ICES Journal of Marine Science
MAY 2005; 62(3) : 528-539
http://dx.doi.org/10.1016/j.icesjms.2005.01.007

Do explicit criteria help in selecting indicators for ecosystem-based fisheries management?

Marie-Joëlle Rochet[a, *]and Jake C. Rice[b]

[a] Laboratoire de Mathématiques Appliquées à l'Exploitation des Ressources Halieutiques et Aquacoles, IFREMER, B.P. 21105 Nantes Cedex 03, France
[b]Canadian Science Advisory Secretariat, Department of Fisheries and Oceans, 200 Kent Street, Ottawa, Ontario, Canada K1A 0E6

*: Corresponding author : M-J. Rochet: tel: +33 240 374121; fax: +33 240 374075

Abstract:

An evaluation framework developed to help select an appropriate suite of indicators to support an ecosystem approach to fisheries management was tested experimentally by asking independent experts to weight the selection criteria provided and to score indicators against those criteria in several ecological settings. The steps in selecting indicators proved to be prone to subjectivity and value judgement, and differences in scores between experts were the main factor contributing to variability in evaluation results. Having to justify scores in a written document did not improve consistency among the experts. The framework, however, did enhance transparency by explicitly stating each issue to be addressed in the selection process, and by giving experts or stakeholders the opportunity to present their values explicitly. For example, using a longer list of simpler selection criteria appeared to provide less controversial results than a shorter list of more complex ones.

Keywords: criteria weights; ecosystems; fisheries management; indicators; selection criteria

Introduction

Rice and Rochet (2005) developed an evaluation framework for selecting an appropriate suite of indicators for supporting an ecosystem approach to fisheries management. This framework was based on experience as science advisers, but its usefulness should be established empirically. As a start, some of the steps were tested by having experts use it, and critically examining the outcomes. As part of the work programme of SCOR-IOC WG 119, an array of size-based (Shin *et al.* This Symposium) and trophodynamic indicators (Cury *et al.* This Symposium) has been evaluated. In addition, two experiments were conducted to evaluate whether the framework ensures consistent responses in selecting indicators, and whether it provides insight into the judgements of the subjects. Although an objective selection is not achievable, identifying which steps of the framework are sensitive to subjectivity could improve transparency in the selection process.

The evaluation framework is structured as a sequence of 8 steps (Rice and Rochet, 2005). This decomposition disentangles the numerous issues to be addressed in the selection process, enhancing efficiency and transparency. However, subjective choices must be made in every step. Step 1 (determining the user needs) will be influenced by the experience of those responsible for management. Steps 2 (listing candidate indicators) and 4 (scoring indicators against criteria) will be sensitive to the background of those involved in the process. Step 3 (determining screening criteria) requires value judgements about the importance of scientific and governance issues to stakeholders. Steps 5 to 7 (summarizing scoring results; deciding how many indicators are needed; and final selection) require interpretation of results of multivariate pattern analysis, which has substantial opportunity for confirmatory biases. Once the suite of indicators has been selected, there will be subjectivity in the way chosen for reporting on them (step 8).

Testing the whole evaluation framework and determining which steps are most sensitive to subjectivity would be interesting, but difficult to achieve. In integrated management settings, every step requires interaction among technical experts, managers, politicians, and community leaders (Belfiore 2003), and cannot be replicated readily. Even were selection left to technical experts in a test environment for a single ecosystem, the costs and labour of obtaining replicate samples of all steps in the proposed framework would be prohibitive.

Instead, our experiments focused on steps 3 and 4 (determining screening criteria and scoring indicators against criteria), not because these steps were considered particularly critical, but rather because experimental testing seemed feasible. Moreover, these steps require participants to state explicitly what values they are applying in making their choices. They are designed to provide a common factual basis for the guiding value-laden interactions during the subsequent steps. Clarifying how participants differ in applying scores should lead to more informed and dispassionate dialogue, and readier consensus, later on in the process.

For our experiments, independent experts were asked to evaluate a common set of indicators in several ecosystem settings. The differences in the weights given to the various criteria and the scores among experts and settings should shed light on factors of variation in the ultimate selection. To avoid confusion arising from terms being interpreted differently by experts from different regions or disciplines (ICES, 2001), the definitions used in the context of the experiments are given in Table 1.

Table 1. Terminology used in the paper, and information about the experimental design (BB = Bay of Biscay, GE = Gambia River Estuary, SS = Eastern Scotian Shelf, BC = Central Coast of British Columbia).

Term (variable type)	Explanation	# levels in E2
Evaluation framework	Framework of 9 steps for evaluating ecosystem indicators (Rice & Rochet This Symposium)	
Indicator (response var)	A characteristic of the marine environment that may be informative about its state and/or the impact of human activities	21 (see table 2)
Criterion (response var)	A characteristic on which an Indicator is evaluated for its information content or status. Some are split into several sub-criteria.	9 (Tables 2 and 3 in Rice & Rochet 2005)
Target group (design var)	Group involved in selecting indicators, and/or interpreting their values	4 – scientists, managers, politicians, community members
Ecosystems (design var)	The areas for which the evaluation framework was tested	4 – BB, GE, SS, BC
Authority (design var)	The level of knowledge of the ecosystem by subjects	2 – Local/non-local expert
Subjects (design var)	The individuals participating in the experiments	16 - BB: 4; GE: 5; SS: 5; BC: 2

Materials and methods

· Experiments

Experiment 1 (E1) deals with individual variation in scoring Indicators and internal consistency of criteria. Experts in the fields of environmental or size-based Indicators were provided with suites of Indicators (Table 2) and asked to score these against the nine Criteria of the Framework, either as a summary list together with their arguments (for environmental Indicators), or as a simple table of scores from 1 to 5 (for size-based Indicators). Having to write an argumentation was hypothesized to reduce subjective judgement in the scoring process: the degree of agreement between two independent evaluations done in this way was used to assess which Criteria are most prone to subjective judgement. The scoring process is a complex one. The property that a criterion is trying to capture can often be viewed from different perspectives (sub-criteria), and the information on these different aspects might be conflicting. Thus, interpretation may not always be straightforward. The evaluation table of size-based Indicators was synthesised by a factor analysis (Mardia *et al.* 1979), a classical tool for analysing questionnaire surveys, which groups similar response profiles and relates them both to sets of questions and sets of respondents. Visualising the scatter of sub-criteria within each criterion helps determining which Criteria have consistent sub-criteria and which do not. The exercise also provides an example of the way complex evaluations can be analysed.

