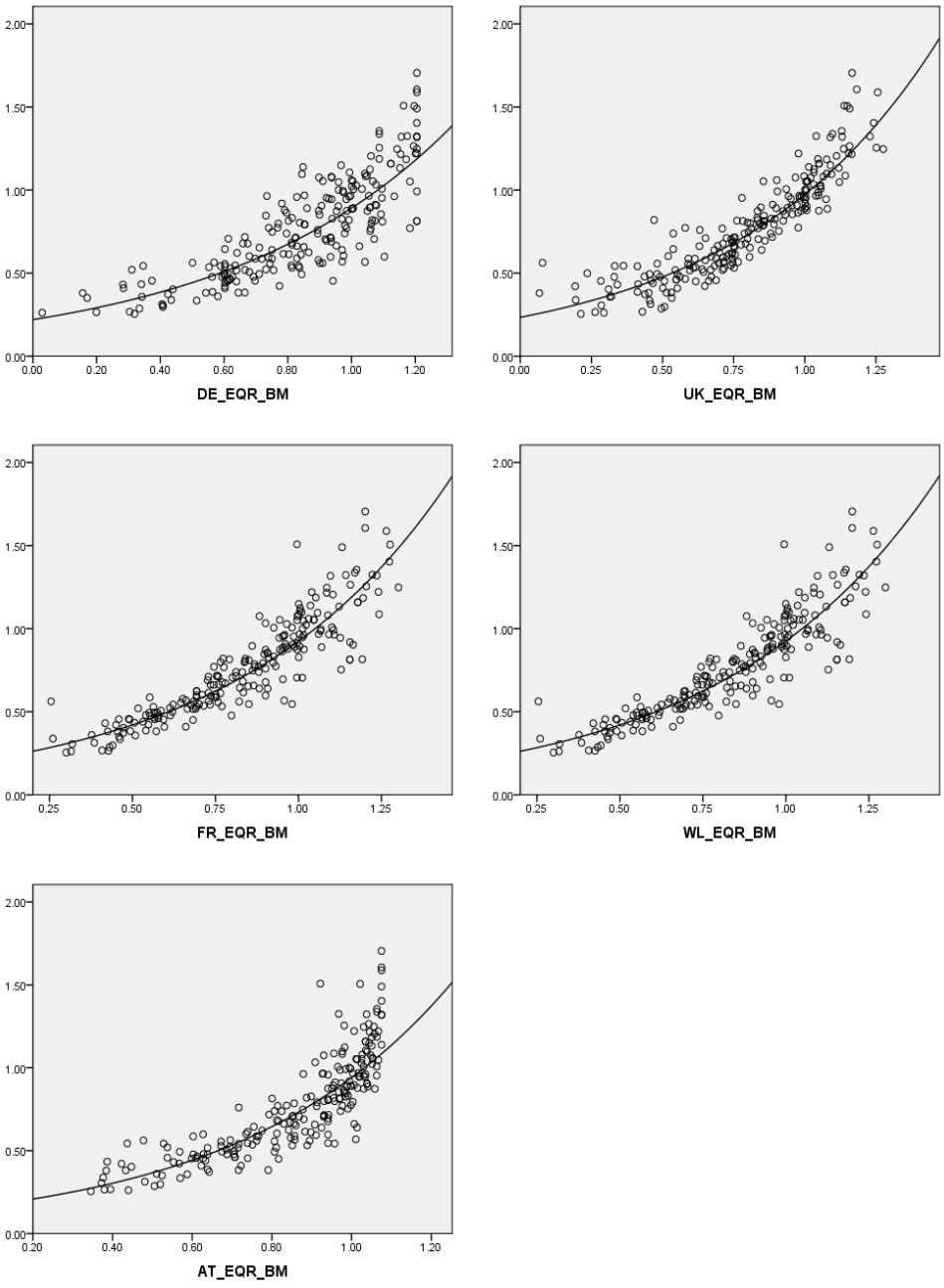


Figure 2. Relationships between mICM_EQR (y) and national benchmarked EQR_B (x) for participating countries.

4.2 Measuring bias between methods

Using the linear model for each method (Table 4) derive where its benchmark normalised HG and GM boundary values should lie on the mICM EQR scale. Thus, for the FR method, for example, the benchmark normalised HG boundary lies at 0.833 (from Table 3). Therefore on the mICM EQR scale this is equivalent to

$$= 0.191*\exp(1.575* EQR_{FR\_B}) = 0.709$$

The values on the mICM EQR scale are then calculated in the same way for each method at the HG and GM boundaries (Table 5). The global average of the values for all methods at a given boundary represents the guideline value for that boundary against which the bias of individual methods is then to be measured. Thus, calculating the HG boundary positions for all methods on the mICM EQR scale produces a global mean of 0.746. The equivalent value at the GM boundary is 0.524. The bias of the DE method at the HG boundary is then equal to 0.776-0.746 = 0.03.

Table 5. Predicted mICM EQR values at the benchmark normalised HG and GM boundaries for five national methods and their associated bias

	DE	UK	FR	WL	AT	mean
HG boundary	0.776	0.693	0.709	0.702	0.847	0.746
GM boundary	0.555	0.528	0.577	0.451	0.510	0.524
HG bias	0.030	-0.052	-0.036	-0.043	0.101	
GM bias	0.031	0.004	0.053	-0.073	-0.014	
HG bias as class eq	0.151	-0.275	-0.234	-0.262	0.449	
GM bias as class eq	0.152	0.023	0.341	-0.445	-0.064	

Note: Cells highlighted in blue indicate potential boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

We have discussed previously (Willby & Birk, 2010ab) the problems of converting bias around a global mean class boundary value into class equivalents. This conversion is important because it allows direct comparison of bias calculated according to the different techniques. One of the main problems is that the width of the classes between the boundaries of interest is unknown and cannot simply be assumed to be 0.2. Therefore deviation from the global mean at a national level cannot be readily transformed to class equivalents. The second problem is that the real meaning of the bias of a method relative to the global mean is contingent on the slope of the regression between the national method and the common metric. The regression slope is critical because it governs the rate of turnover in common metric units in relation to the national EQR.

To convert the deviation on the mICM EQR scale to class equivalents we require two pieces of information. The first is the value on the mICM EQR scale that equates to the benchmarked national reference value used to estimate the national EQR. Provided national EQRs are expressed relative to a value of 1 the benchmarked reference value will simply be 1 divided by the appropriate benchmark. Thus, for the UK, the benchmarked reference value is 1/1.05 = 0.95. Applying the model in Table 4 a benchmarked national EQR value of 0.95 for UK is found to be equivalent to a value of 0.909 on the mICM EQR scale. This anchors the upper part of the mICM EQR scale. The second piece of information required is the lower anchor on the mICM EQR scale. In a simple linear regression we would just take the intercept from the regression model since this equates to a benchmarked national EQR of zero. However, in the present case, due to the exponential form of the models, this approach is likely to result in an underestimation of average class widths in the upper part of the quality gradient in which are interested since there is little variation in the values of mICM EQR in the lower part of the ecological quality gradient. In this instance it is preferable to use the position of the GM boundary as an anchor for deriving class width. Thus for the UK method the effective range of ecological quality on the mICM EQR scale across the upper two classes is from 0.528 to 0.909, a range of 0.381 units. Therefore, an average class width on the mICM EQR scale = 0.381/2 = 0.191 units. Table 6 provides the class widths on the mICM EQR scale for all five methods. The overall average class width across all methods on the mICM EQR scale is 0.188 which compares favourably to the mean width calculated using the approach discussed in Technique 2 (0.173).

Table 6. Determining class equivalent units on the mICM EQR scale to allow estimates of bias around global mean in terms of class equivalent units

mICM EQR scale	**DE**	**UK**	**FR**	**WL**	**AT**
benchmarked reference value	0.957	0.909	0.887	0.780	0.962
Good - Mod. boundary	0.555	0.528	0.577	0.451	0.510
gradient width over upper 2 classes	0.402	0.381	0.310	0.330	0.452
av. class width	0.201	0.191	0.155	0.165	0.226

Table 5 summarises the predicted value for the HG and GM boundary for each method on the mICM EQR scale, using the linear models detailed in Table 4 based on the original (i.e. pre-harmonisation) national class boundaries. The bias relative to the global mean is then converted to class equivalents. Thus, for AT, for example, the HG boundary translates to a mICM EQR value of 0.847, using the model in Table 4, and the class boundaries in Table 3. Thus, relative to the global mean there is a bias of 0.847 – 0.746 = +0.101 mICM EQR units at the HG boundary for the AT method. The average class width on the mICM EQR scale for AT is 0.226 (Table 6).Therefore, in class equivalent terms, the bias for the method of AT at the HG boundary is 0.101/0.226 = 0.449 classes (i.e. approaching half a class). According to Table 5 three countries, UK, FR and WL, have a more relaxed position at the HG boundary and have a bias corresponding to about one quarter of a class. At the GM boundary the FR method is precautionary relative to the global average by about one third of a class. The main

feature is the relaxed position of WL at the GM boundary where the bias expressed in class equivalents indicates that the present boundary is almost half a class too relaxed.

4.3 Boundary harmonisation

Boundaries can be considered to be harmonised when they lie, or are moved to within a small deviation of the global mean value. Within this range of deviation further adjustments to the position of national class boundaries would ideally have little effect on the level of agreement between classifications. What represents a suitably small range of deviation is open to question but, to date, has been accepted as 0.25 class equivalents. At its lower (conservative) end this range should respect the cumulative errors in the overall process of transforming national boundaries in a common dataset to a common scale. At the upper end it is clear that a difference between the most extreme countries of >1 class means they are more likely to belong to different rather than the same class. Similarly, if one method is >0.5 class width from the harmonisation guideline, by inference it must be closer to another boundary than the boundary of interest. Consequently ± 0.5 classes should always be the maximum tolerable harmonisation band width.

In the present example, we can see from Table 5 that three countries would lie outside the 0.25 class range on either the HG (UK, WL, AT) or GM (FR and WL) boundaries, although all five countries are inside within the 0.5 class width band at both boundaries. For both AT and FR the bias indicates that these countries are more precautionary than 0.25 of a class but in this direction both countries can opt not to modify their boundaries. Approaching a bias of -0.5 classes the WL method is clearly too relaxed at the GM boundary relative to the global mean. The bias of -0.26 classes in this method at the HG boundary is sufficiently close to a threshold of -0.25 for no adjustment to be necessary. A minor upward revision of the UK HG boundary would be needed to secure bias of <0.25 classes.

