# Combining Multiple Indicators to Determine Conservation Status Based on Expert Preferences

**by**

**Elysia Brunet**

B.Sc., Carleton University, 2007

Research Project Submitted in Partial Fulfillment
of the Requirements for the Degree of

Master of Resource Management

in the
School of Resource and Environmental Management
Faculty of Environment



**SIMON FRASER UNIVERSITY**

**Summer 2012**

# Approval

**Name:** **Elysia Brunet**

**Degree:** **Master of Resource Management (Planning)**

**Title of Project:** **Combining multiple indicators to determine conservation status based on expert preferences**

**Project Number:** **548**

**Examining Committee:**

**Randall Peterman**
Senior Supervisor
Professor
School of Resource and Environmental Management
Simon Fraser University

**Wolfgang Haider**
Professor
School of Resource and Environmental Management
Simon Fraser University

**Andrew Cooper**
Associate Professor
School of Resource and Environmental Management
Simon Fraser University

**Carrie Holt**
Research Scientist
Pacific Biological Station
Fisheries and Oceans Canada

**Date Defended/Approved:** June 13, 2012

## Partial Copyright Licence

## Abstract

Assessments of the conservation status of species depend on using multiple indicators, and most methods for combining indicators either assume that all indicators are equally important or they use some other pre-determined weighting. This article discusses the case of Canada's Wild Salmon Policy, which requires that the biological status of Conservation Units (CUs) of Pacific salmon (*Oncorhychus* spp.) be assessed by combining the status of several indicators or metrics. We developed a questionnaire for experts based on stated preference methods and found that the status of spawner abundance and trend in spawners metrics had the highest relative importance in assessment of CU status, especially for cases with high data quality and amount (DQA). Without information on metric status, DQA had little influence on CU status ratings. Our study presents a novel method for combining indicators to assess conservation status, and in future could be applied to other species and contexts.



**Keywords**: combining indicators; conservation status; expert opinion; conjoint rating; best-worst scaling; salmon;

# Dedication

To interconnectedness, interdependence, and union.

## Acknowledgements

I would like to thank to Randall Peterman for his enthusiasm, support, and dedication. Thank you Andrew Cooper, Wolfgang Haider, and Carrie Holt for your wisdom, insights, and sharing of your love of research. A special thanks to Ben Beardmore, Don Anderson, Pascal Haegeli, Sue Grant, and all the expert respondents. A big thank you to Fisheries Oceans Canada, whose approval of the project made it a reality. Generous financial support was provided by Randall Peterman from an individual Research Grant and funding from Fisheries and Oceans Canada.

I am so very grateful for my wonderful family and friends. You bring love and inspiration to my every day. Thank you for your constant support, encouragement, and listening. A special thanks to Alston's logic, Chantelle's sweetness, Andres' groundedness, and Lesley's fire. Thank you Noah for your brightness, kindness, and fun.

# Table of Contents

# List of Tables

## List of Figures

# 1. Introduction

Worldwide concern about the conservation status of populations and species (Hoffman et al. 2010), as well as the current state and declining trends of many marine fisheries and ecosystems (e.g., Worm et al. 2009; Hutchings et al. 2010; Shin et al. 2010b; Branch et al. 2010) have stimulated a growing need for assessment of their current status and evidence-based conservation efforts (Sutherland et al. 2004). Any assessment of conservation status requires the selection, aggregation, and evaluation of multiple indicators (Turnhout et al. 2007), often in data-deficient contexts. Therefore, assessments and management decisions for species are highly dependent on indicator-based approaches.

However, a key challenge in the assessment of conservation status is to meaningfully combine multiple indicators into a single aggregate status. To this end, several quantitative and semi-quantitative methods exist, including graphical displays (e.g., amoeba, radar, and kite plots, Traffic Light approach; Caddy et al. 2005), composite indices (e.g., index of biotic integrity; Veselka et al. 2010), multivariate ordination (e.g., Principle Component Analysis, multivariate analysis; Romero et al. 2007), productivity-susceptibility analysis (PSA; Patrick et al. 2010), and numerous opinion-based approaches such as the Delphi method (Okoli and Pawlowski 2004), and the Analytical Hierarchy Process (Ananda and Herath 2002). Aggregate status based on multiple indicators is often reported as (1) a precautionary minimum (or maximum), (2) an average (or median, where all indicators have equal weight), (3) a weighted average (or median) (4) an exceedance of a priori decision rules or thresholds, and/or (5) a result of ad-hoc decision rules. These aggregation and reporting methods are currently used to determine the status of marine and terrestrial species (Hoffman et al. 2010), marine and terrestrial ecosystems (e.g., Coll et al. 2010; Hobday et al. 2011), as well as various fisheries (e.g., Caddy et al. 2005; Trenkel et al. 2007; Patrick et al. 2010).

Methods to aggregate status, such as those that classify species according to their chance of extinction, have generally been criticized because they apply arbitrary weighting schemes and use various ad-hoc protocols to reflect the reliability of available data (Adelman et al. 2004). In addition, there is concern over the presentation and use of the resulting status information (Mehlman et al. 2004, Miller et al. 2007). Furthermore, most methods for combining indicators ignore or do not explicitly distinguish between the importance (or weight) of a given indicator (also known as "attribute weight" or "impact") and the importance of the indicator's level or status (e.g., low, medium, or high, also known as "level scale"; Louviere and Islam 2008). Arbitrary weighting schemes or ad-hoc protocols may lead to unwarranted optimism or pessimism in conservation status, and therefore may be less representative of our understanding of the state of populations. Instead, weights of indicators and levels need to reflect the contribution of each to the overall conservation status, relative to the other indicators and levels.

In the literature, the terms indicators and metric tend to be used interchangeably. In this paper, "indicator" is reserved as a general term representing a dimension of conservation status. In the case of Pacific salmon (*Oncorhynchus spp.*), an indicator could be one of four categories or types of measures of conservation status, such as, spawner abundance, trend in spawners, harvest rate, and extent of spatial distribution (Holt et al. 2009). We reserve the term "metric" to represent a quantifiable measure of an indicator. For instance, different metrics of the indicator "trend in spawners" might be the ratio of mean spawner abundance of the current generation to a historical mean, or the rate of change in log-transformed spawner abundances over 3 generations or 10 years (e.g., Porszt et al. 2012). Because in our case study we selected only a single metric for each indicator, we herein use the terms "spawner abundance", "trend in spawners", "harvest rate", and "distribution" as abbreviations for metrics, as explained further below. To assess conservation status, we measure each metric according to a low/ medium/ high scale that follows the Traffic Light approach using the terms red, amber, or green.

All methods used to assess conservation status will inevitably, at some level, use expert opinion. In order to transparently assess a species or population's conservation status, we need to apply a method that is consistent and reliable, and also generalizable to different contexts. In comparison to other methods, stated preference methods elicit expert opinion through the joint evaluation of the factors (i.e., metrics) in the decision

context. Here we applied two response tasks from stated preference methods, namely conjoint rating (Green and Srinivasan 1990) and best-worst scaling (Flynn et al. 2007) to derive expert preferences to enable us to combine multiple metrics into a single assessment of conservation status and obtain the relative importance of metrics, metric status (red, amber, green), and data quality and amount to permit consistent aggregation of those metrics in the status assessment. In this application of stated preference methods, several metrics are evaluated simultaneously by experts in a series of hypothetical scenarios in a questionnaire (Green and Srinivasan 1990). In the conjoint rating task, respondents are asked to rate scenarios that differ in their combination levels (i.e., metric status (red, amber, green), and data quality and amount (high or low)) for each metric. In other words, respondents are forced to make trade-offs between the levels of different metrics (e.g., in one scenario the metric status of spawner abundance is green but for harvest rate it is amber), which is a method that efficiently elicits their preferences (Cohen 2003) or opinions about the relative importance of metrics. Furthermore, in the best-worst scaling (BWS) task respondents are asked to identify the best (highest) and worst (lowest) metric and metric levels presented in each scenario. Although the term "preferences" is commonly used in stated preference applications, here the term "opinions" is more applicable. Stated preference methods have been applied to various environmental issues, including environmental valuation (Carlsson et al. 2003), park management (Lawson and Manning 2002), conservation management (Sorice et al. 2007; Dorow et al. 2009), and wildlife tourism (Semeniuk et al. 2008). However, these methods are most often used to elicit opinions from the general public, whereas in this article we elicit expert opinion (e.g., Tutsch et al. 2010), and use Pacific salmon (*Oncorhynchus* spp.) as a case study.

In 2005, Canada's federal fisheries department, Fisheries and Oceans Canada (DFO), released its policy for the Conservation of Pacific Wild Salmon ("Wild Salmon Policy" or WSP), which requires the implementation of a several strategies to achieve its goal of "restoring and maintaining healthy and diverse salmon populations". Strategy 1 of that policy is to conduct the standardized monitoring of status of wild salmon populations (DFO 2005). More specifically, Action Step 1.3 of the policy calls for annual assessments of the biological status of Conservation Units (CUs), which are to be based on combining the states of multiple metrics for a given CU. Analogous to the United

States' evolutionarily significant units (ESUs; McElhany et al. 2000), a CU is defined as "a group of wild salmon sufficiently isolated from other groups that, if lost, is very unlikely to recolonize naturally within an acceptable time frame (e.g., a human lifetime or a specified number of salmon generations)" (DFO 2005). Species-specific CUs are now delineated for five salmon species (sockeye (*O. nerka*), chum (*O. keta*), coho (*O. kisutch*), Chinook (*O. tschawytscha*) and pink (*O. gorbuscha*)) across British Columbia and the Yukon (Holtby and Ciruna 2007). The status of metrics in each CU will be determined by evaluating available data against established upper and lower benchmarks that delineate red, amber, and green status zones (Figure 1, Table 1; DFO 2005; Holt et al. 2009; Holt 2009; Grant et al. 2011). The status of these metrics will inform the final biological CU status. At present, numerous methods are used to determine the status of salmon populations (e.g., COSEWIC 2003; COSEWIC 2006; McElhany 2006; Pestal and Cass 2009; Rand 2011, Grant et al. 2011). DFO is currently developing various methods to assess CU status in a generalized assessment framework for all CUs (Holt et al. 2009). One key challenge that DFO faces in implementing Strategy 1 of the WSP is the aggregation of information across multiple metrics into a single assessment of CU status in a consistent and reliable manner (Holt et al. 2009). Our study helps to address this challenge.

There are several key issues to keep in mind when evaluating the conservation status of both marine and terrestrial species. (1) Does it make a difference if we assume that all indicators are equally important in the assessment? (2) Broadly speaking, do we need to account for underlying ecological factors or processes, such as productivity, in the assessment? (3) How do we account for different levels of data quality and amount, given that some species are heavily monitored while others are not at all? (4) Does the determination of conservation status depend on who is doing the assessment? (5) Does the scale (or number of categories) on which one chooses to measure the conservation status matter?

Our objective for this research project was to explore the usefulness of stated preference methods to the process of weighting and aggregating metrics within the context of evaluating the biological status of hundreds of salmon CUs (Holtby and Ciruna 2007). This exploration is also broadly relevant to the other methods and applications mentioned above that assess the conservation status of species and populations. That

broad relevance emerges because we ask several questions that address the issues mentioned above:

Q1. What is the effect of an equal weighting of metrics on the determination of status?

Q2. What is the effect of productivity of salmon populations on the rating of CU status? Which factors influence the effect of productivity?

Q3. What effect does data quality and amount (DQA) have on the rating of CU status? Which factors influence the effect of DQA?

Q4. Do DFO experts differ from other respondents in their assessment of CU status?

Q5. Does the use of different rating scales, 3-point vs. 9-point, affect the CU status?

# 2. Methods

## 2.1. Questionnaire

The top table in Figure 2 represents one of the 49 hypothetical CU scenarios in the questionnaire which simultaneously presents the metrics, metric status, and data quality and amount (DQA). The same four metrics were presented in each CU scenario. We selected one metric for each biological indicator i.e., spawner abundance, trends in spawners, harvest rate, and spatial distribution, from a larger list identified by Holt et al. 2009 (top of table Figure 2). In real-life, the status of each metric is determined by comparing existing data or other knowledge to lower and upper benchmarks that divide the status of each metric status into red, amber, and green zones, and multiple metrics are examined for each indicator (top table of Figure 2). However, in our questionnaire-based study, the metric status presented in each CU scenario was not based on existing data; instead, experts were just provided with definitions of the upper and lower benchmarks for CUs (DFO 2005) and for the four metrics in a supplemental handout (Figure 1, Table 1, and Appendix A). To represent the potential range from near "perfect" information to little or no data, DQA was defined qualitatively as either high or low (top of table Figure 2, and Appendix A). In the questionnaire, productivity was also identified as either high or low and was defined as the intrinsic productivity (number of recruits per spawner at low spawner abundance, i.e., the "*a*" parameter in the Ricker model) of a CU. Unless otherwise specified, we refer to productivity at the CU scale. See Appendix A for more details on the methods.

## 2.2. Experimental design

Of the 55 hypothetical CU scenarios presented to each respondent, 49 CU scenarios represented an orthogonal fractional factorial design (resolution IV; Raghavarao et al. 2011), which is a subset of all possible combinations. The

experimental design (Figure 3) allows for the estimation of main effects and two-way off-diagonal interactions between pairs of metric status when generating answers to Questions 1A and 1B (middle of Figure 2), while accounting for biological constraints between the statuses of different metrics. Specifically, (1) if spawner abundance is red in status, then trend in spawners must also be red in status, and (2) if spawner abundance is amber in status, then trend in spawners must either be amber or red in status. Such interactions are important to consider because they represent potential trade-offs that experts will have to make in their assessment of CU status. These off-diagonal interactions allow the effect of the status of one metric to depend on the status of another metric. For example, the effect on the final status rating for a CU of spawner abundance (Ab) when it is amber (A) may be different when harvest rate (Ha) is red, denoted as AbA-HaR, than when harvest rate is either amber or green. In addition, all two-way interactions between levels of metric status and DQA (e.g., trend in spawners (Tr) is red (R) and DQA is high (H), TrR-DQAH), were taken into account by the design and included in the analysis of Questions 2A and 2B (bottom half of Figure 2). See Appendix B for the complete experimental design.

## 2.3. Response tasks and analysis

Widely used in the field of marketing research for more than 35 years, conjoint rating asks respondents to rate a series of hypothetical scenarios in which attribute levels vary systematically. The other method that we used, best-worst scaling, is increasingly being used in a several fields to elicit opinions by asking respondents to choose the best and worst attribute in each scenario (Flynn 2010). Researchers can then use statistical methods to measure opinions about the attributes and associated attribute levels presented from the trade-offs respondents make when evaluating the scenarios (Caruso et al. 2009). Green and Srinivasan (1990) detail the theoretical development of conjoint analysis, and Alriksson and Öberg (2008) review conjoint analysis methods and their environmental applications. Flynn et al. (2007) provide a best-worst scaling user guide in the field of health. In addition, a brief overview of RUT is given in Appendix C. For this research, we used a conjoint rating task to evaluate the overall biological status of CU scenarios (Questions 1A and 1B in Figure 2) and a best-

worst scaling task to elicit expert opinions about the metrics and metric levels presented (Questions 2A and 2B in Figure 2).

### *2.3.1. Conjoint rating*

To simulate a real-life decision context that stock assessment biologists may face when implementing the WSP, we first asked experts to evaluate all the information presented in each CU scenario and rate the CU's biological status on a 9-point color gradient scale, changing from red to amber to green, separately for high and low productivity (Questions 1A and 1B in Figure 2). Ratings for each CU productivity level were first converted to numerical data (1-9). From the baseline 9-point scale, responses from all experts (All-9), we created two additional data sets composed of only DFO experts' responses (DFO-9) and responses from all experts transformed from a 9-point to a 3-point scale reflecting the three colors (All-3), where 1 = 1, 2 or 3; 2 = 4, 5 or 6; and 3 = 7, 8 or 9. The data sets of experts' responses to the questions Q1A and 1B were used to conduct a conjoint rating analysis to estimate the parameters for the relevant models. According to the data set, the All-9, DFO-9 and All-3 models were used to predict the status of CUs.

For each data set, we estimated the relative importance (or effect) of each rating constant, metric, and metric level (i.e., metric status and DQA) using an adjacent-category ordinal logit model in Latent Gold Choice 4.5 (Vermunt and Magidson 2005) for high and low productivity. As in the marketing research literature (econometrics), the relative importance of the rating constants and metric levels is herein referred to as part-worth utility (PWU), which are the estimated parameters of the following equation:

$$(1) \qquad v_{(q|z_i)} = \beta_q^{con} + y_q \cdot \sum_{m=1}^{M} \beta_m^{att} z_{im}^{att}$$

where $v_{(q|z_i)}$ is the systematic part of the utility or preference for giving a CU scenario $i$ the rating of $q$, where 1≤ $q$ ≤ 9. The parameter $\beta_q^{con}$ is the rating constant or baseline preference for the rating $q$, independent of the metric levels of the CU (defined by the 4 metrics). The parameter $y_q$ is the fixed rating constant and takes on a value equal to $q$ (1≤ $q$ ≤ 9). Superscript $M$ is the number of attributes or metrics, in this case 4, and subscript $m$ refers to a specific metric. Parameters $\beta_m^{att}$ are the part-worth utilities (i.e.,

PWUs or relative importance) for the specific metric levels (including metric level interactions) of the CU scenario $i$, defined by $z_{im}^{att}$. Details of the statistical analysis can be found in Appendix A.

We estimated the main-effects statistical model (no interactions), as well as statistical models with the main effects, plus up to 3 two-way interactions between pairs of metric status (Figure 3, e.g., spawner abundance (Ab) is amber (A) in status and harvest rate (Ha) is red (R), resulting in the acronym AbA-HaR) for a total of 6018 models (i.e., 1 model with no interactions, 33 with 1 interaction, 528 with 2 interactions, 5456 with 3 interactions). All metric levels were retained in all models because together they compose the entirety of the decision context. To account for model selection uncertainty, we used small-sample Akaike Information Criterion ($AIC_c$) model averaging (Burnham and Anderson 2002) across two or more models. All rating constants, metric levels, and two-way interactions PWUs were weighted averages (based on the $AIC_c$ weights) across the models in the top-model set, which included models with a $\Delta AIC_c < 4$ (Burnham and Anderson 2002). To avoid shrinkage of the weighted parameter estimates towards zero, and capture the relative weak effect of some parameters (Grueber et al. 2011), we did not include zeros when a parameter was not present in a model. The above approach was applied to the each of the datasets (All-9, DFO-9, and All-3) separately. The $AIC_c$ relative variable importance (RVI) of the non-main- effects, i.e., interactions, was calculated by summing the relative $AIC_c$ model weights ($w_{jAIC4}$) of all models in the top-model set in which the interaction appeared (Burnham and Anderson 2002, Appendix A).

To evaluate the status of CUs and address each of the five key questions presented at the end of the Introduction section, we generated a new set of hypothetical CUs composed of all possible combinations of the four chosen metrics under the three (red, amber and green) metric status levels, resulting in $3^4 = 81$ CUs. Unlike the CUs presented in the questionnaire, the 81 CUs are characterized by metrics and metric status only, while ignoring data quality and amount (DQA). Of these 81 possible CUs, 27 were removed in order to account for the previously mentioned constraint between the status of spawner abundance and trend in spawners.

For each model, we evaluated the status of the remaining 54 CUs by using the weighted model-averaged PWUs ($\bar{\bar{v}}$) to generate the estimated probability of each rating (1-9) for a given CU scenario $i$ (defined by $z_i$), using the form:

$$(2) \qquad P(x_i = q|z_i) = \frac{\exp\left(\bar{\bar{v}}_{(q|z_i)}\right)}{\sum_{q=1}^{Q} \exp\left(\bar{\bar{v}}_{(q|z_i)}\right)}$$

where $P(x_i = q|z_i)$ is the probability of the respondent rating a CU scenario $x_i$ (e.g., 4) along the 9-point scale. For each of the remaining 54 CUs, the probability of each rating (1-9) was estimated for all combinations of high and low productivity, and high and low DQA. The final CU status of each of the 54 CUs was subsequently determined as the status with the highest summed estimated probability (i.e., red status = summed probability of rating the CU 1, 2 or 3, amber = 4, 5 or 6, and green = 7, 8 or 9). For the All-3 models, we used weighted model-averaged PWUs to estimate the probability of each rating (1-3) for each of the 54 CUs for all combinations of high and low productivity, and high and low DQA (4 models). The CU status was determined as the rating with the highest estimated probability (i.e., red status = probability of rating the CU 1, amber = 2, and green = 3).

We compared the status of the 54 CUs from different models to address each question. Specifically:

Q1. To determine the effects of assuming that the metrics have equal weight, we calculated the status of each of the 54 CUs as the average (mean) metric status, where red = 1, amber = 2, and green = 3 (each metric has a weight of 1). We compared the CU status of the All-9 models to the status of the equal weight method.

Q2. To determine the effects of productivity, we compared the CU status of high productivity models to the status of low productivity models in each data set, for a given level of DQA (either high or low).

Q3. To determine the effects of DQA, we compared the CU status of high DQA models to the status of low DQA models in each data set, for a given level of productivity (either high or low).

Q4. Because of the small number of respondents and the large number of model parameters, we could not segment the rating data into DFO (n=27) and non-DFO (n=10) responses to determine whether these two groups' responses were statistically significantly different. Instead, to determine the effect of group

composition, we compared the CU status of the All-9 models to the status of the DFO-9 models.

Q5. To determine the effects of simplifying the rating scale, we compared the CU status of All-9 models to the status of the All-3 models. We conducted this analysis because the adjacent-category ordinal logit model defined in Equation (1) assumes equal distance between the ratings, i.e., it assumes that the difference in utility gained by rating a CU a 3 versus 4 is equal to rating it as a 1 versus 2 on the 9-point scale. However, in actuality, the transitions between 3 and 4, and 6 and 7 on this rating scale represent actual changes in CU status i.e., the former from red to amber status, and the latter from amber to green status, whereas the change from 2 to 3 does not represent a change in CU status.

For each comparison, we also counted the number of CUs that increased, decreased, and remained the same in status between models. Specifically, we compared in (Q1) high and low productivity models, (Q2) All-9 models and equal weighting method, (Q3) high and low DQA models, (Q4) All-9 and DFO-9 models, and (Q5) All-9 and All-3 models. We chose the CU status (red, amber, or green) as the status in each CU with the highest estimated probability. However, the CU status is uncertain when the difference between the status categories with the highest and second highest probabilities is <5%. For example, say the probability distribution of status for a given CU for model 1 is red = 70%, amber = 30%, green = 0%, and for model 2 is red = 40%, amber = 44%, green = 6%. For model 2, the difference in probability between the CU status (amber) and the status with the next highest probability (red) is less than 5%. The uncertainty in the CU status of model 2 extends to comparisons of CU status between models. Although there is an increase in status from model 1 (red) to model 2 (amber), it is uncertain whether it truly represents an increase. To highlight the degree of uncertainty in each model comparison, we also counted how many CUs were “definitely” increasing, decreasing or the same in status, and “likely” increasing, decreasing or the same in status. CUs that are “likely” increasing, decreasing or the same in status are those where the difference between status categories with the highest and second highest probabilities is less than 5% in one or both models in the comparison ($\Delta$ probability < 5%). The example above is a case of a likely increasing CU. CUs that are “definitely” increasing, decreasing, or the same in status are CUs where the difference between status categories with the highest and second highest probabilities is 5% or greater in both models of the comparison ($\Delta$ probability >= 5%). A modified

version of the example above would show a definitely increasing CU if the probability distribution of status for model 1 is red = 70%, amber = 30%, green = 0%, and model 2 is red = 40%, amber = 46%, green = 4%. Lastly, an example of a CU that is "likely" the same in status is where the probability distribution of status for model 1 is red = 51%, amber = 49%, green = 0%, and for model 2 is red = 60%, amber = 40%, green = 6%. Model 1 has less than a 5% difference between the probability of red and amber status. We use an arbitrary cut-off of 5% to highlight the number of CUs in each model comparison where there is uncertainty in status. In other contexts, the cut-off could be set higher or lower. Alternative cut-offs may change the portion of CUs that are "likely" versus "definitely" increasing, decreasing, or the same in status in each model comparison, but would not affect the overall findings (i.e., number of CUs increasing, decreasing or the same in status).