The second experiment (E2) addressed the robustness of the weighting and scoring process for differences in expert knowledge, based on the H_0: When using the evaluation Framework, there are no differences in the evaluation of Indicators between scientists without and those with detailed knowledge of a particular ecosystem (“non-local” and “local” experts, respectively).

To test this hypothesis, 20 experienced scientists (Subjects) were asked to evaluate a selection of 21 candidate Indicators for four different Ecosystems separately. The Ecosystems were selected to provide contrast among the test cases. Each Subject was familiar with one Ecosystem, but was unlikely to have local knowledge of the other three. Again to provide contrast, not all Indicators necessarily represented sensible choices for a given Ecosystem.

Table 2. Indicators used in the two experiments and their labels as used in the figures (see Shin *et al*., 2005 for a full description of size-based indicators). Categories: E: environmental, Sp: species-based, Sz: size-based and T: Trophodynamics Indicators.

Code	Experiment	Category	Full name
catchR	E2	T	Catch ratios
CPR	E1+E2	E	CPR derived plankton indicators
Divers	E2	Sp	Traditional measures of species diversity
endangR	E2	Sp	Ratio of endangered to non-endangered species
ENSO	E1+E2	E	ENSO or SOI
Exrate	E2	Sp	Exploitation rate
FIB	E2	T	FIB
GenDiv	E2	Sp	Genetic diversity
K	E1+E2	Sz	Fulton’s condition index
Large	E2	Sz	Proportion of large species in assemblage
lbar	E2	Sz	Average length of fish
lbarage	E1	Sz	Mean length at age
lbarcom	E1	Sz	Mean body size of community
lbarpop	E1	Sz	Mean body size of population
lmatpop	E1	Sz	Mean length at maturity of population
lmat	E2	Sz	Mean length at maturity of fish assemblage
lmaxcom	E1	Sz	Mean maximum length of community
lmaxpop	E1	Sz	Maximum length of population
NAO	E1+E2	E	NAO
PDO	E1+E2	E	PDO
ProdI	E2	T	Primary production required to support catches
ProdR	E2	T	Productivity and consumption ratios
Relabund	E2	Sp	Species distribution / relative abundance
SDI	E1	Sz	Stock density indices
Size.div	E1	Sz	Slope and intercept of size diversity spectra
Sizedistr	E2	Sz	Size distribution of species
Spectrum	E1+E2	Sz	Slope and height of the fish assemblage size spectrum
targetR	E2	Sp	Ratio of target to non-target species
TEbar	E2	T	System mean transfer efficiency

However, the experiment was not aimed at finding the best Indicators, but at testing the Framework itself. Although five Subjects were approached from each Ecosystem, three from British Columbia and one from the Bay of Biscay dropped out, leaving 16 respondents.

Each Subject was provided with information about the four Ecosystems, comprising: i) the Target groups and intended uses for the Indicators; ii) the management objectives; iii) a brief description of the Ecosystem and the fisheries. In addition, they received short descriptions of the 21 Indicators including their potential support to management and data requirements, and the nine Criteria and associated considerations (Table 2 in Rice & Rochet This Symposium). Hence, steps 1 and 2 of the Framework were assumed to have been completed.

Each Subject was asked to weigh the screening Criteria for each Target group in each Ecosystem and to score the Indicators against the Criteria. The resultant data were two matrices per Subject. Matrix 1 contained the “weights” (1=little relevance; 2=low; 3=moderate; 4=fair, 5=high) for each Criterion by Ecosystem and Target Group, and matrix 2 the “scores” (na: not applicable; 1=poor; 2=weak; 3= average; 4=good; 5=excellent) for each Indicator by Ecosystem and Criterion.

· Analyses

To obtain general insight in the responses, patterns in the scores assigned by Subjects were investigated at two levels. First, potential differences in strategies applied by individuals in the scoring process were explored, irrespective of Ecosystem and Target group. Second, we explored how Indicators and Criteria are judged relative to each other, across all Subjects, Target groups, and Ecosystems by a factor analysis of the scores (Mardia *et al.* 1979), with missing data scored '0'. By positioning the indicators and criteria (or criteria and target groups) in a common space, and contrasting the scores (or weights) across Ecosystems, Target groups, and Authority, the degree to which experts differentiate these factors can be visualised.

In addressing the issue of objectivity and consistency in the responses, the H_0 stated earlier is difficult to test directly, because there are multiple response variables, and because the two levels in the key design variable (Authority) differ in many ways. Hence, a single overall statistical test for differences in responses between the levels of Authority could obscure more than it revealed. Although large sample sizes could be obtained for analyses by level of Authority or by Ecosystem if scores are pooled across Subjects, differences among individuals increase variance and reduce statistical power. More seriously, if individual differences are inconsistent across Ecosystems, pooling could introduce bias. Because combined tests might confound judgements about weights and scores applied, the two matrices had to be evaluated separately. Two types of contrasts were used: (1) contrasts across Authorities test the hypothesis that non-local and local experts have similar judgements with regard to importance of Criteria or Indicators, when controlling for Ecosystems; (2) contrasts across Ecosystems test the hypothesis that experts give more similar weights or scores to Ecosystems they are not familiar with, compared with the one they know much about, controlling for level of Authority. The latter effect would be superimposed on scoring strategies of individuals, and therefore weights and scores had to be evaluated Subject by Subject.

Because the scores assigned by Subjects were ordinal rather than continuous, ANOVA-type models may not be valid. Analyses of frequencies were used instead. The following tests were performed (for statistical details see annex 1):

Test 1: Did Subjects give consistent weights to the Criteria across Ecosystems?

Test 2: Did Subjects give consistent weights to Criteria for the various Target groups across Ecosystems?

Test 3: Did Subjects apply consistent strategies to assigning weights to the criteria for Target groups across ecosystems, and did the assigned scores differ between the two levels of Authority?