Knowing the level of bias associated with different methods at each class boundary we can then determine which boundaries should be adjusted and by how much. If we consider the GM boundary for WL this boundary needs to be moved to fall within 0.25 classes of the global mean in order to be harmonised. Firstly, determine the mICM EQR equivalent of 0.25 classes for WL. The average class width based on turnover in mICM EQR across the upper two classes is 0.165. Therefore 0.25 classes is equivalent to 0.165/4 or 0.041 units on the mICM EQR scale. Since the method is too relaxed we need to move its GM boundary to within 0.25 classes of the global mean GM value (0.524) which is equivalent to 0.524 – 0.041 = 0.483. We then need to determine the value on the benchmarked WL EQR scale that corresponds to a mICM EQR of 0.483. This requires inverting the regression model to determine where a known value of mICM EQR would lie on the benchmarked national scale. Thus, invert the regression model in Table 4 for WL:

$$\text{mICM EQR} = 0.191 * \exp(1.577 * \text{EQR}_{WL\_B})$$

To give

$$\text{EQR}_{WL\_B} = \text{Ln}(\text{mICM EQR}/0.191)/1.577$$

Thus,

EQR_{WL_B} = Ln(0.483 /0.191)/1.577 = 0.588

Therefore the GM boundary for the WL method needs to be revised upwards from its present position on the benchmarked national EQR scale of 0.544 to a value of 0.588 in order to lie within 0.25 classes of the global mean.

4.4 Boundary synchronisation

Having established the global mean (harmonisation guideline) of all methods on a common scale, a tempting alternative to the concept of a harmonisation band is to simply stipulate that all boundaries should then be fully synchronised with this guideline. The attraction of synchronisation is that it eliminates all bias in classifications and there is no subjectivity in the definition of a band of acceptable bias. We argued for this approach in the first draft of the Guidance Annex V but it was rejected because it implied a need for all countries to make changes to their boundaries, however small. Pešta & Horký (2010) also made this suggestion in some proposed alternatives to the Annex V proposals. We think it is instructive to determine where boundaries would need to lie to achieve complete synchronisation but we are wary of making synchronisation a requirement of boundary harmonisation for several reasons that are covered in the discussion.

To achieve boundary synchronisation we simply need to translate the global mean value on the mICM EQR scale at each boundary to the appropriate value on the national scale, EQR_B, by inverting each regression model in Table 4. Thus, consider the HG boundary, which has a global mean of 0.746. If we wish to synchronise the boundary for FR method;

mICM EQR = 0.191*exp(1.575* EQR_{FR_B})

Therefore,

EQR_{FR_B} = Ln(mICM EQR /0.191)/1.575

Thus,

EQR_{FR_B} = Ln(0.746 /0.191)/1.575 = 0.864

Therefore, to achieve, complete synchronisation, the FR HG boundary would need to move from its starting value on the benchmarked EQR scale of 0.83 to 0.86. Table 7 summarises the appropriate values on the benchmarked national EQR scale for boundary harmonisation or synchronisation.

Table 7. Original, harmonised and synchronised class boundaries on the benchmarked national EQR scales, EQR_{x_B}, derived using Technique 1

Level	Boundary	DE	UK	FR	WL	AT
original benchmarked class boundaries	HG	0.90	0.76	0.83	0.83	0.94
	GM	0.66	0.57	0.70	0.54	0.67
harmonised boundaries (bias ± 0.25 classes)	HG	0.90	0.77	0.83	0.83	0.92
	GM	0.66	0.57	0.68	0.59	0.67
synchronised boundaries (zero bias)	HG	0.87	0.81	0.86	0.86	0.88
	GM	0.62	0.57	0.64	0.64	0.69

4.5 Boundary translation

The benchmarked EQR scale is not a readily recognisable currency for national methods. To compare the harmonised and synchronised class boundaries with the original position of these boundaries in the national methods it is therefore necessary to convert the benchmarked EQRs with which we have worked back to their original national EQR scale. This is achieved simply by multiplying the boundary values on the benchmarked EQR scale by the benchmark corresponding to that method (Table 2). Thus for DE, the continental benchmark (the average value on the DE EQR scale awarded to all continental benchmark sites) is 0.83. Thus to convert the DE boundaries on the benchmarked scale back to their original scaling they must be multiplied by 0.83. For the UK, the atlantic benchmark of 1.05 would function as the multiplier, while for WL a value of 1.12 would be used. Therefore, if we wished to synchronise the DE GM boundary with the global mean across all methods the benchmarked EQR of 0.62 would be multiplied by 0.83 to give a value of 0.52. This compares with the original GM boundary value of 0.55. To harmonise the GM boundary for WL we need to take the value of 0.592 calculated above and multiply it by the WL benchmark of 1.12 to obtain a value of 0.66. Table 8 provides a translation of all harmonised and synchronised boundaries on the benchmarked EQR scale (Table 7) to their original values.

Table 8. Original, harmonised and synchronised class boundaries on national EQR scales derived using Technique 1

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries	HG	0.75	0.8	0.85	0.93	0.875
	GM	0.55	0.6	0.72	0.61	0.625
harmonised boundaries (bias ± 0.25 classes)	HG	0.75	0.81	0.85	0.93	0.85
	GM	0.55	0.6	0.70	0.66	0.625
synchronised boundaries (zero bias)	HG	0.73	0.85	0.89	0.97	0.81
	GM	0.52	0.59	0.66	0.72	0.64

Note: Cells highlighted in blue indicate potential boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

Thus we can say that to achieve harmonisation the GM boundary of the WL method should be raised from 0.61 to 0.66, while at the HG boundary the UK boundary should raised slightly from 0.8 to 0.81. Smaller negative changes to the GM boundary of FR and the HG boundary of AT could also be made but are not obligatory. If for whatever reason one or more countries wish to eliminate the bias in their classifications, the more closely they shift their boundaries to the synchronised value, the less biased they will become. Similarly, if national methods are subject to review at any time in the future, it would certainly be advisable to move class boundaries towards, rather than away from their synchronised position.

4.6 Measuring class agreement

Regression-based approaches are not suitable for measuring change in class agreement since the regression models are insensitive to the repositioning of class boundaries. Attempting to infer class agreement from regressions relies on obtaining an average measure of class width which then presupposes that all classes are of equal width within each classification. An approximate measure of class agreement, which is perhaps more valid for an exercise as a whole, rather than for individual methods, can be obtained by calculating the average absolute regression residual for each method, and then taking the average of these values. Thus for the DE method the average absolute regression residual (mean of absolute difference between observed and predicted values from the mICM EQR v national EQR relationship) is 0.131. The average class width on the mICM EQR scale for DE is 0.201. The average absolute regression residual is therefore equivalent to a class width of 0.131/0.201 = 0.651 classes. Taken across all methods the mean class difference inferred from regression is 0.55 classes (Table 9). In the case of regression of mICM EQR against national EQR_B we might expect that this approach will tend to overestimate class difference since the mICM EQR is not based on a central tendency of other methods and will never be as effective in capturing the global variation across a group of methods as is the pseudo-common metric approach discussed under Technique 2 (see section 5).

Table 9. Comparison of average absolute class difference inferred from regression of mICM EQR against EQR_B (Technique 1) and obtained from individual pairwise comparisons (Technique 3) using the same class boundaries.

	DE	UK	FR	WL	AT	mean
mean abs residual	0.131	0.086	0.093	0.093	0.114	
class equivalent	0.651	0.454	0.599	0.562	0.503	0.554
abs pairwise diff	0.512	0.407	0.387	0.378	0.400	0.417

Technique 3 (see section 6) provides a more effective approach for comparing the difference and agreement between commonly assessed cases since it is sensitive to repositioning of the class boundaries. Thus, we can take the boundaries derived from the Technique 1 analysis and impose

them on Technique 3 to assess how much the agreement between classifications changes as boundaries are first harmonised and then synchronised. Two measures are used here to reflect the level of agreement between classifications (i) the average absolute (i.e. unsigned) difference between all pairs of EQR values and (ii) the proportion of pairwise comparisons that differ by <0.5 class widths. In Table 10 it can be seen that with the original boundaries in place the overall average absolute class difference based on all possible pairwise comparisons between each national set of EQRs is 0.42 classes, and 65% of these comparisons differ by <0.5 classes. Average class difference inferred via regression (Technique 1) suggests an average class difference of 0.55 and may therefore be a reasonable approximation of the true value.

Using the Technique 3 approach, even with fully synchronised boundaries in place there is only a 13% reduction in average absolute class difference to 0.36, while the proportion of comparisons differing by <0.5 classes rises only to 71%. This may reflect the fact that the original class boundaries are relatively well positioned from the outset and all methods are strongly intercorrelated. However, the improvement in agreement between harmonisation to within 0.25 classes and complete synchronisation is clearly trivial and does not provide a persuasive argument for further boundary changes.

Table 10. Comparison of agreement between classifications at different levels of harmonisation of class boundaries using boundaries derived by Technique 1 and measured via the pairwise comparison approach (Technique 3)

Level	Measure	DE	UK	FR	WL	AT	average
original class boundaries	abs_av_diff	0.51	0.41	0.39	0.38	0.40	0.42
	% ± 0.5 class	57	67	66	68	67	65
harmonised boundaries (bias ± 0.25 classes)	abs_av_diff	0.50	0.39	0.35	0.34	0.37	0.39
	% ± 0.5 class	58	70	74	77	71	70
synchronised boundaries (zero bias)	abs_av_diff	0.48	0.38	0.29	0.30	0.36	0.36
	% ± 0.5 class	60	70	77	76	72	71

An alternative to the use of simple metrics to evaluate agreement between classifications is to use the multi-rater kappa coefficient, which offers a chance adjusted measure of agreement. To implement the kappa analysis it is simplest to assess the upper boundaries on a boundary specific basis. Thus to evaluate the HG boundary we simply determine whether cases are classified as High or <High by each method and then calculate the level of agreement between classifications. Similarly to assess the GM boundary the focus is on whether cases are classified as <Good or ≥ Good (Borja et al., 2007; Pont & Delaigue, 2010).

Table 11 provides the results of a boundary-specific kappa analysis for the 215 commonly assessed sites. At the GM boundary there is a very good agreement (kappa >0.7) between classifications even with the original boundaries in place. Although the increasing kappa value illustrates increased agreement between classifications as boundaries converge to the global mean position the increases

are small and based on the overlap in 95% confidence limits do not represent a significant improvement in agreement. At the HG boundary there is also a very good level of agreement between classifications but boundary synchronisation results in a negligible increase in kappa. In general the kappa analysis reiterates the conclusions of the comparison based on simple metrics, namely, that when methods already show a relatively small level of bias from the outset changing the position of these boundaries will have little effect on agreement between classifications.

Table 11. Comparison of agreement between classifications using boundaries determined by Method 1, based on the free multi-rater kappa coefficient

	Good-Moderate			High-Good		
	kappa	L 95CL	U 95CL	kappa	L 95CL	U 95CL
original	0.767	0.725	0.810	0.721	0.685	0.757
harmonised	0.803	0.761	0.845	0.747	0.711	0.783
synchronised	0.829	0.787	0.871	0.749	0.713	0.785

5. Technique 2: Direct approach for boundary comparison using regression

5.1 Mapping methods onto a common metric

The approach described here follows the protocol we outlined in the Guidance Annex V (Willby & Birk, 2010a). The basic principle of this approach is to use the average EQR of a pool of independent countries as the yardstick ('pseudo-common metric', PCM) to determine, via regression, where the test country places its boundaries. By considering all combinations of independent countries a value on the PCM scale is obtained for each method at each boundary. The relative bias between methods at a given boundary can then be assessed with reference to the PCM.

