ICES Journal of Marine Science
November 2008, Volume 65 (8) : Pages 1449-1455
http://dx.doi.org/10.1093/icesjms/fsn124

Evaluating potential indicators for an ecosystem approach to fishery management in European waters

Gerjan J. Piet[1, *], Henrice M. Jansen[1] and Marie-Joëlle Rochet[2]

[1] Wageningen IMARES, PO Box 68, 1970 AB Imuiden, the Netherlands
[2] IFREMER, Département Ecologie et Modèles pour l'Halieutique, BP 21105, 44311 Nantes Cedex 03, France

*: Corresponding author : Gerjan J. Piet, tel: +31 255 564699; fax: +31 255 564644, email address : gerjan.piet@wur.nl

Abstract:

This study describes the process of evaluating potential indicators for an ecosystem approach to fishery management in European waters by evaluating these indicators against existing criteria using questionnaires completed by experts. We (i) compare the use of a longer list of simple criteria with a shorter list of elaborate ones; (ii) compare evaluation results when screening criteria are applied to specific indicators vs. high-level headline indicators; and (iii) examine whether detailed questionnaires, with elaborate indicators and elaborate criteria, result in ranked scores that are less influenced by familiarity with the indicators. The results show that the ranked scores of indicators are affected by the level of detail, both in terms of criteria and indicators, provided in the questionnaires. It appears that adding detail to the questionnaires makes the scoring process more transparent and provides better founded scores; at a certain point, however, more-detailed indicators and/or more-detailed criteria result in decreased performance of the scoring process, reflecting mostly factors that do not determine the suitability of the indicator (e.g. the level of familiarity), while giving the false impression of a more thorough analysis.

Keywords: ecosystem approach, evaluation criteria, headline indicators, questionnaire

1. Introduction

It is generally agreed that an ecosystem approach to fisheries management (EAFM) will have to rely on suites of indicators that track the pressure exercised, the state of the ecosystem and the socio-economic consequences in relation to the management objectives formulated (FAO, 2003; Rice, 2003; Jennings, 2005). While the number of potential metrics that might be tracked is virtually unlimited, the final suite of indicators to be selected needs to provide a 'good coverage' of the human activities that need to be managed as well as of the ecosystem components and attributes affected (Jennings, 2005), but is likely to involve a compromise that would fit the management objectives (Jennings, 2005; Rochet *et al.*, 2007). The selection may be conducted using criteria that ensure that the indicators to be used meet a number of desirable properties (Rice and Rochet, 2005).

The European Union has committed itself to implementing an ecosystem approach into its Common Fisheries Policy (CFP), and has stated some high-level objectives accordingly (CEC, 2002). To this end, it has funded several research and development projects aimed at establishing lists of potential indicators for fisheries management (*e.g.*, FISH/2002/08: "Development of preliminary indicators of environmental integration of the Common Fisheries Policy", and INDECO: "Development of Indicators of Environmental Performance of the Common Fisheries Policy"). While the process of prioritizing issues and developing indicators for these issues is ongoing, there is a need to include methods in this process that can screen potential indicators for their appropriateness. Exercises applying screening criteria to actual lists of indicators should help to identify the potential shortcomings and advantages of different approaches.

Rochet and Rice (2005) tested a framework for indicator selection. In their experiment, a set of 20 candidate indicators that were supposed to have general applicability was evaluated by 16 experts, each familiar with one of four different ecosystems around the world. Outcomes of this exercise were that various steps involved in the selecting process proved to be prone to subjective value judgment, and that differences in scores attributed by the experts were the main factor contributing to the observed variability in the evaluation results for different ecosystems, with little influence of whether the experts were familiar with a particular one. Rochet and Rice (2005) conclude that understanding of the reasons for individual preferences for specific indicators is important to foster dialogue because it helps clarifying the debate. They also suggest that the selection process might be easier if a longer list of simple criteria as provided by Rice and Rochet (2005) is used as opposed to a shorter list of more complex ones.

The issue of complexity is important if such evaluation procedures are to be carried out not only by experts but as part of a wider stakeholder consultation. In the Rochet and Rice (2005) experiment, the framework for selecting indicators proved useful because it gave experts the opportunity to present their values explicitly. However, what level of simplification and/or detail needs to be applied to the criteria to help clarifying the debate without adding confusion? And how many indicators, and at what level of specificity, can be put forward for evaluation by a particular group of stakeholders?

Within the INDECO project, we conducted another screening test that is more specifically directed towards its use in the process of selecting indicators for an EAFM within the CFP. The aim was to i) compare the results obtained with a long list of simple criteria to a shorter list of more elaborate ones; ii) evaluate the effect of applying screening criteria to specific versus more high-level, so-called headline indicators (Jennings 2005); and iii) examine whether familiarity of the experts with a specific indicator influences their evaluation of the importance of that indicator.