Test 4: Did Subjects give the same scores to Indicators for Ecosystems for which they were non-local experts?

Test 5: How similarly did local and non-local experts assign scores to the Indicator-Criteria combinations across Ecosystems?

Test 6: If individual differences in scoring strategies are controlled, did the assigned scores differ between the two levels of Authority across Ecosystems? This test represents a powerful design for detecting the effect of level of Authority directly, by building up from individual contrasts with small sample sizes in the contributing tests.

Test 7: Which criteria differ most in scores between local and non-local experts?

Finally, the sets of Indicators ranking highest for local and non-local experts for each Ecosystem were compared as follows: (1) Target groups were ranked, based on the available information about each ecosystem (Table 3); (2) importance of Criteria was ranked by projecting the Target groups on a factor analysis of Criteria weights by local experts; (3) relevance of Indicators to the most important Criteria was ranked by projecting scores by local experts on a factor analysis; (4) this ranking was compared to a similar ranking based on non-local experts scorings and weightings.

Table 3. Target groups in each Ecosystem (see table 1; British Columbia excluded because too few subjects), ranked from the most (1) to the least important users (4).

Target group	GE	SS	BB
Scientists	1	4	2
Managers	2	1	3
Politicians		2	1
Community members		3	

Results

· EXPERIMENT 1

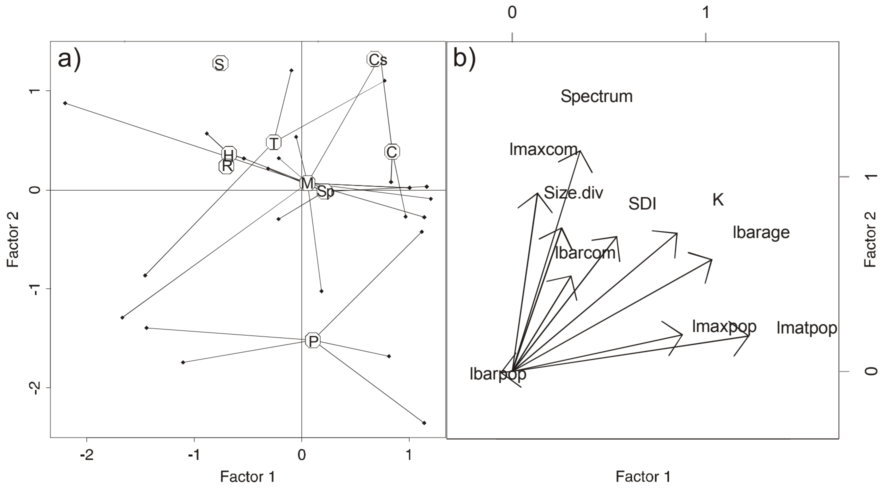
The two independent evaluators for the four environmental Indicators wrote very similar arguments, yet gave different scores to several Indicators on various criteria (Table 4). Agreement between scores was independent on the level of detail in the arguments given for the scores. Hence the hypothesis that writing down the arguments would decrease subjective judgement is rejected. Theoretical basis got no similar scores and is apparently scored with least objectivity. Specificity and Public awareness were also difficult to assess with two disagreements out of four. In contrast, Cost and Availability of historic data received the same scores in all cases, suggesting that quantitative Criteria are easier to score consistently.

Criteria with a high number of sub-criteria are prone to scoring inconsistency. Measurement (11 sub-criteria) and Specificity (3), had scattered sub-criteria scores for size-based Indicators (Figure 1). However, this was not the case for all criteria. All three sub-criteria of Concreteness had consistently high scores for the three most correlated Indicators (Imatpop, Ibarage, and Imaxpop). Similarly, all five sub-criteria of Public awareness had consistently low scores for most Indicators (all have negative scores on Factor 2, Figure 1 A).

Table 4. Scores of the four environmental indicators against the nine evaluation criteria, given by two independent experts after a detailed written examination (u: unknown; n: not feasible; w: weak; m: moderate; f: fair; s: strong). N ≠: number of disagreements.

Criteria Indicators	Concreteness (C)	Theoretical basis (T)	Public awareness (P)	Cost (Cs)	Measurement (M)	Historic data (H)	Sensitivity (S)	Responsiveness (R)	Specificity (Sp)
PDO	w/m	n/w	w/w	s/s	s/s	s/s	n/n	n/n	w/n
ENSO	s/s	n/m	m/s	s/s	s/s	s/s	n/n	n/n	n/n
NAO	s/s	n/m	w/s	s/s	s/s	s/s	n/n	n/n	n/n
CPR	s/s	w/s	m/m	m/m	s/m	s/s	w/m	s/u	w/f
N ≠	1	4	2	0	1	0	1	1	2

Figure 1. Factor analysis of size-based indicators scores (factors 1 and 2, 27 and 26% of total variance, respectively): A) points = sub-criteria connected to their mean: 9 criteria (for codes see Table 4); B) variables: 10 size-based indicators (for codes see Table 2).



· **EXPERIMENT 2**

Despite the small numbers of subjects and unavoidable confounding of Authority and Ecosystems in the design, the high variance in the scores indicate either contrasts among Ecosystems or diverse judgements among experts (Figure 2). Lbar and Sizedistr scored particularly high on Concreteness and Measurement, Proportion of large species in assemblage and Relative abundances scored high on Theory, and the NAO on Availability of historic data. Criteria weights were less variable than Indicator scores, but also had a high dispersion (not shown): Theory and Public awareness yielded the highest contrast among Target groups.

The frequency distributions of both weights and scores by each expert indicate marked individual differences (Figure 3), and experts giving consistently low (or high) scores represented all four ecosystems (Figure 3B). These individual differences complicated the statistical tests, because pooling scores or weights across Subjects would have resulted in insensitive tests.