In addition to accounting for model uncertainty through model averaging, we also incorporate parameter uncertainty in the model-averaged PWUs (from best estimates) by randomly drawing 5,000 parameter estimates for each model in the top-model set from a multivariate normal distribution, taking into account the model's variance-covariance structure (Venables and Ripley 2002). For each model in the top model set, the 5,000 parameter estimates were used to estimate 5,000 rating probabilities (for ratings 1-9) for each of the 54 CUs for all combinations of high and low productivity, and high and low DQA. The 5,000 rating probabilities of each CU were weighted according to the relative $AIC_c$ weight ($w_{jAIC4}$; see Equation A1 in Appendix A) of each model, and then summed across the models in the top model set. For each CU, we summed the rating probabilities (1-9) into respective red (ratings 1-3), amber (ratings 4-6) and green (ratings 7-9) status within each of the 5,000 estimates, and calculated the median probability of each status across all 5,000 estimates. The CU status was determined as the status with the highest estimated median probability. For each comparison in Q1-Q5, we also counted the number of CUs increasing and decreasing in status between models with parameter uncertainty.

Regardless of whether we explicitly account for parameter uncertainty, the PWUs of metric levels estimated from additive conjoint rating models will remain confounded with the PWUs of the metrics. In other words, conjoint rating models do not allow for the separation of metric weight and level scale (metric status and DQA; Lanscar et al. 2007).

Metric weight is the average utility of a metric (without levels) across all of its levels (Flynn 2010), and is relative to other metrics in the questionnaire (e.g., spawner abundance versus trend in spawners). Level scale is a within-metric measure of the utility associated with different metric levels (e.g., green versus amber spawner abundance).

### *2.3.2. Best-worst scaling*

Therefore, to address the issue of separability of metric impact and level scale, and to allow for comparisons of metric and metric levels PWUs in the same units, we used a response task known as "best-worst metric scaling" (BWS) or "maximum-difference scaling". BWS is an extension of decision-choice experiments (DCEs) where respondents choose the "best" metric and "worst" metric presented in each scenario, thereby producing a partial ranking of the metrics (Vermunt and Magidson 2005). In this study, we asked experts to identify the combination (A, B, C or D) of metric, metric status, DQA that pulled their rating of CU status most to the green (best) and red (worst) end of the 9-point scale (Questions 2A and 2B in Figure 2). Since its development by Finn and Louviere (1992), BWS has been used in various research fields, including health care (Louviere and Flynn 2010), health economics (Flynn et al. 2007), consumer ethics (Auger et al. 2007), and wine marketing (Cohen 2003).

To decrease the response burden, the BWS response task (Questions 2A and 2B in Figure 2) was presented in randomized blocks (Appendix B) for only 21 of the 49 CU scenarios because respondents provide both a "best" and "worst" response (twice the amount of information as the conjoint rating task) for each CU scenario. The categorical responses (A, B, C or D) were aggregated across all respondents and a conditional logistical model (Vermunt and Magidson 2005) was used to estimate the metric weights and level scales PWUs:

$$(3) \qquad v_{(k|z_i)} = \beta_k^{met} + \sum_{m=1}^{M} \beta_m^{lev} z_{im}^{lev}$$

where $v_{(k|z_i)}$ is the systematic part of the utility or preference for choosing the categorical response $k$ (i.e., response A, B, C or D), given the metrics and metric levels presented in CU scenario $i$. The parameter $\beta_k^{met}$ is the PWU or relative importance associated with

the metrics only (i.e., metric weight). Parameter $\beta_m^{lev}$ is the PWU for the metric levels only (i.e. level scale; including metric level interactions) of the CU scenario $i$, defined by $z_{im}^{lev}$. Superscript $M$ is the number of attributes or metrics, in this case 4, and subscript $m$ refers to a specific metric. Therefore:

$$(4) \qquad P(x_i = k|z_i, s_i) = \frac{\exp\left(s_i v_{(k|z_i)}\right)}{\sum_{k \in A_i} \exp\left(s_i v_{(k|z_i)}\right)} \quad \text{if } k \in A_i \text{ and 0 if } k \notin A_i$$

where $P(x_i = k|z_i, s_i)$ is the probability of the respondent choosing $k$ (i.e., categorical response A, B, C, or D) as the best (or worst) attributes (i.e., metric and metric levels) for CU scenario $i$. The parameter $s_i$ is a scale factor set to -1 for the worst choice and +1 for the best choice, while $A_i$ is all possible best-worst responses for a given CU scenario. We estimated all main effects, as well as the main effects plus up to 4 two-way interactions between metric levels (e.g., trend in spawners (Tr) is red (R) and DQA is high (H), TrR-DQAH). We calculated the weighted average PWUs of parameters across the top-model set, defined here as we did above, as models with $\Delta AIC_c < 4$ (Burnham and Anderson 2002). Because all main effects appeared in each model, the $AIC_c$ relative variable importance (RVI) of the non-main- effects, i.e., interactions, was calculated by summing the relative $AIC_c$ model weights ($w_{jAIC4}$) of all models in the top-model set in which the interaction appeared (Burnham and Anderson 2002, Appendix A). Details of the statistical analysis can be found in Appendix A.

The main strength of best-worst scaling lies in its ability to separate metric weight from the level scale (Flynn et al. 2007). Unlike DCEs or rating tasks, BWS permits intra- and inter-metric comparison of levels by measuring metric utilities on a common interval scale (Cohen and Neira 2003). Therefore in this analysis, direct comparisons can be made between the model-averaged PWUs of the metric (metric weight), and the PWUs of metric status, DQA, and two-way interactions (level scale). In addition to separating metric weight from level scale values, the best-worst analysis helped address Q2 and Q3. Specifically:

Q2. To further determine the effect of productivity on the assessment of CU status, we compared the metric weight and level scale PWUs of the best-worst models for high and low productivity.

Q3. To further determine the effect of DQA on the assessment of CU status, we compared the level scale PWUs of DQA alone and in the two-way interactions between metric levels (e.g., TrR-DQAH) in each model.

## 2.4. Questionnaire respondents

A total of 64 experts from across British Columbia were invited to participate in this study. Of those, 39 experts completed a questionnaire of 55 hypothetical CU scenarios (e.g., Figure 2). The participants included 27 DFO salmon stock assessment biologists, area chiefs, program heads, and managers, as well as a total of 12 First Nations stock assessment biologists, consultants, and experts from environmental non-governmental organizations (ENGOs). We created an initial list of respondents based on their level of knowledge and experience with salmon stock assessment and management. Invitations to participate in the study were sent by e-mail, followed by either an in-person or phone discussion of the questionnaire before the respondents went through the questions. The survey followed the informed-consent process that was approved by the Simon Fraser University (SFU) Office of Research Ethics. Additional participants were recruited through chain referral sampling (i.e., being referred to by one of our originally targeted experts), which is a method typically used to select key informants from a network (Neuman 2000). We pre-tested a draft questionnaire with five experts from DFO and Simon Fraser University (SFU) in May 2010 (1 of the 5 experts completed the final version of the questionnaire). We also performed additional testing of a pilot questionnaire (5 CU scenarios) with 14 DFO experts on June 17, 2010 (13 of those 14 experts completed the final version of the questionnaire).

# 3. Results

A total of 39 questionnaires were completed for a 61% response rate (two subsequently removed as outliers, n=37). The strength of support for the All-9 and All-3 high and low productivity sub-models in their respective top-model set ($\Delta AIC_c$ <4) was quite similar, indicating no strong preference for one over another (Tables D1 and D4 of Appendix D). Conversely, the strength of support for DFO-9 and best-worst scaling models is very clear (Tables D2 and D3 of Appendix D). Consequently, the All-9 and All-3 models have multiple models in their top models sets, while DFO-9 and best-worst scaling models have only one or two models.

## 3.1. Conjoint rating model

The part-worth utilities (PWUs) of the All-9 models (Table 2) were estimated, and then used to calculate the Conservation Unit (CU) status of the 54 CUs according to Equations (1) and (2). The PWUs of the rating constants, metric levels, and interactions contribute additively to the probability of the CU status being red, amber or green (Equation 1 and 2). Therefore, the magnitude of any PWU in Table 2 represents the degree to which the parameters influence the rating of CU status. A negative PWU pulls the experts' rating of CU status toward the red end of the color scale, while a positive PWU pulls the experts' rating of CU status toward the green end of the color scale. In addition, when rating constant PWUs are large and positive, the rating (1-9) has a greater probability of being selected by respondents (independent of other information) than ratings with low or negative rating constant PWUs. Note that due to effects coding, both the direction (positive/negative) and magnitude of the PWUs in Table 2 are relative to their mean value of zero (sum to zero). Because the metric status PWUs are approximately linearly distributed, the mean of zero approximates the PWUs of the amber status. The PWUs of the DFO-9 and All-3 models are in Tables E1 and E2 of Appendix E. Unless otherwise specified, results presented in Q1-Q5 refer to model-

averaged PWUs obtained from the best parameter estimates (i.e., maximum likelihood value without parameter uncertainty).

### *3.1.1. Overall result*

Overall, the most common estimated status of the 54 hypothetical CUs (with or without parameter uncertainty) is amber, followed by red and then green (Table 3). To visualize the results in an interactive platform, we created a decision support spreadsheet (Appendix F) from the model-averaged PWUs of the All-9 models (Table 2). The user can create hypothetical CU scenarios and observe how changes in metric status and DQA affect the overall CU status and the estimated probability of red, amber, and green status.

For high productivity, experts have the highest probability of rating the CU a 5, and the lowest probability of rating the CU a 1 or 9 (Table 2, “Rating constant” estimates), independent of any additional information. The metric status of spawner abundance has the greatest relative importance in the assessment of CU status, followed by the metric status of trends in spawners, harvest rate, and distribution, the interactions AbA-HaR and DiR-HaA, the DQA of spawner abundance, the interaction AbR-DiA, the DQA of harvest rate, and the interaction AbA-DiG (Table 2).

Similarly, for low productivity, experts have the highest probability of rating the CU a 4, and the lowest probability of rating the CU a 9 on the 9-point scale (Table 2, “Rating constant”). The metric status of spawner abundance has the greatest relative importance, followed by the metric status of trends in spawners, harvest rate, and distribution, the interaction DiR-HaA, the DQA of spawner abundance and harvest rate, and the interactions TrG-HaR, AbA-HaG, and AbA-HaR. For both high and low productivity, the DQA of trend in spawners and distribution has little effect on the rating of CU status (Table 2). The relative importance of the metrics remains confounded with that of its levels.

## 3.2. Best-worst scaling model

The best-worst scaling analyses provide additional information on the relative importance of the metrics, metric status, data quality and amount (DQA), and two-way interactions (e.g., Ab-DQAH) by measuring the metric weight and level scale separately. In the best-worst scaling models (Table 4), the metric weights and level scale (i.e., PWU of metric status, DQA, and interactions) are interpreted much like the PWUs in the conjoint rating model, only they are not used to estimate the CU status of the 54 CUs (although they could be). Here too, due to effects coding, both the direction (positive/negative) and magnitude of the PWUs in Table 4 are relative to their mean value of zero. In this case, a negative PWU indicates that the parameter influences the experts' assessment of CU status toward the red end of the color scale, while a positive PWU indicates that the parameter influences the experts' assessment of CU status toward the green end of the color scale. The magnitude of the PWUs in Table 4 represents the degree to which the parameters influenced the experts' assessment of CU status.

### *3.2.1. Overall result*

Results of the analysis show that the metric weight PWUs are small relative to the level scale PWUs of metric status and two-way interactions (Table 4, "Level scale" estimates). For high productivity, spawner abundance and trend in spawners have equal relative importance (same magnitudes) but have opposite directions of effect, which is an artifact of effects coding. Distribution and harvest rate are relatively unimportant because their PWUs approximate zero. (Table 4, "Metric weight" estimates i.e., top four rows). For low productivity, spawner abundance has greatest relative importance. The three other metrics have lower and about equal importance (Table 4). DQA (by itself) has little or no effect on the assessment of CU status for high and low productivity (Table 4, "Level scale" estimates).

However, as we discuss in Q3 below, DQA affects CU status through interactions between metric status and high DQA. For high productivity, spawner abundance status (green or red) has the highest relative importance (highest level scale), followed by trend in spawners status, distribution status, Ab-DQAH, harvest rate status, Tr-DQAH, and

lastly Di-DQAH (Table 4, “Level scale” estimates). For low productivity, spawner abundance status has the highest relative importance when red in status, whereas trend in spawners status is highest when green in status, followed by Ab-DQAH (when red), distribution status, harvest rate status, and Tr-DQAH (Table 4). The best-worst results complement the conjoint rating findings by separating the influence of the metrics alone (metric weight) from the influence of the metric levels (level scale) on the assessment of CU status.

## Q1. What is the effect of an equal weighting of metrics on the determination of status?

Relative to the CU status estimated by the All-9 models, the equal weighting of metrics leads to an increase in the status of 3 CUs for high productivity, and of 8 or 19 CUs for low productivity (Table 5, Q1 “Number of CUs increasing” column). Relative to the All-9 models, the equal weighting of metrics also leads to a decrease in the status of either 6 or 11 CUs for high productivity, whereas only 1 CU decreases in status for low productivity and high DQA (Table 5, Q1 “Number of CUs decreasing” column). Depending on the case, between 1 and 6 CUs are only likely increasing, likely decreasing, or the same in status ($\Delta$ probability < 5%), which indicates a low level of uncertainty in the effect of an equal weighting of metrics. Similarly, after accounting for parameter uncertainty in the model-averaged PWUs, equal weighting of metrics has the same direction and magnitude of effect on the rating of CU status as the cases described above, which did not consider such uncertainty (Table 5, Q1 “PWUs with parameter uncertainty, Number of CUs increasing and decreasing” columns). Therefore, the results are robust to parameter uncertainty. According to the equal weighting method, the majority of CUs (39 of 54) are amber in status (Table 3, bottom row).

## Q2. What is the effect of productivity of salmon populations on the rating of CU status? Which factors influence the effect of productivity?

### *Conjoint rating*

With everything else held constant, a decrease in productivity leads to a decrease in the status of 15 and 22 in the CUs of the All-9 models, for high and low DQA cases, respectively (Table 5, Q2). No CUs show an increase in CU status. However, between 2 and 9 CUs are only likely decreasing in status and only likely the same status ($\Delta$ probability < 5%; Table 5, Q2), which indicates a moderate degree of uncertainty in the overall effect of productivity on the rating of CU status. Overall, for the All-9 models, a decrease in productivity leads to an increase in the number of CUs with red status, and a decrease in the number of CUs with green status (Table 3). A decrease in productivity also leads to a decrease in the status of CUs for both the DFO-9 and All-3 models. Relative to the All-9 models, the magnitude of the effect of reduced productivity is less for the DFO-9 models, and greater for the All-3 models (Table 5, Q1 "Number of CUs decreasing" column).

The magnitude and direction of the effect of reduced productivity on the lower rating of CU status is robust to parameter uncertainty in the model-averaged PWUs (Table 5, Q2). For example, the All-9 models with parameter uncertainty in PWUs estimate the highest median probability for the status of CU 28 as amber for high productivity and as red for low productivity (Figure 4). As above, where parameter uncertainty is considered, a decrease in productivity led to an increase in the number of CUs with a red status, and a decrease in the number of CUs with green status (Table 3).

How experts rate the status of a CU independent of any additional information, such as the metric status or the level of DQA is also affected by productivity. Based on a comparison of the rating constants for high and low productivity models when all other parameters are held constant, there is a higher probability that experts will rate a CU status lower (towards the red end of the scale) for low productivity CUs than high productivity CUs (Table 2).

In addition, productivity affects which interactions influence the experts' rating of CU status for the All-9 models (Table 2). For high productivity, the interaction between spawner abundance (Ab) when it is amber (A) and harvest rate (Ha) when it is red (R) (AbA-HaR) has the highest relative variable importance (RVI = 0.77), whereas for low productivity, the interaction between distribution (Di) when it is red (R) and harvest rate (Ha) when it is amber (A) (DiR-HaA) contributes most to the experts' assessment of CU status (RVI =0.53; Table 2). The high RVI values indicate that there is a greater weight of evidence that these parameters have non-linear effects (instead of linear) on the assessment of CU status. For high productivity, the effect of spawner abundance (Ab) when it is amber (A) on the overall rating of CU status is more strongly positive when harvest rate (Ha) is red (AbA-HaR). For low productivity, the effect of distribution (Di) when it is red is more negative when harvest rate (Ha) is amber (DiR-HaA). The PWUs of these interactions are similar to those of harvest rate or distribution metric status. No interaction with trend in spawners was included in the top model set for high productivity. All interactions in the top model set for low productivity include harvest rate (Table 2, Table D1 of Appendix D). Similarly, productivity also affects which interactions are included in the All-3 models (Table D4 of Appendix D, Table E2 of Appendix E).

### *Best-worst scaling*

Productivity also affects the results of the best-worst scaling analyses, showing which interactions between the metric status and DQA influence the experts' rating of CU status (Table 4, Table D2 of Appendix D). While the interaction between the distribution (Di) metric status and DQAH (Di-DQAH) is only included in the top-model set for high productivity, it is relatively unimportant in the assessment of status, as shown by the low RVI (RVI =0.38; Table 4, "RVI" column). The interaction between the harvest rate (Hr) metric status and DQAH (Hr-DQAH) is absent from both top-model sets.

## Q3. What effect does data quality and amount (DQA) have on the rating of CU status? Which factors influence the effect of DQA?

### *Conjoint rating*

In the All-9 models, the PWU estimates for DQA are relatively small for spawner abundance and harvest rate, and even smaller for trend in spawners and distribution compared to the PWUs of the metric status (Table 2, “PWU” column for rows with DQA). Despite the relatively small magnitude of the DQA PWUs, a decrease in DQA across all metrics leads to a decrease in the status of 5 CUs for high productivity and 12 CUs for low productivity (Table 5, Q3). No CUs show an increase in CU status when DQA is decreased (Table 5, Q3). A decrease in DQA also leads to an increase in the number of CUs with red status and a decrease in the number of CUs with green status (Table 3, All-9 models). Nonetheless, in the All-9 low-productivity model, the extent to which DQA affects the estimated CU status is moderately uncertain because between 3 and 10 CUs are only likely decreasing in status and only likely the same status, rather than definitively so (Table 5, Q3).

In the DFO-9 models, the DQA PWUs for spawner abundance and harvest rate are relatively small, and even smaller for trend in spawners and distribution compared to the PWUs of the metric status (Table E1 of Appendix E). In the All-3 models, the DQA PWU estimates are also relatively small for spawner abundance, and even smaller for trend in spawners, harvest rate, and distribution (Table E2 of Appendix E). Like the All-9 models, a decrease in DQA leads to a decrease in the status of 5 to 12 CUs estimated by the DFO-9 and All-3 models (Table 5, Q3). With a decrease in DQA, the number of CUs with red status increases, the number of CUs with amber status either increases or decreases (depending on the case), and the number of CUs with green status decreases (Table 3, DFO-9 and All-3 models). The direction and magnitude of the effect of DQA on the rating of CU status is the same for the All-9 models with parameter uncertainty in PWUs compared to cases without parameter uncertainty (Table 5). While not shown in Table 5, there is still some uncertainty in the effect of DQA after accounting for parameter uncertainty. As an example, the All-9 model with parameter uncertainty in PWUs estimates that the status of CU 42 is amber for low productivity and high DQA

and red for low productivity and low DQA (Figure 5). However, these two estimates of CU status are only likely different in status because, within the low productivity and high DQA model, the difference between the highest median probability and the second highest median probability is <5%.

### *Best-worst scaling*

By itself, DQA does not inform the experts' determination of status, as illustrated by the small PWUs (Table 4, "PWU" column). However, the effect of DQA depends on the metric considered. The two-way interactions between the spawner abundance (Ab) metric status and high data quality and amount (DQAH) (Ab-DQAH), and also between the trend in spawners (Tr) metric status and DQAH (Tr-DQAH), are useful predictors of the experts' assessment of status for both high and low productivity, as indicated by their high relative variable importance estimates (RVI) (Table 4, "RVI" column).

For CUs where DQA is high and metric status of spawner abundance and/or trends in spawners is red, the experts' assessment of status will be pulled further towards the red end of the scale, as indicated by the negative PWU estimate (e.g., Table 4, Ab-DQAH "Red" row). Conversely, when metric status is green, the experts' determination of status will be pulled further to the green end of the scale, as shown by the positive PWU estimate (e.g., Table 4, Ab-DQAH "Green" row). The PWUs of interactions for low DQA are not presented in Table 4 because by default of effects coding, they are the same magnitude as for high DQA, but have the opposite sign or direction of effect. As a result, for CUs where DQA is low and metric status is red, the interaction pulls the experts' determination of status further toward the green end of the scale. When DQA is low and metric status is green, the CU status will be pulled further to red end of the scale.

## Q4. Do DFO experts differ from other respondents in their assessment of CU status?

Relative to the status estimated by the All-9 models, the DFO-9 models result in a higher status in 7 CUs for low productivity, and in 2 or 3 CUs for high productivity; only 2 CUs show a decrease in CU status (Table 5, Q4 "Number of CUs increasing and

decreasing in status" columns). For example, the All-9 model with parameter uncertainty in the status estimates CU 19 as amber for low productivity and high DQA, whereas the DFO-9 model estimates the CU status as green (Figure 6). However, depending on the case, between 2 and 7 CUs are only likely increasing and only likely the same in status, which indicates a moderate degree of uncertainty about how different DFO-9 models and All-9 models are in their estimates of CU status. Generally, DFO-9 models estimate fewer CUs with red status, and more CUs with green status, than the All-9 models (Table 3).

With parameter uncertainty in the PWUs, the magnitude and direction of effect remains the same; DFO-9 models lead to an increase in status of CUs relative to the All-9 models. Again, the DFO-9 models with parameter uncertainty generally estimate fewer CUs with red status and more CUs with green status than the All-9 models with uncertainty (Table 3).

## Q5. Does the use of different rating scales, 3-point vs. 9-point, affect the CU status?

Relative to the All-9 models, the analysis on a 3-point scale leads to a decrease in the status of either 6 or 13 CUs for low productivity, and 2 or 3 CUs for high productivity (Table 5, Q5). However, in some cases, the effect of using different rating scales is moderately uncertain because between 2 and 9 CUs are only likely decreasing in status, and between 1 and 7 CUs are only likely the same status rather than definitively so (Table 5, number of CUs depends on the case). Relative to all other models, the All-3 models for low productivity estimate the highest number of CUs as red in status (Table 3). Also note that the All-3 models include numerous interactions in the top-model set (Table D4 of Appendix D, Table E2 of Appendix E).

Relative to a 9-point scale, the direction and magnitude of the effect of rating the CU status on a 3-point scale generally does not change after accounting for parameter uncertainty in the PWUs. The All-3 models for high productivity lead to an increase in the status of only one CU and a decrease in the status of 2 or 3 CUs, whereas the All-3

models for low productivity lead to a decrease in the status of either 6 or 14 CUs, relative to the All-9 models (Table 5).