The evaluation was based on the framework developed by Rice and Rochet (2005), which has eight steps: (1) determining user needs; (2) listing candidate indicators; (3) determining screening criteria; (4) scoring indicators against criteria; (5) summarizing scoring results; (6) deciding how many indicators are needed; (7) final selection; and (8) reporting. We restricted ourselves to steps 2 to 5 because the user needs (step 1) are essentially given by the objectives specified in the revised CFP (Regulation 2371/2002: 'to ensure the long term viability of the fisheries sector through sustainable exploitation of living aquatic resources based on sound scientific advice and on the precautionary approach'), while steps 6 and 7 were considered to be outside the scope of this exercise.

2. Material and methods

The evaluation process differed from the Rice and Rochet (2005) framework in so far that each respondent was supposed to fill in several different questionnaires corresponding to eight scenarios. We aimed at having all experts filling in all questionnaires, however, for practical reasons this could not be achieved, and the actual numbers are reported in table 1. These not only provide an evaluation of the indicators but also allow us to test a number of hypotheses on potential factors affecting the performance of the process. The respondents comprised of 24 experts from 20 research organizations spread over 11 EU Member States with expertise on four marine ecosystems corresponding to the Regional Advisory Council (RAC) areas: Baltic Sea, North Sea, Bay of Biscay (representative for the SW waters), and the Mediterranean. In addition, we had four responses from non-scientist stakeholders, but we have excluded these from the analyses to avoid bias.

For step 2 of the framework, we followed Jennings (2005) and identified a list of candidate indicators for both state and pressure based on a literature review. The state indicators were supposed to cover the entire ecosystem with all its different components and attributes. Given the focus of the user needs as provided by the CFP objectives, the pressure indicators only covered fishing. Starting of from six broad and general issues related to the EAFM, we examined the effect of detail in indicator specificity by introducing a hierarchy of indicators from overall features (‘headline indicator’) to the actual metric (‘specific indicator’). In case no specific indicator has been developed for a particular feature, we used a more general phrasing.

In step 3, we used the list of criteria and sub-criteria proposed by Rice and Rochet (2005). To examine the influence of detailing criteria on the evaluation outcome, either the shorter list of 9 main criteria (concreteness, theoretical basis, public awareness, cost, measurement, availability of historical data, sensitivity, responsiveness and specificity) or the full list of 33 sub-criteria (3, 3, 5, 1, 11, 5, 1, 1, 3 sub-criteria, respectively; see for details Table 2 in Rice and Rochet, 2005) was used.

In step 4, the indicators were scored against the criteria by the respondents for eight different scenarios allowing: (1) an evaluation of either headline (HN) or specific indicators (SN) without explicit criteria; (2) an evaluation of headline indicators (HC) only based on the main criteria; (3) an evaluation of specific indicators (SS) only based on the sub-criteria; (4) an evaluation of the familiarity of the respondents with each headline (HF) and specific (HS) indicator (table 1).

The scoring had two components: an evaluation of the quality of each indicator relative to each criterion (“indicator scoring”) and an evaluation of the relative importance of each criterion (“criteria weighting”). An ordinal scoring of 5 ranks was used to evaluate the performance of each headline indicator against each criterion (1=worst, 5=best), as well as for the weighting of the criteria (1= less important, 5=very important). Weighting of the sub-criteria was done on a relative scale so that the weights of the sub-criteria for each criterion sum to one. For the ‘familiarity scoring’ an ordinal scoring of 3 ranks (1=least familiar, 3=most familiar) was used.

To summarize scoring results (step 5), we created a table in which a mean score for each indicator (H: headline, or S: specific) was derived. Depending on the scenario, these means were calculated differently: for evaluations without explicit criteria (HN, SN), the mean score by indicator was calculated as the mean across all responses; for evaluation involving (sub-)criteria (SS), the scores by (sub-)criterion given by each respondent for each indicator were first weighted by the weighting scores given by the same respondent to derive one indicator score per response and then the mean was calculated across all responses; for evaluations based on specific indicators, we derived a score for headline indicators based on the mean score (SN, SS) of the corresponding specific indicators. For interpretation of the results we also had access to a scoring of the familiarity of the different respondents with each indicator. Overall, this resulted in 6 scenarios (HN, SN, HC, SS, HF, SF, see table 1) providing scorings of indicators. Spearman rank-order correlations (S) were used to investigate the following hypotheses:

(1) there is no difference between the ranked scores of specific and headline indicators (S expected to be 1);

(2) there is no difference between scores using main criteria versus sub-criteria (S expected to be 1); and

(3) longer, more detailed and thus more straightforward questionnaires with specific indicators and sub-criteria result in ranked scores that are less affected by familiarity than shorter and less elaborate questionnaires (S between ranked indicator scores and ranked familiarity scores expected to decrease as the level of detail increases).

As Spearman rank correlation is a non-parametric statistic there is no formal test for it being different from 1, so P-values cannot be provided for the first two tests. Rank correlation is taken as a measure of agreement among the experts' rankings.