In the example used here we first determine the PCM for each country being evaluated. Thus, for AT, for example, the PCM is equivalent to the average benchmarked EQR of the each of the other four countries assessing each site, in this case, DE, UK, FR and WL. The benchmarking of national EQRs is carried out as described in section 3. Thus Technique 2 utilises much of the data already prepared for use in Technique 1 (see section 4). In deriving PCM values, no standardisation of EQRs to a fixed class width is needed (see Technique 3, section 6.1) and it will be advantageous if EQR values are allowed to cover their natural range, rather than being artificially truncated to one.

Secondly, using regression, we establish a simple linear model for relating the appropriate PCM (y) to the national benchmarked EQR (x). The fitted models for the data in this case study are shown in Figure 3. The preference is to use the ordinary least squares estimation technique. If model residuals are not independent of the regressor (x) additional linear transformations can be introduced if these significantly improve the fit to the data. Depending on the availability of supporting information about the sites being assessed, other forms of regression (e.g. mixed linear, non-linear) are acceptable provided that they further improve the fit to the data. In the present example all

methods are highly correlated with the PCM and there is no evidence of non-linearity in the relationships so the regression is restricted to simple linear models. The PCM models for each method are summarised in Table 12.

Table 12. Summary of PCM models for each of the five national methods being compared

	Model Summary				Parameter Estimates	
	R Square	F	Sig.	SE	Constant	Slope
DE	0.733	592	0.000	0.163	0.182	0.767
UK	0.830	1053	0.000	0.127	0.240	0.769
FR	0.917	2374	0.000	0.096	0.085	0.889
WL	0.918	2408	0.000	0.096	0.084	0.891
AT	0.844	1170	0.000	0.163	-0.141	1.131

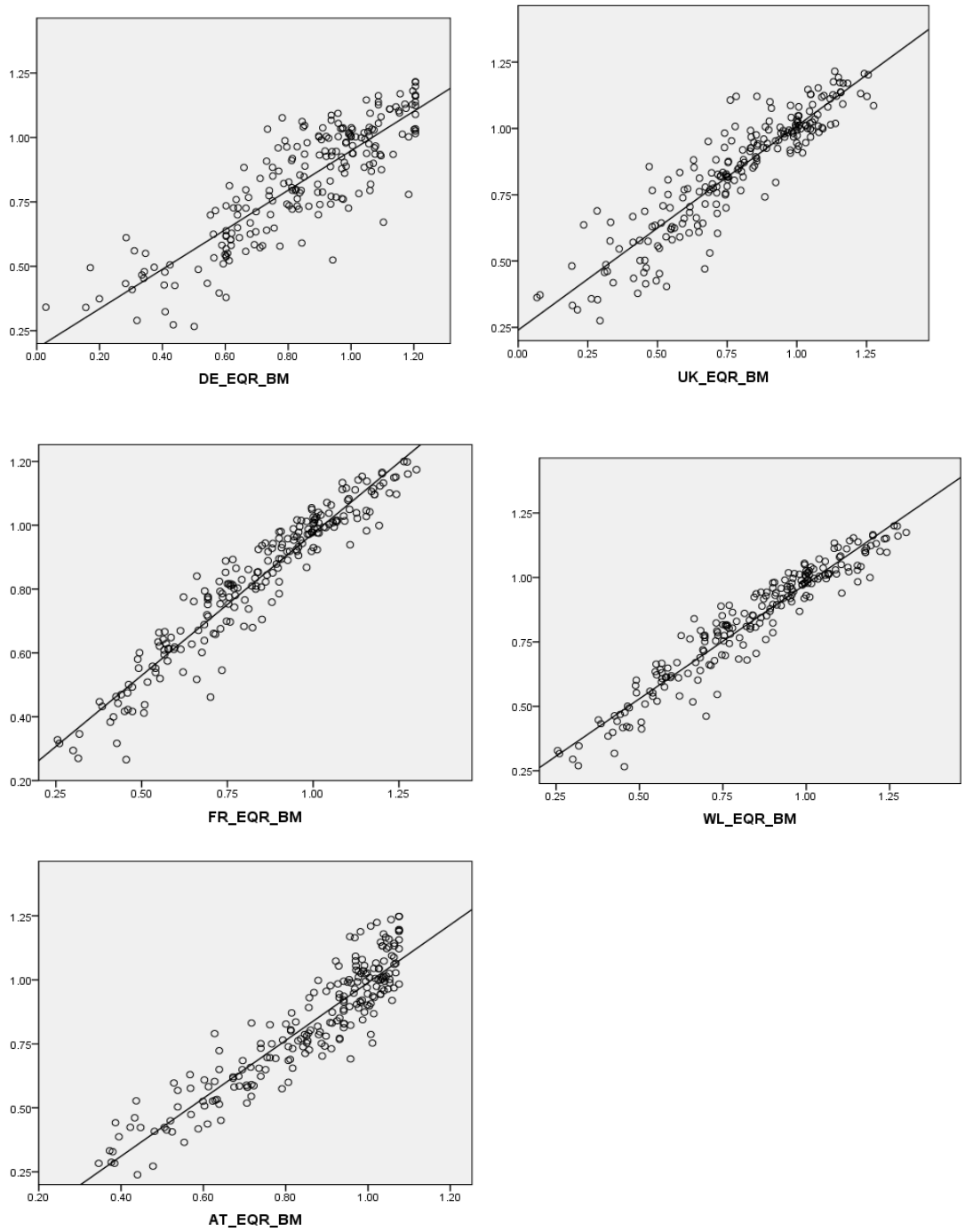


Figure 3. Relationships between a PCM (y), based on the average benchmarked EQR of the remaining four countries and national benchmarked EQR (x) for participating countries.

As with Technique 1 a pre-requisite at this stage is that all methods being compared should be significantly correlated with the PCM. Clearly this requirement is fulfilled in the present example. If it were not, the most deviant method would be removed and the PCM values recalculated without this method until all remaining methods were significantly correlated. Alternatively, those methods lacking a significant correlation with the PCM would be subject to interrogation, the objective being to achieve a satisfactory correlation with the average view of other methods, by, for example, excluding or reweighting component metrics. The same general comments regarding correlation coefficients discussed under Technique 1 (section 4.2) also hold here.

5.2 Measuring bias between methods

Using the linear model for each method (Table 12) we derive where its benchmark normalised HG and GM boundary values should lie on the PCM scale. Thus, for the UK, for example, the benchmark normalised HG boundary lies at 0.762 (from Table 3). Therefore on the PCM_{UK} scale

$$PCM_{UK} = 0.769 * EQR_{UK\_B} + 0.240 = 0.826$$

The values on the PCM scale are then calculated in the same way for each method at the HG and GM boundaries (Table 13). The global average of the values for all methods at a given boundary represents the guideline value for that boundary against which the bias of individual methods is then to be measured. Thus, calculating the HG boundary positions for all methods on the PCM scale produces a global mean of 0.854. The equivalent value at the GM boundary is 0.653.

Table 13. Predicted PCM values at HG and GM boundaries for five national methods and their associated bias

	DE	UK	FR	WL	AT	mean
HG boundary	0.872	0.826	0.826	0.820	0.926	0.854
GM boundary	0.688	0.679	0.709	0.569	0.621	0.653
HG bias	0.018	-0.028	-0.028	-0.034	0.072	
GM bias	0.035	0.026	0.056	-0.084	-0.032	
HG bias as class eq	0.115	-0.193	-0.163	-0.214	0.315	
GM bias as class eq	0.217	0.177	0.321	-0.529	-0.141	

Note: Cells highlighted in blue indicate potential boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

We have already discussed the problems of converting bias around a global mean class boundary value into class equivalents and a potential solution that uses the national regression model to convert turnover in the common metric, in this case the PCM, into class equivalents.

To convert the deviation on the PCM scale to class equivalents we require two pieces of information. The first is the value on the PCM scale that equates to the benchmarked reference value used to estimate the national EQR. Provided national EQRs are expressed relative to a value of 1 the benchmarked reference value will simply be 1 divided by the appropriate benchmark. Thus, for the UK the benchmarked reference value is 1/1.05 = 0.95. On the PCM scale for the UK a benchmarked national EQR value of 0.95 is equivalent to a value of 0.97 (models in Table 12). This anchors the upper part of the PCM scale. The second piece of information required is the lower anchor on the PCM scale equivalent to a national EQR of 0. This is provided by the intercept in the regression model. Thus for the UK method the effective range of ecological quality on the PCM scale is from 0.24 to 0.97, a range of 0.73 units. Therefore, an average class width on the PCM scale = 0.73/5 = 0.15 units. Table 14 provides the class widths on the PCM scale for all five methods. The overall average class width across all methods on the PCM scale is 0.173 which compares favourably to the mean width calculated for Technique 1 (0.188).

Table 14. Determining class equivalent units on the PCM scale to allow estimates of bias around global mean to be expressed in terms of class equivalent units

PCM scale	**DE**	**UK**	**FR**	**WL**	**AT**
benchmarked reference value	0.987	0.972	0.952	0.879	1.002
minimum	0.182	0.240	0.085	0.084	-0.141
gradient width	0.805	0.732	0.867	0.795	1.143
av. class width	0.161	0.146	0.173	0.159	0.229

Table 13 summarises the predicted value for the HG and GM boundary for each method on the PCM scale, using the linear models detailed in Table 12 based on the original (i.e. pre-adjustment) national class boundaries. The bias relative to the global mean is then converted to class equivalents. Thus, for AT, for example, the HG boundary translates to a PCM value of 0.924, using the model in Table 12, and the class boundaries in Table 3. Thus, relative to the global mean there is a bias of 0.924 – 0.853 = +0.071 PCM units at the HG boundary. The average class width on the PCM scale for AT is 0.229 (Table 14). Therefore, in class equivalent terms, the bias for method AT at the HG boundary is 0.071/0.229 = 0.315 classes (i.e. almost a third of a class).

5.3 Boundary harmonisation

In the present example, we can see from Table 13 that three countries would lie outside the 0.25 class range on either the HG (AT) or GM (FR and WL) boundaries, although all five countries are more or less within the 0.5 class width band at both boundaries. In this sense the conclusions

concerning bias are the same as would be made using Technique 1. For both AT and FR the bias indicates that these countries are more precautionary than 0.25 of a class but in this direction both countries can opt not to modify their boundaries. Approaching a bias of -0.5 classes the WL method is clearly too relaxed at the GM boundary relative to the global mean.