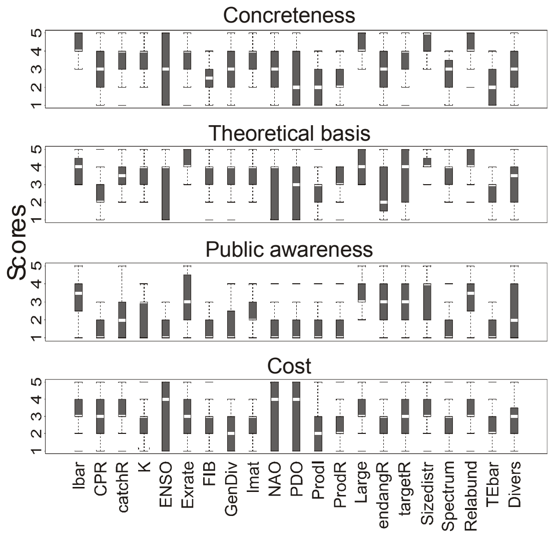


Figure 2. Frequency distributions of scores of the 21 Indicators (for codes see Table 2) on four Criteria, across Subjects, Ecosystems, Authority and Target group. Solid line: median, box ends: upper and lower quartiles. Whiskers: extreme values, excluding outliers (which are plotted individually).

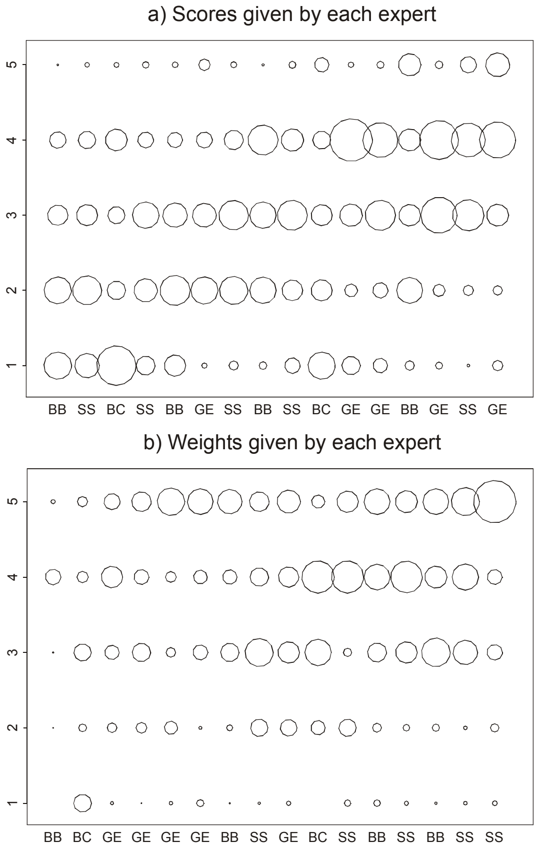
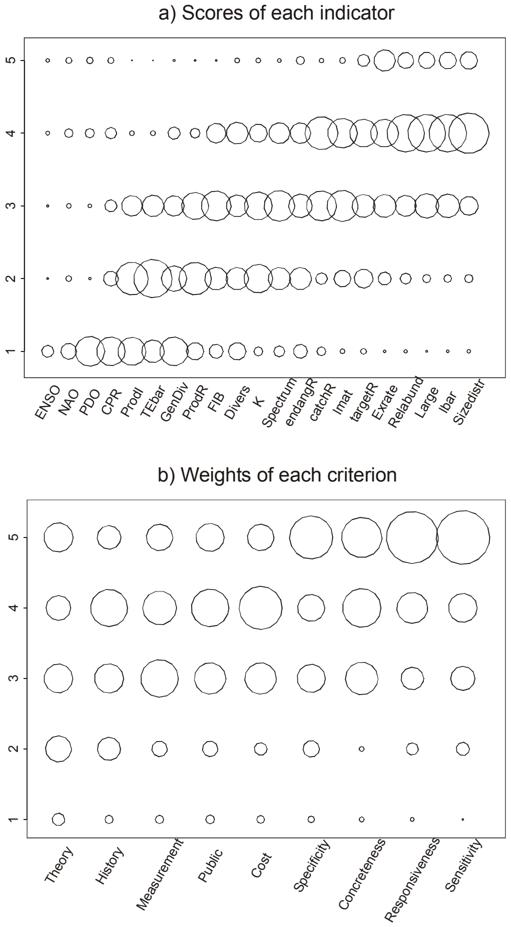


Figure 3. Frequency distributions of scores (A) and weights (B) by all experts (labeled by the ecosystem for which they are local experts), ranked according to the sum of their scores or weights.

The consistency in scores individual subjects assigned to many Indicators (good or poor everywhere) foreshadows the lack of a strong Ecosystem effect. Overall, size-based Indicators received the highest scores, whereas environmental Indicators received the lowest (Figure 4A). Because few Subjects scored ENSO, it was dropped from some tests. Sensitivity and Responsiveness were given the highest weights, whereas History and Theory were given the lowest (Figure 4B).


a) Scores of each indicator

b) Weights of each criterion

Figure 4. Distributions of A) scores of each indicator and B) weights of each criterion, both ranked according to their total value.

Not surprisingly, Responsiveness, Sensitivity and Specificity were most correlated, high for Exploitation rate and low for environmental indicators (Factor 1; Figure 5A). Cost and Measurement formed a second group with high scores for size-based Indicators (Factor 2). The other Criteria were rather independent with high loadings on separate factors: experts considered them to capture different properties of candidate indicators (Figure 5B). The scores on the criteria generally position members of different types of indicators (oceanographic; trophodynamic, species-based, size-based) together in ordination space (Fig 5A-B). However, there is enough scatter within each group that such ordinations should provide guidance in selecting the minimum number of indicators that maximally fill ordination space.