3. Results

Weights of criteria
Ranking of the criteria weights based directly on the main criteria (WC) showed that on average concreteness, public awareness, and cost got the lowest weights (figure 1). Results based on the sub-criteria (WS) showed a somewhat similar pattern, even though the ranks differed slightly, with higher concreteness weights and lower sensitivity weights. Theoretical basis, specificity, responsiveness and specificity got less various scores based on WS than on WC (*e.g.*, sensitivity was scored 4 or 5 in WS but 3, 4 or 5 in WC). $VT looking at the figures it is difficult to tell whether this is true only for specificity and not for example for available data etc. You have to calculate some measure of variation to make a clear statement. Please consider doing this. Variance does not help as there were more respondents to WS (12) than to WC (13), so variance is higher for WS anyway$

Indicator scoring
The scorings and ranks for headline indicators are given in table 2 and figure 2a and for specific indicators in table 3 and figure 2b Overall, indicators associated with traditional fishery management (e.g. various fishing-pressure indicators and status of commercial stocks) scored highest, while indicators of ecosystem functioning and plankton scored lowest.
Hypothesis testing
Spearman rank-order correlation shows that all scenarios are correlated, S-values ranging between 0.31 and 0.88, and thus are reasonably consistent in their evaluation of the indicators (Table 4). $VT I don't understand this, why can you not calculate p-values for the correlation coefficients? In what sense are the tests not independent? Please an understandable statement. See Methods above$. However, the differences between scenarios as reflected in the S-values allow a qualitative investigation of the three hypotheses:
(1) The hypothesis that there is no difference in the ranked scores of headline indicators when the scoring is done for the headline indicators versus a scoring of specific indicators was investigated by comparing the scores of the headline indicators without using criteria (HN) with the scores calculated by taking the mean of the specific indicators. The test was based on the same 15 respondents filling in the questionnaire and resulted in $S_{HN/SN}$=0.82. The deviation from 1 of these S values indicates that the level of detail of the indicators resulted in a different ranking of the headline indicators (Table 4a);
(2) The hypothesis that there is no difference in ranked scores based on main criteria only versus sub-criteria was investigated by comparing the scores of the headline indicators based on main criteria (HC) with the scores based on sub-criteria and calculated by taking the mean of the specific indicators. The test was based on 11 respondents filling in the HC and SS questionnaires (see table 1) of which more than half (6) were filled in by the samerespondents. The value of $S_{HC/SS}$=0.88 indicates that the level of detail of the criteria resulted in a different ranking of the headline indicators (Table 4a);
(3) The hypothesis that longer questionnaires with more concrete indicators and criteria result in ranked scores that are less influenced by familiarity was tested through two comparisons.
Scores of headline indicators not based on criteria (HN) and based on main criteria only (HC) were compared to the familiarity score (HF). The comparison HN/HF was based on the same 15 respondents whereas the HC/HF questionnaires only had 5 respondents in common. These comparisons showed that the use of criteria resulted in scores that were correlated less with the familiarity scores ($S_{HN/HF}$=0.78 versus $S_{HC/HF}$=0.69, Table 4a). A striking result was that scores of specific indicators not based on criteria (SN) and based on the sub-criteria (SS) compared to the familiarity score (SF) showed that a further increase in detail (i.e. specific indicators as opposed to headline indicators and sub-criteria as opposed to main criteria) resulted in scores that were more similar to the familiarity scores ($S_{SN/SF}$=0.58 versus $S_{SS/SF}$=0.67, Table 4b). Finally, a general observation was that familiarity scoring (HF, SF) is highly correlated with all indicator scoring, whatever the method used (Table 4).

4. Discussion and conclusions

This study found that depending on how the question was posed (i.e. different questionnaires, with headline indicators or specific indicators, whether or not criteria and/or sub-criteria are used) the ranking of the indicators will differ independent of whether or not they are filled in by the same experts.

For the ranked scores of headline indicators versus specific indicators. thisdifference may come from the difficulties in the translation from one to the other owing to inherent differences in scores between the specific indicators representing the same headline indicator. For instance, for the headline indicator "physical environment", which is in itself an abstract concept, two specific indicators (temperature and NAO) were applied. In terms of concreteness, the appropriate score for temperature should be higher than for NAO, the latter being a much more abstract concept. Another example is the "Abundance of Commercial Stocks". The headline indicator sounds concrete (score 4 or 5), is something that the public is aware of (score 4 or 5) and is likely to be tightly linked to fishing (score 4 or 5). However, the specific indicator "Proportion of commercial stocks that are within safe biological limits" is based on elaborate assessment models (concreteness score 1 or 2). Accordingly, public awareness may be lower (what are safe biological limits?), and the link with fishing activity may be less (score 2 or 3). Within the headline indicator "Status of marine mammals", differences in scoring against the criterion "historical data" or "measurement" may be considerable. For the specific indicator "seal population in the Wadden Sea", the score should be high because of a wealth of information available, while the "North Sea porpoise population" should give a much lower score because of a very restricted data set. So how should one score the headline indicator when being aware of large differences between specific indicators?

Similarly, the scoring results differed when using sub-criteria versus main criteria, but these differences leveled off across experts and criteria, and the final rankings were highly correlated.