# 4. Discussion

In this case study, we present a novel method for eliciting expert opinions about the relative importance of metrics, metric status, and data quality and amount (DQA) in the evaluation of conservation status. We can account for numerous two-way interactions between pairs of metric status, and metric status and DQA, and combine the part-worth utilities (PWUs) into a final Conservation Unit (CU) status. We build on the works by Melham (2004), Regan et al. (2005), and Goodenough (2012), among others, which explicitly look at the consistency of methods for determining conservation status by evaluating the similarity and biases in the results of each method.

## Q1. What is the effect of an equal weighting of metrics on the determination of status?

Relative to the estimates from the All-9 models, equal weighting of the metrics leads to optimistic assessments of CU status, especially for low productivity. Although changes in CU status are not associated with specific management actions (i.e., limit-reference points (LRPs), DFO 2005), unwarranted optimism about the biological status of salmon CUs could lead to the maintenance of status quo harvest rates and escapement targets, and an eventual reduction in abundance in the CU (if the optimism is not valid). Therefore, optimism in the biological status, especially for low productivity cases, is not precautionary because it may result in management action not being taken when it should be, with the long-term and ultimate cost of CU depletion, quasi-extinction, or extinction. However, equal weighting of metrics also leads to a decrease in status for high productivity; more so in cases with high DQA (11 of the 54 CUs) than low DQA (6 of the 54 CUs). Pessimism in the CU status may also have associated short- term socio-economic costs related to unnecessary management actions that result in socio-economic losses due to foregone fishing opportunities. As applied here, the equal

weighting method does not explicitly account for the effect of productivity on the assessment of biological status.

The equal weighting method assumes that all metrics are equally important to the assessment of CU status (all metric weights = 1), and that the same level scale applies across all metrics (i.e., red = 3, amber = 2, green = 1). Our best-worst scaling results show that the metric weight is unimportant relative to the level scale of the metric status and two-way interactions (Table 4). However, our results also clearly indicate that the level scale depends on the metric. The level scale is highest for abundance, followed by trend in spawners, and either harvest rate or distribution, and is subject to nonlinearities through two-way interactions (Table 4). Although metrics are considered to be equally important to the assessment of CU status in this case study, the assumption of equal weighting may not hold for the assessment of other species or ecosystems (e.g., Pestal and Cass 2009), and may lead to biases in conservation status. The level scale of metrics will also change in other decision-contexts.

In a recent study, Patrick et al. (2010) applied a productivity and susceptibility analysis (PSA) to determine the vulnerability of United States fish stocks. Although a PSA allows users to customize the weight of metrics (range 0-4) and their levels (1-3), the metrics used in the analysis of most fish stocks were assigned the default weight of 2. The metric weights of some fish stocks were later changed through a consensus process involving two or more fishery scientists (Patrick et al. 2009). Similarly, metric levels were assigned weights (i.e., level scale) by consensus following categorical criteria, where low = 1, medium = 2, high = 3. Intermediate weights when spanning two categories were only occasionally applied. By assigning default weights to both metrics and metric levels, this application of a PSA is not unlike the equal weighting method presented here.

However, assigning the relative importance of metrics in isolation can be highly misleading because it cannot account for complex interactions that determine an expert's opinions. In contrast, stated preference methods elicit opinions over the full range of metrics and metric levels that define the decision context (Louviere et al. 2000), as reflected in the questionnaire's scenarios. Furthermore, by considering metrics and their levels simultaneously, interactions that may be vital to the assessment of

conservation status can be accounted for. Additionally, best-worst scaling allows for the separation of weight and level scale by asking experts to consider the importance of metrics and their levels relative to one another (Flynn et al. 2007).

## Q2. What is the effect of productivity of salmon populations on the rating of CU status? Which factors influence the effect of productivity?

Our results indicate that experts rely on productivity as a general guide to their assessment of the conservation status of CUs; they will rate the status of a CU lower for low productivity than for high productivity, independent of any other information. In this case study, we considered the effects of productivity because it determines the rate of recovery or resilience of a stock and is therefore of primary importance in stock assessment (Patrick et al. 2009). In other situations, other ecological factors such as colonization rates, indices reflecting environmental conditions (e.g., Pacific Decadal Oscillation (PDO)), or resource-limitations need to be considered because they too may have important effects on how experts assess conservation status.

Productivity also governs the relative importance (or PWUs) of the metrics when different two-way interactions are considered. For instance, in the All-9 model for high productivity, none of the interactions involved trend in spawners – the effect of the status of trend in spawners on the assessment of CU status is independent of the status of other metrics. In contrast, for low productivity, the effect of the status of trend in spawners is dependent on the status of harvest rate. The effect of green trend in spawners on the assessment of CU status is dampened when harvest rate is red (TrG-HaR, Table 2). For low productivity, each interaction (between pairs of metric status) included harvest rate, indicating the additional importance of this metric beyond the main effects.

Not surprisingly, results from the analysis of experts' opinions on the CU status suggest that if and when CUs experience declines in productivity, the status may either be equal to or lower than when its productivity is high (never higher in CU status), particularly in cases of low DQA. Therefore, we can expect fewer CUs with green status,

and more CUs with red status, for low productivity than high productivity. In contrast, an increase in productivity from low to high should result in an increase in the status of some CUs. In reality, most Fraser River sockeye salmon CUs have experienced a decreasing trend in productivity since the mid-1980's (Grant et al. 2011). Similarly, some coho, pink, chum, and Chinook salmon stocks (although not yet assessed at the CU level) have also experienced large decreases in productivity (Bradford and Irvine 2000; Dorner et al. 2008).

The large effect of productivity (as shown by the number of CUs decreasing in status when productivity decreases) may in part be an artifact of the questionnaire's layout (Figure 2), because experts rated the CU status for high and low productivity side-by-side, explaining the difference. In addition, we presented productivity as known and stable, either high or low, whereas in reality, productivity may be increasing, decreasing, or both in different time periods. Further research is needed to identify how productivity or other fundamental ecological factors affect the experts' conservation status assessment of other species, communities, and ecosystems.

## Q3. What effect does data quality and amount (DQA) have on the rating of CU status? Which factors influence the effect of DQA?

The evaluation of conservation status often lacks an acknowledgement or explicit characterization of the uncertainty underlying the data (Regan et al. 2005, Lukey et al. 2010). Furthermore, interpretation of the available data can cause inconsistent assessments of conservations status (IUCN 2001; Regan et al. 2005). In this study, we asked experts to consider the extremes of the DQA spectrum. However, unlike Lukey et al. (2010), we found that DQA does not greatly influence the rating of CU status. When DQA decreased from high to low in the All-9 models, experts rated a lower status for less than a tenth of the 54 CUs for high productivity and less than a quarter of the 54 CUs for low productivity (Table 5). Furthermore, for low productivity, the effect of DQA on the rating of CU status is somewhat uncertain due to the 10 CUs that are merely likely decreasing in status rather than definitively so (Table 5). However, our results

show that the effect of DQA on the assessment of biological status is robust to parameter uncertainty.

Similarly, in the best-worst scaling analysis, DQA by itself has relatively little influence on the experts' assessment of CU status. In other words, if we only know the DQA of each metric (i.e., metric status is not known), DQA does not inform experts about the CU status. However, DQA becomes vital to the assessment of CU status in a two-way interaction with spawner abundance, trend in spawners, and distribution (the latter only for high productivity). High DQA will augment the effects of the metric status according to the direction of effect (e.g., green metric status with high DQA will influence the experts' assessment of CU status further toward the green end of the scale), while low DQA will dampen the effect of the metric status. Specifically, when DQA is low and metric status is red, experts are not precautionary; they discount the negative effect of a red metric status on their evaluation of CU status, which influences their assessment of CU status towards the green end of the scale (more optimistic). Under a precautionary approach, the negative effect of a red metric status in data-limited CUs would not be discounted, but instead would influence the experts' assessment of CU status further towards the red end of the scale.

To our knowledge, this case study is the first to report the importance of interactions between two metric levels in a best-worst analysis. We therefore emphasize that methods used to assess conservation status must be capable of measuring interactions (two-way and possibly higher-order) in order to appropriately specify models and gain insights into what drives experts' opinions (Louviere 2006).

Although experts did not rate the status of CUs in which no data exists for one or more metrics explicitly, our definition of low DQA spanned the range from little or no data (Appendix A). Consistent with the precautionary approach, data-deficient species should be assigned a status of threatened until sufficient information is available for further assessment (Mace et al. 2008, Lukey et al. 2010). Uncertainty in the conservation status of data-deficient species may no longer be grounds for inaction. Accordingly, data-deficient CUs could be assigned a red status, pending additional information. Because the amount of data is known to be generally positively related to the economic and social value of a fishery (Chen et al. 2003), non-exploited CUs are more likely to be data-

deficient. Although a policy in which data-deficient CUs are automatically assigned a red status would create a more pessimistic outlook on the biological status of Pacific salmon populations, it would help avoid the potential loss of CUs if followed up with precautionary management actions. Consequently, data-deficient CUs would be regarded as high priority for monitoring and research in order to better inform the rating of their status. However, with finite resources, monitoring of data-deficient CUs will come at the expense of monitoring CUs with high socio-economic value. The balance is the key for management agencies to find.

The methods presented in this case study are useful for data-poor situations, because comparisons between metrics are made possible by linking what is known about the metric to pre-determined color-coded categories (red, amber, and green), which extends the methods detailed by the Traffic Light approach (Halliday et al. 2001; Caddy 2005). Furthermore, through relatively simple response tasks, conjoint rating and best-worst scaling analyses provide the extent to which DQA contributes to assessments of conservation status. In this and other contexts, future research is needed on how to best represent and communicate the levels of DQA associated with metrics to clarify the influence of uncertainty on assessment of conservation status. Also, identifying what aspect of DQA is more important (quality or amount), or if relevant at all, in other species contexts could contribute to understanding the role of data uncertainty in determining conservation status of other biological units.

## Q4. Do DFO experts differ from other respondents in their assessment of CU status?

Due to the small number of DFO expert respondents (n=27) and the large number of model parameters (20-23 depending on the model), our results may reflect either actual differences in how DFO experts rate CUs or may be due to the small number of degrees of freedom in DFO-9 models. Similarly, the lack of interactions in the DFO-9 models indicates that, in this case study, DFO experts view the relative importance of metrics on CU status as independent of one another. While the questionnaire's experimental design explicitly allowed for the estimation of interactions, it

is impossible to know whether the lack of interactions in DFO-9 models is an artifact of too few DFO respondents, or a “true” representation of DFO experts’ opinions.

Assuming that the results represent actual differences in opinions between all respondents and only DFO experts, then DFO experts’ are slightly more optimistic in their determination of CU status than all respondents, particularly in low productivity cases. Not surprisingly, given that the majority (27 of 37 experts) of the responses included in the All-9 models are from DFO experts, DFO-9 models lead to an increase in the status of only 2 to 7 CUs relative to the All-9 models (Table 5). Because the data sets behind the All-9 and DFO-9 models are not mutually exclusive, our results may be either over- or under-estimating the differences in CU status between DFO experts and all respondents.

Upon further examination of CUs that increase in status (both high and low productivity models), the CU status estimated by the DFO-models is equal to the metric status of spawner abundance (except CU 42 for low productivity low DQA; Table G1 of Appendix G). For example, the metric status for spawner abundance of CU 19 is green. Accordingly, for low productivity high DQA, the status of CU 19 is amber as estimated by the All-9 models and green according to the DFO-9 models. Therefore, in cases where there is an increase in the estimated CU status between the All-9 and DFO-9 models, DFO experts’ assessment of CU status seems to be driven by the metric status of spawner abundance.

Our results highlight the continuing need to explicitly account for existing differences as well as the lack of differences in opinions among groups of experts. It likely does matter who determines conservation status (Regan et al. 2005). However, in this case study, there is no definitive answer as to whether DFO experts differ from other respondents in their assessment of CU status because of the small sample size in both groups of experts. We caution against averaging across multiple expert opinions because it, as well as other collaborative methods, may hide valid differences in interpretations that are based on different experiences. Such methods do not tend to account for or communicate the range of opinions within and among groups of experts.

## Q5. Does the use of different rating scales, 3-point vs. 9-point, affect the CU status?

For this question, we investigated whether there is a difference in the estimated CU status if experts rated the status of CUs according to 3-point scale (red, amber, and green status zones) instead of using the 9-point scale (Figure 2). Although respondents did not actually rate the CU status along a 3-point scale, they may still perceive that they rated the CU status in terms of the three status zones, given the emphasis placed on those zones in the rating task. The 9-point scale may be relatively unimportant or secondary to completing the rating task (i.e., the location within a given colored status zone does not affect the CU status). In fact, because there are 3 status zones superimposed on the 1-9 color scale, we don't actually know how respondents perceived the color scale and completed the rating task. There are likely many ways a respondent could approach such a task. For example, the rating task could be a sequential process. Initially, the respondent decides which of the three status zones to rate the CU, and then adjusts their rating up or down within that zone or always use the middle rating within the zone (i.e., ratings 2, 4, and 7). Alternatively, the 9-point scale essentially had 9 colors. Respondents may have simply rated the CU by the color that best represented the status.

Comparing the use of different rating scales is particularly relevant because the All-9 models may violate the model assumption of equal distance between ratings (Vermunt and Magidson 2005). Rating a CU a 3 or 4 on the 9-point color scale infers a change in status (from red to amber status) whereas rating it a 1 or 2 does not infer a change in status (stays red). Therefore, the assumption of equal distance may only be valid within each status zone (e.g., between 1 and 2, and between 2 and 3 for red status). In the All-3 case, we assume that the distance between red and amber, and amber and green status zones is the same, and thereby do not violate the conjoint rating model assumption.

Relative to the All-9 models, assessments of status by the All-3 models are somewhat pessimistic, leading to a decrease in status for only 2 or 3 of the 54 CUs for high productivity and 6 or 13 of the 54 CUs for low productivity (Table 5, Q5). With parameter uncertainty in the PWUs, the All-3 models' estimates of CU status remain

equally precautionary for low productivity as first anticipated by the estimates of CU status without parameter uncertainty. Because we cannot know which rating scale is the "true" scale, we can only highlight discrepancies between the results of the All-9 and All-3 models.

According to the literature, the use of different rating scales can affect the survey findings. Rating scales with small numbers categories are generally considered to provide less valid and less discriminating results than those with six or more categories (Preston and Colman, 2000). Although in this case study we use rating scales, other methods to assess conservation status use a set number of categories (e.g., IUCN has 8 threat categories (IUCN 2001), and COSEWIC has 7 status categories (COSEWIC 2010)). What are the implications of changing the number of conservation status categories? Additional research is needed to uncover how differences or changes (creation or deletion) in rating scales or in the number of categories may affect assessments of conservation status.

## 4.1. Advantages

The methods presented in this article allow researchers to determine the relative importance of metrics used to assess CU status when these metrics are considered jointly by respondents. Assessments of conservation status rarely (if ever) result from the evaluation of a single indicator. Therefore, the relative importance of indicators will be poorly measured if evaluated separately (Alriksson and Öberg 2008). By presenting multiple hypothetical scenarios, conjoint rating and best-worst scaling are methods well suited for evaluating respondents' trade-offs between the metrics presented. The experimental design, underlying both of these response tasks, efficiently captures unbiased and precise statistical information on opinions without respondents having to rate all possible scenarios (Bridges et al. 2011). For example, only 49 of the possible 1296 scenarios (i.e., full factorial design with 4 metrics and 6 levels $6^4$ =1296) were presented to experts in the questionnaire.

In addition, the conjoint rating and best-worst scaling methods are based on rigorous statistical foundations (Alriksson and Öberg 2008; Flynn et al. 2007) and can

therefore be used to generate quantitative data that illustrate the level of uncertainty in the estimated CU status, the relative importance of different metrics, metric levels, and two-way interactions (between metric levels; Tables 4 and 5, Appendix G). Interactions represent occurrences where the opinion about one metric depends on the level of another metric. Without taking into account such interactions, like most other methods do, there is a high likelihood of obtaining biased estimates from linear additive models (Louviere 2006), or other models for that matter. Therefore, and as our study shows, the use of interactions in the experimental design and statistical analysis of these methods is warranted (De Bekker-Grob et al. 2012).

Through the separation of metric weight and level scale in the best-worst scaling analysis, we were able to confirm that, in this case study, metrics have little to no effect on the assessment of salmon CU status without information on their status or level of DQA. Nevertheless, there likely are cases where metric weight is of greater relative importance. Without a best-worst scaling analysis, the relative importance of a metric is confounded with that of its levels in a conjoint rating task or choice experiment; in such cases, the PWU of the metric levels cannot be interpreted as indicating the PWU of a metric (Lanscar et al. 2007). Separation of metric weight and level scale is therefore useful in determining what drives assessments of conservation status, and in helping decision-makers consider whether to improve the monitoring of key metrics.

A major strength of stated preference methods is that the response tasks and experimental design are flexible, and should simulate the decision context as closely as possible (Lanscar and Louviere 2008). In this case study, the methods chosen, conjoint rating and best-worst scaling, allowed us to mimic the real-life decision context that experts would face when implementing the Wild Salmon Policy (WSP) for some CUs. Future assessments during the implementation of the WSP may be the outcome of a small-group process (Dr. C. A. Holt, Pacific Biological Station, Nanaimo, British Columbia, V9T 6N7, 2010 pers. comm.), and the results here may help to inform that process. However, group processes may hide the full range of expert opinions (Dalkey 1972). Assessments of conservation status should not be made by a single individual, but instead by several individuals separately before a peer review meeting (Regan et al. 2005).

Furthermore, many researchers recommend using a collaborative process, such as the Delphi method, to determine the relative importance of metrics and their levels (e.g., Patrick et al. 2010). The Delphi method depends on group dynamics to reach consensus among experts (Okoli and Pawlowski 2004). We do not repeat here the well-known merits and critiques of the Delphi method (Hasson and Keeney 2011; Landeta 2006); we argue instead that assigning the relative importance of metrics and metric levels should not come from a consensus-making process where differing opinions may not be represented explicitly. Rather, it is essential to clearly characterize the range of expert opinion and to identify the degree to which these opinions converge. In our study, the PWUs of metrics, metric levels, and interactions reflect the pooling of the full range of individuals' opinions, instead of the outcome of individual or group decision-making, and probabilities for the categories of biological status are estimated; this represents an improvement over a single-category outcome of conservation status. Unlike collaborative methods to assess conservation status, the types of questionnaires we applied can be completed by any number of respondents (even hundreds, although unlikely to have so many experts) or at different times to capture learning (e.g., before and after workshops). Models can be updated over time as the responses from additional participants are included, and separate models can be estimated for different groups or periods.

In addition, methods used to assess conservation status rarely acknowledge different sources of uncertainty (Regan et al. 2005). In our research study, we account for model uncertainty by using $AIC_c$ model averaging, as well as parameter uncertainty, by sampling 5,000 parameter estimates from each model's variance-covariance matrix. Estimates of CU status from model-averaged PWUs were not compared to estimates from single best models. However, we did compare the CU status with and without incorporating uncertainty in the parameter estimates. Both the direction and magnitude of the various effects (Q1-Q5) were quite robust to parameter uncertainty (Table 5). Although we did not, perform a sensitivity analysis on the 5% cut-off, accounting for these two sources of uncertainty helps us examine the robustness of our findings.

## 4.2. Limitations

Rating-based conjoint analysis has been a mainstay in marketing research and has been subjected to external validation (Green and Srinivasan 1990). However, extending the conclusions of our results beyond the context of our study on salmon CU status should not be considered at the moment. The direction and magnitudes of the PWUs most likely depend on the decision-context (framing effects), and are susceptible to change in the presence or absence of metrics (Johnson 1987). In practice, the specific results of this case study cannot inform the assessment of CU status when metrics or metric levels differ from those used here in some future evaluation (e.g., a category of "unknown" for metric status or addition of a second metric of trend in spawner abundance such as percent decline in the most recent three generations of salmon). The results here merely provide a snapshot of experts' current opinions for the metrics and metric levels presented for evaluation, and may not predict future behavior and outcomes of this same group of respondents with same or different set of metrics (Bridges et al. 2011). The methods applied in this study do not encompass the entire process of determining the biological status of a salmon CU, but nevertheless can serve as a useful tool to elicit details on the structure of experts' opinions.

The number of scenarios presented to respondents depends on the type of response task, the number and complexity of the metrics presented, number of respondents, and the experimental design (Bridges et al. 2011). Since we were interested in eliciting the opinions of a small pool of experts (64 of them were invited to participate), we needed to ask them to respond to the entire fractional factorial design (49 scenarios plus a few extra, Appendix A), which produced a substantive response burden. Despite a high response rate, comparisons of the DFO-9 results to the All-9 model remain tentative because of the relatively small number of DFO (n=27) and non-DFO (n=10) experts. The best-worst scaling analysis provides complimentary information supporting the interpretation of results from the conjoint rating analysis by separating metric weight and level scale. The best-worst scaling results cannot be directly compared to the conjoint analysis results because the response tasks (and therefore data sets) differ. The rating and best-worst data sets could be combined and analyzed simultaneously for one joint model (Magidson et al. 2009).

In this case study, we use $AIC_c$ model averaging in order to address model uncertainty and make inferences by quantifying the degree of relative support for each model in the top-model set (Johnson and Omland 2004). Although this method has numerous advantages over traditional hypothesis testing (Grueber et al. 2011), model averaging has rarely been used in stated preference applications (Layton and Lee 2006; Rose et al. 2009). Instead, the majority of stated conjoint rating and best-worst scaling applications use the statistical significance of parameters estimates as the basis for their results. Traditional frequentist methods ignore model uncertainty, and instead assume the existence of a single model that best explains the phenomena under research (Grueber et al. 2011). Among the practical issues associated with model averaging, the interpretation of the model main effects is problematic when interactions are present, unless parameters are centralized (i.e., interpretation of the main effects must also consider the influence of interactions; Grueber et al. 2011)

## 4.3. Future research

The stated preference methods presented here could be used to holistically assess the conservation status of other fisheries, species and ecosystems, and provide the relative importance of metrics and metric levels used in the assessment. For example, a similar questionnaire could easily be applied to assess the status of Salmon Management Areas on Canada's east coast for implementing the Wild Atlantic Salmon Conservation Policy (DFO 2009). In the context of the (Pacific) Wild Salmon Policy (WSP), the questionnaire design could readily be changed in numerous ways by (1) using species- or CU-specific metrics, (2) using multiple metrics from one indicator, e.g., short- and long-term trend in total spawners, (3) altering the number of metrics (to determine its effect on PWUs; similar to Islam et al. 2007), (4) or adding an 'unknown' metric status for CUs where no information exists for a given metric. Additional questionnaires could be used at later periods in time to evaluate the degree to which experts' opinions changed over time, after either additional workshops or experience implementing the WSP. Additional response tasks could be used to gain further information on experts' opinions when making trade-offs between two metrics (e.g., spawner abundance and trend in spawners) or metric status.

In this case study, only one model structure (conditional logit model) was used in both the conjoint and best-worst scaling analysis. Future research could perform model averaging across multiple model structures in order to explicitly address the effects of structural uncertainty on the assessment of conservation status. For example, the nested logit model allows researchers to partition choices into groups (Hensher et al. 2005), while the mixed effect model allows for opinions to vary across individuals through a respondent-specific random parameter (Lanscar and Louviere 2008). In addition, with a large number of respondents, latent class models could be used to segment the respondents into different groups depending on different preference patterns (Hensher et al. 2005). However, assessments of conservation status rely heavily on a relatively small number of experts.

