Annex

Summary of comments received to the draft "Guidelines for indicator development", 9. February 2010. Page numbers refer to the original document (version 2 of the guidance)

No.	Person/Institute	Comment	Reply
1	Alberto Basset	Email comment: I would suggest to leave the door open in the guidance to the selection of metrics describing non-taxonomic 'indicative parameters' to be included in the biological assessment methods. I can add a paragraph on this point at page 7 in the section Compliance to... if you think that it is relevant	We added on page 7: "Common metrics derived from taxonomical as well as non-taxonomical data shall cover all required parameters ..."
2	Alberto Basset	Email comment: in many sections of the guidance a paragraph or a table listing weaknesses of the proposed solution, warnings and recommendations would be very useful for the Guidelines Stakeholders. Specific recommendations according to the water body category may also be required and useful	We see this guidance mainly as a cook book for WISER scientists and don't think it should be used by stakeholders.
3	Alberto Basset	Page 10, B16: i.e, that the dose response paterns are not significantly different? If it is the meaning it is better to state it more clearly. Otherwise it need probably to be added.	We added:" , i.e. the dose-response patterns are not significantly different"
4	Alberto Basset	Page 11, B17: It is probably useful to anticipate here that multimetric index must have an higher sensitivity and robustness combined (larger proportion of explained variance) than simpler metrics. It must hold	We added: "In general, the multimetric index is more sensitive and robust than the single metrics."
5	Alberto Basset	Page 6, B2: It would be useful to add a section specifying when intercalibration is not possible; country specific types? i.e., Italian crater lakes, ? Venice lagoon? Different habitat type sampled (littoral vegetation vs benthic sediments? And how overcome these limitations	We added: "Among assessment methods that are conceptually different, or focus on dissimilar pressures or water types intercalibration cannot be accomplished. In these cases the guidance requests the use of alternative approaches such as on-site comparisons, i.e. comparing the classification results of the various national methods applied to the same water bodies. A calibration of national class boundaries against comparable gradients of pressure may also be a feasible option."
6	Alberto Basset	Page 7, B5: I think that the unique focus on a taxonomic is limiting, both for the assessment and for the large scale intercalibration. Composition has better to be given as general being referred to taxonomic, functional or size composition. The focus is in contrast with examples given in a later paragraph dealing with metrics based on functional groups	We agree with this view, however the WFD states "taxonomic composition" in its normative definitions. This is why we have used this term in the Table.
7	Alberto Basset	Page 8, B11: and ecological theory and modelling. (The focus on metrics ecologically sounded has to be reflected in all section of the guideline).	We changed accordingly.
8	Alberto Basset	Page 9, B13: A second general comment is that I think that we should quote in the guidelines the weaknesses of these 'practical agreements'. It is general and in my opinion required to most sections of the guidelines.	We added: "However, the implications on the quality and precision of the assessment needs to be thoroughly evaluated."
9	Angel Borja	Email comment: Another important issue is that of the differences between 'multimetric' and 'multivariate' methods, regarding the guidelines. Both approaches are different, and this is why I tried to highlight this in the text, including the word 'multivariate' in some of the paragraphs.	We have taken over your additions on the "multivariate" indices at the beginning of the guidance. As far as we know this approach is only used in the M-AMBI, but it would be worthwhile that other experts test it for their BQEs.
10	Angel Borja	Email comment: Finally, in the intercalibration exercise we are comparing methods which have been published in peer-review journals, compared against pressures, used by scientists other than the original authors, with	Yes, this is true. However, the IC guidance implements the "IC feasibility checks" that aim at displacing methods that are not fully compliant/badly tested.

No.	Person/Institute	Comment	Reply
		unpublished methods (or published as reports in local languages), without being scrutinized externally. In my opinion this can be dangerous, because probably we are trying to compare things that are not comparable	
11	Angel Borja	Email comment: I don't understand why to use these 'common metrics' that, finally, are another 'new' method not national. If they are the mirror in which the national methods must be compared (it seems that they are 'the law') they should be taken as the European method for the BQE and not the others, if not we are doing double (even triple) work in developing and intercalibrating methods.	The new IC guidance also regards the direct pair-wise approach for boundary comparison. However, WISER was asked to support in common metric development, not the carrying out the entire IC exercise.
12	Angel Borja	Email comment: Other problems can come from the different method, surfaces, replicates, sieves, etc., used in the sampling.	Sure, and it needs to be tested in how far these differences hamper the use of common metrics.
13	Angel Borja	Email comment: Regarding the document itself, in my opinion, most of the disagreements between methods probably come from the reference conditions selection, referred to the type studied. Nothing (or very few) is said in the document.	We have added a passage about the harmonized approaches to derive reference conditions for the intercalibration exercise.