If divergence from familiarity is an appropriate way to assess the performance of the evaluation process, then the results indicate that for the evaluation of a relatively few headline indicators a longer list of simpler selection criteria indeed improves the process as Rice and Rochet (2005) suggested. However, the opposite trend is observed when an extensive list of many specific indicators is applied. This suggests that there is a point on the gradient from short lists of complex criteria and headline indicators to long lists of simple criteria and specific indicators when the performance of the evaluation process starts to decline.

The strong correlations between familiarity and any of the scorings suggest that to some degree experts will give higher scores to indicators they are more familiar with, and/or lower scores to indicators they do not know well. Experts may not have sufficient information (Rochet and Rice, 2005) and if this information is lacking the level of familiarity may largely determine the outcome of the scoring. Among the non-scientists, several declined to fill in the questionnaires, because they did not consider themselves informed well enough to score the indicators against criteria. Thus, for non-scientist stakeholders the effect of familiarity may apply even more.

Clearly, the process of indicator selection for an EAFM in the EU should involve enough respondents from different stakeholder groups and nationalities with sufficient expertise to ascertain commitment to the evolving suite of indicators. While scoring is a convenient aid in summarizing the evaluations by different people, there may be no need to score indicators against criteria in the actual selection process. An indicator might just pass or fail against each criterion, or might be evaluated more qualitatively with 'pros' and 'cons', while the final selection could be the result of a negotiation rather than of some numerical scoring. As all scientific activity needs to be balanced against the resources available, our experience has been that asking a large group of respondents to go through extensive questionnaires may not be the best way to use these resources.

The results of this evaluation and the concerns expressed so far may be discussed against the background of the 8-step evaluation framework of Rice and Rochet (2005):

(1) *Determining user needs*. The current objectives of the CFP are not specific enough to allow a proper scoring of the indicators, because these are restricted to commercial stocks only. Specific operational objectives for other ecosystem components need to be formulated at the appropriate scale, *e.g.* according to Jennings' (2005) framework which includes an additional step to identify those activities that are most likely to compromise the broad objectives presently formulated.

Listing candidate indicators. Since too many indicators will aggravate the evaluation process, we would advise to start with a limited suite of indicators. Concrete indicators have been developed for some ecosystem features, while none exist for others. We addressed this problem by distinguishing two hierarchical levels of indicators: headline indicators and specific indicators. While this distinction was intended to resolve discrepancies between types of indicators available, the feedback of (notably the non-scientific) respondents showed that for an evaluation by different stakeholders it may be more appropriate to have them evaluate headline indicators as specific indicators are often meaningless to them and could obfuscate the evaluation. The evaluation and selection of specific indicators for a particular headline indicator should be done by individuals that are sufficiently familiar with their merits. This may be ascertained by providing respondents with all the relevant information prior to the actual scoring.

Determining screening criteria. Although the respondents considered the criteria and sub-criteria appropriate, the use of sub-criteria did not affect or improve the scoring to any large extent. The obvious requirement is that the level of detail in the criteria should balance the level of scientific information available, and hence the expertise of the respondents. Adding sub-criteria with increasingly more subtle differences is expected to hamper the scoring process as soon as their evaluation requires more expertise than is present within the respondent group.

Scoring indicators against criteria. Using explicit criteria can make the scoring process more transparent and moving from short lists of complex criteria to longer lists of simpler sub-criteria is expected to provide better-founded scores (Rochet and Rice, 2005), providing the level of detail is tuned to the level of expertise of the respondents. If this is not the case, extended questionnaires scoring many specific indicators against sub-criteria with an elaboration of their respective weighting factors does not improve the evaluation process, while giving the suggestion of a more thorough exercise. Making all relevant information available prior to the scoring and allowing an exchange of viewpoints should reduce variation and bias, and hence result in the scores converging. Following this logic, the implication would be that if all available information is discussed within the group of respondents, then more sub-criteria can be used in the evaluation.

Summarizing scoring results. We observed marked differences in weighting of the different (sub-)criteria between individual scientists. These differences may be assumed to become even larger when more stakeholder groups are involved in the process (e.g., NGOs are more likely to give the highest weights to public awareness, managers to responsiveness and politicians to costs). Specifying the weightings of the different criteria given by each stakeholder group may facilitate the process of selecting indicators by making it more transparent, but how these weightings should be applied in obtaining the final scoring and thus the preferred indicators needs to be resolved. Instead of the most obvious choice of preferring the indicators with the highest average value, an alternative approach could be to select indicators that meet some minimum level of acceptance for all criteria.

Deciding how many indicators are needed. Several considerations determine the choice of the number of selected indicators. The first choice is that we need indicators for both state and pressure (Jennings, 2005). A minimum requirement for the ecosystem state indicators would be that for each ecosystem component and attribute for which operational objectives are formulated at least one headline indicator with a specific indicator is selected. This minimum selection may be expanded by also including indicators that are not necessarily affected by the fishery themselves but that should be considered in the management of the core ecosystem components (e.g. environmental indicators). Finally, there is the choice to have more than one specific indicator for one or more of the headline indicators. Again, this should be determined by how much additional information this new specific indicator provides. In the end, however, the number of indicators that are selected and how they are combined will not only be determined on scientific grounds but also by the requirements of the manager who needs to work with them or the costs involved in collecting the necessary data.