Knowing the level of bias associated with different methods at each class boundary we can then determine which boundaries should be adjusted and by how much. If we consider the GM boundary for WL this boundary needs to be moved to within 0.25 classes of the global mean in order to be harmonised. Firstly, determine the PCM equivalent of 0.25 classes for WL. The average class width based on turnover in PCM is 0.159. Therefore 0.25 classes is equivalent to 0.159/4 or 0.04 units on the WL PCM scale. Since the method is too relaxed we need to move it to within 0.25 classes of the global mean GM value (0.653) which is equivalent to 0.653 – 0.04 = 0.613. We then need to determine the value on the benchmarked WL EQR scale that corresponds to a PCM of 0.609. Thus, invert the regression model for WL

$PCM_{WL} = 0.891 * EQR_{WL\_B} + 0.084$

To give

$EQR_{WL\_B} = (PCM_{WL} - 0.084)/0.891$

Thus,

$EQR_{WL\_B} = (0.613 - 0.084)/0.891 = 0.594$

Therefore the GM boundary for the WL method needs to be revised upwards from its present position of 0.544 on the benchmarked national EQR scale to a value of 0.594 in order to lie within 0.25 classes of the global mean.

5.4 Boundary synchronisation

To achieve boundary synchronisation we simply need to translate the global mean value on the PCM scale at each boundary to the appropriate value on the national scale. This requires inverting the regression model to determine where a known value of PCM would lie on the benchmarked national scale. Thus, if we take the HG boundary, which has a global mean of 0.853, for the FR method

$PCM_{FR} = 0.889 * EQR_{FR\_B} + 0.085$

Therefore,

$EQR_{FR\_B} = (PCM_{FR} - 0.085)/0.889$

Thus,

$EQR_{FR\_B} = (0.853 - 0.085)/0.889 = 0.864$

Therefore, to achieve, complete synchronisation, the FR HG boundary would need to move from its starting value on the benchmarked EQR scale of 0.83 to 0.86.

Table 15. Original, harmonised and synchronised class boundaries on the benchmarked national EQR scales, EQR_{x_B}, derived using Technique 2

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries	HG	0.90	0.76	0.83	0.83	0.94
	GM	0.66	0.57	0.70	0.54	0.67
harmonised boundaries (bias ± 0.25 classes)	HG	0.90	0.76	0.83	0.83	0.93
	GM	0.66	0.57	0.69	0.60	0.67
synchronised boundaries (zero bias)	HG	0.88	0.80	0.86	0.86	0.88
	GM	0.61	0.54	0.64	0.64	0.70

Note: Cells highlighted in blue indicate boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

5.5 Boundary translation

The approach to translating boundaries on the benchmarked national EQR scale to their original scaling is identical to that described for Technique 1 in section 4.5. Thus, for example, to harmonise the GM boundary for WL we need to take the value of 0.594 calculated above and multiply it by the WL benchmark of 1.12 to obtain a value of 0.67 (Table 16). This compares with the original GM boundary value of 0.61 and a value of 0.66 calculated via Technique 1 as the appropriate value for harmonisation.

Table 16. Original, harmonised and synchronised class boundaries on national EQR scales derived using Technique 2

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries	HG	0.75	0.8	0.85	0.93	0.875
	GM	0.55	0.6	0.72	0.61	0.625
harmonised boundaries (bias ± 0.25 classes)	HG	0.75	0.8	0.85	0.93	0.86
	GM	0.55	0.6	0.71	0.67	0.625
synchronised boundaries (zero bias)	HG	0.73	0.84	0.89	0.97	0.82
	GM	0.51	0.56	0.66	0.72	0.65

Note: Cells highlighted in blue indicate boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

Thus we can say that to achieve harmonisation the GM boundary of the WL method should be raised from 0.61 to 0.67. Smaller negative changes to the GM boundary of FR and the HG boundary of AT

could also be made but are not obligatory. The pattern of suggested adjustments to class boundaries agrees closely with that obtained using Technique 1.

5.6 Measuring class agreement

As discussed for Technique 1 regression-based approaches are not suitable for measuring change in class agreement since regressions are insensitive to the repositioning of class boundaries. An approximate measure of class agreement can however, be obtained by calculating the average absolute regression residual for each method and converting to this to a class equivalent. Thus for the DE method the average absolute regression residual (mean of absolute difference between observed and predicted values from the EQR_B v PCM relationship) is 0.092. The average class width on the PCM scale for DE is 0.161. The average absolute regression residual is therefore equivalent to a class width of 0.092/0.161 = 0.573 classes. Taken across all methods the mean class difference inferred from regression is 0.38 classes (Table 17). The Technique 2 regression approach should generally tend to underestimate class difference because a national method is being compared with the central tendency of other national methods, not their individual values. Thus, the approach based on pairwise comparisons (Technique 3, described in section 6) produces an overall mean class difference of 0.42 using the same class boundary values. Thus, average class difference inferred via regression appears to offer a reasonable approximation of the true value.

Table 17. Comparison of average absolute class difference inferred from regression and obtained from individual pairwise comparisons.

	DE	UK	FR	WL	AT	mean
mean abs residual	0.092	0.067	0.049	0.049	0.075	
class equivalent	0.573	0.457	0.283	0.306	0.328	0.384
abs pairwise diff	0.512	0.407	0.387	0.378	0.400	0.417

Technique 3 provides a more effective approach for comparing the difference and agreement between commonly assessed cases since it is sensitive to repositioning of the class boundaries. Taking the boundaries derived from Technique 2 we can impose these on Technique 3 to assess how agreement between classifications changes as boundaries are first harmonised and then synchronised. Given the close similarity between the boundaries at harmonised and synchronised levels in Technique 1 and Technique 2 we would expect the results of an analysis of class agreement to also be very similar. This is confirmed by Table 18, with the results virtually replicating those shown in Table 10 for Technique 1.

Table 18. Comparison of agreement between classifications at different levels of harmonisation of class boundaries using Technique 2

Level	Measure	DE	UK	FR	WL	AT	average
original class boundaries	av. abs. diff	0.51	0.41	0.39	0.38	0.40	0.42
	% ± 0.5 class	57	67	66	68	67	65
harmonised boundaries (bias ± 0.25 classes)	av. abs. diff	0.51	0.40	0.36	0.34	0.38	0.40
	% ± 0.5 class	58	69	74	77	70	70
synchronised boundaries (zero bias)	av. abs. diff	0.48	0.38	0.29	0.30	0.35	0.36
	% ± 0.5 class	60	70	76	76	72	71

Table 19 provides the results of a boundary-specific kappa analysis for the 215 commonly assessed sites. At the GM boundary there is a very good agreement (kappa >0.7) between classifications even with the original boundaries in place. Although the increasing kappa value illustrates increased agreement between classifications as boundaries converge to the global mean position the increases are small and based on the overlap in 95% confidence limits do not represent a statistically significant improvement in agreement. At the HG boundary there is also a very good level of agreement between classifications but boundary synchronisation results in a negligible increase in kappa. In general the kappa analysis reiterates the conclusions of the comparison based on simple metrics, namely, that when methods already show a high level of relatedness modifications to the position of class boundaries will have a trivial effect on agreement between classifications.

Table 19. Comparison of agreement between classifications, based on the free multi-rater kappa coefficient

	Good-Moderate			High-Good		
	kappa	L 95CL	U 95CL	kappa	L 95CL	U 95CL
original	0.767	0.725	0.810	0.721	0.685	0.757
harmonised	0.793	0.751	0.836	0.740	0.704	0.775
synchronised	0.827	0.785	0.869	0.745	0.709	0.781

6. Technique 3: Boundary comparison based on direct pairwise comparisons

6.1 Establishing a basis for comparison

An alternative to the use of regression against a common axis to compare class boundaries is to directly compare classifications for all commonly assessed sites in a dataset. Thus, if there are m methods in an exercise and n commonly assessed cases, there are (m-1)*n possible comparisons to

be made for each method. In the present case, this means there are (5-1)*215 = 860 comparisons per method. One could simply compare classes, rather than EQRs, as suggested in the NL proposal, but this would sacrifice useful numerical data and could be a significant constraint on the comparison of methods where only small amounts of data are available. Also, two cases separated by only 0.02 EQR units may belong to different classes in different methods, a view which clearly inflates the actual differences between them, while, in another situation, two cases separated by 0.19 EQR units may belong to the same class in different methods. In large datasets such instances will be averaged out but they will become increasingly influential in small datasets. In this approach we compare EQR values themselves.

There are two prerequisites for the direct comparison of EQRs. Firstly, the national EQRs must be significantly intercorrelated, either with each other, or with a PCM, calculated as described in Technique 2. This method requires no assumptions regarding normality and it is acceptable to use non-parametric tests to establish the significance of correlation between methods (e.g. Spearman's Rank Correlation or Kendall's tau). The same comments regarding correlation coefficients made in regard to Technique 1 apply here. Thus the size of the correlation coefficient (equivalent to a Pearson's correlation of 0.5-0.6) is more critical than the level of significance, provided a minimum of $p < 0.05$ has been met. Secondly, all EQR values must be standardised to a common scale so that differences between values at commonly assessed sites can be interpreted as class equivalents. This is because a difference in EQR value of 0.1 units at a commonly assessed site between one method and another, may not represent the same class equivalent in all the methods being compared. Comparisons at an EQR level will therefore require the piecewise linear transformation of all EQRs to a common scale and class boundary system where all classes are 0.2 EQR units in width. The transformation is computationally simple and effectively maps each site within a national class where all the classes have the same width. Thus in method A, if the HG and GM boundaries lie at 0.84 and 0.66 respectively, a site with an EQR of 0.75 lies exactly in the middle of Good status, or, on a 0.2 unit system, half way between 0.6 and 0.8 (i.e. 0.7). In method B, if the HG and GM boundaries lie at 0.82 and 0.62 respectively, a site with an EQR of 0.72 lies exactly in the middle of Good status, or, on a 0.2 unit system, again half way between 0.6 and 0.8 (i.e. 0.7). Thus, despite the raw EQR values differing, they are identical on a transformed scale. Conversely, a site with an EQR of 0.83 in each methods would be classified as Good status, according to the class boundaries of method A, but high status according to the class boundaries of method B.

In previous iterations of this approach classifications have been fully transformed to reflect the location of all four class boundaries (HG, GM, MP, PB). However, this approach gives the lower class boundaries equal weight to the upper boundaries whereas in IC we are only concerned with the upper class boundaries; the focus is on whether sites are classified as High, Good or <Good. This implies that it would be more fitting for IC purposes to simply map national classifications onto a system where the EQR units 0.8-1.0 represent high status, units 0.6-0.8 represent good status, and 0-0.6 represent <Good status. We are not concerned with how national methods subdivide the lower 0.6 units of the EQR gradient between Moderate, Poor and Bad status. Using this same principle we can create a greater focus on specific class boundaries. Therefore, if we are most concerned with the bias and disagreement between classifications in relation to the GM boundary we would simply remap a classification onto a scale where EQR units 0.6-1.0 represent ≥ Good and units 0-0.6 represent <Good. Similarly, if we wish to focus on the HG boundary, we should remap

classifications onto a scale where EQR units 0.8-1.0 represent High and units 0-0.8 represent <High. The added advantage of this approach is that we can manipulate the position of the HG or GM boundaries without worrying about the potential distortion caused by leaving the MP or PB boundaries fixed.