14	Angel Borja	Page 10, A15: This requires a good knowledge of the pressure history and evolution: this is not static, hence, changes in time from metrics could (should) be due to changes in pressure	We added: "This may require knowledge of the pressure history and evolution to avoid that changes in time from metrics are not due to changes in pressure."
15	Angel Borja	Page 11, A18: What does it happen with methods which are not multimetric but multivariate? In this case, the weight of each metric is not the same always: it changes depending on the case. This is why, normally, it can be more robust, because it tends to capture the highest variability	If we understand your approach correctly, the weighting is always depending on the underlying dataset. This would mean a differing common metric depending on the dataset, which seems not to be feasible. However, some additional explanation for you would help in understanding the implications of your suggestion.
16	Angel Borja	Page 12, A21: "control natural variability". Comment: I don't understand this sentence: in theory, any metric used should be as more independent as possible from natural variability.	The sentence refers to prediction systems. We have rephrased the sentence for clarification ("to correct metric results for the effects of natural environmental variability").
17	Angel Borja	Page 6, A3: Is this possible? E.g. comparing richness when taking only one replicate of 0.1 m2 or when taking 3 replicates of the same surface	The richness cannot be compared directly, but as EQR, if the reference conditions are comparable (derived from reference data sampled with the same technique, screened by harmonized reference criteria).
18	Angel Borja	Page 7, A6: This is confusing to me: if common metrics shall cover all metrics within the multimetric indices, finally any metric used is a common metric. To me this is nonsense	We re-formulated: "Common metrics shall cover all required parameters combined to a multimetric index ..."
19	Angel Borja	Page 8, B10: Evaluation of robustness using signal/noise ratios needs to take into account (be weighed according to) the intrinsic spatial-temporal heterogeneity of different water body categories.	We added: "The evaluation of robustness using the s/n ratios generally needs to take into account the intrinsic spatial-temporal heterogeneity of different water types."
20	Angel Borja	Page 9, A12: I agree with the comment of Alberto to this sentence: this is an important weakness, because in some cases (e.g. opportunistic/sensitive spp. Ratios) the differences can be radically high. Hence, the common metric at family level could be not useful	We added: "However, the implications on the quality and precision of the assessment need to be evaluated thoroughly."
21	Dorte Krause-Jensen	Page 17, dkj6: I suggest using the term 'boundary' (instead of threshold) in order not to confuse it with 'ecological thresholds representing the level of a driver that cause abrupt changes in an ecosystem,	done
22	Dorte Krause-Jensen	Page 4, dkj1: It is not necessarily a taxa list. For marine vegetation it can also be e.g. the depth limit of a key species or of a characteristic community. table 1 also shows that there is no demand for the parameter 'taxonomic composition' for macroalgae and angiosperms in coastal waters.	We have changed the sentence into "Assessment methods (often referred to as "classification method") translate biological information of a water body into an ecological status class (ranging from high status to bad status). "

No.	Person/Institute	Comment	Reply
23	Dorte Krause-Jensen	Page 9, dkj2: Same comment as on p. 4 – since for marine vegetation there is no demand for collecting species lists.	We changed accordingly.
24	Dorte Krause-Jensen	Page 9, dkj3: In the Baltic GIG and NEA GIG and in collaboration with a project called 'MOPODECO' we currently compile data on 'total macroalgal cover' and associated information on water chemistry from Nordic countries and countries surrounding the Baltic Sea. These data will be stored in a separate database. We hope to be able to do the same exercise for the variable 'eelgrass depth limit'. I think both these metrics can be considered common metrics.	Great! We changed accordingly.
25	Dorte Krause-Jensen	Page 9, dkj5: Comment on section "development of assessment systems": I find this section confusing and to a large extent repetitive of the first section. I suggest shortening it considerably so that it just mentions which elements are similar to the first section and which elements are necessary additional components. For the work I'm involved in (marine vegetation) I see most of the work in the two sections as being completely identical. Or else I have missed a point?? For example both sections include a description of 'normalisation of metrics'. - I think it would be more logic to provide the full description in the first section	We agree that the division into two main chapters (dealing with common metrics and with the development of assessment systems) is quite arbitrary and overlapping. However, both tasks will be performed at different times by most workpackages (common metric development: spring 2010; assessment system development: beginning in summer 2010). We therefore think that two separate methodological descriptions are useful. Although both descriptions are overlapping we think it is more practical to have them separate and leave both descriptions complete, so that they can be used as a "cook book". We have added an introduction (an extension of the original chapter "aims") to the deliverable to clarify the structure and the intentions.
26	EMU	Page 4, K1: How is the common assessment method connected with the national methods and types? How to overgo from national type to intercalibration type?	These questions are out of the scope of this guidance. Please refer to the new IC guideline.