Final selection. The information needed to guide the final selection of indicators can be derived from the scoring of indicators against screening criteria, assuming that the shortfalls mentioned previously are resolved. A possible refinement of the approach could be to conduct this in two stages: the first stage involving different stakeholders where headline indicators are scored against (a subset of) the criteria and weightings of the criteria per stakeholder group are identified and a second stage involving a more restricted group where for each headline indicator one or more specific indicators are evaluated against (a more detailed or extended set of) screening criteria.

(2) *Reporting*. This study has not provided any relevant information for the reporting process.

Acknowledgements

The work was funded by the European Commission through two projects: “Development of Indicators of Environmental Performance of the Common Fisheries Policy” (INDECO, FP6- 513754) and “Indicators for fisheries management in Europe (IMAGE, FP6 - 044227). We thank in particular our respondents without whose perseverance to go through the questionnaires this evaluation would not have been possible: E. Andrulewicz, M. Appelberg, R. Aps, A. Borja, J. Brown, F. Colloca, N. Daan, O. Giovanardi, S. Greenstreet, M. Gristina, S. Jennings, S. Libralato, I. Lutchman, E. Meeuwsen, H. Ojaveer, P. Orr, M. Pommarede, F. Pranovi, P. Pelusi, S. Raicevich, M. Romanelli, N. Streftaris, S. Sverdrup-Jensen, M. Tasker, P. Tomasik and D. Wilson.

References

CEC. 2002. Council Regulation 2371/2002 of 20 December 2002 on the conservation and sustainable exploitation of fisheries under the Common Fisheries Policy. OJ L 358/59 31.12.2202.

FAO. 2003. The ecosystem approach to fisheries. FAO Technical Guidelines for Responsible Fisheries, 4, Suppl. 2, FAO, Rome. 112 pp.

Jennings, S. 2005. Indicators to support an ecosystem approach to fisheries. Fish and Fisheries, 6: 212-232.

Piet, G. J., Quirijns, F. J., Robinson, L., and Greenstreet, S. P. R. 2007. Potential pressure indicators for fishing, and their data requirements. ICES Journal of Marine Science, 64: 110-121.

Rice, J. C., and Rochet M.-J. 2005. A framework for selecting a suite of indicators for fisheries management. ICES Journal of Marine Science, 62: 516-527.

Rice, J. 2003. Environmental health indicators. Ocean & Coastal Management, 46: 235-259.

Rochet, M.-J., and Rice J. C, 2005 Do explicit criteria help in selecting indicators for ecosystem-based fisheries management? ICES Journal of Marine Science, 62: 528-539.

Rochet, M.-J., Trenkel, V. M., Forest, A., Lorance, P., and Mesnil, B. 2007. How could indicators be used in an ecosystem approach to fisheries management? ICES CM 2007/R: 05. 15 pp.

Tables

Table 1. Coding of the eight questionnaires that were put to respondents to fill in scores for two types of indicators (H: headline; S: specific) and weighting factors (W) for four types of value judgements (C: criteria, S: sub-criteria; N: none; F: familiarity). SN and SS were derived as mean of the relevant set of specific indicators. The number of respondents per questionnaire is indicated.

Type of scoring	Code	Number of respondents	Type of indicator		Type of judgment			
			H	S	C	S	N	F
Indicator	HN	15	X	-	-	-	X	-
	SN	15	X	X	-	-	X	-
	HC	11	X	-	X	-	-	-
	SS	11	X	X	-	X	-	-
	HF	15	X	-	-	-	-	X
	SF	15	-	X	-	-	-	X
Weighting	WC	13	-	-	X	-	-	-
	WS	12	-	-	-	X	-	-

Table 2. Mean scoring (with rank order in brackets) for headline indicators (with their short names as used in Figure 2a given in brackets) based on three different questionnaires (for explanation code see table 1). Scores range from 1 (worst) to 5 (best).