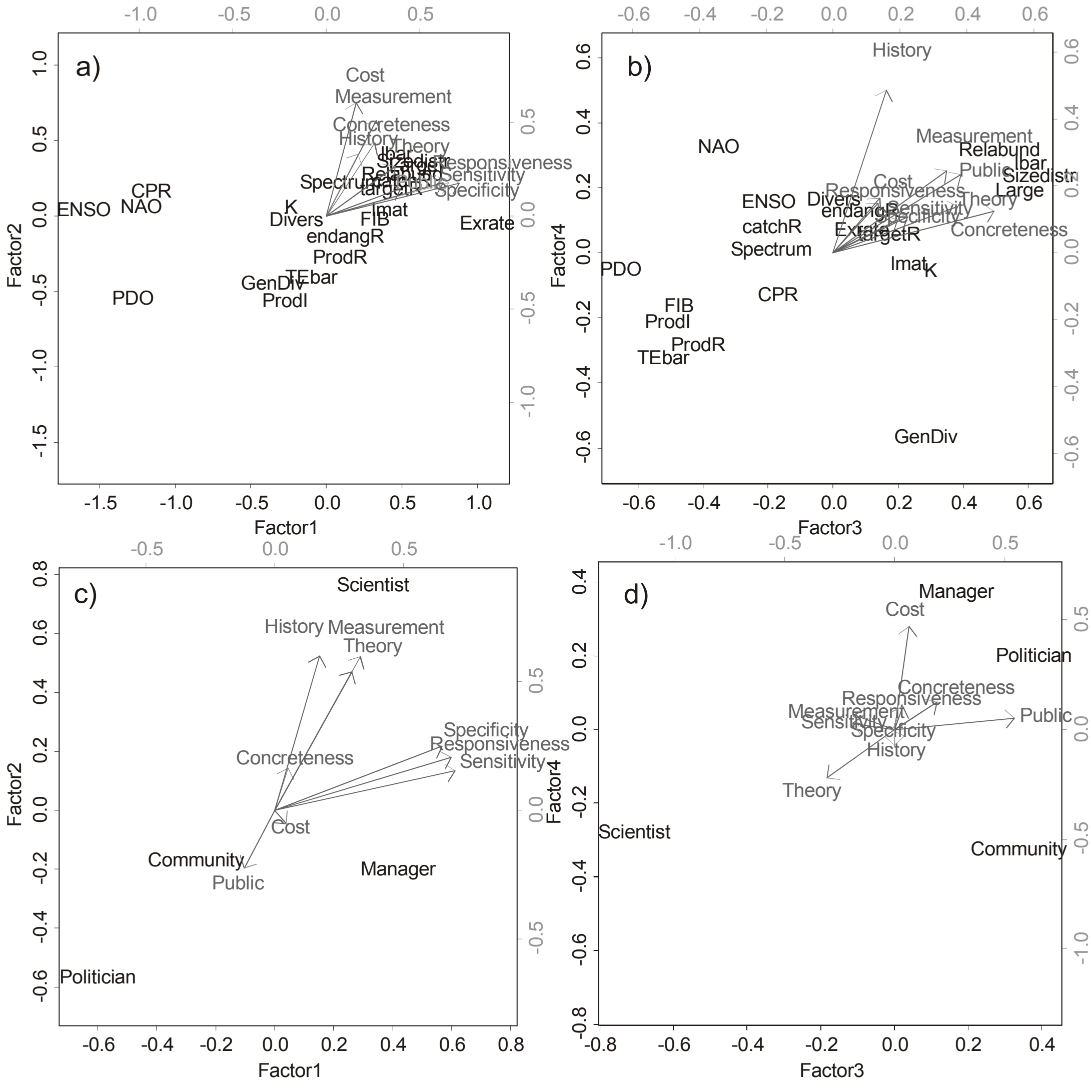


Figure 5. Factor analyses of scores and weights: common plots on factors 1 to 4 of A-B) indicators and criteria (variance explained: 0.31, 0.19, 0.14 and 0.08, respectively; sum= 72%), and C-D) criteria and targets (variance explained: 0.28, 0.21, 0.09 and 0.05, respectively; sum=64%).

Performance criteria (Sensitivity, Responsiveness and Specificity) were also given similar weights (Figure 5C; highest for managers and scientists and lowest for politicians). History, Measurement and Theory formed a second group (high for scientists and low for politicians); Public awareness was considered important to politicians and the community, but not to scientists, whereas Cost received higher weights for managers (Figure 5D).

The following test results were obtained:

Test 1 - Of the eight Subjects having assigned weights to all criteria for all ecosystems and Target groups, six treated all Ecosystems consistently (combined P>0.90 in five cases). The other two assigned very different weights for different Ecosystems (combined P<0.0001 in both cases). This effect was due to different weights to Historic data for the Gambia Estuary (both Subjects), and for Specificity and Responsiveness (one case each; Figure 6). Both were non-local experts, and their weights for all other Ecosystems were similar.

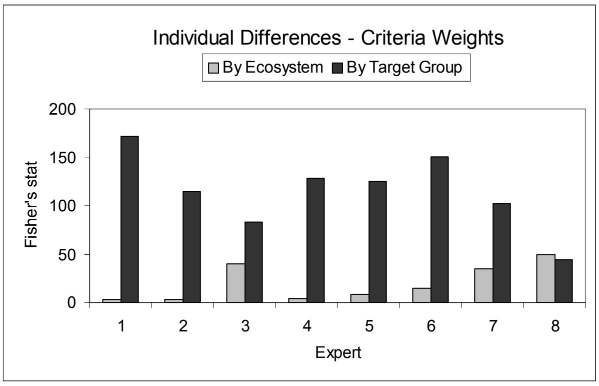


Figure 6. Fisher's statistic (P=0.05: critical value=19.4) for combined G-tests of weights assigned to criteria by 8 subjects, for tests of Authority by Ecosystem (diagonals) and Authority by Target Group (bars).

Test 2 - All Subjects assigned significantly different weights to the different Target groups (Figure 6; probabilities that response profiles were similar ranged from 0.002 to 10^{-25}). Combined with the results of test 1, with two exceptions for the Gambia estuary, local and non-local experts made similar judgements about the weights specific Target groups assign to criteria.

Test 3 - The interpretation of test 2 results is upheld when all response profiles are examined, including those of eight Subjects who did not weight criteria for politicians and community members in some Ecosystems (and hence could not be included in Tests 1 and 2). We applied an arbitrary standard that weight profiles are "similar" across Ecosystems as long as the weights differed by not more than one for any Target group. Nine out of sixteen Subjects always gave "similar" profiles to the Target groups across Ecosystems. Of the remaining seven, four gave dissimilar profiles for at least seven, one for four and two for three of the nine Criteria. However, the weights assigned by these seven were so inconsistent across Ecosystems that overall effects only showed up for Historical Data, and either Responsiveness or Specificity. Hence, the scoring profiles indicate that Subjects were either very consistent (the majority) or very inconsistent in the weights they assigned to Target groups across ecosystems.