27	EMU	Page 7, K2: Only macrophytes, because phytobenthos is one part of macrophytes.	No, the definition of macrophytes is "visible by the naked eye", and phytobenthos taxa are mostly not.
28	EMU	Page 7, K3: We added disturbance sensitive taxa (Lobelia, Isoetes) in LCB3 type of lakes.	Though certainly useful, it is not especially required by the WFD normative defs. Therefore, we did not enter it in the table as a separate column; basically it falls under the term "disturbance sensitive taxa".
29	EMU	Page 7, K4: In this context what does the pollution mean? Is it a toxic pollution or also a nutrient loading? In the last case we have an indicative parameter (filamentous macroalgae).	The definition is up to the method developers.
30	EMU	Page 7, K5: What does this word mean? How should the heteroscedasticity be visually inspected?	We added: "heteroscedasticity, i.e. an inhomogeneous variance of the residuals that can simply be observed in the residuals' plot.
31	EMU	Very few pressure variables are available. Time distance between stressor impact and metric chance may be very long.	However, the demonstration of the pressure-impact-relation is one of the important requirements also highlighted by the IC guidance.
32	Gwendolin Porst	Page 12, P38: Pressure? Depending on terminology adopted. This one I think is borrowed from chemistry: stressors are the env. parameters reflecting the intensity of a pressure. Otherwise you need to state that the terms pressure and stressor are interchangeable.	We sticked to the term "stressor" to be consistent with the IC guidance.
34	Gwendolin Porst	Page 4, P1: You use the term "system" in the rest of the document. I would adopt a consistent terminology to avoid confusion	harmonized to the term "methods"
35	Gwendolin Porst	Page 4, p2: This WFD terminology is somehow biologically odd (angiosperms are also macrophytes). I prefer the use of the term BQEs	We would like to stick with the WFD terminology, differentiating between macrophytes (freshwater) and angiosperms (salt and brackish water).
36	Gwendolin Porst	Page 5, p4: Provocative question to you: or a common political compromise?	Well, science develops the options that political decision making is opting for.
37	Gwendolin Porst	Page 5, p5+p7: Types based on very broad categories of very basic parameters	Added "broadly defined common intercalibration types"

No.	Person/Institute	Comment	Reply
38	Gwendolin Porst	Page 7, p11: needs explanation on what you mean here	We added "intercalibrated at least for some indicative parameters" instead of "(partly)".
39	Gwendolin Porst	Page 7, p13: This is a sensitive point. Would you honestly refer to a correlation of R2=0.5 as high? Of course not. But it might be taken by agencies as a rule forgetting that another 50% of the variance is not explained by the relationship.	There are exercises where 0.5 is easily met, others do not reach this level. The general intercalibration criteria for comparability should flag the quality of this relationship.
40	Gwendolin Porst	Page 7, p14: Ideally you want to include a formal inference about regression (analysis of residuals). But we know that this will invalidate almost all the work done so far….	Why will this invalidate the work done so far?
41	Gwendolin Porst	Page 8, p15: ..are equally well reproduced along the whole region where the metric is applied	We changed accordingly.
42	Gwendolin Porst	Page 8, p16: Low compared to what? Maybe lower than spatial (inter sites) variability?	This is meant in absolute terms.
43	Gwendolin Porst	Page 8, p18: Not clear what you mean here	We added ", i.e. constantly increasing or decreasing across the gradient of stressors."
44	João Carlos Marques	On the other hand, I feel that the whole guideline is clearly focused on metric types and metric able to provide ‘‘snapshots” of given ecosystem structural properties, but providing little or no information at all about ecosystem functioning	We clarified that functional metrics (such as feeding type composition, body sizes) are most welcome metrics. We added on page 7: "Common metrics derived from taxonomical as well as non-taxonomical data shall cover all required parameters ..." and amended Table 1 accordingly.
45	João Carlos Marques	Page 8, A8+A9: The question of robustness in relation spatial variation is correct, but the scale deserves here a reference. A comment should be introduced regarding this problem of spatial scale. Fort instance, in transitional waters the spatial scale of variation is very much dependent on shape, hydraulics etc. The same applies to temporal scales, which must be reflected in the methods applicable in transitional waters.	We added: "The evaluation of robustness using the s/n ratios generally needs to take into account the intrinsic spatial-temporal heterogeneity of the different water types. If the BQE shows strong spatial and/or temporal variability (e.g. phytoplankton) the common metric may be water type-specific and/or calculated from fixed sampling season data only."