Studies which compare assessment methods for conservation status (e.g., Adelman et al. 2004; Brito et al. 2010; Goodenough 2012; Melham et al. 2004; Regan et al. 2005) are essentially already investigating the effects of such structural uncertainty on conservation status, because each method represents an independent interpretation. Goodenough (2012) even recommends that different methods or protocols be considered simultaneously when determining conservation status (e.g., IUCN, NatureServe, regional assessments). Incongruencies in status between methods can generally undermine their credibility (Brito et al. 2010). Further comparisons would aid in evaluating the underlying biases, and potential implications of the use of each method. The overarching issue remains unresolved: How do we weight and combine the status of different methods? There is no single agreed upon “right” way, but there are certainly some ways that will lead to biased results, as we found for the case of the equal weighting method’s optimistic assessments of CU status for low productivity.

## 4.4. Conclusions

Not surprisingly, considering the historical emphasis by salmon fisheries management agencies on spawner abundance estimates, we found that the metric status of spawner abundance and trend in spawners have the highest relative importance in the experts’ assessment of CU status, especially where there is high DQA (through two-way interaction terms Ab-DQAH and Tr-DQAH). The metric status of

distribution and harvest rate are relatively less important. Without information on the metric status, both metrics and DQA are least important in the experts' rating of CU status.

Although numerous methods exist to weight and combine metrics in the assessment of conservation status, stated preference methods (in this case study, conjoint rating and best-worst scaling (BWS)) are especially useful for eliciting opinions and thereby the relative importance of metrics because they (1) require respondents to consider all metrics and metric levels simultaneous and make trade-offs, (2) efficiently capture the range of expert opinions, (3) use rigorous statistical analysis to generate PWUs, error estimates, and the level of consensus for each conservation status, (4) permit the separation of metric weight and level scale (BWS only), and (5) allow for the estimation of two-way interactions.

Our results highlight the need to account for underlying ecological factors or processes, such as productivity, because they can affect the determination of conservation status, as well as the relative importance of metric levels through the inclusion of interaction terms. Our results also highlight the importance of using methods capable of estimating two-way interactions, both between metrics and metric levels and between different metric levels. By modeling only the main effects and ignoring relevant interactions, studies assume that the PWU of a metric level is independent of the level of another, potentially leading to sub-optimal model predictions (Hensher et al. 2005). According to the expert respondents in this case study, the equal weighting of the metrics primarily leads to a less precautionary assessment of conservation status for low productivity cases, and should be used carefully, because metric weight and level scale may be different in other situations. In addition, DQA needs to be effectively communicated and accounted for because it can have large effects on assessments of conservation status through interaction terms. Similarly, who is doing the assessment can potentially lead to different assessments of status. Therefore, it is paramount to communicate the full range of expert opinion. Likewise, the rating scale used to determine the conservation status can in some cases lead to different outcomes. Although methods used to assess conservation status should consider all these factors, many possible applications of stated preference methods in the assessment and management of populations, species, and ecosystems remain unexplored.

# References


Ananda, J., and G. Herath. 2002. Assessment of wilderness quality using the Analytic Hierarchy Process. Tourism Economics 8(2): 189–206.

Auger, P., T. M. Devinney, and J. J. Louviere. 2007. Measuring the importance of ethical consumerism: A multi-country empirical investigation in Hooker, J., Hulpke, JF., Madsen, P. (eds), Controversies in International Corporate Responsibility, Carnegie Mellon University - Philosophy Documentation Center, Charlottesville, USA, pp. 207-221.

Andelman, S. J., C. Gross, and H. M. Regan. 2004. A review of protocols for selecting species at risk in the context of U.S. Forest Service viability assessments. Acta Oecologica 26(2): 75–83.

Alriksson, S., and T. Öberg. 2008. Conjoint analysis for environmental evaluation – A review of methods and applications. Environmental Science & Pollution Research 15(3): 244–257.

Ben-Akiva, M., and S. R. Lerman. 1985. Discrete Choice Analysis: Theory and Application to Travel Demand. The MIT Press, Cambridge/Massachusetts.

Bradford, M. J., and J. R. Irvine. 2000. Land use, fishing, climate change and the decline of Thompson River, British Columbia, coho salmon. Canadian Journal of Fisheries and Aquatic Sciences 57: 13–16.

Bradlow, E. T. 2005. Current issues and a 'wish list' for conjoint analysis. Applied Stochastic Models in Business and Industry 21: 319–323.

Branch, T. A., R. Watson, E. A. Fulton, S. Jennings, C. R. McGilliard, G. T. Pablico, and D. Ricard. 2010. The trophic fingerprint of marine fisheries. Nature 468: 431–435.

Bridges, J. F. P., A. B. Hauber, D. Marshall, A. Lloyd, L. A. Prosser, D. A. Regier, F. R. Johnson, and J. Mauskopf. 2011. Conjoint Analysis Applications in Health—a Checklist: A Report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. Value in Health 14: 403–413.

Brito, D., R. C. Ambal, T. Brooks, N. de Silva, M. Forster, W. Hao, et al. 2010. How similar are national red lists and the IUCN Red List? Biological Conservation 143: 1154–1158.

Burnham, K. P., and D. R. Anderson. 2002. Model selection and multimodel inference: a practical information-theoretic approach. 2nd Edition. Springer-Verlag, New York, New York, USA.

Caddy, J.F., E. Wade, T. Surette, M. Hebert, and M. Moriyasu. 2005. Using an empirical traffic light procedure for monitoring and forecasting in the Gulf of St. Lawrence fishery for the snow crab, *Chionoecetes opilio*. Fisheries Research 76: 123–145.

Carlsson, F., P. Frykblom, and C. Liljenstolpe. 2003. Valuing wetland attributes: an application of choice experiments. Ecological Economics 47: 95–103.

Caruso, E. M., D. A. Rahnev, and M. R. Banaji. 2009. Using Conjoint Analysis to Detect Discrimination: Revealing Covert Preferences From Overt Choices. Social Cognition 27(1): 128–137.

Chen, Y., L. Chen, and K. I. Stergiou. 2003. Impacts of data quantity on fisheries stock assessment. Aquatic Sciences 65: 92–98.

Cohen, S. H. 2003. Maximum difference scaling: improved measures of importance and preference for segmentation. Sequim, WA: Sawtooth Software, Inc.

Cohen, S. H., and L. Neira. 2003. Measuring preference for product benefits across countries: overcoming scale usage bias with maximum difference scaling. Paper presented at ESOMAR 2003 Latin America Conference, Punta del Este, Uruguay.

Coll, M., L. J. Shannon, D. Yemane, J. S. Link, H. Ojaveer, S. Neira, D. Jouffre, et al. 2010. Ranking the ecological relative status of exploited marine ecosystems. ICES Journal of Marine Science 67: 769–786.

Cook, R. D., and S. Weisber. 1982. Residuals and Influence in Regression. Chapman & Hall, New York.

COSEWIC. 2003. COSEWIC assessment and status report on the sockeye salmon *Oncorhynchus nerka* (Cultus population) in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. ix + 57 pp.

COSEWIC. 2006. COSEWIC assessment and status report on the Chinook salmon *Oncorhynchus tshawytscha* (Okanagan population) in Canada. Committee on the Status of Endangered Wildlife in Canada. Ottawa. vii + 41 pp. (www.sararegistry.gc.ca/status/status_e.cfm).

Dalkey, N. C. 1972. The Delphi method: An experimental study of group opinion. In N C. Dalkey, D. L. Rourke, R. Lewis, & D. Snyder (Eds.). Studies in the quality of life: Delphi and decision-making (pp. 13-54). Lexington, MA: Lexington Books.

De Bekker-Grob, E. W., M. Ryan, and K. Gerard. 2012. Discrete choice experiments in health economics: a review of the literature. Health Economics 21(2): 145–172.

Dorner, B., R. M. Peterman, and S. L. Haeseker. 2008. Historical trends in productivity of 120 Pacific pink, chum, and sockeye salmon stocks reconstructed by using a Kalman filter. Canadian Journal of Fisheries and Aquatic Sciences 65(9):1842–1866.

Dorow, M., B. Beardmore, W. Haider, and R. Arlinghaus. 2009. Using a novel survey technique to predict fisheries stakeholders' support for European eel (*Anguilla anguilla L.*) conservation programs. Biological Conservation 142(12): 2973–2982.

Finn, A., and J. J. Louviere. 1992. Determining the appropriate response to evidence of public concern: the case of food safety. Journal of Public Policy and Marketing 11(1): 12–25.

Fisheries and Oceans Canada, 2005. Canada's Policy for the Conservation of Wild Pacific Salmon. pp. 57.

Flynn, T. N., J. J. Louviere, T. J. Peters, and J. Coast. 2007. Best-worst scaling: What it can do for health care research and how to do it. Journal of Health Economics 26: 171–189.

Flynn, T. N. 2010. Valuing citizen and patient preferences in health: recent developments in three types of best-worst scaling. Expert Review of Pharmacoeconomics and Outcomes Research 10(3): 259–267.

Goodenough, A. E. 2012. Differences in two species-at-risk classification schemes for North American mammals. Journal for Nature Conservation 20: 117–124.

Grant, S. C. H., B. L. MacDonald, T. E. Cone, C. A. Holt, A. Cass, E. J. Porszt, J. M. B. Hume, and L. B. Pon. 2011. Evaluation of Uncertainty in Fraser Sockeye (*Oncorhynchus nerka*) Wild Salmon Policy Status using Abundance and Trends in Abundance Metrics. DFO Canadian Science Advisory Secretariat Research Document 2011/nnn. vi + xx p.

Green, P. E., and V. Srinivasan. 1990. Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice. Journal of Marketing 54: 3–19.

Grueber, C. E., S. Nakagawa, R. J. Laws, and I. G. Jamieson. 2011. Multimodel inference in ecology and evolution: challenges and solutions. Journal of Evolutionary Biology 24: 699–711.

Halliday, R. G., L. P. Fanning, and R. K. Mohn. 2001. Use of the traffic light method in fishery management planning. DFO Canadian Science Advisory Secretariat Research Document 2001/108.

Hasson, S., and S. Keeney. 2011. Enhancing rigour in the Delphi technique research. Technological Forecasting and Social Change 78(9): 1695–1704.

Hensher, D. A., J. M. Rose, and W. H. Greene. 2005. Applied choice analysis: A primer. Cambridge University Press, pp.717.

Hobday, A. J., A. D. M. Smith, I. C. Stobutzki, C. Bulman, R. Daley, J. M. Dambacher, R. A. Deng, et al. 2011. Ecological risk assessment for the effects of fishing. Fisheries Research 108(2-3): 372–384.

Hoffman M., C. Hilton-Taylor, A. Angulo, M. Böhm, T. M. Brooks, et al. 2010. The impact of conservation on the status of the world's vertebrates. Science 330: 1503–1509.

Holt, C., A. Cass, B. Holtby, and B. Riddell. 2009. Indicators of status and benchmarks for Conservation Units in Canada's Wild Salmon Policy. DFO Canadian Science Advisory Secretariat Research Document 2009/058.

Holt, C. A. 2009. Evaluation of Benchmarks for Conservation Units in Canada's Wild Salmon Policy: Technical Documentation. DFO Canadian Science Advisory Secretariat Research Document 2009/059.

Holtby, L. B., and K. A. Ciruna. 2007. Conservation Units for pacific salmon under the Wild Salmon Policy. DFO Canadian Science Advisory Secretariat Research Document 2007/070.

Hutchings, J. A., C. Minto, D. Ricard, J. K. Baum, and O. P. Jensen. 2010. Trends in abundance of marine fishes. Canadian Journal of Fisheries and Aquatic Sciences 67: 1205–1210.

IUCN. 2001. IUCN Red List Categories and Criteria. Version 3.1. IUCN Species Survival Commission. Switzerland/UK: IUCN.

Johnson, J. B., and K. S. Omland. 2004. Model selection in ecology and evolution. TRENDS in Ecology and Evolution 19(2): 101–108.

Landeta, J. 2006. Current validity of the Delphi method in social sciences. Technological Forecasting and Social Change 73(5): 467–482.

Lanscar, E., J. Louviere, and T. Flynn. 2007. Several methods to investigate relative attribute impact in stated preference experiments. Social Science & Medicine 64:1738–1753.

Lanscar, E., and J. Louviere. 2008. Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics 26(8): 661–677.

Lawson, S. R., and R. E. Manning. 2002. Tradeoffs among social, resource, and management attributes of the Denali wilderness experience: a contextual approach to normative research. Leisure Science 24: 297–312.

Layton, D. F., and S. T. Lee. 2006. Embracing model uncertainty: strategies for response Pooling and model averaging. Environmental and Resource Economics (34): 51–85.

Lee, J. A., G. N. Soutar, and J. Louviere. 2007. Measuring values using best-worst scaling: The LOV example. Psychology and Marketing 24(12): 1043–1058.

Louviere, J. L. 2006. What you don't know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. Environmental and Resource Economics 34: 173–188.

Louviere, J. L., and T. N. Flynn. 2010. Using best-worst scaling choice experiments to measure public perceptions and preferences for healthcare reform in Australia. The Patient: Patient-Centered Outcomes Research 3(4): 275–283.

Louviere, J. L., D. Hensher, and J. Swait. 2000. Conjoint preference elicitation methods in the broader context of random utility theory preference elicitation methods. In Conjoint Measurement: Methods and Applications, Gustafsson, A., A. Herrmann and F. Huber (eds). Springer: Berlin. 279–318pp.

Louviere, J. L., and T. Islam. 2008. A comparison of importance weights/measures derived from choice-based conjoint, constant sum scales and best worst scaling. Journal of Business Research, 61(9): 903–911.

Lukey, J. R., Crawford, S. S., and D. J. Gillis. 2010. Effect of information availability on assessment and designation of species at risk. Conservation Biology 24: 1398–1406.

Mace, G. M., N. J. Collar, K. J. Gaston, C. Hilton-Taylor, H. R. Akçakaya, N. Leader-Williams, E. J. Milner-Gulland, and S. N. Stuart. 2008. Quantification of extinction risk: IUCN's system for classifying threatened species. Conservation Biology 22(6): 1424–1442.

Magidson, J., D. Thomas, and J.K. Vermunt. 2009. A new model for the fusion of maxdiff scaling and ratings data. 2009 Sawtooth Software Conference Proceedings, 83–103.

McElhany, P., C. Busack, M. Chilcote, S. Kolmes, B. McIntosh, J. Myers, D. Rawding, A. Steel, C. Steward, D. Ward, T. Whitesel, and C. Willis. 2006. Revised Viability Criteria for Salmon and Steelhead in the Willamette and Lower Columbia Basins. Willamette/Lower Columbia Technical Recovery Team and Oregon Department of Fish and Wildlife.

McElhany, P., M. H. Rucklelshaus, M. J. Ford, T. C. Wainwright, and E. P. Bjorkstedt. 2000. Viable Salmonid Populations and the Recovery of Evolutionary Significant Units. U.S. Dept. Commer., NOAA Technical Memorandum NMFS-NWFSC-42.

McFadden, D. 1974. Conditional logit analysis of qualitative choice behaviour. In P. Zarembka (ed.), Econometrics. New York, NY: Academic Press.

Mehlman, D. W., K. V. Rosenberg, J. V.,Wells, and Robertson, B. 2004. A comparison of North American avian conservation priority ranking systems. Biological Conservation 120: 383–390.

Miller, R. M., J. P. Rodríguez, T. Aniskowicz-Fowler, C. Bambaradeniya, R. Boles, M. A. Eaton, U. Gärdenfors, V. Keller, S. Molur, S. Walker, and C. Pollock. 2007. National threatened species listing based on IUCN criteria and regional guidelines: Current status and future perspectives. Conservation Biology 21(3): 684–696.

Morgan, M. G., and M. Henrion. 1992. Uncertainty: A Guide to Dealing with Uncertainty in Quantitative Risk and Policy Analysis. Cambridge University Press. 346pp.

NatureServe. 2011. NatureServe explorer: An online encyclopedia of life, version 7.1.Arlington, VA: NatureServe. Available http://www.natureserve.org/explorer

Neuman, W. L. 2000. Social Research Methods Qualitative and Quantitative Approaches, 4th Edition. Allyn & Bacon: Boston.

Okoli, C., and S. D. Pawlowski. 2004. The Delphi method as a research tool: an example, design considerations and applications. Information & Management 42: 15–29.

Patrick, W. S., P. Spencer, O.Ormseth, J. Cope, J. Field, D. Kobayashi, T. Gedamke, E. Cortés, K. Bigelow, W. Overholtz, J. Link, and P. Lawson. 2009. Use of productivity and susceptibility indices to determine stock vulnerability, with example applications to six U.S. fisheries. U.S. Dep. Commerce., NOAA Tech. Memo. NMFS-F/SPO-101, 90 p.

Patrick, W. S., P. Spencer, J. Link, J. Cope, J. Field, D. Kobayashi, P. Lawson, T. Gedamke, E. Cortés, O. Ormseth, K. Bigelow, and W. Overholtz. 2010. Using productivity and susceptibility indices to assess the vulnerability of United States fish stocks to overfishing. Fishery Bulletin 108(3): 305–322.

Paulhus, D. L. 1991. Measurement and control of response bias. In J.P. Robinson, P.R. Shaver, L.S Wright (eds.), Measures of personality and social psychological attitudes. Academic Press, San Diego, CA.

Peacock, S. J., and C. A. Holt. 2010. A review of metrics of distribution with application to Conservation Units under Canada's Wild Salmon Policy. Can. Tech. Rep. Fish. Aquat. Sci. 2888: xii + 36 p.

Peacock, S. J., and C. A. Holt. 2012. Metrics and sampling designs for detecting trends in the distribution of spawning Pacific salmon (*Oncorhynchus* spp.). Canadian Journal of Fisheries and Aquatic Sciences 69(4): 681–694.

Pestal, G., and A. Cass. 2009. Fraser Sockeye Resource Assessment Framework: Using Qualitative Risk Evaluations to Prioritize Resource Assessment Activities for Fraser River Sockeye. Report submitted to Fisheries and Oceans Canada, pp. 74.

Peterman, R. M., B. J. Pyper, and J. A. Grout. 2000. Comparison of parameter estimation methods for detecting climate-induced changes in productivity of Pacific salmon (*Oncorhynchus* spp.). Canadian Journal of Fisheries and Aquatic Sciences 57: 181–191.

Porszt, E. J., R. M. Peterman, N. K. Dulvy, A. B. Cooper, and J. R. Irvine. 2012. Reliability of Indicators of Decline in Abundance. Conservation Biology. doi: 10.1111/j.1523-1739.2012.01882.x

Preston, C. C., and A. M. Colman. 2000. Optimal number of response categories in rating scales: reliability, validity, discriminating power, and respondent preferences Acta Psychologica 104(1): 1–15.

R Development Core Team. 2011. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Raghavarao, D., J. B. Wiley, and P. Chitturi. 2011. Choice-based conjoint analysis; models and designs. Chapman & Hall/CRC.

Rand, P. S. 2011. *Oncorhynchus nerka (SKEENA R, LOWER: Alastair)*. In: IUCN 2011. IUCN Red List of Threatened Species. Version 2011.2. <www.iucnredlist.org>. Downloaded on 09 March 2012.

Regan, H., M. Burgman, M. A. McCarthy, L. L. Master, D. A. Keith, G. M. Mace, et al. 2005. The consistency of extinction risk classification protocols. Conservation Biology 19: 1969–1977.

Rice, J., and M. Rochet. 2005. A framework for selecting a suite of indicators for fisheries management. ICES Journal of Marine Science 62(3): 516–527.

Romero, J. B. Martínez-Crego, T. Alcoverro, and M. Pérez. 2007. A multivariate index based on the seagrass *Posidonia oceanica* (POMI) to assess ecological status of coastal waters under the water framework directive (WFD). Marine Pollution Bulletin 55: 196–204.

Rose, J. M., R. Scarpa, and M. C. J. Bliemer. 2009. Incorporating model uncertainty into the generation of efficient stated choice experiments: A model averaging approach, International Choice Modelling Conference, March 30-April 1, Yorkshire U.K.

Semeniuk, C. A. D., W. Haider, B. Beardmore, and K. D. Rothley. 2008. A multi-attribute trade-off approach for advancing the management of marine wildlife tourism: a quantitative assessment of heterogeneous visitor preferences. Aquatic Conservation, 19: 194–208.

Shin, Y. J., A. Bundy, L. J. Shannon, M. Simier, M. Coll, E. A. Fulton, et al. 2010b. Can simple be useful and reliable? Using ecological indicators to represent and compare the states of marine ecosystems. ICES Journal of Marine Science, 67(4): 717–731.

Sorice, M. G., C. Oh, and R. B. Ditton. 2007. Managing scuba divers to meet ecological goals for coral reef conservation. Ambio 36(4): 316–322.

Sutherland, W. J., A. S. Pullin, P. M. Dolman, and T. M. Knight. 2004. The need for evidence-based conservation. Trends in Ecology and Evolution 19: 305–308.

Thurston, L. L. 1927. A law of comparative judgment. Psychological Review 34: 273–286.

Train, K. E. 1986. Qualitative choice analysis: Theory econometrics, and an application to automobile demand. MIT Press.

Trenkel, V. M., M.-J. Rochet, and B. Mesnil. 2007. From model-based prescriptive advice to indicator-based interactive advice. ICES Journal of Marine Science 64(4): 768–774.

Turnhout, E., M. Hisschemoller, and H. Eijsackers. 2007. Ecological indicators: between the two fires of science and policy. Ecological Indicators 7: 215–228.

Tutsch, M., W. Haider, B. Beardmore, K. Lertzman, A. B. Cooper, and R. Walker. 2010. Estimating the consequences of wildfire for wildfire risk assessment. Canadian Journal of Forest Research 40(11): 2104–2114.

Veselka,W., J. T. Anderson, and S. Kordek. 2010. Using dual classifications in the development of avian wetland indices of biological integrity for wetlands in West Virginia, USA. Environmental Monitoring and Assessment 164: 533–548.

Venables, W. N., and B. D. Ripley. 2002. Modern applied statistics with S. Fourth Edition. Springer, New York. ISBN 0-387-95457-0. Available: http://www.stats.ox.ac.uk/pub/MASS4

Vermunt, J. K., and J. Magidson. 2005. Technical Guide for Latent GOLD Choice 4.0: Basic and Advanced, Belmont, Massachussetts: Statistical Innovations Inc.

Worm, B., R. Hilborn, J. K. Baum, T. A. Branch, J. S. Collie, C. Costello, M. J. Fogarty, E. A. Fulton, J. A. Hutchings, S. Jennings, O. P. Jensen, H. K. Lotze, P. M. Mace, T. R. McClanahan, C. Minto, S. R. Palumbi, A. Parma, D. Ricard, A. A. Rosenberg, R. Watson, and D. Zeller. 2009. Rebuilding Global Fisheries. Science 325: 578–585.