To achieve the piecewise linear transformation of national EQRs we need to know the position of reference, HG and GM boundaries on each national scale and their benchmarked normalised equivalent. To remap a national classification onto a common scale the following formula is used, in which EQR_{A_B} represents the EQR of country A on the benchmarked scale, and REF, HG and GM represent the position of reference, High-Good and Good-Moderate boundaries on the same scale.

$$\begin{aligned}&=IF(EQR_{A\_B} >= REF_{A\_B},1,\\&IF(EQR_{A\_B} >= HG_{A\_B},((EQR_{A\_B} - HG_{A\_B})/(REF_{A\_B} - HG_{A\_B}))*0.2+0.8,\\&IF(EQR_{A\_B} >= GM_{A\_B},((EQR_{A\_B} - GM_{A\_B})/(HG_{A\_B} - GM_{A\_B}))*0.2+0.6,\\&IF(EQR_{A\_B} < GM_{A\_B},((EQR_{A\_B} / GM_{A\_B})*0.6)))))\end{aligned} \quad (1)$$

Thus, if we consider a raw national EQR for FR of 0.77 applied to the assessment of a French site. The continental benchmark for FR is 1.03. Therefore the benchmarked national EQR is 0.75. The REF, HG and GM boundaries for the FR method lie at 1, 0.85 and 0.72 respectively, or on the benchmarked scale, 0.98, 0.83 and 0.7. Thus, since the observed value lies between the HG and GM boundaries, its transformed equivalent is:

$$((0.75-0.7)/(0.83-0.7))*0.2+0.6 = 0.68$$

If our interest was only in intercalibrating classifications in terms of <Good or ≥Good status we would modify expression 1 above so that

$$\begin{aligned}&=IF(EQR_{A\_B} >= REF_{A\_B},1,\\&IF(EQR_{A\_B} >= GM_{A\_B},((EQR_{A\_B} - GM_{A\_B})/(REF_{A\_B} - GM_{A\_B}))*0.4+0.6,\\&IF(EQR_{A\_B} < GM_{A\_B},((EQR_{A\_B} / GM_{A\_B})*0.6))))\end{aligned} \quad (2)$$

Thus, based on an EQR_B of 0.75 for FR, in relation to a <Good or ≥Good division the EQR would be transformed to

$$((0.75-0.7)/(0.98-0.7))*0.4+0.6 = 0.67$$

If our interest was only in intercalibrating classifications in terms of <High or High status we would modify the expression so that

$$\begin{aligned}&=IF(EQR_{A\_B} >= REF_{A\_B},1,\\&IF(EQR_{A\_B} >= HG_{A\_B},((EQR_{A\_B} - HG_{A\_B})/(REF_{A\_B} - HG_{A\_B}))*0.2+0.8,\\&IF(EQR_{A\_B} < HG_{A\_B},((EQR_{A\_B} / HG_{A\_B})*0.8))))\end{aligned} \quad (3)$$

Thus, based on an EQR_B of 0.75 for FR, in relation to a <High or High status division the EQR would be transformed to

$$(0.75/0.83)*0.8 = 0.72$$

6.2 Measuring bias between methods

We start with a set of benchmark normalised EQR values for each method. We then transform these according to expression 2 above so that the emphasis is on the distinction between <Good or ≥Good. We then calculate the differences between each method and all other methods. Thus, if the appropriately transformed EQR values for methods A, B and C for a given site are 0.77, 0.65 and 0.80 respectively, the differences for A are +0.12 (versus B) and -0.03 (versus C), for B are -0.12 (versus A) and -0.15 (versus C), and for C are +0.03 (versus A) and +0.15 (versus B). The bias between each method and every other method for all commonly assessed sites is then calculated and an average of the overall population of pairwise differences is determined for each method (i.e. the mean of all 860 pairwise comparisons between that method and the other four methods). This provides a measure of the bias in EQR values between one method and all other methods, which can be transformed directly to class equivalent units (divide by 0.2). The same approach is followed if we wish to determine the average absolute (i.e. unsigned) difference between one method and another. Thus, if two sites are assessed by two methods and the bias between method A versus B at sites 1 and 2 is +0.05 and -0.05, the average bias is zero, while the average absolute difference is 0.05 (or 0.05/0.2 =0.25 classes).

6.3 Boundary harmonisation

In Method 3, the approach to boundary harmonisation differs from Techniques 1 and 2. This is because there is not a fixed line against which all methods can be compared in relation to a given boundary. Moreover, changing the position of the class boundaries of one country will directly affect the bias in other classifications even though their class boundaries are untouched. Thus, if method A appears too relaxed relative to methods B and C by half a class, we cannot simply raise the class boundary of A by half a class since this will potentially render either B or C too relaxed. Hence we require an iterative approach where we seek to change the class boundaries of the most outlying methods by 0.01 EQR units at a time, until all methods achieve the necessary level of bias. By restricting the analysis to a single boundary at a time the question of whether to manipulate the HG or GM boundary to adjust the bias is avoided.

Consider the GM boundary in the worked example. The existing national boundaries have been compared and indicate that the most relaxed method, WL, on average differs from the remaining four countries by 0.35 classes (Table 20, 0 iterations). This means that across the 215 sites assessed the WL EQR was on average 0.35 classes higher than the EQR assigned to those sites by the other four countries. Consequently, the GM boundary should be raised for WL to render it less precautionary. To test the effect of iteratively modifying the class boundaries of the different methods on the resulting bias it will be simplest to set up some linked formulas such that the class boundaries needed for the piecewise linear transformation are read from a table of benchmark transformed class boundaries which are in turn read from the class boundaries on the original national scale. Manipulating the latter will thus cause the classifications and associated bias to update automatically. Continue manipulating the national class boundary, 0.01 units at a time, until all methods achieve a bias within 0.25 class equivalents. In the present example, this requires a further six iterations, raising the GM boundary for WL to 0.68 and reducing its bias relative to all

other countries to 0.24 classes (Table 20, Figure 4). At this point, no changes to the GM boundary of any other country are needed for boundary harmonisation.

6.4 Boundary synchronisation

In the same way that boundaries can be manipulated to reduce bias to acceptable limits, we can continue to manually adjust the position of the boundary of the most biased method in each cycle until the bias across all methods cannot be reduced any further (Table 20, Figure 4). At this point we could consider that the boundaries have been synchronised. If we continue to reduce the bias in the worked example, no changes to the boundary of other methods is required until the 14th iteration, where after an increase in the position of the AT GM boundary, and small reductions in the position of the DE and FR boundaries are required. In practice it could be argued that the DE and FR boundaries are acceptable as they stand and we could therefore decide to halt the iterations at the point where a reduction in the class boundary of these countries is required. Exactly how to deal with more precautionary countries in IC needs to be resolved, although for the purposes of boundary synchronisation it is a relatively minor concern.

The final stages of boundary synchronisation using this method are relatively tedious and there is a risk that they may lead to over-adjustment of boundaries. One approach is to consider where the 95% confidence limits of the average bias in the different methods (here ~0.04 classes) would indicate that the bias does not differ significantly from zero (i.e. it is <0.04 classes) and to stop the iterations at this point (Figure 4).

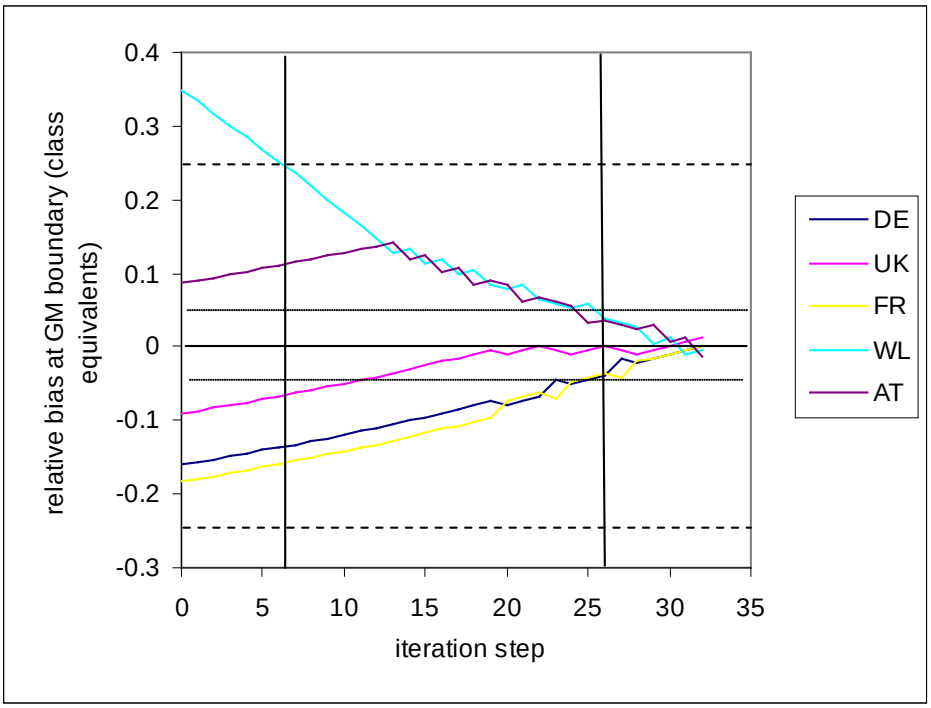


Figure 4. Convergence in bias at GM boundary through iterative adjustment of national boundaries. Dashed lines indicate bias of one quarter of a class. Dotted line represents 95% confidence limits of zero bias. Vertical lines represent potential stopping points.

To establish the comparability of classifications in relation the HG boundary we would the repeat steps outlined in 6.2 to 6.4, this time commencing with a set of EQR values transformed according to expression 3 above. This exercise indicates that with the original HG boundaries in place the bias between all methods is within 0.25 classes indicating that no change to the position of the HG boundary is required for harmonisation. The harmonised boundaries in Table 21 would then be substituted into expression 1 to measure the bias across High, Good and <Good classes. Combining the synchronised values for the HG and GM boundaries into expression 1 in fact results in a small increase in class bias between methods (if compared to the absence of bias when these boundaries are assessed independently) which requires a further series of iterations to remove. This suggests that full synchronisation at a boundary specific level is liable to over-correct class boundaries. In preference to this it is simpler to take the boundaries equivalent to the point where bias does not differ significantly from zero (Table 21). Merging HG and GM boundaries extracted at this level of synchronisation does not require any further iteration to adjust the resulting bias.