46	João Carlos Marques	Page 8, B7: Although this constitutes the theoretical perspective, something should be said about practical difficulties. For instance, in transitional waters, due to natural variability, noise in relation to the stressor to be assessed is generally quite high.	We added: ". However, level of noise differs between water categories and types and is, for instance, relatively high for transitional waters that show large natural variability. "
47	Ken Irvine	Page 13, K40: Replace "watershed" by "catchment".	done
48	Ken Irvine	Page 13, K42: "The gradient may be a continuous measure or may be classified into five classes or even into the two classes “unstressed” and “stressed”, only". Comment: In my view this should not be endorsed as an acceptable approach. It smacks of appeasement to some already established views.	We clarified (here and in other sections of the document) that (1) this approach is the "last exit" and should only be used if no sufficient data on environmental gradients are available; (2) the classification into "stressed" and "unstressed" must be based on environmental data (to avoid "appeasement").
49	Ken Irvine	Page 13, K44: "Analysis of the gradient may be restricted to a single stressor or may include the impact of multiple stressors". Comment: Do multiple stressors not have multiple gradients? Any good examples where multiple pressures have been applied robustly?	We have added: "Analysis of the gradient may be restricted to a single stressor or may include the impact of multiple stressors *if stressors cannot be separated (i.e., if sites are affected by more than one stressor simultaneously).*"
50	Ken Irvine	Page 14, K47: "Metrics have to be considered as inappropriate if they: (1) are less than robust and have a high temporal and/or spatial variability". Comment: Can there be at least some indication of what amount of variability is acceptable, or, for e.g, a matrix of “confidence of metrics"?	We agree this would be ideal, but variability will greatly differe between BQEs and ecosystem types. We have therefore rephrased the sentence: "(1) are less than robust and have a temporal and/or spatial variability exceeding variability caused by anthropogenic influences".
51	Ken Irvine	Page 15, K50: While it seems reasonable, on what basis is r > 0.8 the cut off point?	There is no underlying calculation; as the cutoff point will depend on the dataset, we have added "e.g.".
52	Ken Irvine	Page 15, K51: In case a multimetric index is targeted, it should preferably contain at least one metric from each type (Table 1) and, therefore,reflect	Not necessarily; we have added some references addressing this point.

No.	Person/Institute	Comment	Reply
		multiple dimensions of biological systems. Comment: Is linearity assumed?	
53	Ken Irvine	Page 16, K52 and K53: Generation of a multimetric index. Comments: Do not more metrics mean less confidence by chance alone? Surely there is a need for some statistical justification	We have added: "Different combinations of metrics (always including the relevant metric types) should be correlated against the stress gradients as done earlier for the selection of candidate metrics. The finally selected multimetric index should be among those metric combinations best correlating to the stress gradient."
54	Ken Irvine	Page 16, K52: What about non-linearity and class boundaries? Or maybe need to refer to section on setting boundaries below.	We believe this issue is sufficiently addressed in the section "setting of class boundaries".
55	Ken Irvine, Gweldolin Porst	Page 16, K56, P57: Because the metrics are scaled to reference conditions and expectations for the stream classes, any decision on subdivision should reflect the distribution of the scores for the reference sites. Comment: Don't understand this. Does it relate to possibility of high variation among reference sites? This then brings into question reliability of either type-specific approach or that the reference sites are not really at reference state.	We decided to delete this sentence, as its meaning has been described in the section "metric normalization". We refer to this section.
56	Kevin Irvine	Page 10, K24: Still means 50% is not. See various papers and books by Hakensen on this.	We shall strive for highest correlation to decrease the error when predicting the national class boundaries in the regression analysis to be conducted in intercalibration. A small error will give us highest confidence in the predicted boundary value on the common metric scale. However, pilot exercises have shown that national methods are often less well correlated to each other (and the common metrics).
57	Kevin Irvine	Page 10, K25+K26: Not clear. Also, how are stress continua addressed here.	This is based on a qualitative scheme to distinguish stressed from unstressed (two classes). This is in fact inferior to five-class or quantitative approaches.
58	Kevin Irvine	Page 10, K27: I doubt it. Metrics in use should be adequately evaluated and this is best done through published peer reviewed literature.. Expert knowledge /common practice provides the hypothesis.	With regard to the evaluation of metric robustness it will be indispensable to refer to "expert judgment". Some colleagues are dealing with the design of bioassessment methods for a long time and have made experiences that have not been published in every detail. At least we should consider this within the WISER work as a potential source of knowledge.
59	Kevin Irvine	Page 10, K28: Not sure if this makes sense. Check meaning.	We changed into "may not guarantee that in the intercalibration analyses the boundary values of the ecological status classes are sufficiently comparable."
60	Kevin Irvine	Page 10, K29: Would this not better be a power analysis as correlation relates to n	We have mentioned power analysis as an alternative suited technique.
61	Kevin Irvine	Page 11, K30: Such as? Is this the same as mentioned in next section "Criteria" (page 11, second bullet point)	There is the concept of alternative benchmarks introduced in the IC guidance. For example, the benchmark can be based on least disturbed conditions that are clearly linked to the reference state (demonstration of the level of deviation). The concept will be explained in the missing annex of the IC guidance to be drafted this spring.