# Tables and Figures

***Table 1. Questionnaire's definitions of the lower and upper benchmarks for metrics of spawner abundance, trend in spawners, harvest rate, and spatial distribution indicators.***

| Metric | Lower Benchmark | Upper Benchmark | Comments |
|---|---|---|---|
| Spawner Abundance: Mean spawner abundance over the most recent generation. | $S_{gen}$: The abundance of spawners that will result in rebuilding to $S_{MSY}$ within one generation under average conditions in the absence of fishing. | $0.8S_{MSY}$: 80% of spawners at maximum sustained yield ($S_{MSY}$). | |
| Trend in Spawners: Ratio of mean spawner abundance of current generation to historical mean. | 0.25 | 0.5 | Identified by Pestal and Cass (2009) through qualitative evaluation of expert opinion. |
| Harvest rate: Percent harvest rate relative to productivity in most recent generation. | $F_{MSY}$: Fishing mortality that produces the maximum sustainable yield. | $0.7F_{MSY}$: 70% of the fishing mortality that produces the maximum sustainable yield ($F_{MSY}$). | |
| Distribution: Areal extent of spawners within the CU relative to historical mean. | LOW = Contraction in spatial distribution relative to historical mean where there is concern due to a lack of CU resilience or increased vulnerability to stochastic events. | HIGH = Small or no change in spatial distribution relative to historical mean; it refers to cases in which there is no concern about CU resilience. | No theoretical basis or data to identify upper and lower benchmarks. In this questionnaire, distribution benchmarks are qualitative comparisons relative to historical levels. |

***Table 2. Results of analyses of CU status ratings from all respondents (n=37; All-9). Model-averaged part-worth utilities (PWUs, Equation (1)) of the rating constants, metric status, and data quality and amount (DQA) of the top model set (∆AICc<4) for high and low productivity. Rating constants are rating specific (1-9) intercepts in Equation (1) (where red status = ratings 1-3, amber status= 4-6, and green status = 7-9). Each model includes interactions between the status of two metrics. Symbols in those interactions, red (R), amber (A), and green (G), spawner abundance (Ab), trend in spawners (Tr), harvest rate (Ha), distribution (Di). Also shown is the associated unconditional standard error ($SE_{nc}$, Equation (A3) of Appendix A), 95% confidence interval (CI). Because all metric levels were retained in all models, the relative variable importance (RVI) is shown for only the two-way interactions, as calculated from the Akaike weights. Blanks in some columns in interaction rows indicate that the interaction was absent from the top models set.***

| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{v}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{v}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| Rating constant | | | | | | | | |
| 1 | -4.355 | 0.105 | (-4.561,-4.149) | | -0.257 | 0.042 | (-0.339, -0.174) | |
| 2 | -0.477 | 0.038 | (-0.552,-0.402) | | 2.001 | 0.033 | (1.936, 2.065) | |
| 3 | 2.059 | 0.026 | (2.008, 2.110) | | 2.590 | 0.026 | (2.539, 2.641) | |
| 4 | 2.471 | 0.020 | (2.431, 2.511) | | 3.040 | 0.015 | (3.009, 3.070) | |
| 5 | 3.068 | 0.014 | (3.041, 3.095) | | 2.255 | 0.010 | (2.236, 2.275) | |
| 6 | 2.225 | 0.013 | (2.199, 2.251) | | 0.789 | 0.012 | (0.764, 0.813) | |
| 7 | 0.775 | 0.019 | (0.737, 0.813) | | -0.327 | 0.019 | (-0.365, -0.289) | |
| 8 | -1.166 | 0.037 | (-1.238, -1.095) | | -3.258 | 0.043 | (-3.343, -3.173) | |
| 9 | -4.599 | 0.084 | (-4.765, -4.434) | | -6.832 | 0.123 | (-7.073, -6.592) | |
| Spawner abundance status | | | | | | | | |
| Red | -1.290 | 0.013 | (-1.316, -1.264) | | -1.217 | 0.007 | (-1.231, -1.203) | |
| Amber | 0.147 | 0.007 | (0.134, 0.160) | | 0.248 | 0.003 | (0.242, 0.254) | |
| Green | 1.143 | 0.005 | (1.134, 1.152) | | 0.969 | 0.003 | (0.962, 0.975) | |
| Spawner abundance DQA | | | | | | | | |
| High | 0.134 | 0.001 | (0.132, 0.136) | | 0.106 | 0.001 | (0.105, 0.107) | |
| Low | -0.134 | 0.001 | (-0.136, -0.132) | | -0.106 | 0.001 | (-0.107, -0.105) | |
| Trend in spawners status | | | | | | | | |
| Red | -0.805 | 0.002 | (-0.810, -0.801) | | -0.723 | 0.002 | (-0.727, -0.719) | |
| Amber | -0.050 | 0.002 | (-0.054, -0.047) | | -0.013 | 0.002 | (-0.017, -0.010) | |
| Green | 0.856 | 0.003 | (0.851, 0.861) | | 0.736 | 0.003 | (0.730, 0.743) | |
| Trend in spawners DQA | | | | | | | | |
| High | -0.010 | 0.001 | (-0.011, -0.009) | | -0.008 | 4.8E-04 | (-0.009, -0.007) | |
| Low | 0.010 | 0.001 | (0.009, 0.011) | | 0.008 | 4.8E-04 | (0.007, 0.009) | |
| Harvest rate status | | | | | | | | |

| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{\nu}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{\nu}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| Red | -0.427 | 0.003 | (-0.432, -0.422) | | -0.361 | 0.002 | (-0.366, -0.356) | |
| Amber | 0.073 | 0.002 | (0.068, 0.077) | | 0.046 | 0.003 | (0.041, 0.051) | |
| Green | 0.354 | 0.003 | (0.349, 0.360) | | 0.315 | 0.003 | (0.310, 0.320) | |
| Harvest rate DQA | | | | | | | | |
| High | 0.085 | 0.001 | (0.083, 0.086) | | 0.065 | 4.8E-04 | (0.064, 0.066) | |
| Low | -0.085 | 0.001 | (-0.086, -0.083) | | -0.065 | 4.8E-04 | (-0.066, -0.064) | |
| Distribution status | | | | | | | | |
| Red | -0.416 | 0.006 | (-0.429, -0.404) | | -0.346 | 0.004 | (-0.354, -0.337) | |
| Amber | 0.028 | 0.002 | (0.023, 0.033) | | 0.028 | 0.002 | (0.025, 0.031) | |
| Green | 0.388 | 0.004 | (0.381, 0.395) | | 0.318 | 0.002 | (0.313, 0.322) | |
| Distribution DQA | | | | | | | | |
| High | -0.021 | 0.001 | (-0.022, -0.020) | | 0.005 | 4.9E-04 | (0.004, 0.006) | |
| Low | 0.021 | 0.001 | (0.020, 0.022) | | -0.005 | 4.9E-04 | (-0.006, -0.004) | |
| Interactions | | | | | | | | |
| AbA-HaR | 0.352 | 0.019 | (0.314, 0.390) | 0.77 | 0.028 | 0.009 | (0.011, 0.046) | 0.08 |
| DiR-HaA | -0.165 | 0.030 | (-0.224, -0.105) | 0.40 | -0.204 | 0.023 | (-0.250, -0.158) | 0.53 |
| AbA-DiG | -0.046 | 0.016 | (-0.077, -0.016) | 0.12 | | | | |
| AbR-DiA | -0.098 | 0.050 | (-0.196, 0.001) | 0.15 | | | | |
| AbA-HaG | | | | | -0.033 | 0.010 | (-0.053, -0.014) | 0.10 |
| TrG-HaR | | | | | -0.063 | 0.016 | (-0.094, -0.031) | 0.19 |

***Table 3. Number of CUs (out of 54) with each estimated status (red, amber, green) from model-averaged part-worth utilities (Model-Averaged PWUs), and after incorporating parameter uncertainty (PWUs with Uncertainty) in each conjoint rating model, as well as when the equal weighting method was used (where CU status = average of metric status in a given CU, red = 1, amber = 2, and green = 3). Analyses were performed for combinations of high and low productivity, and high and low DQA.***

| Model | | Number of CUs with red status | | Number of CUs with amber status | | Number of CUs with green status | |
|---|---|---|---|---|---|---|---|
| | | Model-Averaged PWUs | PWUs with Uncer-tainty | Model-Averaged PWUs | PWUs with Uncer-tainty | Model-Averaged PWUs | PWUs with Uncer-tainty |
| All-9 | High productivity, high DQA | 10 | 11 | 33 | 32 | 11 | 11 |
| | High productivity, low DQA | 13 | 14 | 32 | 31 | 9 | 9 |
| | Low productivity, high DQA | 18 | 17 | 32 | 33 | 4 | 4 |
| | Low productivity, low DQA | 27 | 28 | 26 | 25 | 1 | 1 |
| DFO-9 | High productivity, high DQA | 10 | 10 | 30 | 30 | 14 | 14 |
| | High productivity, low DQA | 12 | 12 | 34 | 34 | 8 | 8 |
| | Low productivity, high DQA | 15 | 17 | 31 | 29 | 8 | 8 |
| | Low productivity, low DQA | 22 | 23 | 29 | 28 | 3 | 3 |
| All-3 | High productivity, high DQA | 11 | 11 | 33 | 33 | 10 | 10 |
| | High productivity, low DQA | 14 | 15 | 33 | 32 | 7 | 7 |
| | Low productivity, high DQA | 28 | 28 | 25 | 25 | 1 | 1 |
| | Low productivity, low DQA | 32 | 33 | 22 | 21 | 0 | 0 |
| Equal weighting method | | 11 | | 39 | | 4 | |

***Table 4. Same as Table 2 except the results are for analyses of best-worst scaling responses (Equation (3) and (4)) from all respondents (n=37).***

| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{v}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{v}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| **Metric weight** | | | | | | | | |
| Spawner abundance | 0.121 | 0.005 | (0.111, 0.132) | | 0.094 | 0.076 | (-0.054, 0.242) | |
| Trend in spawners | -0.129 | 0.004 | (-0.137, -0.122) | | -0.03 | 0.062 | (-0.150, 0.091) | |
| Harvest rate | 0.007 | 0.004 | (-0.001, 0.015) | | -0.037 | 0.063 | (-0.162, 0.087) | |
| Distribution | 0.001 | 0.004 | (-0.007, 0.008) | | -0.027 | 0.062 | (-0.149, 0.094) | |
| **Level scale** | | | | | | | | |
| Spawner abundance status | | | | | | | | |
| Red | -2.340 | 0.043 | (-2.425, -2.255) | | -2.132 | 0.204 | (-2.532, -1.731) | |
| Amber | 0.224 | 0.026 | (0.172, 0.275) | | 0.07 | 0.159 | (-0.241, 0.381) | |
| Green | 2.116 | 0.024 | (2.069, 2.163) | | 2.062 | 0.155 | (1.758, 2.366) | |
| Spawner abundance DQA | | | | | | | | |
| High | -0.010 | 0.010 | (-0.029, 0.009) | | -0.21 | 0.101 | (-0.409, -0.012) | |
| Low | 0.010 | 0.010 | (-0.009, 0.029) | | 0.21 | 0.101 | (0.012, 0.409) | |
| Trend in spawners status | | | | | | | | |
| Red | -1.923 | 0.023 | (-1.968, -1.877) | | -1.991 | 0.153 | (-2.290, -1.692) | |
| Amber | -0.118 | 0.020 | (-0.158, -0.078) | | -0.218 | 0.143 | (-0.499, 0.062) | |
| Green | 2.041 | 0.026 | (1.990, 2.091) | | 2.209 | 0.161 | (1.894, 2.525) | |
| Trend in spawners DQA | | | | | | | | |
| High | 0.136 | 0.007 | (0.123, 0.149) | | 0.144 | 0.08 | (-0.013, 0.301) | |
| Low | -0.136 | 0.007 | (-0.149, -0.123) | | -0.144 | 0.08 | (-0.301, 0.013) | |
| Harvest rate status | | | | | | | | |
| Red | -1.234 | 0.015 | (-1.263, -1.206) | | -1.338 | 0.121 | (-1.576, -1.100) | |
| Amber | 0.096 | 0.011 | (0.075, 0.118) | | 0.153 | 0.105 | (-0.053, 0.359) | |
| Green | 1.138 | 0.014 | (1.110, 1.165) | | 1.185 | 0.12 | (0.950, 1.419) | |
| Harvest rate DQA | | | | | | | | |
| High | 0.008 | 0.006 | (-0.005, 0.020) | | -0.033 | 0.08 | (-0.189, 0.123) | |
| Low | -0.008 | 0.006 | (-0.020, 0.005) | | 0.033 | 0.08 | (-0.123, 0.189) | |
| Distribution status | | | | | | | | |
| Red | -1.458 | 0.029 | (-1.516, -1.400) | | -1.512 | 0.12 | (-1.748, -1.277) | |
| Amber | 0.005 | 0.013 | (-0.019, 0.030) | | -0.055 | 0.105 | (-0.261, 0.151) | |
| Green | 1.453 | 0.036 | (1.381, 1.524) | | 1.568 | 0.12 | (1.333, 1.802) | |
| Distribution DQA | | | | | | | | |
| High | 0.080 | 0.006 | (0.068, 0.092) | | 0.062 | 0.078 | (-0.091, 0.215) | |
| Low | -0.080 | 0.006 | (-0.092, -0.068) | | -0.062 | 0.078 | (-0.215, 0.091) | |

| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| Ab-DQAH | | | | 1 | | | | 1 |
| Red | -1.362 | 0.110 | (-1.578, -1.146) | | -1.744 | 0.35 | (-2.429, -1.058) | |
| Amber | 0.080 | 0.067 | (-0.052, 0.211) | | 0.355 | 0.26 | (-0.154, 0.865) | |
| Green | 1.282 | 0.060 | (1.164, 1.400) | | 1.388 | 0.245 | (0.908, 1.869) | |
| Tr-DQAH | | | | 0.87 | | | | 1 |
| Red | -1.074 | 0.055 | (-1.182, -0.966) | | -0.932 | 0.227 | (-1.378, -0.487) | |
| Amber | -0.349 | 0.052 | (-0.450, -0.248) | | -0.092 | 0.225 | (-0.533, 0.350) | |
| Green | 1.423 | 0.056 | (1.314, 1.532) | | 1.024 | 0.228 | (0.577, 1.471) | |
| Di-DQAH | | | | 0.13 | | | | |
| Red | -0.089 | 0.053 | (-0.193, 0.016) | | | | | |
| Amber | -0.029 | 0.011 | (-0.050, -0.008) | | | | | |
| Green | 0.118 | 0.088 | (-0.055, 0.290) | | | | | |

***Table 5. Comparison of the estimated status (red, amber, green) of 54 hypothetical CUs from models with model-averaged part-worth utilities (Model-averaged PWUs), and after incorporating parameter uncertainty (PWUs with parameter uncertainty) in each conjoint rating model under various conditions (Q1-Q5). For each comparison between models with model-averaged PWUs, we counted the number of CUs increasing, decreasing or with the same in status (first number in the column), as well as the number of CUs that are "definitely" (first number in parentheses) and "likely" (second number in parentheses) increasing, decreasing or the same in status. CUs that are "definitely" increasing or decreasing in status or have the same status are CUs where the difference between status categories with the highest and second highest probabilities is 5% or greater in both models of the comparison (∆ probability ≥= 5%). CUs that are "likely" increasing or decreasing in status or have the same status are those where the difference between status categories with the highest and second highest probabilities is less than 5% in one or both models in the comparison (∆ probability < 5%). See Table G1 of Appendix G. We also counted the number of CUs increasing or decreasing in status between models with parameter uncertainty. See Table G2 of Appendix G.***

| Model | | Model-averaged PWUs | | | PWUs with parameter uncertainty | |
|---|---|---|---|---|---|---|
| | | Number of CUs increasing in status (definitely, likely) | Number of CUs decreasing in status (definitely, likely) | Number of CUs with same status (definitely, likely) | Number of CUs increasing in status | Number of CUs decreasing in status |
| Q1. Change from All-9 to equal weighting model | | | | | | |
| | High productivity, high DQA | 3(3,0) | 11(10,1) | 40(38,2) | 3 | 10 |
| | High productivity, low DQA | 3(3,0) | 6(5,1) | 46(42,4) | 4 | 6 |
| | Low productivity, high DQA | 8(7,1) | 1(0,1) | 46(43,3) | 7 | 1 |
| | Low productivity, low DQA | 19(13,6) | 0 | 35(33,2) | 20 | 0 |
| Q2. Change from high to low productivity | | | | | | |
| All-9 | High DQA | 0 | 15(13,2) | 39(33,6) | 0 | 13 |
| | Low DQA | 0 | 22(13,9) | 32(28,4) | 0 | 22 |
| DFO-9 | High DQA | 0 | 11(6,5) | 43(39,4) | 0 | 13 |
| | Low DQA | 0 | 15(9,6) | 39(37,2) | 0 | 16 |
| All-3 | High DQA | 0 | 26(21,5) | 28(28,0) | 0 | 26 |
| | Low DQA | 0 | 25(23,2) | 29(25,4) | 0 | 25 |
| Q3. Change from high to low data quality and amount (DQA) | | | | | | |
| All-9 | High productivity | 0 | 5(2,3) | 49(44,5) | 0 | 5 |
| | Low productivity | 0 | 12(2,10) | 42(39,3) | 0 | 14 |
| DFO-9 | High productivity | 0 | 8(5,3) | 46(42,4) | 0 | 8 |
| | Low productivity | 0 | 12(6,6) | 42(38,4) | 0 | 11 |

| Model | | Model-averaged PWUs | | | PWUs with parameter uncertainty | |
|---|---|---|---|---|---|---|
| | | Number of CUs increasing in status (definitely, likely) | Number of CUs decreasing in status (definitely, likely) | Number of CUs with same status (definitely, likely) | Number of CUs increasing in status | Number of CUs decreasing in status |
| All-3 | High productivity | 0 | 6(3,3) | 48(47,1) | 0 | 7 |
| | Low productivity | 0 | 5(4,1) | 49(43,6) | 0 | 6 |
| Q4. Change from All-9 to DFO-9 model | | | | | | |
| | High productivity, high DQA | 3(1,2) | 0 | 51(49,2) | 4 | 0 |
| | High productivity, low DQA | 2(0,2) | 2(1,1) | 50(48,2) | 2 | 1 |
| | Low productivity, high DQA | 7(3,4) | 0 | 47(40,7) | 4 | 0 |
| | Low productivity, low DQA | 7(0,7) | 0 | 47(42,5) | 7 | 0 |
| Q5. Change from All-9 to All-3 model | | | | | | |
| | High productivity, high DQA | 0 | 2(2,0) | 52(49,3) | 1 | 2 |
| | High productivity, low DQA | 0 | 3(1,2) | 51(47,4) | 0 | 3 |
| | Low productivity, high DQA | 0 | 13(4,9) | 41(40,1) | 0 | 14 |
| | Low productivity, low DQA | 0 | 6(3,3) | 48(41,7) | 0 | 6 |

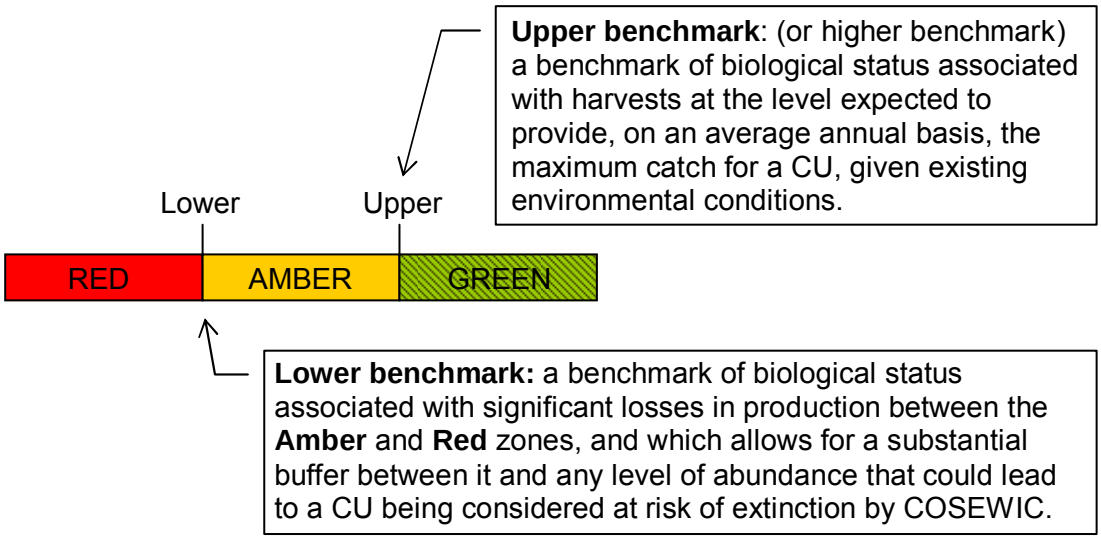



***Figure 1. General definitions of lower and upper benchmarks for a CU that were provided to expert respondents in this study's questionnaire (DFO 2005).***

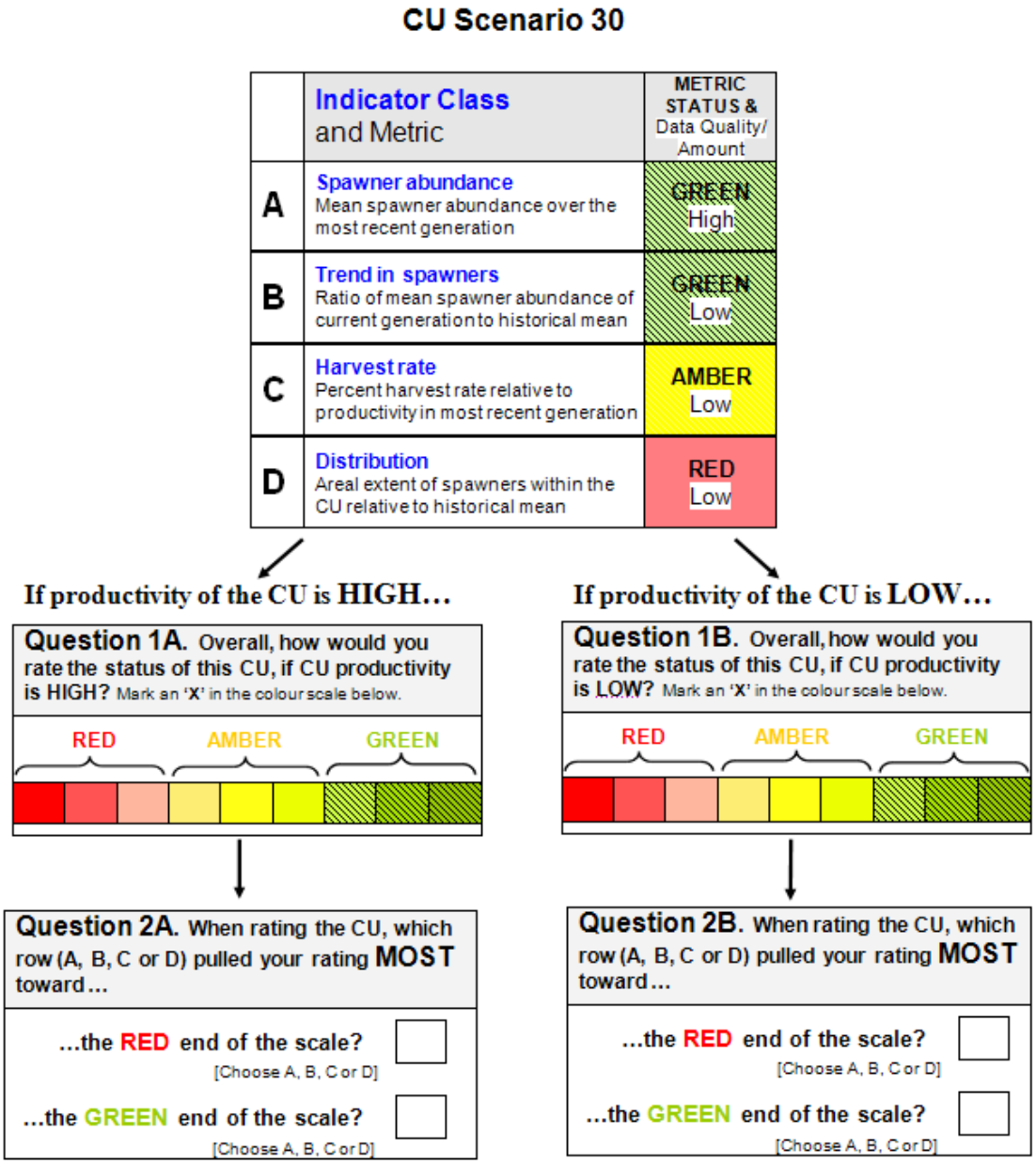



***Figure 2. Example of a hypothetical CU scenario (top table) that simultaneously presents the four metrics, metric status (red, amber or green), and data quality and amount (high or low), and the response tasks (in bottom half) presented in the questionnaire used in this study. Under high and low productivity, expert respondents first rated the status of each CU scenario along the 9-point color scale (Questions 1A and 1B) and then provided additional information on what pulled their rating of the CU to either end of the color scale (Questions 2A and 2B). Each expert was given 55 of these CU scenarios to rate.***