Table 20. Reduction in bias at GM boundary through iterative changes in position of national class boundaries

iteration	bias as class equivalents					position of GM boundary				
	DE	**UK**	**FR**	**WL**	**AT**	**DE**	**UK**	**FR**	**WL**	**AT**
0	-0.161	-0.091	-0.183	0.348	0.087	0.55	0.6	0.72	0.61	0.625
1	-0.157	-0.087	-0.179	0.333	0.090	0.55	0.6	0.72	0.62	0.625
2	-0.153	-0.083	-0.175	0.317	0.094	0.55	0.6	0.72	0.63	0.625
3	-0.149	-0.079	-0.171	0.301	0.098	0.55	0.6	0.72	0.64	0.625
4	-0.145	-0.075	-0.167	0.285	0.102	0.55	0.6	0.72	0.65	0.625
5	-0.141	-0.071	-0.163	0.268	0.106	0.55	0.6	0.72	0.66	0.625
6	-0.136	-0.067	-0.159	0.252	0.111	0.55	0.6	0.72	0.67	0.625
7	-0.132	-0.063	-0.155	0.235	0.115	0.55	0.6	0.72	0.68	0.625
8	-0.128	-0.059	-0.151	0.218	0.119	0.55	0.6	0.72	0.69	0.625
9	-0.124	-0.054	-0.146	0.201	0.123	0.55	0.6	0.72	0.70	0.625
10	-0.119	-0.050	-0.142	0.183	0.128	0.55	0.6	0.72	0.71	0.625
11	-0.115	-0.045	-0.137	0.165	0.132	0.55	0.6	0.72	0.72	0.625
12	-0.110	-0.041	-0.133	0.146	0.137	0.55	0.6	0.72	0.73	0.625
13	-0.105	-0.036	-0.128	0.127	0.142	0.55	0.6	0.72	0.74	0.625
14	-0.100	-0.030	-0.122	0.133	0.120	0.55	0.6	0.72	0.74	0.635
15	-0.095	-0.026	-0.118	0.113	0.125	0.55	0.6	0.72	0.75	0.635
16	-0.089	-0.020	-0.112	0.119	0.103	0.55	0.6	0.72	0.75	0.645
17	-0.084	-0.015	-0.107	0.099	0.108	0.55	0.6	0.72	0.76	0.645
18	-0.079	-0.009	-0.101	0.104	0.085	0.55	0.6	0.72	0.76	0.655
19	-0.074	-0.004	-0.096	0.084	0.090	0.55	0.6	0.72	0.77	0.655
20	-0.079	-0.010	-0.074	0.078	0.085	0.55	0.6	0.71	0.77	0.655
21	-0.073	-0.004	-0.068	0.084	0.061	0.55	0.6	0.71	0.77	0.665
22	-0.068	0.001	-0.063	0.064	0.067	0.55	0.6	0.71	0.78	0.665

Table 20 (cont.). Reduction in bias at GM boundary through iterative changes in position of national class boundaries

iteration	bias as class equivalents					position of GM boundary				
	DE	**UK**	**FR**	**WL**	**AT**	**DE**	**UK**	**FR**	**WL**	**AT**
23	-0.044	-0.005	-0.069	0.058	0.061	0.54	0.6	0.71	0.78	0.665
24	-0.050	-0.010	-0.047	0.052	0.055	0.54	0.6	0.70	0.78	0.665
25	-0.044	-0.005	-0.041	0.058	0.032	0.54	0.6	0.70	0.78	0.675
26	-0.039	0.000	-0.036	0.038	0.037	0.54	0.6	0.70	0.79	0.675
27	-0.015	-0.005	-0.042	0.032	0.031	0.53	0.6	0.70	0.79	0.675
28	-0.020	-0.011	-0.020	0.026	0.025	0.53	0.6	0.69	0.79	0.675
29	-0.015	-0.006	-0.015	0.005	0.030	0.53	0.6	0.69	0.80	0.675
30	-0.009	0.000	-0.009	0.011	0.006	0.53	0.6	0.69	0.80	0.685
31	-0.004	0.006	-0.003	-0.010	0.012	0.53	0.6	0.69	0.81	0.685
32	0.002	0.012	0.003	-0.004	-0.013	0.53	0.6	0.69	0.81	0.695

Note: Threshold after seven iterations represents bias <0.25 classes. Threshold after 26 iterations represents bias amongst methods within 95% confidence limits of zero.

The harmonised and synchronised national class boundaries are shown translated into their original scaling in Table 22. Compared with Techniques 1 and 2 the overall level of change needed to the original class boundaries to achieve boundary synchronisation is proportionally less using Technique 3, but the amount of adjustment that would be required to the WL GM boundary is significantly greater.

Table 21. Original, harmonised and synchronised class boundaries on the benchmarked national EQR scales, EQR_{x_B} derived using Technique 3

Level	**Boundary**	**DE**	**UK**	**FR**	**WL**	**AT**
original class boundaries	HG	0.90	0.76	0.83	0.83	0.94
	GM	0.66	0.57	0.70	0.54	0.67
harmonised boundaries (bias ± 0.25 classes)	HG	0.90	0.76	0.83	0.83	0.94
	GM	0.66	0.57	0.70	0.61	0.67
synchronised boundaries (zero bias)	HG	0.85	0.77	0.84	0.87	0.92
	GM	0.65	0.57	0.68	0.71	0.73

Note: Cells highlighted in blue indicate boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

Table 22. Original, harmonised and synchronised class boundaries on national EQR scales derived using Technique 3

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries	HG	0.75	0.80	0.85	0.93	0.875
	GM	0.55	0.60	0.72	0.61	0.625
harmonised boundaries (bias ± 0.25 classes)	HG	0.75	0.8	0.85	0.93	0.875
	GM	0.55	0.6	0.72	0.68	0.625
synchronised boundaries (zero bias)	HG	0.71	0.81	0.86	0.97	0.855
	GM	0.54	0.60	0.70	0.79	0.675

Note: Cells highlighted in blue indicate boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

6.5 Measuring class agreement

The process described above allows bias to be assessed on a boundary-specific basis. We could measure associated class agreement (as <High v High or <Good or ≥Good) separately in relation to the position of each boundary, or we could combine the information on the harmonised and synchronised position of the HG and GM boundaries and then measure the corresponding class agreement. The latter approach is used here and is more compatible with the approach followed in Techniques 1 and 2. Thus to measure the class agreement with harmonised boundaries we take the GM boundaries from Table 21 (where the only change is a rise in the GM boundary for WL to 0.61) and the HG boundaries (where the assessment of bias indicated that no change was required since, in relation to the High or <High split the bias was within 0.25 classes for all methods). We then apply the piecewise linear transformation according to expression 1 above incorporating the harmonised HG and GM boundaries. To test the effect of boundary synchronisation on class agreement we then substitute the harmonised boundaries with the synchronised boundaries in Table 21. Class agreement, assessed in terms of the absolute average class difference, and as the proportion of comparisons differing by <0.5 classes is summarised in Table 23. The results obtained are very similar to those using Techniques 1 or 2 and the gain in class agreement associated with boundary synchronisation is small.

Table 23. Comparison of agreement between classifications at different levels of harmonisation of class boundaries using Technique 3

Level	Measure	DE	UK	FR	WL	AT	average
original class boundaries	abs_av_diff	0.51	0.41	0.39	0.38	0.40	0.42
	% ± 0.5 class	57	67	66	68	67	65
harmonised boundaries (bias ± 0.25 classes)	abs_av_diff	0.51	0.40	0.37	0.35	0.40	0.41
	% ± 0.5 class	58	67	72	76	67	68
synchronised boundaries (zero bias)	abs_av_diff	0.49	0.39	0.31	0.33	0.38	0.38
	% ± 0.5 class	61	69	77	74	72	70

Table 24 provides the results of a boundary-specific kappa analysis for the 215 commonly assessed sites applying the national class boundaries associated with varying degrees of harmonisation. As with Techniques 1 and 2 there is very good agreement between classifications at both class boundaries with harmonisation and synchronisation providing quite marginal gains. In fact, in Technique 3 the fully synchronised class boundaries indicate a degree of over adjustment when compared to class boundaries that are adjusted only to the point where the mean pairwise difference for each method lies within the 95% confidence limits of zero bias.

Table 24. Result of comparing agreement between national methods at GM and HG boundaries in Technique 3 using a multi-rater kappa coefficient.

	Good-Moderate			High-Good		
	kappa	L 95CL	U 95CL	kappa	L 95CL	U 95CL
original	0.767	0.725	0.810	0.721	0.685	0.757
harmonised	0.784	0.742	0.826	0.721	0.685	0.757
within 95%CL of 0	0.812	0.770	0.854	0.758	0.722	0.794
synchronised	0.801	0.759	0.843	0.753	0.717	0.788

7. Discussion

7.1 Comparison of comparability of national methods based on different assessment techniques

All three Techniques of comparison discussed here lead to similar conclusions, namely that all the national methods are well correlated, there is a high level of class agreement that is relatively insensitive to the manipulation of class boundaries, and the GM boundary of the WL method is most strongly out of place and would need raising in order for all methods to display a comparable level of

ambition (Table 25). In terms of boundary synchronisation the different Techniques for comparison do not yield identical results. In particular, when the greatest change to a class boundary implied by any one of three approaches discussed is < c.0.2 class equivalents compared to its original position it is likely that another approach will suggest that an even smaller change is required, or should even be in the opposite direction. However, when any one technique implies a change to a class boundary of >c.0.2 classes compared to its starting position it appears safe to assume that other techniques for measuring comparability would indicate the same magnitude and direction of change. When there is an interest in boundary synchronisation it may be best to treat implied boundary changes of <c.0.2 classes in any one method relative to its original position as being specific to the Technique of comparison that is employed and to disregard changes this small. Technique 3 also differs from 1 and 2 in concentrating the changes in a single misplaced boundary whilst leaving most other boundaries with small changes. Potentially this may lead to over adjustment of one boundary to achieve a low overall level of bias.

Table 25. Summary of harmonised boundaries (bias within 0.25 classes of global mean) in the five methods assessed as revealed by three Techniques of comparison. The original national class boundaries are given for reference.

Level	Boundary	DE	UK	FR	WL	AT
original class boundaries	HG	0.75	0.80	0.85	0.93	0.875
	GM	0.55	0.60	0.72	0.61	0.625
Technique 1	HG	0.75	0.81	0.85	0.93	0.85
	GM	0.55	0.60	0.70	0.66	0.625
Technique 2	HG	0.75	0.80	0.85	0.93	0.86
	GM	0.55	0.60	0.71	0.67	0.625
Technique 3	HG	0.75	0.80	0.85	0.93	0.875
	GM	0.55	0.60	0.72	0.68	0.625

Note: Cells highlighted in blue indicate boundary reduction (too precautionary); cells highlighted in red indicate boundary increase (too relaxed).