Ecosystem component	**Headline indicator**	**HN**	**HC**	**HF**
Physical/Chemical	Physical environment (Physical)	2.6 (20)	3.6 (4)	2.1 (12)
	Chemical environment (Chemical)	2.7 (19)	3.5 (7)	2.0 (14)
Plankton	Phytoplankton (Phytoplankton)	2.9 (16)	3.0 (15)	1.6 (18)
	Zooplankton (Zooplankton)	3.1 (15)	2.9 (17)	1.7 (17)
Fish	Abundance commercial stocks (Commercial)	4.8 (1)	3.8 (3)	2.7 (2)
	Abundance other populations (Other)	3.9 (8)	3.1 (12)	2.4 (7)
	Size/age structure species (Size/age)	4.3 (4)	3.6 (5)	2.6 (3)
	Genetic composition species (Genetic)	2.9 (17)	2.9 (17)	1.4 (19)
	Size structure community (Size structure)	3.9 (7)	3.2 (8)	2.6 (3)
	Species composition (Biodiversity)	3.9 (8)	3.1 (13)	2.4 (7)
	Abundance community (Community)	3.6 (13)	3.1 (9)	2.5 (6)
Other components	Status marine mammals (Mammals)	3.8 (10)	3.1 (10)	1.9 (15)
	Status seabirds (Seabirds)	3.5 (14)	3.1 (11)	1.7 (16)
	Status marine reptiles (Reptiles)	2.9 (18)	2.9 (16)	1.4 (20)
	Status benthos (Benthos)	3.7 (12)	2.8 (19)	2.3 (10)
	Status sensitive habitat (Habitat)	4.1 (6)	3.0 (14)	2.1 (11)
Ecosystem	Ecosystem functioning (Functioning)	3.7 (11)	2.6 (20)	2.1 (12)
Fishing Pressure	Fleet capacity (Fleet)	4.1 (5)	4.3 (1)	2.7 (1)
	Fishing effort by métier (Effort)	4.7 (2)	3.9 (2)	2.6 (3)
	Fishing impact (Impact)	4.5 (3)	3.5 (6)	2.4 (7)

Table 3. Mean specific indicator scoring (with rank order in brackets) for specific indicators (with their short names as used in Figure 2b given in brackets) based on three different questionnaires (for explanation code see table 1). Scores range from 1 (worst) to 5 (best).

Headline indicator	Specific indicator	SN	SS	SF
Physical environment	Temperature (Temperature)	2.9 (41)	3.7 (2)	2.2 (23)
	NAO (NAO)	2.6 (46)	3.0 (25)	1.7 (45)
Chemical environment	Salinity (Salinity)	2.7 (44)	3.5 (4)	2.2 (25)
	Oxygen concentration (Oxygen)	2.8 (43)	3.5 (5)	2.1 (28)
	N and P levels (Eutrophication)	2.9 (40)	3.2 (18)	1.8 (40)
Phytoplankton	Primary production (Prim Prod)	3.1 (37)	2.9 (31)	1.8 (37)
	Water transparency (Wat transparency)	2.1 (51)	3.3 (14)	1.8 (40)
	Chl. *a* (Chlorophyll a)	2.6 (47)	3.1 (21)	1.8 (37)
Zooplankton	CPR-derived plankton indicators (CPR)	2.6 (45)	2.5 (43)	1.4 (51)
	Zooplankton biomass (zooplankton)	3.2 (36)	2.7 (39)	1.8 (40)
Abundance commercial stocks	Proportion within safe biological limits (Safe Biol Limit)	4.5 (3)	3.4 (8)	2.5 (10)
Abundance other populations	Numerical abundance selected species (Abundance)	4.2 (5)	3.3 (10)	2.3 (17)
	Biomass selected species (Biomass)	3.9 (16)	3.3 (14)	2.4 (16)
	Measure of decline (Meas Decline)	4.1 (8)	3.0 (26)	2.2 (23)
Size/age structure species	Average length selected species (Average length)	4.2 (6)	3.5 (6)	2.6 (5)
	Average weight selected species (Average weight)	3.9 (14)	3.4 (7)	2.6 (5)
	Average age selected species (Average age)	3.7 (21)	3.3 (9)	2.5 (10)
Genetic composition species	Maturation norm (Maturation norm)	2.8 (42)	2.6 (40)	1.7 (45)
Size structure community	Mean weight (Mean weight)	3.4 (29)	3.3 (13)	2.6 (2)
	Mean length (Mean length)	3.4 (29)	3.3 (12)	2.6 (2)
	Proportion of large fish (% large fish)	3.5 (25)	3.3 (10)	2.6 (5)
Species composition community	Mean maximum length (Mean max len)	3.3 (32)	3.2 (20)	2.6 (5)
	Biodiversity - Hill's N0 (Biodiversity N0)	2.4 (50)	2.4 (45)	1.8 (37)
	Biodiversity - Hill's N1 (Biodiversity N1)	2.4 (48)	2.4 (46)	1.9 (34)
	Biodiversity - Hill's N2 (Biodiversity N2)	2.4 (48)	2.4 (46)	1.9 (34)
	Proportion of target species (% target spcs)	3.3 (32)	3.1 (23)	2.5 (10)
Abundance community	Total numbers (Total numbers)	3.5 (24)	2.9 (31)	2.5 (10)
	Total biomass (Total biomass)	3.5 (27)	2.9 (29)	2.5 (10)
Status marine mammals (40)	Abundance selected marine mammal species (Mammals)		3.9 (16)	2.8 (33)1.8
Status seabirds	Abundance selected seabirds species (Seabirds)	3.6 (22)	2.8 (37)	1.7 (48)
Status marine reptiles	Abundance selected marine reptile species (Reptiles)	3.1 (37)	2.8 (35)	1.5 (49)
Status benthos	Abundance sensitive benthic species (Sens. Benthic)	3.9 (16)	2.8 (33)	2.3 (20)
	Epibenthos community (Epibenthos)	3.3 (31)	2.6 (41)	2.1 (31)
	Infauna community (Infauna)	3.0 (39)	2.4 (44)	1.9 (34)
Status sensitive habitat	Area coverage sensitive habitats (Habitats)	3.5 (25)	3.1 (22)	2.2 (25)
Ecosystem functioning	Ecosystem functioning (Ecosystem funct)	3.8 (20)	2.1 (50)	2.1 (28)
	Primary Production Required (PPR)	3.6 (23)	2.6 (42)	2.0 (32)
	Catch ratios (Catch ratios)	3.9 (19)	3.1 (24)	2.3 (17)
	Mean transfer efficiency (Transfer eff)	3.2 (34)	2.2 (49)	1.8 (40)
	Trophic level (Trophic level)	3.9 (14)	2.7 (38)	2.3 (20)
	Fishing in Balance index (FIB)	3.2 (34)	2.3 (48)	1.7 (45)
	Finn Cycling Index (Finn Cycling)	3.4 (28)	2.0 (51)	1.4 (50)
Fleet capacity	Fleet capacity (Number vessels)	4.2 (6)	3.9 (1)	2.8 (1)
Fishing effort	Fishing effort (Hours fishing)	4.5 (2)	3.7 (2)	2.6 (2)
Fishing impact	Mortality commercial species (Mort Commercial)	4.6 (1)	3.2 (19)	2.6 (5)
	Mortality other fish species (Mort Other fish)	4.1 (8)	3.0 (28)	2.3 (20)
	Mortality benthic species (Mort Benthic)	4.1 (11)	2.8 (35)	2.1 (28)
	Mortality marine mammals (Mort Mammals)	4.1 (11)	3.0 (27)	1.9 (33)
	Mortality vulnerable species (Mort vulnerable)	4.5 (3)	2.9 (30)	2.2 (25)
	Proportion catch discarded (Catch discarded)	4.1 (11)	3.2 (16)	2.4 (15)
	Proportion area affected (Area affected)	4.1 (8)	3.2 (16)	2.3 (17)