62	Kevin Irvine	Page 11, K31: This seems quite arbitrary. An inherent difficulty with multimetrics is exactly the contributing reliability of each individual metric. The WFD requires a number of features to be included, but some of these may have very weak relationships with a pressure gradient or be highly influenced by spatial or temporal variability, Unless there is a process to test this the final multimetric could bring with it the high variance of individual metrics. Page 11, K32: See previous comment. What criteria will this weighting factor have to meet.	... It could for example reflect the normative definitions in equal proportions (see common intercalibration metric used in the CB riv GIG invertebrate exercise).
63	Kevin Irvine	Page 11, K33: Could be seen as a means to circumvent objectivity	What is meant here?
64	Kevin Irvine	Page 11, K34: Is there a need for Bon feroni type correction when more than one metric is used. Are there any issues with conforming here to WFD normative definitions?	The WFD normative defs are not specifying how much the different indicative parameters have to influence the overall score ...

No.	Person/Institute	Comment	Reply
65	Kevin Irvine	Page 6, K9: Not clear. The analysis of what? Reliability, representativeness, statistical power? Maybe rephrase after "and", "are critically important for successful intercalibration".	We changed accordingly.
66	Kevin Irvine	Page 7, K10: Why in brackets, Were they partly or fully calibrated? In any case, what does "partly" mean?	We added "intercalibrated at least for some indicative parameters" instead of "(partly)".
67	Kevin Irvine	Page 8, K19: Presumably, therefore, of limited use for WFD, but of some use in setting up hypothesis or suggestions for likely useful metrics that can then be evaluated further.	We are supporting the GIG work, so they have to validate our proposals in their official exercises.
68	Kevin Irvine	Page 9, K20: These may for pragmatic basis be defined at family, but could be of limited use in assessment., and hence not cost effective. This is of course not a new debate, and see O'Toole et al (2008) from the REBECCA project for lake inverts. A thorough review of this across biological elements and water body types could be an important WISER output.	This is one of the most relevant arguments when explaining the difference between common intercalibration metrics and bioassessment metrics. One is for boundary comparison, the other's for monitoring the ecol. qual.
69	Kevin Irvine	Page 9, K21: Critical assessment of national metrics would be useful. A potential problem is that national metrics may be based initially on "expert judgement" with a subsequent and understandable tendency to defend these judgements. The STAR project provided a good example of such testing.	A least we asked in the questionnaire if the pressure-impact-relationships were validated by data analysis.
70	Kevin Irvine	Page 9, K22: Can some statistical criteria for this, given the 5-point scale of WFD classification, be recommended?	It is almost impossible to give numbers here, as the differences between BQEs and water types are great.
71	Laurence Carvalho	Email comment: Having discussed this at our recent WP3.1 WISER/GIG meeting, the biggest disagreement we have in relation to your guidance is that our common metrics in 3.1 are being derived at the sub-BQE level - separate common metrics for composition and blooms and MS metrics will be compared with these two. This comment came from a couple of GIG people who had read your guidance. That does not necessarily rule out then having a "common multi-metric" that MS can compare with too - but we do not plan to develop that by April 2010.	We extended the relevant sentence to "Common metrics shall cover all required parameters in a multimetric index to allow for the intercalibration of the full BQE (Schmedtje et al. 2009), but may also reflect single parameters only if the BQE can currently not be fully intercalibrated. "
72	Laurence Carvalho	Email comment: I also agree with some of Peter's comments. For lake phytoplankton we have no strong intention to have common metrics across GIGs or across very different lake types. The common metric approach may be the same, but the metrics will be calibrated differently e.g. TP optima and tolerances of phytoplankton in low and medium alkalinity lake types in Northern GIG will not be the same as those for the same taxa in high alkalinity lakes being inter-calibrated in CB GIG. Similarly we do not plan to have metrics that are robust across seasons. Seasonal variability within the phytoplankton community is too great and so we aim to restrict the season that our metrics apply to and to examine variability in the metric within this season.	We have now clarified that (1) a common metric is not necessarily applicable for all GIGs - it might be specific for a GIG or even a type.