7.2 Strengths and weaknesses of different techniques

Technique 1 allows indirect comparison of methods via an independent common metric that acts as a yardstick against which the position of national class boundaries can be assessed. This technique is the only option that can be considered in the absence of a set of commonly assessed sites. Thus if the sampling methods of countries A and B exclude the application of the method of B to data of A and vice versa there will be no commonly assessed sites. Note that if a dataset contains small numbers of gaps because certain national methods cannot be applied to some sites the imputation technique proposed by Pešta & Horký (2010) can be used to fill these gaps and increase the availability of commonly assessed sites. Following this Techniques 2 or 3 may be applied.

Technique 1 provides a solution to the absence of commonly assessed sites. The regression approach allows the bias of different national methods at individual boundaries to be established quickly, and for the level of adjustment of boundaries that is needed for harmonisation or synchronisation to be easily determined. On the negative side it is difficult to accurately assess class agreement other than through the use of the precision in the regression models which can only offer an estimate of average class difference. In this approach there is no possibility to compare EQRs directly unless a synthetic dataset is constructed as described in Annex V which takes a synthetic set of common metric values and uses the regression model and its precision to simulate a corresponding set of realistic EQRs for each national method. The major obstacle to this approach is likely to be the need to develop a common biological metric(s) with which all national methods are satisfactorily correlated. Birk & Willby (2010b) offer a possible solution to this, although it is reliant on the ability of all methods to assess some sites, even if only approximately. A common metric is effectively a compromise between a number of separate methods and it is likely that the common metric will include derivatives of one or more national methods. If the sampling methods used by different countries are very divergent there is the risk that estimates of values of a common metric are influenced by the data source unless common metrics can be expressed in a dimensionless form that is independent of sampling method or effort.

Although Technique 1 can be applied in situations where sites have been commonly assessed there would need to be a clear justification for doing so since this would require the construction of a common metric before classifications could be compared. However, if a simple common metric is readily available it may prove instructive to establish where national boundaries lie on this metric. Thus, for example, we might chose to determine where a series of countries locate their GM boundary in relation to light transparency or maximum depth of colonisation when assessing the status of lake macrophytes.

Technique 2 was proposed for use in Annex V because it follows the same basic mechanism of Technique 1 but allows the direct comparison of classifications via an independent pseudo-common metric (PCM). Technique 2 is ideal where there is a large common dataset, covering a wide gradient of ecological quality, and where all sites are assessed by at least four different methods. In this situation it is straightforward to establish the magnitude and direction of deviation from the global mean boundary on the PCM scale and to determine the adjustments needed to harmonise or synchronise the boundaries of each national method. The disadvantages of Technique 2 are common with those of Technique 1 in the sense that class agreement can only be inferred through the precision associated with the regression model. To derive accurate measures of class agreement through direct comparison, the commonly assessed data used in Technique 2 must be analysed according to the pairwise comparison approach described under Technique 3. In contrast to Technique 1 Technique 2 is amenable to assessment of class agreement by kappa analysis because it utilises a set of commonly assessed sites. The main constraint on the use of Technique 2 is those situations involving either small datasets of variable quality and where data is not normally distributed, or exercises involving only two or three countries. In such instances simple linear regression cannot be used to establish a model for comparing class boundaries, while the pseudo-common metric is either meaningless or strongly biased by the small number of countries from which it must be calculated. In such cases methods should be compared according to Technique 3.

Technique 3, based on pairwise comparisons of standardised EQRs, has the clear advantage that it will function satisfactorily regardless of the amount of available data and the number of countries being compared. Thus it can be applied in cases where there are only two or three countries (which would preclude the use of Technique 2) and can be used when the available range of ecological quality is narrow. Within Technique 3, it is also simple to calculate metrics describing class agreement, such as the average absolute difference between pairs of values, which can only be estimated in Techniques 1 and 2. Essentially Technique 3 is a universal approach that can be applied in any situation where sites have been commonly assessed.

The disadvantages of this technique are that harmonisation of boundaries must be achieved through an iterative process that might prove more time consuming than harmonisation of boundaries via regression models. In general the initial process of benchmark normalisation ought to restrict the incidence of extreme bias in the position of class boundaries. This should allow boundaries to be adjusted to provide bias of <0.25 classes within relatively few iteration steps. However, achieving synchronisation of class boundaries to eliminate bias could prove time consuming unless it is automated by a simulation procedure. Synchronisation also appears to be heavily concentrated on a single boundary rather than being distributed across several boundaries as in Techniques 1 and 2 and might therefore lead to over-adjustment. The other disadvantage of this Technique is that it is more sensitive to the proportion of commonly assessed sites awarded the same class, especially if these sites belong to the lower classes. If the difference in EQR between pairs of values is assessed across the full dataset and that data contains a large proportion of sites that are commonly assessed as Poor or Bad status this could dominate the measure of class agreement. One option to avoid this would be to systematically exclude those sites commonly assessed by all countries as belonging to the lowest two classes before undertaking the analysis. In the case of the example considered here no sites were commonly assessed as being at Poor or Bad status and only 14% of sites were commonly assessed as being at Moderate or worse status.

7.3 Correcting bias between national classifications

All three Techniques clearly reveal the level of bias between national methods whether in relation to a global mean boundary (Techniques 1 and 2), or via a comparison of the pairwise differences between one method and all other methods (Technique 3). In Technique 3 the bias is less specific to a given boundary because it is measured across all sites, but is most dependent on where either the HG or GM boundary is located. In Technique 1 and 2 the measure of bias is specific to a given point but the relationship between the national EQR and the biological- or pseudo-common metric is shaped by sites covering the range of available ecological quality.

Bias reflects the level of ambition in national classifications. At some position between their original locations and their synchronised positions (where bias is eliminated in respect to that Technique) we could conclude that the level of ambition was sufficiently similar between all methods for it to be regarded as effectively equal. At this stage class boundaries could be considered as having been harmonised. Bias of 0.25 classes either side of the global mean was considered to represent an acceptable level of difference in bias between national methods in Round 1 of IC based on the various uncertainties associated with different steps in the process of comparison. While a bias of

±0.25 classes is somewhat subjective there is nothing revealed by our present analysis to support changing this value. Indeed our attempts to synchronise class boundaries through the different Technique that are discussed here suggest that when class boundaries start off within ~0.2 classes of the global mean the decision on how much such boundaries should move in order to eliminate bias is peculiar to the Technique used for comparison. On such grounds it seems reasonable to treat any boundaries that already lie within 0.25 classes of the global mean as being as reasonably harmonised as the process of comparison will permit.

We think it is instructive to determine where boundaries would need to lie to achieve complete synchronisation but we are wary of making synchronisation a requirement of boundary harmonisation for several reasons. Firstly, the national EQR v common metric regressions are specific to a given common dataset, combination of countries, and range of ecological quality. Unless a common dataset is already very large (thousands of commonly assessed cases) it seems likely that re-sampling of national datasets to generate a new common dataset, adding further countries or changing the range of ecological quality that is represented will change the national EQR v common metric regressions sufficiently to provide a new and different target for boundary synchronisation. At a coarser level, however, countries that look precautionary or relaxed at the level of 0.25 class units in one dataset will retain this status in any dataset. Secondly, tolerating a small level of bias recognises that there are various additive sources of uncertainty at preceding steps in the IC process, which, while definable, would probably be difficult to quantify accurately, and are specific to a given dataset, and the type of comparison being undertaken. Note that these uncertainties are considered to be separate from the prediction error associated with the national EQR v common metric regression (which in most cases will be much larger than 0.25 class equivalents). Thirdly, when implied adjustments to the position of a boundary are small it appears that such adjustments are often specific to a particular Technique of comparison, rather than being common to all approaches. Finally, the evidence from the present study is that full boundary synchronisation leads to a trivial increase in agreement amongst classifications when measured using direct pairwise comparisons or by the use of a multi-rater kappa coefficient. This is because a very small number of misaligned boundaries contribute the majority of the bias. To achieve complete synchronisation the additional boundary changes that would be required to other methods generally amount to 0.02-0.03 EQR units (~10-15% of a class width on the national scales) and it is questionable whether changes this small can be justified.

7.4 Class agreement

In quantifying class agreement we are effectively trying to obtain a measure of confidence that two or more national methods will award the same class to a given site. Having reduced, or even eliminated the bias in classifications by harmonisation or synchronisation of boundaries, the level of agreement between classifications will still depend primarily on the initial relatedness of EQR values. Thus we might envisage a situation where several methods are highly correlated but, with the original class boundaries in place producing a modest bias, various measures still indicate greater class agreement than in a second comparable situation where any bias between methods has been eliminated by synchronisation but those methods are only weakly (but still significantly) correlated.

We can also say that for a given initial level of bias between methods adjustment of class boundaries will have a proportionally greater effect on class agreement when methods are highly correlated, because the bias is likely to be large relative to the error, than when methods are only moderately correlated and the bias will be small relative to the error. However, it is arguably also the case that when methods are highly correlated at the outset the likelihood of significant bias between them is then smaller than when methods are weakly correlated.

It is problematic to obtain measures of class agreement through regression-based approaches (Techniques 1 and 2). One option suggested here is to use a measure of precision (in this case the mean absolute residual) transformed to a class equivalent. In Technique 2 where there is a set of commonly assessed cases, the raw material exists for establishing class agreement through the approach established in Technique 3. A kappa analysis can also be conducted directly within Technique 2 to establish agreement between classifications at different levels of boundary harmonisation. In Technique 1 it will be difficult to establish class agreement, other than through the use of estimates of model precision, when this method is applied in situations where data from one country cannot be assessed by all other countries (i.e. there are no commonly assessed sites). A solution to this problem we proposed in Annex V was to generate a test dataset using a synthetic set of benchmarked common metric values and to then simulate a realistic set of national EQR values for these sites based on the national EQR versus common metric relationship and its associated precision. This would provide a set of values which could be treated as commonly assessed sites and then subjected to Technique 3 and kappa analysis. Technique 3 provides the best approach for obtaining various simple measures of class agreement, such as the average absolute difference between classifications or the proportion of classifications differing by a fixed amount. Other similar metrics could be devised but all will be highly inter-correlated. These measures should be reported alongside a statement indicating the maximum and minimum level of bias among methods being compared.