Test 4 - Only the Indicator scores given by non-local experts to PDO were significantly different across Ecosystems (Test statistic = 77.02, P<0.001; Figure 7), while values of the test statistic for CPR, Sizedistr, and NAO were almost significant. Overall, the heterogeneity among non-local experts can be largely explained by the scores given to Historic data: when this Criterion was dropped, scoring variance was reduced by at least 75% in 14 out of the 20 Indicators (Figure 7). However, Historical data accounted for less than 10% of the heterogeneity for PDO and about 20% for NAO.

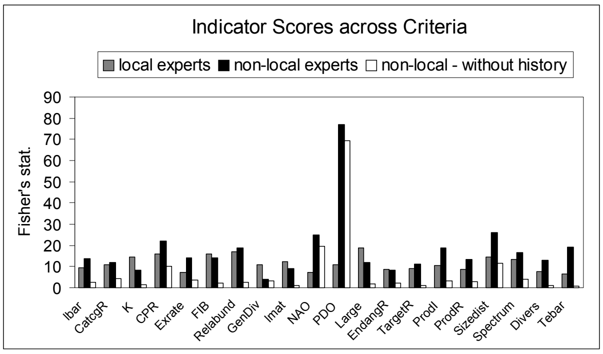


Figure 7. Fisher's statistic for combined G-tests of the scores given to each indicator across 9 criteria (P=0.05: critical value=28.9), by local experts (diagonals), non-local experts (bars), and non-local experts on 8 criteria (leaving out Historic Data – solid bars; critical value=23.0).

Test 5 - Because Subjects tended to assign similar scores across the Criteria for most Indicators and all Ecosystems with which they were not familiar (test 4), and because the design did not allow having Authority and Ecosystem as crossed factors, direct tests of the Authority effect would be inconclusive. Therefore, Fisher's statistics were calculated for each Indicator exactly as in Test 4, but for local experts only (Figure 7). Contrasting test statistics for local and non-local experts by Indicator provides only a heuristic index of their scoring patterns; the differences themselves are not a statistical property. However, the test statistics for differences among local experts are lower than those for non-local experts for 80% of the Indicators. This indicates that local experts made more similar judgements about a given Indicator than non-local experts, irrespective of Ecosystem, compared to when they were scoring the same Indicator for other systems. This effect was particularly strong for the physical oceanographic indicators (test statistic PDO = 10.86, NAO = 7.26, both P>0.5).

Test 6 - This test revealed a significant difference between local and non-local experts in the scores they assigned across all Ecosystems (pairwise t = 1.8; df = 19, P=0.04). Because the skewness in the Fisher statistics from tests 4 and 5 was not completely eliminated by the transformation, the less stringent assumption was made that these statistics have only rank-order accuracy. The non-parametric Wilcoxon ranks test for the two levels of Authority across all indicators (statistic=274; n=40; P=0.02) indicates that the hypothesis of a common sampling distribution for the two levels of Authority must be rejected.

Test 7 - Examining the G-statistic scores, the differences between local experts and non-local experts appeared to be largest in the scores assigned to Historic Data. When the paired t-test was recalculated without this criterion, the difference was no longer significant (paired t = 0.372, df = 19, P = 0.38), supporting the conjecture that the main difference was linked to Historic Data. Local experts tended to assign lower scores to Historic Data than non-local experts for all Ecosystems (Figure 8). This suggests that experts were inclined to trust data for unfamiliar Ecosystems, while emphasising the problems in the data they knew well.

· Indicators ranking

Ordinations of weights and scores by local and non-local experts were similar, but not identical (see Figure 9 for the Scotian Shelf), leading to slightly different rankings of Criteria, but substantial differences between the final ranks of Indicators (Table 5). Overall, the evaluation suggests that Exploitation rate, Relative abundance of species, Proportion of large species or Size distribution of species, and Ratio of target to non-target species should be selected because they are considered as sensitive, specific and responsive (important to managers) and concrete (matters to politicians), provided that Historic data are available (relevant to scientists). Environmental Indicators were selected for scientists because they scored high on the availability of historic data. The overlap between selected suites by local and non-local experts was limited (1/4 to 2/4; Table 5).

Table 5. Suites of indicators selected by local / non-local experts by ecosystem.

Ecosystem	Local experts	Non-local experts	Overlap
BB	1. Large 2. endangR 3. NAO 4. Exrate	1. Exrate 2. Relabund 3. lbar 4. Sizedistr	1/4
GE	1. Large 2. Relabund 3. targetR 4. Sizedistr	1. NAO 2. Relabund 3. targetR 4. Exrate	2/4
SS	1. Exrate 2. targetR 3. Large 4. lbar	1. Exrate 2. Sizedistr 3. Relabund 4. lbar	2/4

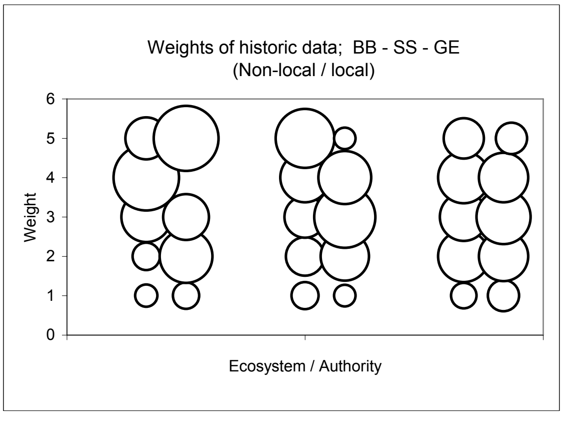


Figure 8. Frequency distribution of weights assigned to Historic data by 8 subjects, by Ecosystem (left: BB; middle: SS; right: GE) and by non-local and local experts (stacked bubbles left and right, respectively).