73	Laurence Carvalho	Email comment: My other general comment on the guidance is that I don't really see many differences in the "process" sections under "development of draft common metrics" and "development of assessment schemes". I understand they're not necessarily the same thing - but isn't the process virtually identical? The guidance seems somewhat repetitive for this reason and I wondered whether just the additional parts, or differences, could be highlighted for the latter. To me the main difference is that	We agree that the division into two main chapters (dealing with common metrics and with the development of assessment systems) is quite arbitrary and overlapping. However, both tasks will be performed at different times by most workpackages (common metric development: spring 2010; assessment system development: beginning in summer 2010). We therefore think that two separate methodological descriptions are useful. Although both descriptions are overlapping we think it is more practical to have them separate and leave both descriptions complete, so that they can be used as a "cook book". We have added an introduction (an extension of the original

No.	Person/Institute	Comment	Reply
		recommended assessment schemes do not necessarily have to follow the least common denominator approach in taxonomy that common metrics may be forced to. Anyway, just a thought if you do want to shorten it.	chapter "aims") to the deliverable to clarify the structure and the intentions.
74	Laurence Carvalho	Page 10, L13: For lake phytoplankton we are considering more than one per metric type – can we not adopt more than one if it gives a better or more rounded representation of the impact of that pressure?	Yes, you can. We changed this.
75	Laurence Carvalho	Page 7, L3: Lake phytoplankton don't agree with this (especially some GIG people) – our common metrics are being derived at the sub-BQE level – separate common metrics for composition and blooms and MS metrics will be compared with these two. That does not rule out then having a "common multi-metric" that MS can compare with too – but that will not be developed by April 2010	We extended the relevant sentence to "Common metrics shall cover all required parameters in a multimetric index to allow for the intercalibration of the full BQE (Schmedtje et al. 2009), but may also reflect single parameters only if the BQE can currently not be fully intercalibrated. "
76	Laurence Carvalho	Page 8, L6 (and others): But only a restricted typology & biogeography. Common metrics may use the same approach across GIGs and types but they may be calibrated differently, e.g. taxa optima and tolerances to stress may be different between GIGs and/or some types (e.g. TP optima in N GIG low alkalinity lakes vs CB GIG high alkalinitylakes)	We added: "If the BQE shows strong spatial and/or temporal variability (e.g. phytoplankton) the common metric may be water type-specific and/or calculated from fixed sampling season data only."
77	Martin Sondergaard	Page 14, ms10: What if no metrics for a specific BQE is found appropriate ?	We have added: "If none of the calculated metrics fulfils the criteria it should be considered generating "new" metrics, e.g. based on species predominantly occurring in stressed or unstressed sites following an indicator species analysis."
78	Martin Søndergaard	Page 1, ms1: First of all I think we need an introductory paragraph which emphasizes some of the difficulties, for example that it is not always easy to put "nature" into fixed boxes and numbers. You could for example use and refer to some of the points made by Moss in several papers. Maybe also some points regarding the risk of misclassification should be inserted here.	An introduction has been added.
79	Martin Søndergaard	Page 1, ms1: It might also be idea to put some example (showing calculations) into an appendix. This would be very helpful for managers, but maybe that will come in further steps within WISER.	We see this guidance mainly as a cook book for WISER scientists and don't think it should be used by stakeholders. If we in a later stage will extent the scope of the guidance it should be amended with examples.
80	Martin Søndergaard	Page 10, ms8: I think we should suggest that we try avoid the use "experts" as much as possible, because there is great risk that they are influenced (without being aware) by their own experience, which might not be the complete view.	With regard to the evaluation of metric robustness it will be indispensable to refer to "expert judgment". Some colleagues are dealing with the design of bioassessment methods for a long time and have made experiences that have not been published in every detail. At least we should consider this within the WISER work as a potential source of knowledge.
81	Martin Søndergaard	Page 10, ms9: Would twice really be enough for testing? I think it need to be at least three.	Of course three times would be better to estimate the temporal robustness. However, a site sampled twice may already give a suitable estimate of robustness.
82	Martin Søndergaard	Page 8, ms4: Probably it would always be difficult to use a metric which does not have a monotonous response. At least it makes it more complicated.	Sure, the most desirable option is ideally the monotonous responding metric.
83	Martin Søndergaard	Page 9, ms7: And if that is not "well correlated" what then – delete the metric ?	We shall strive for highest correlation to decrease the error when predicting the national class boundaries in the regression analysis to be conducted in intercalibration. A small error will give us highest confidence in the predicted boundary value on the common metric scale.
84	Nuria Marba	Page 11, NM10: Only it can be a multimetric index and not a single parameter? A metric that is a single parameter may be sensitive to pressures and robust in space and time too.	That is true, but can it cover all indicative parameters of the WFD normative definitions? For Angiosperms "tax. Comp." and "Abd." need to be addressed (may be put into one weighted average metric) ...