Table 4. Similarity in rank-order of scores for (a) headline and (b) specific indicators based on different scoring scenarios using Spearman correlation coefficients (for explanation code see table 1).

a. Headline indicators

	SN	HC	SS	HF
HN	0.82	0.48	0.38	0.78
SN	-	0.48	0.37	0.62
HC	-	-	0.88	0.69
SS	-	-	-	0.60

b. Specific indicators

	SS	SF
SN	0.31	0.58
SS	-	0.67

Figures

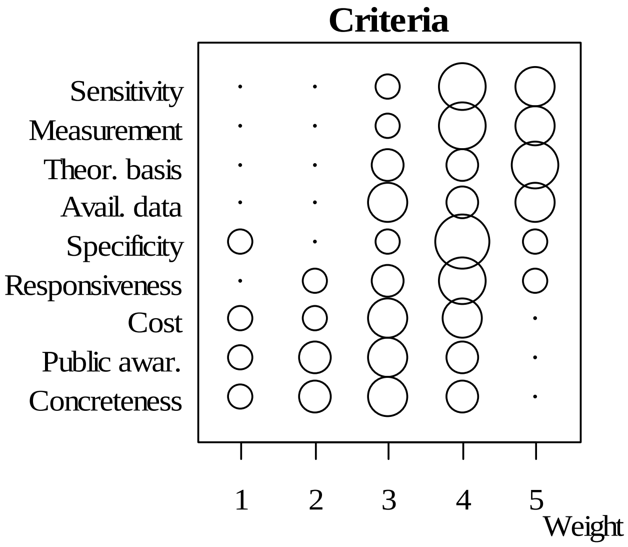

Figure 2a.

Figure 1. Frequency distribution of weights given by all respondents against main criteria WC (top) and against sub-criteria WS (bottom). Bubble area is proportional to the frequency of allocation of each weight, circles of radius zero are plotted as dots. $VT How is zero represented, as a dot? Might be better to use a different symbol. Or is there no zero?$ Criteria are ordered by increasing sum of weights obtained in the WC questionnaire.