***Figure 3. All two-way interactions between the status (red, amber, or green) of any two of the four metrics included in the questionnaire's experimental design, denoted by an "X". Blanks represent interactions that were not included, i.e., all of those on the diagonals.***

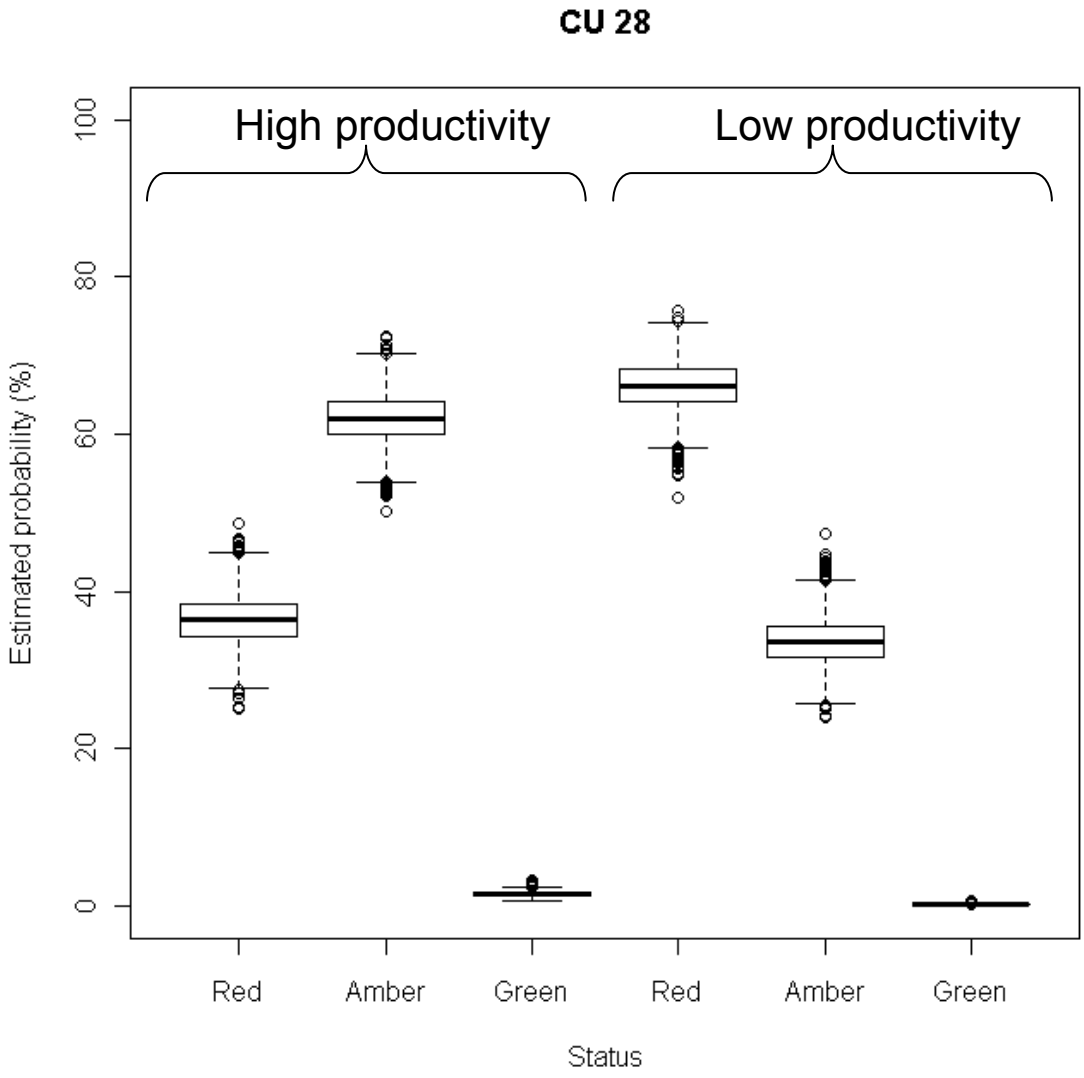



***Figure 4. Box and whisker plots showing the estimated median probability (thick line), first and third quartiles (thin box outline), 1.5 times the inner quartile range (lower and upper whiskers), and outliers (open circles) of each status for CU 28 according to the All-9 models under high and low productivity, and constant high DQA. CU 28: Spawner abundance = amber, trend in spawners = amber, harvest rate = red, distribution = red. The median probability of each status is calculated across 5,000 probability estimates.***

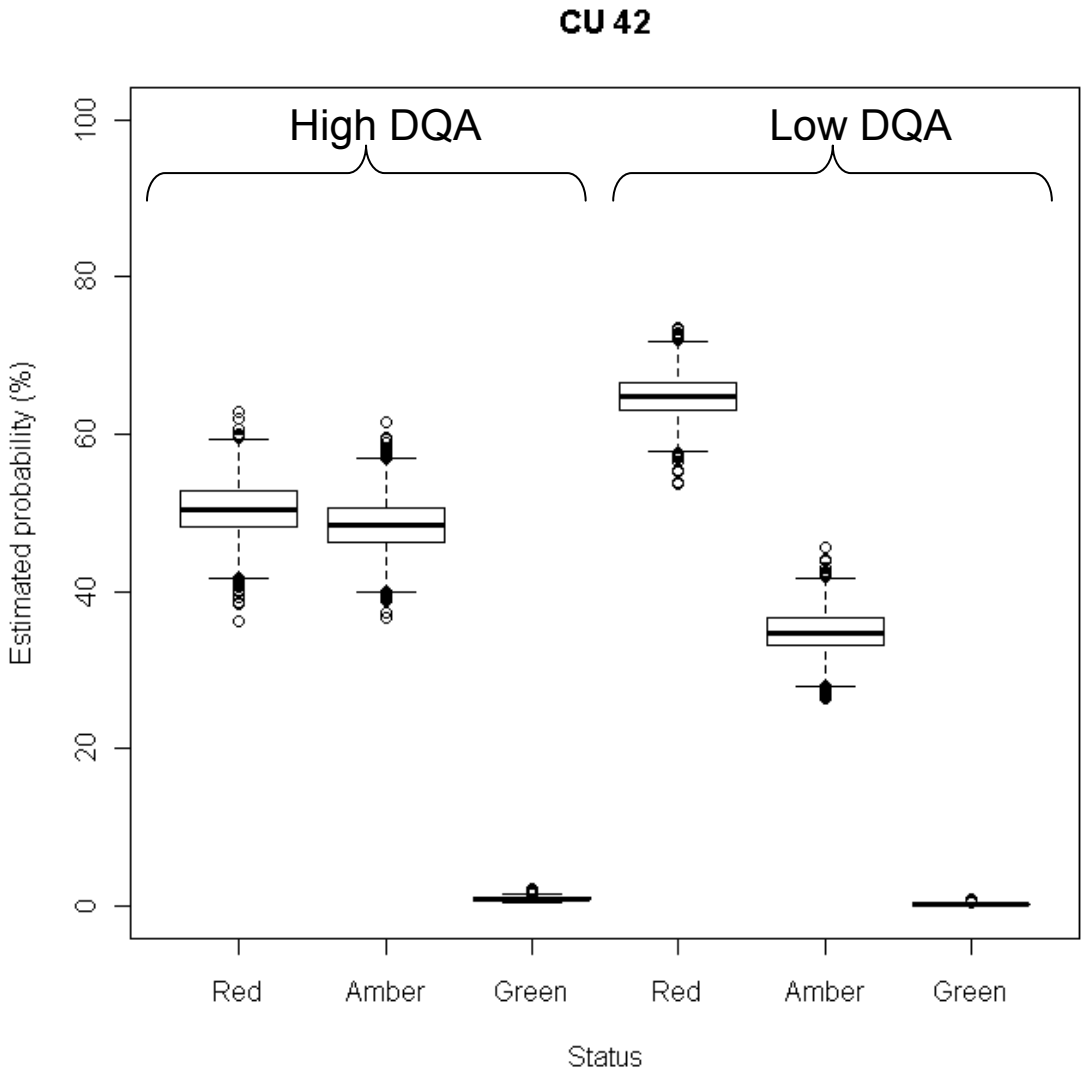



***Figure 5. Box and whisker plots showing the estimated median probability (thick line), first and third quartiles (thin box outline), 1.5 times the inner quartile range (lower and upper whiskers), and outliers (open circles) of each status for CU 42 according to the All-9 models under constant low productivity, and high and low data quality and amount (DQA). CU 42: spawner abundance = green, trend in spawners = red, harvest rate = red, distribution = amber. The median probability of each status is calculated across 5,000 probability estimates.***

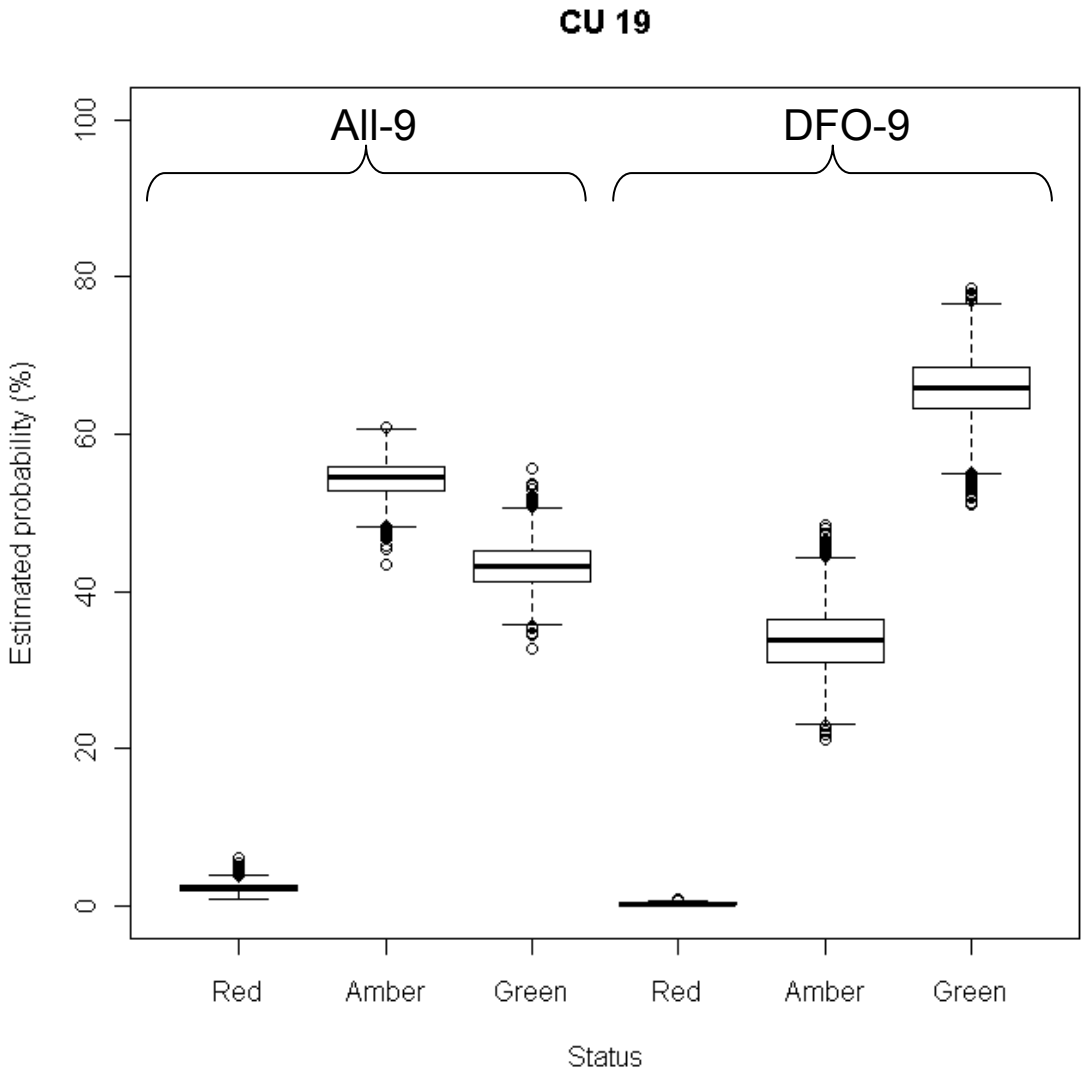



***Figure 6. Box and whisker plots showing the estimated median probability (thick line), first and third quartiles (thin box outline), 1.5 times the inner quartile range (lower and upper whiskers), and outliers (open circles) of each status for CU 19 according to All-9 and DFO-9 models under low productivity and high DQA. CU 19: spawner abundance = green, trend in spawners = amber, harvest rate = green, distribution = green. The median probability of each status is calculated across 5,000 probability estimates.***

# Appendices

# Appendix A

# Details to the Methods

### Questionnaire

Each scenario for a salmon Conservation Unit (CU) in the questionnaire simultaneously presented the metrics, metric status, and data quality and amount (top half of Figure 2). To minimize biases associated with experts' prior knowledge (e.g., historical spawner abundance data or current trend in spawner abundance of a particular CU), the CU scenarios were unnamed and generic, and did not represent a particular management area, CU, or salmon species. Likewise, to minimize the variability among respondents in the outcomes, the decision-making context (i.e., the framing of the problem or situation) was specified as unambiguously as possible (Morgan and Henrion 1992, Louviere 2006). To do this, we provided respondents with background information on the objectives of the project and definitions of the questionnaire's components, similar to the following descriptions.

### Defining metrics

For implementing the Wild Salmon Policy (WSP), Fisheries and Oceans Canada (DFO) is interested in tracking four biological indicators over time, spawner abundance, trends in spawners, harvest rate, and spatial distribution. For each of these classes of indicator, we selected one metric from a larger list identified by Holt et al. 2009, and modified the definitions to better suit the generic nature of the questionnaire (top half of Figure 2). These same four metrics were presented in each CU scenario, and are currently considered to be the most widely available and applicable to all salmon species.

### Defining metric levels: Metric status

Metric status is determined by comparing existing data to lower and upper benchmarks that divide the status of each metric status into red, amber, and green zones (top half of Figure 2). While no numerical references were presented to experts, they were provided with definitions of the upper and lower benchmarks for CUs (DFO 2005), and each of four metrics in a supplemental handout (Figure 1, Table 1). The quantitative benchmarks for metrics of spawner abundance, trend in spawners, and harvest rate were presented in Holt (2009) and Holt et al. (2009), discussed during workshops on the implementation of WSP Strategy 1, and await further revisions (pers. comm. C. A. Holt). Further analyses of the reliability of metric for detecting shifts in distribution are still required to identify the upper and lower benchmarks for metrics of spatial distribution (Peacock and Holt 2010).

### Defining metric levels: Data quality and amount

The data quality and amount (DQA) available to determine the biological status of each metric is highly variable across salmon species and CUs (Grant et al. 2011). To represent the potential range from near "perfect" information to little or no data, DQA was defined qualitatively as either high or low (top half of Figure 2). High DQA represents (1) consistent estimates every year from reliable methods, such as spawner abundance estimates based on mark-recapture and stream surveys, and (2) data on more than 3 generations or 10 years. In contrast, low DQA represents (1) no estimates or inconsistent estimates (e.g., only every other year) from less reliable methods,

such as tower counts (for estimates of spawner abundance), or (2) either no data or data that exist for less than 3 generations or 10 years.

## Global variable: CU productivity

Multiple factors contribute to temporal changes in the productivity of salmon populations and the differences among them (Peterman et al. 2000). In turn, productivity may affect the experts' evaluation of CU status. In the questionnaire, productivity is identified as either high or low and is defined as the intrinsic productivity (number of recruits per spawner at low spawner abundance, i.e., the "*a*" parameter in the Ricker model) of a CU. Herein, the CU spatial scale is implied when referring to productivity.

## Experimental Design

Each of the four metrics has six possible levels, representing the different combinations of metric status and DQA, i.e., red-high, red-low, amber-high, amber-low, green-high, green-low (top half of Figure 2). Of the 55 CU scenarios presented to respondents, 49 were generated through the experimental design, and 2 as warm-ups, 2 as holdouts, and 2 repeated CU scenarios (not included in the analysis). To control for effects of the order in which a person fill out questionnaires (Cohen and Neira 2003), seven versions of the questionnaire were generated, systematically alternating the order in which CU scenarios and indicators/metrics appeared on the page. See Appendix B for the experimental design.

## Response tasks and analysis

### *Conjoint rating*

Before conducting any analyses, the rating responses of two of the expert respondents were excluded from the original dataset because their Cook's Distance values stood out well in excess of others and exceeded the recommended cut-off criterion (Cook and Weisber 1982). In addition, all rating data passed general quality-control rules, i.e., final CU status ratings cannot be (1) lower or "worse" than the status of the lowest status metric, and (2) higher or "better" than the highest metric status in each scenario.

The adjacent-category ordinal logit model described in Equations (1) and (2) assumes that the ratings (1-9) are equidistant from each other, with a distance of one (Vermunt and Magidson 2005). Effects coding was used to scale the rating constants, metrics, and metric levels, by coding their part-worth utilities (PWUs) such that they sum to zero (Hensher et al. 2005). For example, metrics with two levels are assigned -1 and 1, while those with three levels are assigned -1, 0, and 1 (Hensher et al. 2005).

We estimated parameters for the main-effects statistical model (no interactions), as well as statistical models with the main effects plus up to 3 two-way interactions between pairs of metric status (Figure 3) for a total of 6018 models. All metric levels were retained in the main effects model (and all subsequent models thereafter) because together they compose the entire decision context.

Without having a priori knowledge of which interactions influence an expert's determination of CU status, we used model averaging of parameters across the top models (defined below) to account for model selection uncertainty (Johnson and Omland 2004). The small sample Akaike Information Criterion ($AIC_c$) was calculated for each model, and the $\Delta AIC_c$ of each model was calculated as the difference between the $AIC_c$ of that model and the $AIC_c$ of the best model (i.e., the model with lowest $AIC_c$) (Burnham and Anderson 2002). The top-model set is the set of models within $\Delta AIC_c < 4$ (Burnham and Anderson 2002). Consideration of both lower and higher

$AIC_c$ cut-offs (including considering models that account for 95% of the absolute Akaike weights) resulted in either the selection of single models with low absolute weight, or very large model sets (e.g., > 30 models) that could potentially lead to spurious results. The relative $AIC_c$ weight ($w_j$) of each model *j* in the top-model set (R) was calculated as,

(A1) $$w_j = \frac{\exp(-1/2\Delta AIC_{cj})}{\sum_{j=1}^{R}\exp(-1/2\Delta AIC_{cj})}$$

The relative weight $\Delta AIC_c$<4 ($w_{jAIC4}$) was calculated as in Equation (A1) but only across models in the top-model set. All rating constants and metric levels PWUs ($\bar{v}$) in the top-model set were calculated as a weighted averaged over the models in which the parameters appeared (Burnham and Anderson 2002) such that,

(A2) $$\bar{\bar{v}} = \sum_{j=1}^{R} w_{jAIC4}\,\bar{v}_j$$

where $\bar{v}_j$ is the estimate of $\bar{v}$ in the *j*th model. The associated unconditional standard error estimates ($\overline{SE_{nc}}(\bar{\bar{v}})$) were calculated as:

(A3) $$\overline{SE_{nc}}(\bar{\bar{v}}) = \sum_{j=1}^{R} w_{jAIC4}\left[\left(\overline{SE}(\bar{v}_j)\right)^2 + \left(\bar{\bar{v}} - \bar{v}_j\right)^2\right]$$

where $\overline{SE}(\bar{v}_j)^2$ is the conditional variance (standard error squared) of $\bar{v}$ of the $j^{th}$ model, and $\left(\bar{\bar{v}} - \bar{v}_j\right)^2$ is the model selection variance. The resulting confidence intervals account for both model selection uncertainty and metric PWU variance (Johnson and Omland 2004).

The conjoint rating task cannot provide information to what extent experts are using the metric status or data quality and amount alone to assess the CU status or whether the metrics used in the assessment of CU status are driving the CU ratings (Lanscar et al. 2007). Therefore, in additive conjoint rating models, the PWUs of metrics remain confounded with the metric level PWUs.

***Best-worst scaling***

The best-worst scaling (BWS) analyses addresses the above limitation of the conjoint rating analysis by providing additional information on the relative importance of the metrics, metric status, data quality and amount (DQA) and two-way interactions (e.g., Ab-DQAH). There are actually three types of BWS: the object case (case 1), the profile or attribute case (case 2) and the multi-profile case (case 3) (Flynn, 2010). We used the profile case (case 2) to ask expert respondents to consider one CU scenario at a time and within it choose the one "best" and the one "worst" metric based on the metric levels presented.

The statistical model underlying BWS assumes that the relative probability of choosing a given pair of metrics is proportional to the distance between the metrics' levels on the utility scale (Flynn et al. 2007). The model also assumes that respondents are simultaneously comparing each metric/level combination relative to others found in the same scenario, and then choosing the pair that exhibits the maximum perceptual difference (Cohen and Neira 2003). By forcing respondents to consider only the extremes of the utility scale, best-worst scaling provides more information than "pick one" choice tasks commonly associated with DCEs (Flynn et al. 2007). BWS prevents respondents from choosing the middle or one end of the utility scale (Lee et al. 2007) thereby minimizing response biases that may occur in other methods such as rating scales (Paulhus 1991).

The BWS task is modeled (see Equation (3) and (4)) as a sequential choice process where the "worst" choice (to the red end of the rating scale) is equivalent to a first choice (Vermunt and Magidson 2005). Subsequently, the "best" choice (i.e., toward the green end of the rating scale) is equivalent to a first choice out of the remaining categorical response (A, B, C or D), but where the choice probability is inversed (by using a scale factor). By identifying the response task as a

ranking, the Latent Gold 4.5 software automatically removes the “worst” response from the set available for the “best” choice.

Analogous to in the conjoint rating analysis, we estimated parameters for the main-effects statistical model (no interactions), as well as statistical models with the main effects plus up to 4 two-way interactions between metric status and DQA for a total of 16 models. Here too, metric levels were retained in the main effects model (and all subsequent models thereafter). The small sample Akaike Information Criterion ($AIC_c$) was calculated for each model and the difference between that value and the next best model’s $AIC_c$ ($\Delta AIC_c$) was determined (Burnham and Anderson 2002). The top-model set was identified using $\Delta AIC_c < 4$ (Burnham and Anderson 2002). The relative $AIC_c$ weight ($w_j$) of each model *j* in the top-model set (R) was calculated as in Equation (A1). All metric weights and level scale PWUs ($\bar{\bar{v}}$) in the top-model set were calculated as a weighted averaged over the models in which the parameters appeared using the Equation (A2). The associated unconditional standard error estimates ($\overline{SE_{nc}}$) of the metric weights and level scale PWUs were calculated using Equation (A3).