85	Nuria Marba	Page 4, NM2: By "taxa list"o you mean Biological Quality Elements? I agree with Dorte that "taxa list" is not appropriate.	We have changed the sentence into "Assessment methods (often referred to as "classification method") translate biological information of a water body into an ecological status class (ranging

No.	Person/Institute	Comment	Reply
			from high status to bad status). "
86	Nuria Marba	Page 6, NM3: Not all BQEs are classified based on taxa lists. Eg for angiosperms in coastal (and transitional) waters often metrics are based on features of a single species	We changed into "Data acquisition, i.e. the field sampling and sample processing to yield biological information;"
87	Nuria Marba	Page 8, NM7: Regarding angiosperms and macroalgae for marine waters the indicator taxa varies across GIGs –due to the biogeography of the taxa. So, a common metric across all marine/coastal GIGs will not be available.	... Unless it is based on non-taxonomical (functional) information ... But here we added: "If the BQE shows strong spatial and/or temporal variability (e.g. phytoplankton) the common metric may be water type-specific and/or calculated from fixed sampling season data only."
88	Peeter Noges	Page 4, P1: Isn't this definition oversimplified? Even for a taxonomic index just a taxa list is in most cases not sufficient and quantitative data are needed as well.	We have changed the sentence into "Assessment methods (often referred to as "classification method") translate biological information of a water body to an ecological status class (ranging from high status to bad status). "
89	Peeter Noges	Page 6, P2: Could the meaning of 'benchmarks' be explained a bit more? Is it a method, metric, value, range or site?	We included a footnote with the explanation: "Definition of trans-national (absolute) reference points in intercalibration based on data from near-natural reference sites or sites impacted by similar levels of impairment."
90	Peeter Noges	Page 7, P4: Is there any alternative if such common metric cannot be found . Can we be sure that a suitable common metric exists and it is only a question of effort made to find it?	The intercalibration processes also allows for the definition of pseudo-common metrics, i.e. the average of all national EQRs at a site excluding the country to be compared against. However, this is only possible in IC Option 3.
91	Peeter Noges	Page 7, P5: Is this really achievable and if not, does it mean that the national method is bad? National data cover only a fraction of the full range of a common metric. Within this narrow range the relationship may become too blurred or will the national methods be tested on the full GIG range? In this case the national method will be applied to areas for which it was originally not created.	We shall strive for highest correlation to decrease the error when predicting the national class boundaries in the regression analysis to be conducted in intercalibration. A small error will give us highest confidence in the predicted boundary value on the common metric scale.
92	Peeter Noges	Page 7, P6: Slope depends on units. Here the ranges are obviously given for standardised variables.	We added: "(standardized values)" after the slope criteria.
93	Peeter Noges	Page 9, p7: For example, proportional metrics ranging from 0.1% to 0.5% or the number of sensitive taxa ranging from 0 to 3 along the whole stressor gradient should be avoided. Comment: Would 1 to 5% be better? Is there any criterion?	It is almost impossible to define a criterion here, as BQEs and water types are too different.
95	Peeter Noges	Page 9, p8: in case of correlation coefficients > 0.7 (…) comment: In the long-term data from L. Võrtsjärv (Estonia) the correlation between phytoplankton biomass and Chl a is 0.7. These expressions of abundance are mostly considered equivalent. I cannot imagine that any other couple of phytoplankton based parameters either single or multimetric, quantitative or qualitative could give a comparable match.	This section is dealing with the correlation of assessment systems (i.e. a common metric and a national assessment systems). Here , the correlations are often very strong. We are NOT dealing here with the correlations of environmental parameters and biotic metrics (here, in fact a correlation of >0.7 would be unusually strong). We shall strive for highest correlation to decrease the error when predicting the national class boundaries in the regression analysis to be conducted in intercalibration. A small error will give us highest confidence in the predicted boundary value on the common metric scale.
97	Peter Hendriksen	Mentioned in the email accompanying the corrected guidance: Definition of reference conditions is pivotal to assessments, boundary setting and development of metrics. I think this deserves more attention in the guidelines.	We have added a passage about the harmonized approaches to derive reference conditions for the intercalibration exercise.
98	Peter Hendriksen	Mentioned in the email accompanying the corrected guidance: However, as previously stated I have my serious doubts about the possibility of coming up with a common coastal/transitional phytoplankton metric which fulfils all the requirements. While it may be possible to use a common metric within a GIG I don't think we can find a universal metric applicable to all GIGs. You describe the product as a multimetric index	In your case, some "all-GIG" metric is not feasible, even an "all-types" metric within one GIG seems not to be likely. Perhaps it is possible to derive type-specific common metrics for the different salinity levels covered in the exercise. Our guidance is describing the ideal case, so pragmatic solutions are to be found anyway.