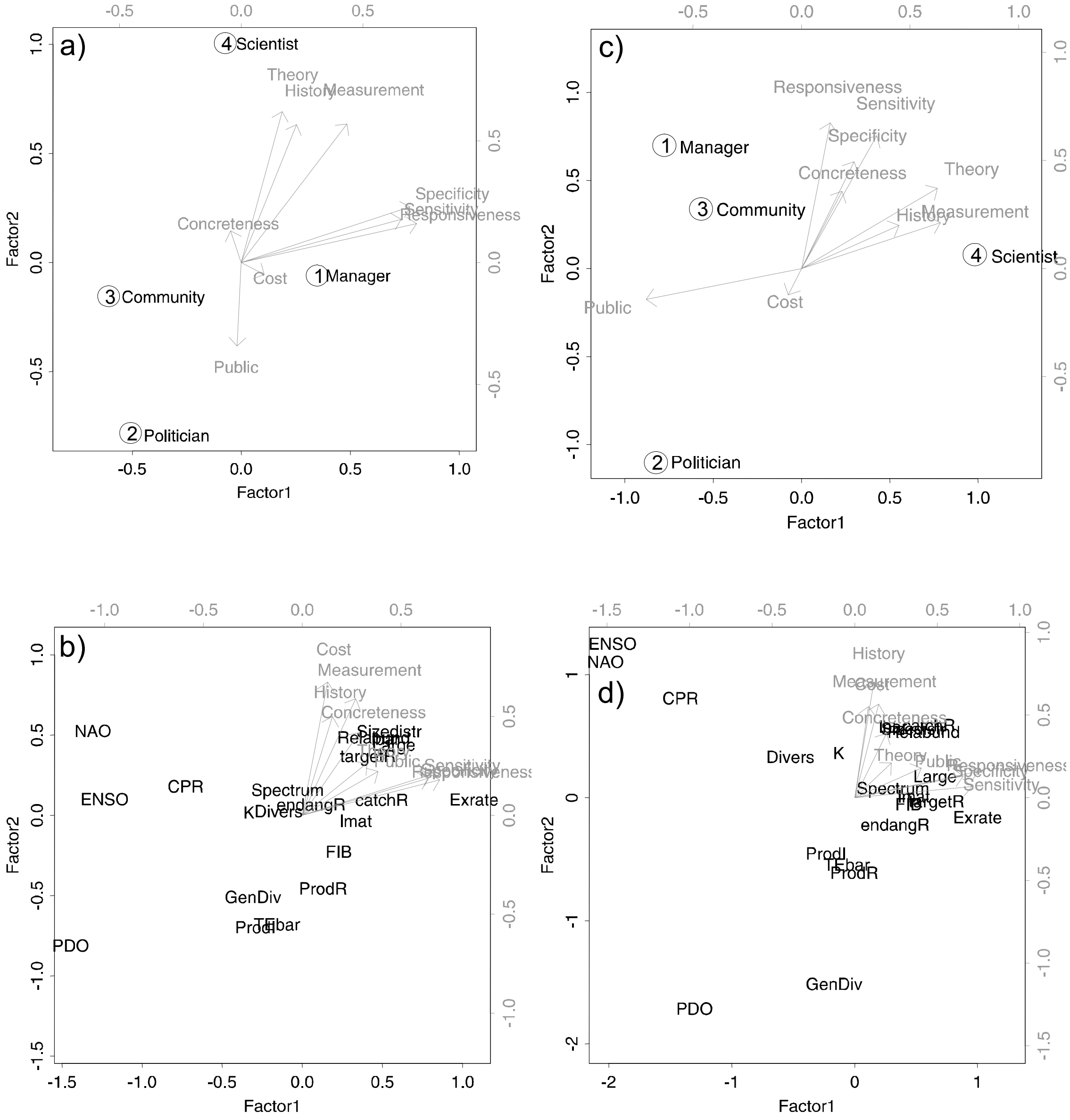


Figure 9. Factor analyses of criteria weights (top) and indicators scores (bottom) by non-local (left) and local (right) experts for the Scotian Shelf.

Discussion

Experiment 1 clearly shows that written arguments do not lead to consistent scores delivered by different experts, even though they may agree on the broad characteristics of a criterion. This may be partly ascribed to a poor description of the scoring procedure required by our framework. More generally, there is a serious need to develop scoring methods for many purposes (see also Rice & Rochet This Symposium).

Criteria based on multiple sub-criteria may be difficult to score consistently. Experiment 1 with size-based indicators, which are a rather homogeneous groups with regards to many of the nine criteria, clearly shows that a larger number of simpler criteria would be more easy to interpret than a smaller list of criteria with sub-criteria. For example, Measurement could be split into Variance, Bias, Measurement error and Representativeness, each of which may be easier to score than if they are integrated in a single score. There is a cost, however, in that a longer list of scoring criteria makes subsequent steps of consolidating response profiles more complicated.

Consistent with experiment 1, the main effect found in experiment 2 was the dominance of individual differences in scoring and weighting strategies over effects of Authority or Ecosystem. Although subjects assigned differential weights to the criteria for different Target groups, most subjects tended to assign similar weights to a given Target group across ecosystems, irrespective of their familiarity with these. The only major difference was in the value assumed for Historic data, where local experts tend to give lower scores. Formally the null hypothesis of Experiment 2 cannot be rejected but the large individual differences made our tests weak.

Regarding the experimental design, many subjects commented that they did not have enough relevant information to feel confident assigning scores. This is inherent in experimental testing and fundamentally different from a true application, when much more Ecosystem information is available. Some subjects also commented that environmental indicators did not deserve the same treatment as the others. Indeed, they got extraordinary scorings and in the multivariate descriptions accounted for a large part of observed variance. This would be an incentive to differentiate between criteria weights for separate groups of indicators: indicators of fishing impacts have to be sensitive to fishing and responsive to management actions, but these criteria are irrelevant for environmental indicators. Scoring was felt the most difficult and arbitrary step and users would benefit from clear guidance.

Many subjects wondered why they were requested to weight criteria for target users for each ecosystem separately. Indeed, the results suggest that experts judge that the importance of criteria to Target groups hardly varies across ecosystems. However, in a true application, Target Groups will decide for themselves what weights to apply to criteria and indicators. This may challenge the present finding. If the importance of the various criteria to individual Target groups is indeed uniform across systems, then using globally established weightings may save time and effort in real applications. Differential scoring of indicators were also difficult for criteria like Theory or Concreteness. Again, these are general properties of a specific Indicator, and may need to be established only once. Measurement, Sensitivity, Responsiveness and Specificity were found the criteria most likely to vary between ecosystems.