a) Headline indicators

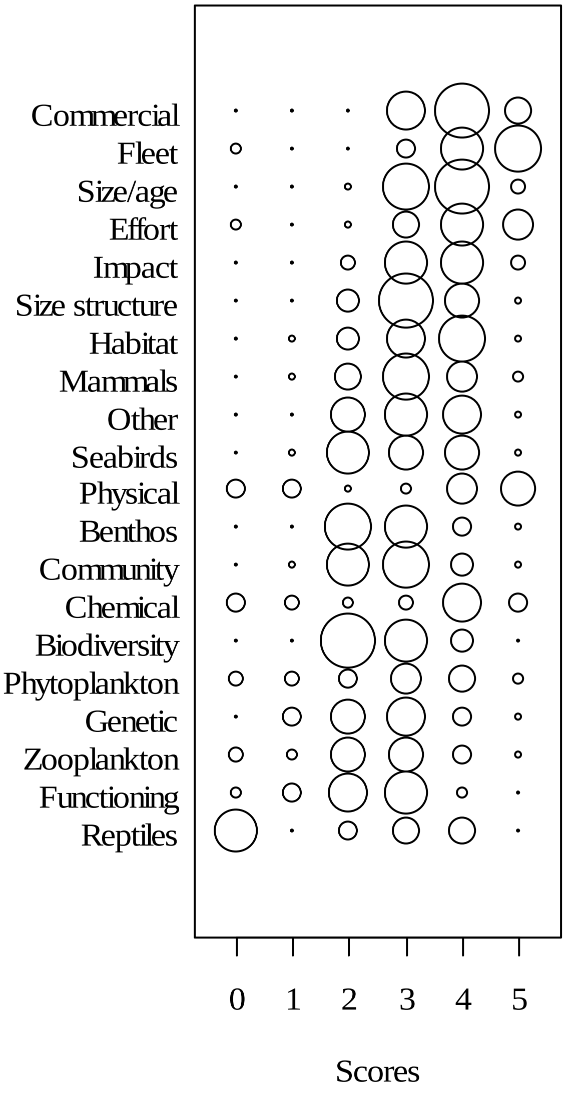

Figure 2b.

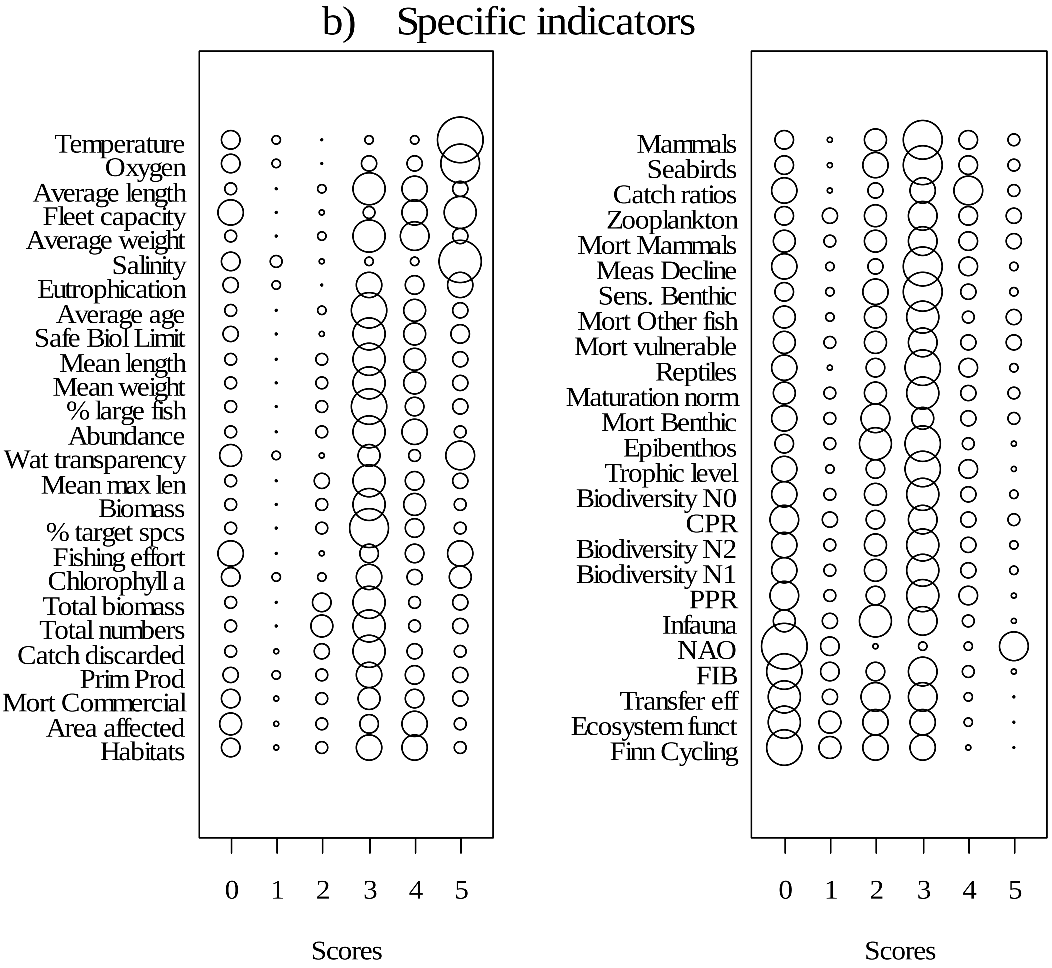

Figure 2. Frequency distribution of scores, ranked by average score, given by all experts to (a) headline indicators (HC evaluation) and (b) specific indicators (SS evaluation). Bubble radius is proportional to the frequency of each score (for abbreviations of indicator names see tables 2 and 3, respectively).