No.	Person/Institute	Comment	Reply
		composed of one metric per metric type. One of the missing indicative parameters for phytoplankton is the taxonomical composition which is at least as dependent on the salinity as on our common stressor eutrophication within the gradient from almost fresh waters in the Baltic Sea to full oceanic salinity in the Atlantic and Mediterranean. At the same time this salinity gradient almost represents an "oligotrophication" gradient from the Baltic area without any reference sites to the much less eutrophic geographical regions with sites representing reference conditions. Thus the response of such a common metric to human activities/impairments may be completely masked by natural environmental differences. I find it very hard to imagine that we can find a taxonomical index with an unambiguous dose response relationship to eutrophication across this salinity range.	
99	Peter Hendriksen	Page 12, D14: Move metric definition (from the section "development of assessment methods" to the beginning of the document.	done
100	Peter Hendriksen	Page 15, D20: I hope that reference site classification is not uncertain – this is crucial for boundary setting.	We have added a comment in which cases reference conditions should be regarded as uncertain.
101	Peter Hendriksen	Page 4, D1: Assessments may be based on several other features than taxa (e.g. chlorophyll, biomass etc.). Why not introduce the concept of “metrics” here?	We added a definition of the term "metric" in the introduction.
102	Peter Hendriksen	Page 6, D2: Is it truly a common metric if data acquisition differs?	If the data acquisition differs, the use of common metrics is the only tool to intercalibration (no direct comparison possible). However, the level of this difference must not be too high (otherwise even a common metric is useless).
103	Peter Hendriksen	Page 7, D4: Why only lakes? (biomass as indicative parameter of phytoplankton)	according to the normative defs only specified for lakes
104	Torben Lauridsen	It does not reflect the simple fact that building indicators must follow a well defined method but will not result as a common recipe. Some of this “precautionary spirit” should, in my opinion, be reflected in the document, in principle in the introductory sections.	This has been included into the introduction.
105	Torben Lauridsen	Page 11, tll7: An explanatory variable for eutrophication could be Chl a. I suggest Chla since this is correlated to TP but addition to this also a variable in which all other biological elements are integrated.	We added Chl-a to the example.
106	Torben Lauridsen	Page 11, tll8: Is this supposed to be based on expert judgements at this stage?	Yes, and in regard to the activities done in the GIG (if any).
107	Torben Lauridsen	Page 13, tll9: Metric calculation, exclusion of numerically unsuited metrics. Comment: At this stage I guess these two bullets will include expert judgments. if so I suggest to mention this in the text.	We have added comments on the decisions required by the workpackage scientists.
108	Jannicke Moe	Page 4, JMO1: I think "water body types" is the correct expression (throughout the document).	The WFD uses the term “water body type”, but we prefer the shorter term “water type” which is meaning the same and should be as clear as the WFD term.
109	Jannicke Moe	Page 5, JMO3: Beginning or end of March? (I thought end, but others have said beginning...)	We added “end of March”
110	Jannicke Moe	Page 5, JMO4: More specifically, for comparing the boundary values of the ecological classes - ?	We added: “Common metrics are a common yardstick for comparing national assessment systems and their classification of ecological status.”
111	Jannicke Moe	Page 6, JMO5: "Comparable" is used often but I'm not always sure what it means in the different contexts. Same? Similar? Consistent? (Things can be inconsistent but still be comparable...)	We changed “comparable” into “similar”.

No.	Person/Institute	Comment	Reply
112	Jannicke Moe	Page 7, JMO6: This is a good metaphor, it could even be stretched further ("exchange rates", ...). I suggest that you use this also in the non-technical summary. I have the impression that some project partners still have the wrong idea of common metrics (that we all must use the same common metrics also nationally), so this may be an important point.	We put the term “international currencies” in the non-technical summary.
113	Jannicke Moe	Page 7, JMO7: upper and/or lower boundary?	When writing about “boundaries of good ecological status”, this implies the upper and lower boundary. So it’s not necessary to extent the phrase here.
114	Jannicke Moe	Page 8; JMO8: Are these really used for rivers too?	Yes, in the normative definitions for rivers it is written: “Planktonic blooms occur at a frequency and intensity which is consistent with the type-specific physicochemical conditions.”
115	Jannicke Moe	Page 10, JMO12: I know that "explained by" is a common way of putting it, but it doesn't seem right to regard the common metric as an explanatory variable.	However, we want to keep this expression because it is immediately clear what is meant by it.
116	Jannicke Moe	Page 11, JMO13: What is a comparable relationship in this case? F.ex. slope and/or intercept are not significantly different?	We added “(e.g. regression model is similar)”.
117	Jannicke Moe	Page 16, JMO14: What about methods for identifying non-linear relationships?	We added: “Non-parametric regression models can also be applied for explorative analyses of the pressure–response relationships to identify non-linear patterns (e.g. Schartau et al. 2007).”
118	Jannicke Moe	Page 18, JMO16: How can you identify discontinuities in the relationship if only using a linear regression/correlation method? (See recommendations in Schartau et al 2007).	Well spotted! But with the above made change this shall be okay now.