# Appendix B.

# Experimental Design

***Table B1. Experimental design of the questionnaire for both the conjoint rating and best-worst scaling response tasks. For each of the 49 scenarios, the 4 metrics (spawner abundance, trend in spawners, harvest rate, and distribution) are given a metric status, where 1 = Red, 2 = Amber, 3 = Green and level of data quality and amount (DQA), where 1 = High, 2 = Low. The experimental design accounts for a biological constraint: (1) if spawner abundance is red in status, then trend in spawners must also be red in status, and (2) if spawner abundance is amber in status, then trend in spawners must either be amber or red in status. Seven versions of the questionnaire were generated by systematically alternating the order of the scenario blocks.***

| | Spawner abundance | | Trend in spawners | | Harvest rate | | Distribution | | Scenario block |
|---|---|---|---|---|---|---|---|---|---|
| Scenario | Status | DQA | Status | DQA | Status | DQA | Status | DQA | |
| 1 | 2 | 2 | 2 | 2 | 3 | 2 | 1 | 1 | 1 |
| 2 | 3 | 1 | 3 | 1 | 2 | 1 | 3 | 2 | 1 |
| 3 | 3 | 2 | 3 | 2 | 1 | 1 | 2 | 2 | 1 |
| 4 | 1 | 2 | 1 | 1 | 3 | 1 | 3 | 1 | 1 |
| 5 | 2 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 1 |
| 6 | 3 | 1 | 1 | 2 | 1 | 2 | 1 | 2 | 1 |
| 7 | 3 | 2 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| 8 | 2 | 2 | 1 | 2 | 3 | 1 | 2 | 1 | 2 |
| 9 | 3 | 1 | 1 | 1 | 1 | 1 | 3 | 1 | 2 |
| 10 | 1 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 2 |
| 11 | 2 | 1 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 12 | 3 | 2 | 3 | 2 | 2 | 2 | 3 | 2 | 2 |
| 13 | 3 | 2 | 3 | 1 | 1 | 2 | 2 | 2 | 2 |
| 14 | 3 | 1 | 2 | 2 | 3 | 2 | 1 | 2 | 2 |
| 15 | 3 | 1 | 3 | 2 | 1 | 2 | 1 | 1 | 3 |
| 16 | 2 | 1 | 1 | 1 | 3 | 2 | 3 | 2 | 3 |
| 17 | 3 | 2 | 3 | 2 | 3 | 1 | 2 | 2 | 3 |
| 18 | 3 | 2 | 1 | 2 | 2 | 1 | 2 | 1 | 3 |
| 19 | 3 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 3 |
| 20 | 1 | 2 | 1 | 1 | 2 | 2 | 1 | 2 | 3 |
| 21 | 2 | 1 | 2 | 1 | 2 | 2 | 3 | 1 | 3 |
| 22 | 3 | 1 | 3 | 1 | 1 | 1 | 2 | 1 | 4 |
| 23 | 3 | 2 | 3 | 1 | 3 | 2 | 2 | 2 | 4 |
| 24 | 2 | 1 | 2 | 2 | 3 | 1 | 3 | 2 | 4 |
| 25 | 3 | 2 | 2 | 2 | 2 | 2 | 3 | 1 | 4 |
| 26 | 1 | 1 | 1 | 2 | 1 | 2 | 2 | 2 | 4 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 27 | 3 | 2 | 1 | 1 | 2 | 2 | 1 | 1 | 4 |
| 28 | 2 | 2 | 1 | 2 | 2 | 1 | 1 | 2 | 4 |
| 29 | 3 | 2 | 3 | 2 | 3 | 2 | 3 | 1 | 5 |
| 30 | 3 | 2 | 3 | 1 | 3 | 1 | 1 | 2 | 5 |
| 31 | 2 | 1 | 1 | 2 | 1 | 1 | 1 | 1 | 5 |
| 32 | 1 | 1 | 1 | 2 | 2 | 2 | 2 | 2 | 5 |
| 33 | 2 | 2 | 2 | 1 | 1 | 2 | 2 | 1 | 5 |
| 34 | 3 | 2 | 1 | 2 | 2 | 2 | 3 | 2 | 5 |
| 35 | 3 | 1 | 2 | 1 | 2 | 1 | 2 | 2 | 5 |
| 36 | 3 | 1 | 3 | 2 | 2 | 1 | 3 | 1 | 6 |
| 37 | 3 | 2 | 2 | 1 | 1 | 2 | 3 | 2 | 6 |
| 38 | 1 | 1 | 1 | 2 | 3 | 2 | 2 | 1 | 6 |
| 39 | 2 | 2 | 2 | 1 | 1 | 1 | 1 | 2 | 6 |
| 40 | 3 | 2 | 3 | 2 | 2 | 2 | 1 | 1 | 6 |
| 41 | 3 | 1 | 1 | 2 | 3 | 1 | 2 | 2 | 6 |
| 42 | 2 | 2 | 1 | 1 | 2 | 2 | 2 | 2 | 6 |
| 43 | 3 | 2 | 1 | 1 | 3 | 2 | 2 | 2 | 7 |
| 44 | 2 | 2 | 2 | 2 | 2 | 1 | 2 | 2 | 7 |
| 45 | 3 | 1 | 3 | 1 | 2 | 2 | 2 | 1 | 7 |
| 46 | 3 | 1 | 3 | 2 | 2 | 2 | 1 | 2 | 7 |
| 47 | 2 | 2 | 1 | 2 | 1 | 2 | 3 | 1 | 7 |
| 48 | 3 | 1 | 2 | 1 | 3 | 1 | 1 | 1 | 7 |
| 49 | 1 | 2 | 1 | 2 | 1 | 1 | 3 | 2 | 7 |

# Appendix C.

# Random Utility Theory

Choice modeling and by extension best-worst scaling is rooted in random utility theory (RUT) – a paradigm first proposed by Thurston (1927) as a means to understand and model paired comparisons of choice alternatives. RUT posits that decision-making or choice behaviour by individuals is composed by both deterministic (observable) and random (unobservable or error) components, that when added give an overall utility (McFadden 1974), such that:

(C1) $$U_i = V_i + \varepsilon_i \ ,$$

where $U_i$ is the unobserved, latent utility of an alternative $i$ (e.g., rating a CU scenario as 4 along the 9-point color scale), $V_i$ is the deterministic, quantifiable part of utility made up of the metrics that explain differences in choice alternatives. The random or error component $\varepsilon_i$ represents all unaccounted metrics affecting choices, and other factors describing the variability in choices across individuals (Train 1986; Ben-Akiva and Lerman 1985). In multinomial logit regressions the error term is assumed to follow the Gumbel or Type I Extreme Value distribution. RUT implies that 'utility' is inherently stochastic, therefore we can predict the probability that an individual will choose an alternative $i$ as:

(C2) $$P(i|C) = P[(V_i + \varepsilon_i) > (V_j + \varepsilon_j)] \ ,$$

where $C$ is the set of all possible alternatives, and $j$ is any other alternative. An individual will choose alternative $i$ if the deterministic and random components of that alternative are larger than the deterministic and random components of all other alternatives. RUT assumes that some underlying subjective dimension, such as the 'utility' or 'relative importance', can be quantified by assigning numerical values that reflect the relative position of the questionnaire metrics on that underlying scale (Cohen 2003). RUT also assumes that individuals seek to maximize their utility (degree of satisfaction) when making choices.

# Appendix D.

## Top-model Sets

**Table D1. Summary of top-model set ($\Delta AIC_c$<4) resulting from adjacent-category ordinal logit models of CU status ratings (9-point scale) from all respondents (n=37), for high and low productivity. Baseline (no interactions) denotes models that include only the rating constants and metric levels. All other models include the baseline model and one or more interactions (xx-xx) between pairs of metric status where red (R), amber (A), and green (G), spawner abundance (Ab), trend in spawners (Tr), harvest rate (Ha), and distribution (Di). $AIC_c$ is the Akaike information criterion corrected for small sample size, $\Delta AIC_c$ refers the difference in $AIC_c$ between a given model and the model with the lowest $AIC_c$, $AIC_c$ relative weight ($w_j$) is the weight of each model relative to all statistical models (total of 6018), and relative weight $\Delta AIC_c < 4$ ($w_{jAIC4}$) is the relative support for each model in top-model set.**

| Models in Top-model Set | Sample size (n) | Number of parameters | Log-likelihood | $AIC_c$ | $\Delta AIC_c$ | $AIC_c$ relative weight ($w_j$) | Relative weight $\Delta AIC_c$<4 ($w_{jAIC4}$) |
|---|---|---|---|---|---|---|---|
| *For high productivity* | | | | | | | |
| AbA-HaR | 37 | 21 | -2583.362 | 5255.258 | 0.000 | 0.273 | 0.331 |
| DiR-HaA | 37 | 21 | -2580.287 | 5256.461 | 1.203 | 0.150 | 0.181 |
| AbA-HaR, DiR-HaA | 37 | 22 | -2570.151 | 5256.588 | 1.330 | 0.141 | 0.170 |
| AbA-HaR, AbA-DiG | 37 | 22 | -2570.507 | 5257.300 | 2.042 | 0.098 | 0.119 |
| AbA-HaR, AbR-DiA | 37 | 22 | -2570.612 | 5257.510 | 2.252 | 0.089 | 0.107 |
| AbA-HaR, AbR-DiA, DiR-HaA | 37 | 23 | -2564.139 | 5259.202 | 3.944 | 0.038 | 0.046 |
| Baseline (no interactions) | 37 | 20 | -2583.371 | 5259.241 | 3.983 | 0.037 | 0.045 |
| *For low productivity* | | | | | | | |
| DiR-HaA | 37 | 21 | -2665.062 | 5433.724 | 0.000 | 0.309 | 0.404 |
| Baseline (no interactions) | 37 | 20 | -2671.205 | 5434.911 | 1.187 | 0.171 | 0.223 |
| DiR-HaA, TrG-HaR | 37 | 22 | -2659.867 | 5436.019 | 2.295 | 0.098 | 0.128 |
| AbA-HaG | 37 | 21 | -2666.444 | 5436.488 | 2.764 | 0.078 | 0.101 |
| AbA-HaR | 37 | 21 | -2666.634 | 5436.867 | 3.143 | 0.064 | 0.084 |
| TrG-HaR | 37 | 21 | -2666.986 | 5437.573 | 3.849 | 0.045 | 0.059 |

**Table D2. Same as Table D1 except results are for analyses of best-worst scaling responses. Models include the baseline model and one or more interactions between a metric, spawner abundance (Ab), trend in spawners (Tr), distribution (Di), and high data quality and amount (DQAH).**

| Models in Top-model Set | Sample size (n) | Number of para-meters | Log-likelihood | $AIC_c$ | $\Delta AIC_c$ | $AIC_c$ relative weight ($w_j$) | Relative weight $\Delta AIC_c<4$ ($w_{jAIC4}$) |
|---|---|---|---|---|---|---|---|
| *For high productivity* | | | | | | | |
| Ab-DQAH, Tr-DQAH | 37 | 19 | -1182.819 | 2448.344 | 0.000 | 0.854 | 0.871 |
| Ab-DQAH, Tr-DQAH, Di-DQAH | 37 | 21 | -1174.283 | 2452.165 | 3.821 | 0.126 | 0.129 |
| *For low productivity* | | | | | | | |
| Ab-DQAH, Tr-DQAH | 37 | 19 | -1186.693 | 2456.091 | 0.000 | 0.975 | 1.000 |

**Table D3. Same as Table D1 except results are for analyses of CU status ratings by DFO respondents (n=27; DFO-9).**

| Models in Top-model Set | Sample size (n) | Number of para-meters | Log-likelihood | $AIC_c$ | $\Delta AIC_c$ | $AIC_c$ relative weight ($w_j$) | Relative weight $\Delta AIC_c<4$ ($w_{jAIC4}$) |
|---|---|---|---|---|---|---|---|
| *For high productivity* | | | | | | | |
| Baseline (no interactions) | 27 | 20 | -1871.999 | 3923.997 | 0.000 | 0.999 | 1.000 |
| *For low productivity* | | | | | | | |
| Baseline (no interactions) | 27 | 20 | -1942.49 | 4064.987 | 0.000 | 0.999 | 1.000 |

**Table D4. Same as Table D1 except results are for analyses of CU status ratings using a 3-point scale (n=27; All-3).**

| Models in Top-model Set | Sample size (n) | Number of para-meters | Log-likelihood | $AIC_c$ | $\Delta AIC_c$ | $AIC_c$ relative weight ($w_j$) | Relative weight $\Delta AIC_c<4$ ($w_{jAIC4}$) |
|---|---|---|---|---|---|---|---|
| *For high productivity* | | | | | | | |
| AbA-HaR, AbR-DiG, AbA-DiG | 37 | 17 | -1024.515 | 2115.240 | 0.000 | 0.140 | 0.223 |
| AbA-HaR, AbA-DiG | 37 | 16 | -1028.093 | 2115.386 | 0.145 | 0.131 | 0.208 |
| AbA-HaG, AbA-DiG | 37 | 16 | -1028.424 | 2116.047 | 0.807 | 0.094 | 0.149 |
| AbA-HaR, AbA-DiG, DiR-HaG | 37 | 17 | -1024.922 | 2116.054 | 0.814 | 0.093 | 0.149 |
| AbA-HaR, AbR-DiA, AbA-DiG | 37 | 17 | -1025.689 | 2117.588 | 2.348 | 0.043 | 0.069 |
| AbA-HaR, AbA-HaG, AbA-DiG | 37 | 17 | -1025.784 | 2117.779 | 2.538 | 0.039 | 0.063 |
| AbA-HaG, AbG-HaA, AbA-DiG | 37 | 17 | -1026.017 | 2118.244 | 3.003 | 0.031 | 0.050 |
| AbA-HaR, AbG-DiR, AbA-DiG | 37 | 17 | -1026.087 | 2118.385 | 3.144 | 0.029 | 0.046 |
| AbA-HaG, AbR-DiG, AbA-DiG | 37 | 17 | -1026.155 | 2118.521 | 3.281 | 0.027 | 0.043 |
| *For low productivity* | | | | | | | |
| AbG-HaR, DiR-HaA | 37 | 16 | -1074.434 | 2208.069 | 0.000 | 0.092 | 0.211 |
| AbA-HaR, DiR-HaA | 37 | 16 | -1074.547 | 2208.295 | 0.226 | 0.082 | 0.188 |
| DiR-HaA | 37 | 15 | -1078.077 | 2209.010 | 0.941 | 0.057 | 0.132 |
| AbA-HaG, DiR-HaA | 37 | 16 | -1075.012 | 2209.224 | 1.156 | 0.051 | 0.118 |
| AbG-HaR, AbG-DiA, DiR-HaA | 37 | 17 | -1072.562 | 2211.335 | 3.266 | 0.018 | 0.041 |
| AbA-HaR, AbG-DiA, DiR-HaA | 37 | 17 | -1072.690 | 2211.590 | 3.521 | 0.016 | 0.036 |
| AbG-TrR, AbG-HaR, DiR-HaA | 37 | 17 | -1072.796 | 2211.802 | 3.733 | 0.014 | 0.033 |
| AbG-TrA, AbG-HaR, DiR-HaA | 37 | 17 | -1072.796 | 2211.802 | 3.733 | 0.014 | 0.033 |
| AbA-TrR, AbG-HaR, DiR-HaA | 37 | 17 | -1072.796 | 2211.802 | 3.733 | 0.014 | 0.033 |
| AbG-TrR, DiR-HaA | 37 | 16 | -1076.387 | 2211.974 | 3.905 | 0.013 | 0.030 |
| AbG-TrA, DiR-HaA | 37 | 16 | -1076.387 | 2211.974 | 3.905 | 0.013 | 0.030 |
| AbA-TrR, DiR-HaA | 37 | 16 | -1076.387 | 2211.974 | 3.905 | 0.013 | 0.030 |
| AbG-TrR, AbA-HaR, DiR-HaA | 37 | 17 | -1072.911 | 2212.032 | 3.964 | 0.013 | 0.029 |
| AbG-TrA, AbA-HaR, DiR-HaA | 37 | 17 | -1072.911 | 2212.032 | 3.964 | 0.013 | 0.029 |
| AbA-TrR, AbA-HaR, DiR-HaA | 37 | 17 | -1072.911 | 2212.032 | 3.964 | 0.013 | 0.029 |

# Appendix E.

## DFO-9 and All-3 Models

**Table E1. Results of analyses of CU status ratings from DFO respondents along a 9-point scale (n=27; DFO-9). Model-averaged part-worth utilities (PWUs) of the rating constants, metric status, and data quality and amount (DQA) of the top model set (∆AICc<4) for high and low productivity. Rating constants are rating specific (1-9) intercepts in Equation (1) (where red status = rating=1-3, amber status= 4-6, and green status = 7-9). Also shown is the associated unconditional standard error ($SE_{nc}$, Equation (A3) of Appendix A), and 95% confidence interval (CI).**

| | *For high productivity* | | | *For low productivity* | | |
|---|---|---|---|---|---|---|
| | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** |
| Rating constant | | | | | | |
| 1 | -4.158 | 0.28 | (-4.706, -3.610) | -0.049 | 0.172 | (-0.385, 0.287) |
| 2 | -0.416 | 0.165 | (-0.739, -0.093) | 2.041 | 0.179 | (1.689, 2.392) |
| 3 | 2.134 | 0.175 | (1.791, 2.476) | 2.652 | 0.172 | (2.315, 2.990) |
| 4 | 2.524 | 0.17 | (2.192, 2.857) | 3.011 | 0.141 | (2.735, 3.288) |
| 5 | 3.088 | 0.139 | (2.815, 3.360) | 2.286 | 0.117 | (2.056, 2.516) |
| 6 | 2.175 | 0.122 | (1.936, 2.414) | 0.803 | 0.113 | (0.581, 1.025) |
| 7 | 0.702 | 0.123 | (0.461, 0.943) | -0.285 | 0.126 | (-0.532, -0.038) |
| 8 | -1.216 | 0.162 | (-1.535, -0.898) | -3.26 | 0.212 | (-3.675, -2.846) |
| 9 | -4.832 | 0.276 | (-5.372, -4.291) | -7.199 | 0.426 | (-8.033, -6.364) |
| Spawner abundance status | | | | | | |
| Red | -1.504 | 0.11 | (-1.720, -1.288) | -1.323 | 0.097 | (-1.513, -1.133) |
| Amber | 0.261 | 0.061 | (0.142, 0.379) | 0.281 | 0.054 | (0.175, 0.387) |
| Green | 1.243 | 0.076 | (1.095, 1.392) | 1.043 | 0.066 | (0.914, 1.171) |
| Spawner abundance DQA | | | | | | |
| High | 0.14 | 0.028 | (0.086, 0.194) | 0.131 | 0.026 | (0.081, 0.181) |
| Low | -0.14 | 0.028 | (-0.194, -0.086) | -0.131 | 0.026 | (-0.181, -0.081) |
| Trend in spawners status | | | | | | |
| Red | -0.753 | 0.051 | (-0.852, -0.653) | -0.664 | 0.046 | (-0.754, -0.574) |
| Amber | -0.015 | 0.042 | (-0.096, 0.066) | 0.008 | 0.037 | (-0.065, 0.081) |
| Green | 0.768 | 0.056 | (0.659, 0.877) | 0.656 | 0.049 | (0.561, 0.751) |
| Trend in spawners DQA | | | | | | |
| High | -0.009 | 0.027 | (-0.061, 0.044) | -0.007 | 0.025 | (-0.055, 0.042) |
| Low | 0.009 | 0.027 | (-0.044, 0.061) | 0.007 | 0.025 | (-0.042, 0.055) |

| | *For high productivity* | | | *For low productivity* | | |
|---|---|---|---|---|---|---|
| | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** |
| Harvest rate status | | | | | | |
| Red | -0.368 | 0.042 | (-0.450, -0.287) | -0.367 | 0.039 | (-0.443, -0.290) |
| Amber | 0.016 | 0.037 | (-0.056, 0.088) | 0.01 | 0.034 | (-0.056, 0.077) |
| Green | 0.353 | 0.042 | (0.271, 0.434) | 0.356 | 0.039 | (0.280, 0.432) |
| Harvest rate DQA | | | | | | |
| High | 0.085 | 0.028 | (0.031, 0.140) | 0.061 | 0.026 | (0.011, 0.111) |
| Low | -0.085 | 0.028 | (-0.140, -0.031) | -0.061 | 0.026 | (-0.111, -0.011) |
| Distribution status | | | | | | |
| Red | -0.449 | 0.043 | (-0.533, -0.364) | -0.402 | 0.04 | (-0.481, -0.323) |
| Amber | 0.02 | 0.036 | (-0.052, 0.091) | 0.033 | 0.034 | (-0.034, 0.099) |
| Green | 0.429 | 0.043 | (0.344, 0.514) | 0.369 | 0.039 | (0.292, 0.447) |
| Distribution DQA | | | | | | |
| High | 0.002 | 0.027 | (-0.051, 0.056) | 0.013 | 0.025 | (-0.036, 0.062) |
| Low | -0.002 | 0.027 | (-0.056, 0.051) | -0.013 | 0.025 | (-0.062, 0.036) |

**Table E2. Same as Table E1 except these results are for analyses of CU status ratings from all respondents along a 3-point scale (n=37; All-3). Rating constants are rating specific (1-3) intercepts in Equation (1) (where red status = rating=1, amber status= 2, and green status = 2). Each model includes interactions between pairs of metric status. Symbols in those interactions, red (R), amber (A), and green (G), spawner abundance (Ab), trend in spawners (Tr), harvest rate (Ha), distribution (Di). Because all metric levels were retained in all models the relative variable importance (RVI) is shown for only the two-way interactions, as calculated from the Akaike weights. Blanks in some columns in interaction rows indicate that the interaction was absent from the top models set.**