A pertinent question is whether we should manipulate class boundaries though harmonisation so as to reduce bias (as suggested here), or so as to maximise class agreement. Pont & Delaigue (2010) proposed changing boundaries in order to optimise the kappa value across a series of methods at different class boundaries in turn. We agree that kappa offers a useful measure of class agreement but consider that the primary aim should be to reduce bias between classifications to achieve a comparable level of ambition, and to then report the kappa value (and other measures of agreement) associated with those boundaries. In the present example, despite a significant variation in bias at the GM boundary between different methods (revealed by all three potential Techniques), a kappa analysis provided no justification to alter the boundaries from their original values, since the kappa value with the original boundaries in place was not significantly different from the value with boundaries completely synchronised. In all cases, despite the differences in bias, the level of agreement between classifications distinguishing between <Good and ≥Good status, or <High and High status achieved values of kappa equivalent to 'very good' agreement. Nominally, regardless of where the class boundaries are positioned (at least between their original and synchronised positions), there is a very good chance that all five methods would agree whether a site was <Good or ≥Good. This is because kappa is most sensitive to the general relatedness of methods and will only respond to changes in the status of the very small proportion of sites whose EQRs lie on opposing sides of the <Good or ≥Good boundary. When methods are already highly related, only a

small proportion of sites will be affected by a boundary change. This points to a further weakness in the use of kappa analysis as a basis for manipulating class boundaries. As dataset size decreases the average interval between ranked EQR values in different methods will increase. In order to raise the kappa value sites must be relocated across a class boundary. Thus, when dataset size is small, this implies making a proportionally greater change in a class boundary in order to capture those sites in the 'wrong' class, than when dataset size is large. This could serve to further increase the level of bias between methods measured at a class boundary. It is also a feature of kappa analysis that the increase in kappa that is achieved through changing a class boundary will depend on the relative composition of sites that are commonly assessed by all methods as being, for example, <Good or ≥Good. If a common dataset contains a high proportion of sites commonly assessed as <Good a change in the GM boundary will effect a proportionally smaller number of sites and will have a proportionally smaller effect on the kappa value than in a dataset where the proportion of sites commonly assessed as <Good is small. As a basic precaution when applying this analysis it would be advisable to ensure that within those sites being compared the proportion of sites commonly assessed by all methods as being <Good or ≥Good status is roughly equal.

As part of the present analysis we sought to optimise the value of kappa across HG and GM boundaries independently by iteratively changing the national class boundary values, ignoring the associated bias. Through this approach it was possible to achieve values of kappa that were slightly higher than those obtained using the class boundaries that minimised the bias between methods. However, when the kappa-optimising boundaries were imposed on Techniques 1 to 3 (Table 26) this was at the expense of an increased bias at the upper class boundaries. In the case of Techniques 1 and 2, applying the boundaries that maximised the kappa coefficient resulted in one method deviating by more than 0.25 classes from the global mean at the GM boundary while several methods fell on the edge of the harmonisation band for one or other boundary. In the case of Technique 3 the resulting bias using the kappa-optimising boundaries was much smaller and only narrowly exceeded the 95% confidence limits of zero bias. Inspection of Table 26 shows that the solution for the synchronised boundaries in Technique 3 is on average much closer to the kappa-optimising boundaries than are the synchronised boundaries from Techniques 1 and 2. Crucially, all three approaches share the same conclusion that the GM boundary in the WL method is by far the most deviant, requiring an increase of between 0.11 – 0.18 EQR units for boundary synchronisation or kappa optimisation.

It is not clear what would represent an equivalent to boundary harmonisation in an approach based directly on manipulating class boundaries in order to change the value of kappa. Possibly a suitable approach would be to derive the optimum value of kappa and its associated confidence limits following the method of Pont & Delaigue (2010). A set of boundaries closer to the initial values could then be established which first yielded a kappa value lying within the confidence limits of the optimum value. At this stage boundaries could be considered harmonised in the sense that further changes would not result in statistically greater class agreement as measured by kappa. In general we conclude that it will be preferable to use kappa, where possible, as part of a reporting suite of variables to reflect the confidence that classifications will agree, rather than as a tool for directly manipulating class boundaries to secure maximum agreement since this carries the risk of over-adjusting class boundaries in some Techniques with a subsequent increase in bias.

Table 26. Summary of application of Kappa-optimising boundaries and bias minimising boundaries from Methods 1, 2 and 3 on class agreement and bias metrics

Method	DE	UK	FR	WL	AT	kappa	max. ± bias	abs. class diff	% ± 0.5 class
Kappa-optimising boundaries									
HG	0.700	0.800	0.880	0.960	0.835	0.851	-0.258	0.374	69.7
GM	0.520	0.620	0.700	0.760	0.675	0.769	0.309		
Method 1 - synchronised									
HG	0.726	0.854	0.887	0.968	0.812	0.808	0.000	0.363	71.0
GM	0.516	0.594	0.657	0.717	0.639	0.749	0.000		
Method 2 - synchronised									
HG	0.727	0.832	0.880	0.960	0.812	0.827	0.000	0.363	70.6
GM	0.522	0.575	0.664	0.725	0.658	0.745	0.001		
Method 3 - synchronised									
HG	0.710	0.810	0.860	0.970	0.855	0.800	-0.010	0.379	70.5
GM	0.540	0.600	0.700	0.790	0.675	0.753	0.013		

7.5 Reporting and judging the acceptability of values for comparability metrics

The Techniques discussed here allow values to be generated for the two key components of comparability;

- **Bias** – the magnitude and direction of deviation by one method relative to the global mean of all methods at each upper class boundary, expressed in class equivalents.
- **Class agreement** – the confidence that two or more classifications will report the same class for a given site, as assessed by the average absolute difference between pairs of values, the proportion of classifications differing by an agreed amount (e.g. half a class), or the multi-rater kappa coefficient.

In some approaches presented here (Technique 1) it is not possible to generate all potential metrics relating to class agreement without resorting to synthetic datasets. In this case it is assumed that the various metrics are sufficiently correlated for a partial assessment of class agreement to suffice.

We suggest that a global set of standards could be used to judge acceptability, as illustrated in Table 27. Bias and class agreement form the basic criteria for judging comparability. There are two overriding thresholds in this approach:

(i) Bias **must** be <0.5 class equivalents regardless of the type of comparison. Beyond this level a national class boundary is closer to another class boundary than the one being compared. Hence this is the maximum acceptable bias in any type of comparison. The 'desirable' level for bias is <0.25 class equivalents. Within this range we would conclude that national methods exhibited a similar level of ambition subject to the overall constraints imposed by

the comparison exercise itself. Arguably, if indirect comparisons are being undertaken (Method 1) in which there is a lower inherent level of uncertainty this should form the maximum acceptable bias.

(ii) Mean average absolute class difference across the methods being compared **must** be within one class equivalent (or an appropriate proportion of pairwise comparisons must differ by less than 0.1 EQR units). A minimum kappa value of 0.4 must be achieved, signifying a moderate level of agreement between classifications. These metric values reflect class agreement and represent the lowest tolerable level of coherency of methods. Arguably at lower levels of agreement between methods the uncertainty in the estimate of bias is so large as to place the overall result in question (i.e. the bias may appear low simply due to the poor relatedness of methods). The 'desirable' level for class agreement is an average absolute class difference of <0.5 units for all methods, although whether this can be routinely achieved in direct comparisons for some BQEs is currently uncertain.

The overall acceptability would be judged on both indicators, with failure of either leading to rejection. Failure on the basis of bias would be redeemable by harmonisation of class boundaries. Failure on the basis of class agreement would require a more fundamental assessment of the methods being compared and an analysis of the causes of random error in the relationship between methods. In principle, the initial requirements of a minimum level of relatedness between methods before comparison is attempted should reduce the likelihood of an exercise failing on the basis of poor class agreement.

Table 27. Scheme for judging acceptability of comparisons based on levels of class bias and class agreement

		Class agreement		
mean av. abs. class diff*		<0.5	<1.0	>1.0
mean % ±0.5 classes$^{\phi}$		>50	30-50	<30
kappa coefficient†		>0.6	0.6-0.4	<0.4
Bias #	< ±0.25 classes			
	< ±0.5 classes			
	> ±0.5 classes			

* Measure applicable to Techniques 1, 2 and 3. $^{\bullet}$ Measure applicable only in Technique 3. † Measure applicable in Techniques 2 and 3. $^{\#}$ Refers to the national method exhibiting the greatest bias

In the scheme presented in Table 27 an exercise will pass ('green light') if both bias and class agreement metrics fall inside the desirable range. At this point we would conclude that all methods share a similar level of ambition and there is a good to very good chance that all methods would classify any site the same. Exercises where either (not but both) bias or class agreement fall outside desirable but inside acceptable limits could be considered to achieve a qualified pass ('amber light') subject to a supporting statement justifying the reasons why an exercise cannot achieve the desirable thresholds. In the case of class agreement, for example, datasets may be very small or data

of limited quality for some water body types, or, for some BQEs, the associated taxonomy may be poorly resolved or the tradition of biomonitoring poorly established. Thus, such grounds might be appropriate to justify a qualified pass for phytoplankton in lakes but not for benthic macroinvertebrates in streams. Particular scrutiny should be applied when one or more methods in an exercise state that they cannot achieve desirable limits for bias. Note that this applies only to those methods whose boundaries are too relaxed. Having benchmarked national EQRs before assessing their comparability, typological differences between countries cannot be invoked as grounds for explaining differences in bias between methods. An exercise that did not achieve acceptable limits for bias and/or class agreement would be considered to have failed and for the national methods to not be sufficiently comparable.

8. References

Birk S, Willby NJ (2010a) The definition of alternative benchmarks in intercalibration – outline of general procedure and case study on river macrophytes. Final report. University of Duisburg-Essen, University of Stirling. Essen and Stirling: 22pp.

Birk S, Willby NJ (2010b) Towards harmonization of ecological quality classification: establishing common grounds in European macrophyte assessment for rivers. Hydrobiologia 652:149-163.

Borja A, Josefson AB, Miles A, Muxika I, Olsgard F, Phillips G, Rodríguez JG, Rygg B (2007) An approach to the intercalibration of benthic ecological status assessment in the North Atlantic ecoregion, according to the European Water Framework Directive. Marine Pollution Bulletin 55:42-52.

Horký P (2010) Czech comments on Annexes III and IV of the Intercalibration Guidance. April 2010. T.G. Masaryk Water Research Institute. Prague: 6pp.

Pešta M, Horký P (2010) Alternative approach in Annex V. June 2010. T.G. Masaryk Water Research Institute. Prague: 6pp.

Pont D, Delaigue O (2010) Intercalibration process: kappa method. Version 1.2. Cemagref, Unité de recherche hydrosystèmes et bioprocédés. Antony Cedex: 10pp.

Van den Berg MS (2010) Alternative comparability criteria for intercalibration phase 2. RIZA. Lelystat: 4pp.

Willby NJ, Birk S (2010a) Definition of comparability criteria for setting class boundaries - Outline of the procedure to compare and harmonise the national classifications of ecological status according to the WFD intercalibration exercise. University of Stirling, University of Duisburg-Essen. Stirling and Essen: 25pp.

Willby NJ, Birk S (2010b) Report on the theoretical evaluation of the proposed alternative approach of the Netherlands. University of Stirling, University of Duisburg-Essen. Stirling and Essen: 21pp.