Experiment 2 had two statistical shortcomings. First, although the design was fully balanced, four Subjects did not complete the questionnaires. Hence, sample sizes are too small for some case-specific statistical tests, and alternatives with less statistical power had to be used. Second, the large individual differences found in scoring strategies (Figures 3 and 4) mean that the Subject effect needs to be retained in the models. This is not easy when the key design variables – level of Authority and Ecosystem - are inherently confounded. The ideal analyses would have been fully crossed with Authority, Ecosystem, and Target group, looking at higher-order interactions. However such a design would have required an impractically large number of subjects from each ecosystem, and each subject can only be a local expert or a non-local expert with regard to each Ecosystem.

The design also had some strengths. The multiple Criteria and multiple Indicators serve as replicate tests, as long as results are tested pair-wise across Authority, Ecosystems, and Target groups. The possibility to test these effects for 21×9 combinations means that the frequency distribution of outcomes can be used, without giving undue importance to the significance of individual tests. The logic is similar as for meta-analysis. Although small expected values in some cells affect the sensitivity of each test, the pattern across many tests is informative. The occasionally small expected values make the individual probabilities of the test statistics unreliable, which a meta-analysis approach mitigates substantially.

Both experiments, in different ways, highlight the importance of differences in perspective among experienced scientists. The conclusion must be that Steps 3 and 4 do not ensure objective and consistent evaluations of ecosystem indicators in individual applications. Even though the indicators are strongly science-based, their evaluation inherently involves value judgements and is disconnected from scientific rigour. The available methods to scale the scores and weights to some consistent standard would not help. Because personal interactions feature prominently in essentially every other step in the framework, individual differences should be confronted rather than scaled away.

This situation gives the framework an added, and unintended value. Because it forces experts to present their values explicitly, it provides a body of factual information from which to commence dialogue among participants. Examining the sources of different preferences and concerns may allow consensus on a final suite of indicators to be reached more quickly and more amicably. When compromises are finally reached, it becomes easier for groups representing contrasting values to see where their interests have been served. The large individual differences also highlight the importance of the evaluation method used (cf Table 2 in Rice and Rochet, 2005): in both experiments, evaluations were based on ‘judgement within team’, which was ranked as the method that would provide least confidence in the outcome. However, in real applications states and agencies will have to invest considerable resources in the monitoring of the indicators to be selected, often basing decisions with far-reaching ecological, social, and economic consequences on them. Therefore, those who wish to have a science basis for their ecosystem approach to management would be wise to invest in advance testing of indicator performance rather than just asking experts to select a suite of them.

The ultimate test of the framework is if it leads to useful choices for particular ecosystems. This cannot be evaluated until there is experience from real-world applications. The performance testing needs to employ some formal evaluation method, e.g., retrospective tests based on signal detection theory, or rule-based management with monitoring and feedback controls. Such tests lie in the future.

Acknowledgements

We are indebted to Keith Brander, Lynne Shannon, Franz Mueter, Yunne Shin and Kees Zwanenburg, the members of SCOR WG 119 who provided materials for our experiments. We are grateful to the subjects of Experiments 1 and 2 for their patience in trying to understand and apply our instructions: Jean-Jacques Albaret, Joe Arbour, Benoît Beliaeff, Jacques Bertrand, Bob O'Boyle, Keith Brander, Alida Bundy, Jean-Dominique Durand, André Forest, Glen Jamieson, Paul Keiser, Raymond Laë, Franz Mueter, Jacques Panfili, Jean-Charles Poulard, Yunne Shin, Guy Vidy, Kees Zwanenburg, and an anonymous expert. We thank Verena Trenkel for comments on our experimental design, and Kjellrun Hauge and Niels Daan for valuable comments about a previous version of this manuscript.

References

Belfiore, S. 2003. The growth of integrated coastal management and the role of indicators in integrated coastal management: introduction to the special issue. Ocean and Coastal Management 46: 225-234.

Cury, P., L.J. Shannon, J.-P. Roux, G. Daskalov, A. Jarre, D. Pauly & C.L. Moloney. This Symposium. Trophodynamic Indicators for an Ecosystem Approach to Fisheries.

Mardia, K.V., J.T. Kent & J.M. Bibby. 1979. Multivariate analysis. New York: Academic Press.

Rice, J. & M.J. Rochet. This Symposium. A framework for selecting a suite of indicators for fisheries management.

Shin, Y.-J., M.-J. Rochet, S. Jennings, J. Field & H. Gislason. This Symposium. Using size-based indicators to evaluate the ecosystem effects of fishing.

Sokal, R.R. & F.J. Rohlf. 1995. Biometry. New York: W.H. Freeman and Company.

Annex 1. Analytical details for the statistical tests reported for Experiment 2. Methods for Tests 3 and 7 explained in text.

Test 1.

Goodness-of-fit tests (G-test) of frequency distribution of weight (1-5) by ecosystem (1-4), for subjects and criteria separately; replication through level of authority and target groups.

Probabilities of G-statistics calculated directly. G-tests based on 20 observations in either 8 or 12 cells, depending on number of scoring levels used by individual subjects.

Fisher's method for combining probabilities from a set of independent tests (Sokal & Rohlf 1995) was used to sum probabilities for individual subjects across 9 criteria; Fisher's statistic indicates the overall probability that individual subjects assigned the same scores to a given criterion for all ecosystems, regardless of level of authority.

Test 2. As in test 1 but for Target Group as factor with Ecosystem as source of replication.

Tests 4 and 5.

G-test for frequency distribution of scores (1-5) by ecosystem (1-4) for each indicator and criterion separately, using scores only when the subject was not a local expert.

Fisher's statistic (calculated as in tests 1 and 2) indicates the overall probability that subjects gave the individual variables different scores in different ecosystems for which they were non-local experts.

Test 6.

Fisher's statistics from Test 4 and 5 were transformed to make their distributions approximately normal across indicators and two levels of authority. Pair-wise t-test contrasts of the transformed Fisher statistics between the two levels of Authority were then conducted across the twenty indicators.

A one-tailed pair-wise t-test was used, because scores assigned by local experts to each indicator-ecosystem combination were taken as the “correct” scores.