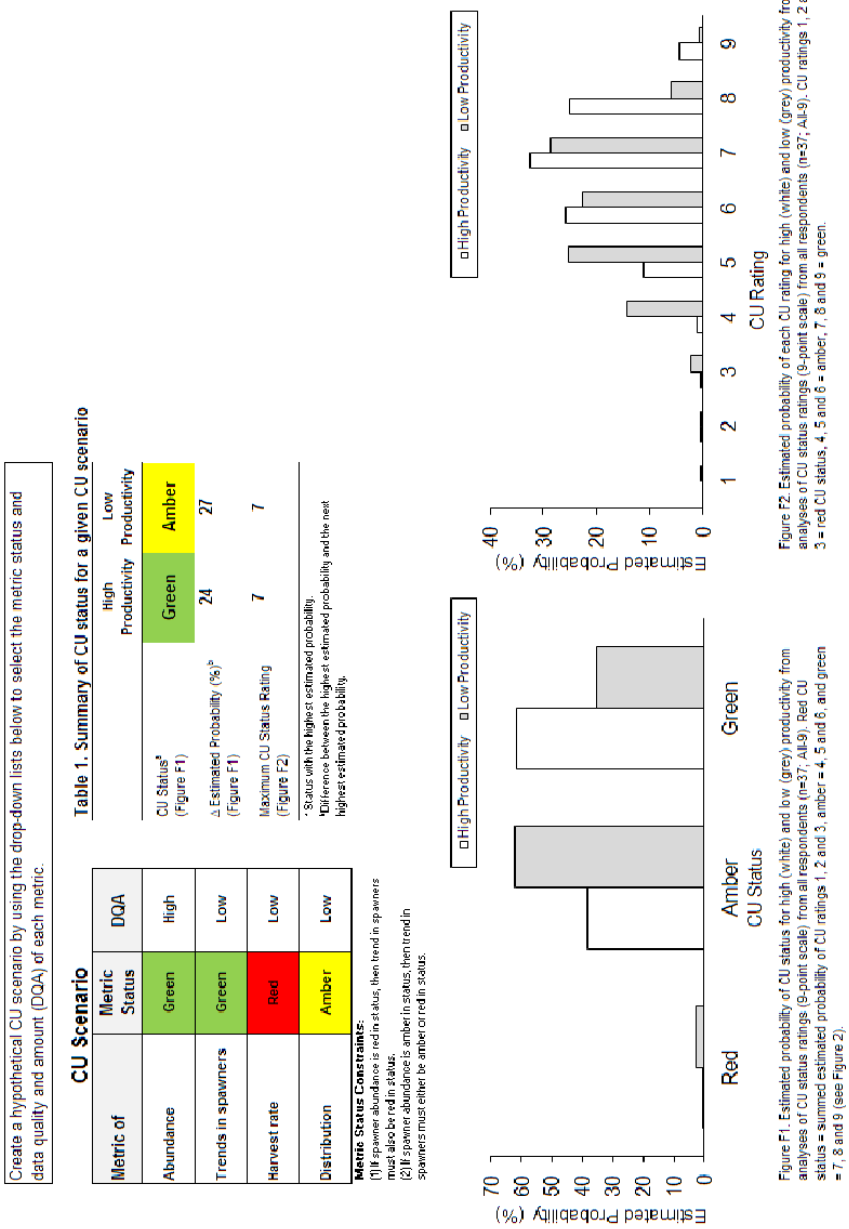
| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| Rating constant | | | | | | | | |
| Red (1) | -0.135 | 0.043 | (-0.220, -0.051) | | 0.993 | 0.038 | (0.919, 1.068) | |
| Amber (2) | 1.505 | 0.005 | (1.496, 1.515) | | 1.288 | 0.003 | (1.282, 1.295) | |
| Green (3) | -1.370 | 0.043 | (-1.454, -1.286) | | -2.281 | 0.040 | (-2.360, -2.203) | |
| Spawner abundance status | | | | | | | | |
| Red | -2.996 | 0.115 | (-3.222, -2.771) | | -2.999 | 0.127 | (-3.248, -2.749) | |
| Amber | 0.679 | 0.067 | (0.547, 0.811) | | 0.779 | 0.050 | (0.682, 0.876) | |
| Green | 2.317 | 0.047 | (2.225, 2.410) | | 2.220 | 0.050 | (2.123, 2.317) | |
| Spawner abundance DQA | | | | | | | | |
| High | 0.331 | 0.004 | (0.322, 0.339) | | 0.200 | 0.004 | (0.193,0.207) | |
| Low | -0.331 | 0.004 | (-0.339, -0.322) | | -0.200 | 0.004 | (-0.207, -0.193) | |
| Trend in spawners status | | | | | | | | |
| Red | -1.569 | 0.011 | (-1.590, -1.547) | | -1.537 | 0.015 | (-1.567, -1.507) | |
| Amber | -0.045 | 0.008 | (-0.061, -0.029) | | -0.055 | 0.013 | (-0.080, -0.030) | |
| Green | 1.614 | 0.011 | (1.592, 1.635) | | 1.592 | 0.017 | (1.559, 1.624) | |
| Trend in spawners DQA | | | | | | | | |
| High | -0.098 | 0.003 | (-0.105, -0.091) | | -0.070 | 0.003 | (-0.076, -0.064) | |
| Low | 0.098 | 0.003 | (0.091, 0.105) | | 0.070 | 0.003 | (0.064, 0.076) | |
| Harvest rate status | | | | | | | | |
| Red | -0.695 | 0.013 | (-0.721, -0.669) | | -0.651 | 0.052 | (-0.754, -0.549) | |
| Amber | -0.136 | 0.031 | (-0.196, -0.076) | | 0.137 | 0.020 | (0.098, 0.175) | |
| Green | 0.831 | 0.023 | (0.786, 0.876) | | 0.515 | 0.021 | (0.473, 0.556) | |
| Harvest rate DQA | | | | | | | | |
| High | 0.074 | 0.004 | (0.066, 0.082) | | 0.155 | 0.003 | (0.148, 0.161) | |
| Low | -0.074 | 0.004 | (-0.082, -0.066) | | -0.155 | 0.003 | (-0.161, -0.148) | |

| | *For high productivity* | | | | *For low productivity* | | | |
|---|---|---|---|---|---|---|---|---|
| | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** | **PWU ($\overline{\overline{v}}$)** | **$SE_{nc}$** | **95% CI** | **RVI** |
| Distribution status | | | | | | | | |
| Red | -1.138 | 0.019 | (-1.175, -1.101) | | -0.617 | 0.013 | (-0.643, -0.590) | |
| Amber | 0.100 | 0.010 | (0.080, 0.120) | | 0.052 | 0.012 | (0.030, 0.075) | |
| Green | 1.038 | 0.019 | (1.000, 1.076) | | 0.564 | 0.009 | (0.548, 0.581) | |
| Distribution DQA | | | | | | | | |
| High | -0.052 | 0.004 | (-0.061, -0.044) | | 0.014 | 0.003 | (0.008, 0.020) | |
| Low | 0.052 | 0.004 | (0.044, 0.061) | | -0.014 | 0.003 | (-0.020, -0.008) | |
| Interactions | | | | | | | | |
| AbA-HaR | 0.788 | 0.114 | (0.566, 1.011) | 0.76 | 0.218 | 0.094 | (0.034, 0.403) | 0.31 |
| AbR-DiG | 0.370 | 0.362 | (-0.339, 1.080) | 0.27 | | | | |
| AbA-DiG | -1.192 | 0.095 | (-1.379, -1.006) | 1 | | | | |
| AbA-HaG | -0.307 | 0.306 | (-0.907, 0.293) | 0.3 | -0.076 | 0.046 | (-0.167, 0.015) | 0.12 |
| DiR-HaG | 0.111 | 0.073 | (-0.032, 0.253) | 0.15 | | | | |
| AbR-DiA | -0.082 | 0.106 | (-0.290, 0.126) | 0.07 | | | | |
| AbG-HaA | 0.032 | 0.024 | (-0.014, 0.079) | 0.05 | | | | |
| AbG-DiR | 0.028 | 0.019 | (-0.010, 0.066) | 0.05 | | | | |
| AbG-HaR | | | | | -0.246 | 0.097 | (-0.437, -0.056) | 0.35 |
| DiR-HaA | | | | | -1.071 | 0.076 | (-1.220, -0.921) | 1 |
| AbG-DiA | | | | | -0.036 | 0.018 | (-0.072, 3.9E-04) | 0.08 |
| AbG-TrR | | | | | 0.043 | 0.023 | (-0.002, 0.088) | 0.09 |
| AbG-TrA | | | | | -0.043 | 0.023 | (-0.088, 0.002) | 0.09 |
| AbA-TrR | | | | | -0.043 | 0.023 | (-0.088, 0.002) | 0.09 |

# Appendix F.

## CU Status Decision Support Spreadsheet

**Figure F1. Example results are from model-averaged PWUs analyses for high and low productivity (All-9; Table 2). Experts can use the spreadsheet to visualize how changes in metric status and data quality and amount (DQA) affect the final CU status (Figure F1) and the estimated probability for each status or CU rating (Figure F2), for high and low productivity.**



# Appendix G.

## Estimated CU Status of 54 Hypothetical CUs

**Figure G1. Estimated status of 54 hypothetical CUs (R = red, A = amber, G = green) from model-averaged part-worth utilities (PWUs) of all models under high and low productivity, and high and low DQA, along with the difference between the highest probability (i.e., CUstatus) and the next highest probability (∆ Probability (Status - Next Highest)). CU status of the equal weighting method is the average status of multiple metric in a given CU, where red = 1, amber = 2, and green = 3.**

| Metric | CUs | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| Spawner abundance | R | G | R | R | G | G | R | G | G | G | G | G | A | A | A | G | G | A | G | G | G | G | G | A | A | A | R |
| Trend in spawners | R | G | R | R | R | G | R | R | R | R | G | G | A | A | A | A | G | A | A | A | A | G | G | A | A | R | R |
| Harvest rate | R | G | R | G | R | R | G | G | R | G | G | R | A | A | G | A | A | G | G | A | G | G | A | A | R | A | A |
| Distribution | R | G | G | R | R | R | G | G | G | R | R | G | A | G | A | A | A | G | G | G | A | A | G | R | A | A | A |
| All-9, High productivity, High DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 92 | 92 | 69 | 70 | 48 | 3.5 | 21 | 10 | 66 | 67 | 61 | 63 | 70 | 53 | 57 | 0.0 | 70 | 32 | 55 | 33 | 26 | 82 | 85 | 63 | 60 | 44 | 63 |
| All-9, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 96 | 82 | 83 | 83 | 19 | 37 | 48 | 42 | 70 | 69 | 34 | 36 | 66 | 70 | 72 | 34 | 47 | 58 | 25 | 1.3 | 9 | 65 | 70 | 40 | 35 | 14 | 78 |
| All-9, Low productivity, High DQA | R | G | R | R | R | A | R | A | A | A | A | A | A | A | A | A | G | A | A | A | A | G | G | A | A | R | R |
| Δ Probability (highest-next highest) | 99 | 52 | 96 | 96 | 30 | 60 | 82 | 64 | 29 | 30 | 6 | 8 | 37 | 55 | 54 | 58 | 4.0 | 65 | 15 | 39 | 41 | 29 | 31 | 5 | 2.3 | 25 | 94 |
| All-9, Low productivity, Low DQA | R | G | R | R | R | A | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R | R |
| Δ Probability (highest-next highest) | 100 | 25 | 98 | 98 | 56 | 58 | 91 | 52 | 0.9 | 0.2 | 37 | 38 | 9 | 33 | 32 | 59 | 28 | 52 | 44 | 60 | 61 | 2.5 | 0.5 | 25 | 28 | 52 | 97 |
| DF0, High productivity, High DQA | R | G | R | R | A | A | R | G | A | A | G | G | A | A | A | G | G | A | G | G | G | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 94 | 92 | 73 | 79 | 57 | 3.6 | 33 | 3.8 | 56 | 64 | 58 | 67 | 69 | 47 | 52 | 3.0 | 66 | 19 | 64 | 40 | 34 | 81 | 84 | 64 | 67 | 49 | 74 |
| DFO-9, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | A | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 97 | 80 | 86 | 90 | 26 | 42 | 60 | 36 | 73 | 69 | 23 | 37 | 65 | 70 | 72 | 36 | 34 | 53 | 31 | 0.3 | 7 | 60 | 64 | 37 | 43 | 15 | 87 |
| DFO-9, Low productivity, High DQA | R | G | R | R | R | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | G | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 99 | 73 | 92 | 93 | 0.4 | 45 | 69 | 31 | 57 | 55 | 21 | 25 | 55 | 60 | 59 | 32 | 29 | 38 | 32 | 0.2 | 0.7 | 55 | 54 | 25 | 30 | 4.2 | 92 |
| DFO-9, Low productivity, Low DQA | R | G | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | G | R | R | R | R |
| Δ Probability (highest-next highest) | 100 | 51 | 97 | 97 | 36 | 63 | 85 | 58 | 32 | 29 | 17 | 12 | 28 | 52 | 52 | 59 | 8 | 62 | 5 | 37 | 36 | 24 | 23 | 10 | 5 | 32 | 97 |
| All-3, High productivity, High DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 98 | 90 | 86 | 92 | 27 | 36 | 48 | 13 | 67 | 72 | 36 | 60 | 74 | 57 | 56 | 30 | 47 | 18 | 56 | 15 | 16 | 76 | 75 | 45 | 65 | 34 | 90 |
| All-3, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| Δ Probability (highest-next highest) | 99 | 84 | 91 | 95 | 2.5 | 55 | 65 | 36 | 74 | 63 | 12 | 42 | 66 | 69 | 68 | 50 | 25 | 41 | 36 | 10 | 9 | 63 | 62 | 24 | 51 | 10 | 94 |
| All-3, Low productivity, High DQA | R | G | R | R | R | A | R | A | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R | R |
| Δ Probability (highest-next highest) | 100 | 18 | 99 | 99 | 55 | 66 | 97 | 48 | 3.2 | 3.9 | 37 | 37 | 12 | 35 | 30 | 63 | 25 | 49 | 53 | 62 | 64 | 7 | 0.6 | 20 | 26 | 55 | 99 |
| All-3, Low productivity, Low DQA | R | A | R | R | R | A | R | A | R | R | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | R | R |
| Δ Probability (highest-next highest) | 100 | 11 | 99 | 99 | 73 | 53 | 98 | 24 | 32 | 33 | 57 | 57 | 17 | 8 | 1.6 | 48 | 49 | 26 | 66 | 61 | 58 | 35 | 29 | 47 | 51 | 72 | 99 |
| Equal weighting method | R | G | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | G | A | A | G | G | A | A | A | R |

| Metric | CUs | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 |
| Spawner abundance | A | A | R | R | A | A | A | A | A | A | A | G | G | G | G | G | G | G | G | G | G | R | R | A | A | G | G |
| Trend in spawners | A | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | A | A | A | A | A | R | R | A | A | G | G |
| Harvest rate | R | A | R | A | R | R | G | G | A | R | G | A | G | A | R | A | R | R | A | G | R | A | G | R | G | A | R |
| Distribution | R | R | A | R | A | R | G | A | G | G | R | A | A | G | A | R | R | A | R | R | G | G | A | G | R | R | A |
| All-9, High productivity, High DQA | A | A | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| Δ Probability (highest-next highest) | 30 | 9 | 82 | 80 | 3.8 | 32 | 73 | 62 | 65 | 34 | 32 | 60 | 41 | 35 | 70 | 72 | 68 | 43 | 39 | 15 | 13 | 42 | 47 | 73 | 73 | 41 | 37 |
| All-9, High productivity, Low DQA | R | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| Δ Probability (highest-next highest) | 1.7 | 23 | 90 | 89 | 27 | 57 | 61 | 38 | 43 | 2.6 | 0.6 | 73 | 64 | 60 | 53 | 56 | 68 | 65 | 63 | 46 | 44 | 64 | 67 | 59 | 58 | 8 | 3.0 |
| All-9, Low productivity, High DQA | R | R | R | R | R | R | A | R | A | R | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| Δ Probability (highest-next highest) | 31 | 54 | 98 | 98 | 57 | 77 | 25 | 0.8 | 1.1 | 35 | 34 | 38 | 55 | 56 | 3.3 | 6 | 32 | 56 | 58 | 63 | 63 | 89 | 90 | 28 | 29 | 32 | 34 |
| All-9, Low productivity, Low DQA | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | R | R | A | A |
| Δ Probability (highest-next highest) | 57 | 74 | 99 | 99 | 75 | 88 | 5.0 | 31 | 29 | 60 | 59 | 10 | 33 | 34 | 27 | 24 | 3.3 | 35 | 38 | 55 | 54 | 95 | 95 | 1.9 | 0.8 | 56 | 58 |
| DF0, High productivity, High DQA | A | A | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | R | R | A | A | G | G |
| Δ Probability (highest-next highest) | 41 | 12 | 86 | 88 | 19 | 20 | 68 | 67 | 69 | 50 | 39 | 57 | 33 | 27 | 73 | 72 | 63 | 32 | 39 | 9 | 5 | 55 | 59 | 68 | 72 | 32 | 39 |
| DFO-9, High productivity, Low DQA | A | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| Δ Probability (highest-next highest) | 4.9 | 24 | 93 | 94 | 18 | 51 | 66 | 42 | 47 | 17 | 3.4 | 73 | 63 | 59 | 58 | 54 | 70 | 62 | 66 | 46 | 34 | 75 | 78 | 66 | 59 | 9 | 0.7 |
| DFO-9, Low productivity, High DQA | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| Δ Probability (highest-next highest) | 9 | 35 | 97 | 97 | 30 | 62 | 55 | 34 | 33 | 0.5 | 4.0 | 59 | 55 | 55 | 37 | 33 | 52 | 58 | 60 | 39 | 35 | 83 | 83 | 53 | 50 | 12 | 7 |
| DFO-9, Low productivity, Low DQA | R | R | R | R | R | R | A | R | R | R | R | A | A | A | A | R | A | A | A | A | A | R | R | A | A | A | A |
| Δ Probability (highest-next highest) | 43 | 63 | 99 | 99 | 60 | 81 | 29 | 0.4 | 1.3 | 35 | 39 | 36 | 57 | 56 | 3.2 | 2.2 | 24 | 54 | 51 | 62 | 60 | 93 | 93 | 25 | 21 | 46 | 42 |
| All-3, High productivity, High DQA | A | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | G |
| Δ Probability (highest-next highest) | 22 | 25 | 94 | 97 | 8 | 49 | 69 | 65 | 65 | 49 | 22 | 73 | 52 | 53 | 68 | 49 | 72 | 52 | 69 | 42 | 13 | 77 | 76 | 69 | 71 | 11 | 23 |
| All-3, High productivity, Low DQA | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| Δ Probability (highest-next highest) | 3.2 | 47 | 96 | 98 | 17 | 66 | 73 | 51 | 50 | 28 | 3.1 | 69 | 66 | 66 | 55 | 29 | 63 | 66 | 73 | 59 | 36 | 85 | 85 | 73 | 60 | 34 | 2.3 |
| All-3, Low productivity, High DQA | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | R | R | A | A |
| Δ Probability (highest-next highest) | 54 | 74 | 99 | 99 | 76 | 87 | 17 | 40 | 34 | 64 | 64 | 10 | 28 | 34 | 28 | 22 | 12 | 41 | 45 | 57 | 57 | 98 | 98 | 1.1 | 1.8 | 51 | 54 |
| All-3, Low productivity, Low DQA | R | R | R | R | R | R | R | R | R | R | R | R | R | A | R | R | R | A | A | A | A | R | R | R | R | A | A |
| Δ Probability (highest-next highest) | 72 | 85 | 100 | 100 | 86 | 93 | 44 | 62 | 58 | 78 | 78 | 19 | 0.4 | 6 | 53 | 48 | 18 | 15 | 21 | 37 | 38 | 99 | 99 | 30 | 31 | 64 | 66 |
| Equal weighting method | R | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A |

**Table G2. Estimated status of 54 hypothetical CUs (R = red, A = amber, G = green) from model-averaged PWUs for each model (like in Table G1), and after incorporating uncertainty in the PWUs (with uncertainty).**

| Metric | CUs | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 | 22 | 23 | 24 | 25 | 26 | 27 |
| Spawner abundance | R | G | R | R | G | G | R | G | G | G | G | G | A | A | A | G | G | A | G | G | G | G | G | A | A | A | R |
| Trend in spawners | R | G | R | R | R | G | R | R | R | R | G | G | A | A | A | A | G | A | A | A | A | A | G | G | A | A | R |
| Harvest rate | R | G | R | G | R | R | G | G | R | G | G | R | A | A | G | A | A | G | G | A | G | G | A | A | R | A | A |
| Distribution | R | G | G | R | R | R | G | G | G | R | R | G | A | G | A | A | A | G | G | G | A | A | G | R | A | A | A |
| All-9, High productivity, High DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| All-9, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| All-9, Low productivity, High DQA | R | G | R | R | A | A | R | A | A | A | A | A | A | A | A | A | G | A | A | A | A | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | A | A | A | A | A | A | G | A | A | A | A | G | G | A | A | A | R |
| All-9, Low productivity, Low DQA | R | G | R | R | A | A | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R |
| DF0, High productivity, High DQA | R | G | R | R | A | A | R | G | A | A | G | G | G | A | A | A | G | G | A | G | G | G | G | G | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | G | A | A | G | G | G | A | A | A | G | G | A | G | G | G | G | G | A | A | R |
| DFO-9, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | A | G | A | A | G | G | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | A | G | A | A | G | G | A | A | R |
| DFO-9, Low productivity, High DQA | R | G | R | R | R | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | G | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | R | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| DFO-9, Low productivity, Low DQA | R | G | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | G | R | R | R | R |
| With Parameter Uncertainty | R | G | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | G | R | R | R | R |
| All-3, High productivity, High DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | G | G | G | G | A | A | A | R |
| All-3, High productivity, Low DQA | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | A | A | G | G | A | A | A | A | G | A | G | A | A | G | G | A | A | A | R |
| All-3, Low productivity, High DQA | R | G | R | R | A | A | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R | R |
| With Parameter Uncertainty | R | G | R | R | A | A | R | A | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | R | R |
| All-3, Low productivity, Low DQA | R | A | R | R | A | A | R | A | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | R | R | R | R |
| With Parameter Uncertainty | R | A | R | R | A | A | R | A | R | R | A | A | R | A | R | A | A | A | A | A | A | A | A | R | R | R | R |
| Equal weighting method | R | G | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | A | A | G | G | A | A | A | R |

| Metric | CUs | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | 28 | 29 | 30 | 31 | 32 | 33 | 34 | 35 | 36 | 37 | 38 | 39 | 40 | 41 | 42 | 43 | 44 | 45 | 46 | 47 | 48 | 49 | 50 | 51 | 52 | 53 | 54 |
| Spawner abundance | A | A | R | R | A | A | A | A | A | A | A | G | G | G | G | G | G | G | G | G | G | R | R | A | A | G | G |
| Trend in spawners | A | R | R | R | R | R | R | R | R | R | R | R | R | R | R | R | A | A | A | A | A | R | R | A | A | G | G |
| Harvest rate | R | A | R | A | R | R | G | G | A | R | G | A | G | A | R | A | R | R | A | G | R | A | G | R | G | A | R |
| Distribution | R | R | A | R | A | R | G | A | G | G | R | A | A | G | A | R | R | A | R | R | G | G | A | G | R | R | A |
| All-9, High productivity, High DQA | A | A | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| With Parameter Uncertainty | A | A | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| All-9, High productivity, Low DQA | R | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| With Parameter Uncertainty | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | G | G |
| All-9, Low productivity, High DQA | R | R | R | R | R | R | A | R | A | R | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| All-9, Low productivity, Low DQA | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | R | R | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | R | A | A | A | A | R | R | A | R | A | A |
| DF0, High productivity, High DQA | A | A | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | R | R | A | A | G | G |
| With Parameter Uncertainty | A | A | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | G | R | R | A | A | G | G |
| DFO-9, High productivity, Low DQA | A | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| With Parameter Uncertainty | A | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | G |
| DFO-9, Low productivity, High DQA | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| DFO-9, Low productivity, Low DQA | R | R | R | R | R | R | A | R | R | R | R | A | A | A | A | R | A | A | A | A | A | R | R | A | A | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | A | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | A | A | A | A |
| All-3, High productivity, High DQA | A | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | G |
| With Parameter Uncertainty | A | R | R | R | A | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | G |
| All-3, High productivity, Low DQA | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | A | A | A | A | R | A | A | A | A | A | A | A | A | A | A | R | R | A | A | A | A |
| All-3, Low productivity, High DQA | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | R | R | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | R | R | R | R | R | A | A | A | R | R | A | A | A | A | A | R | R | R | R | A | A |
| All-3, Low productivity, Low DQA | R | R | R | R | R | R | R | R | R | R | R | R | R | A | R | R | R | A | A | A | A | R | R | R | R | A | A |
| With Parameter Uncertainty | R | R | R | R | R | R | R | R | R | R | R | R | R | A | R | R | R | A | A | A | A | R | R | R | R | A | A |
| Equal weighting method | R | R | R | R | R | R | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A | A |