UNEP

CBD

Convention on Biological Diversity

Distr.
GENERAL

UNEP/CBD/ID/AHTEG/2015/1/INF/6
11 August 2015

ENGLISH ONLY

AD HOC TECHNICAL EXPERT GROUP
MEETING ON INDICATORS FOR THE
STRATEGIC PLAN FOR BIODIVERSITY
2011-2020
Geneva, Switzerland, 14-17 September 2015

ANALYTICAL OPTIONS FOR AICHI NATIONAL INDICATORS REPORTED TO THE CBD

Note by the Executive Secretary

1. The Executive Secretary is circulating herewith, for the information of participants in the meeting of the Ad Hoc Technical Expert Group on Indicators for the Strategic Plan for Biodiversity 2011-2020, a note on analytical options for national indicators for progress in the achievement of the Aichi Biodiversity Targets reported under the Convention. The present note complements document UNEP/CBD/ID/AHTEG/2015/1/2.

2. The fourth edition of the *Global Biodiversity Outlook* included an overview of progress towards the Aichi Biodiversity Targets made by Parties, based on information provided in the fifth national reports. The possibility of using this type of information to generate indicators of progress is discussed in document UNEP/CBD/ID/AHTEG/2015/1/2 and refers to an initial assessment which demonstrates that it is feasible to develop a statistically sound indicator based on such information (see paragraphs 20 and 26-28 of UNEP/CBD/ID/AHTEG/2015/1/2). The present note contains that assessment.

3. The assessment was prepared by the United Nations Environment Programme World Conservation Monitoring Centre in consultation with the Secretariat of the Convention on Biological Diversity, and is presented in the form and language in which it was received by the Secretariat.

Analytical Options for Aichi National Indicators Reported to the CBD

A report to: (i) assess the feasibility of creating a single metric of relative national progress for each Aichi Target based on national reports, (ii) test the significance of differences in progress between Targets and across years for each Target, and (iii) determine whether there are any socioeconomic factors consistently associated with progress.

UNEP WCMC

Analytical options for Aichi National Indicators reported to

the CBD

Authors
Derek Tittensor

In collaboration with
The Secretariat of the Convention on Biological Diversity

Published
August 2015

Citation
UNEP-WCMC (2015). Analytical Options for Aichi National Indicators Reported to the CBD. Report, 39 pages.

Copyright 2015 United Nations Environment Programme

The United Nations Environment Programme World Conservation Monitoring Centre (UNEP-WCMC) is the specialist biodiversity assessment centre of the United Nations Environment Programme (UNEP), the world's foremost intergovernmental environmental organization. The Centre has been in operation for over 30 years, combining scientific research with practical policy advice.
This publication may be reproduced for educational or non-profit purposes without special permission, provided acknowledgement to the source is made. Reuse of any figures is subject to permission from the original rights holders. No use of this publication may be made for resale or any other commercial purpose without permission in writing from UNEP. Applications for permission, with a statement of purpose and extent of reproduction, should be sent to UNEP-WCMC, 219 Huntingdon Road, Cambridge, CB3 0DL, UK.

The contents of this report do not necessarily reflect the views or policies of UNEP, contributory organizations or editors. The designations employed and the presentations of material in this report do not imply the expression of any opinion whatsoever on the part of UNEP or contributory organizations, editors or publishers concerning the legal status of any country, territory, city area or its authorities, or concerning the delimitation of its frontiers or boundaries or the designation of its name, frontiers or boundaries. The mention of a commercial entity or product in this publication does not imply endorsement by UNEP.

UNEP WCMC

UNEP World Conservation Monitoring Centre (UNEP-WCMC)
219 Huntingdon Road,
Cambridge CB3 0DL, UK
Tel: +44 1223 277314
www.unep-wcmc.org

UNEP promotes environmentally sound practices globally and in its own activities. Our distribution policy aims to reduce UNEP's carbon footprint

Contents

Executive Summary

This report examines national level progress towards the 2020 Aichi Targets as derived from 146 5th national reports to the CBD that were submitted in 2014/2015. This represents 74% of the 196 Parties to the CBD. A subset of these data were visualised in a figure on page 131 of the Global Biodiversity Outlook 4 report, but were not analysed in detail as it was unclear whether there was an appropriate way to assess relative progress for this type of data.

This report presents the results of a preliminary analysis to investigate the range of approaches for assessing the results of the 5th national reports. In particular, it aims to answer the following questions:

(1) Is it feasible and statistically defensible, given the ordinal (ordered category) nature of the data, to generate a single metric of progress towards each Target that summarizes the results of each national report, and if so, what should this metric be?

(2) If such a metric can be developed, what is the best way of using it assessing the significance of (a) differences in aggregated progress between Targets, and (b) temporal differences (i.e. through time) in aggregated progress for each Target?

(3) What factors are associated or correlated with national progress towards each Target? Can we develop a predictive model identifying those factors?

We note that differences in national targets (different levels of ambition) and different approaches to assessing progress (outcome vs. process) mean that the approach is not suitable for comparison of progress made by one Party to another; therefore, we only compare Targets. Furthermore, there are a number of additional complicating factors for an analysis of the 5th national reports, such as missing data, variable levels of confidence in each national report progress assessment, and non-independence of data. Currently, data are only available for a single set of national reports, but this report looks ahead to applicable methods when multi-year data are available.

We find that it is possible to generate a single metric of progress (Question 1) towards each Target that summarizes the results of each national report. There are two ways of doing this: treating the data as ordinal (ordered categories), and using a mode or a median (along with the interquartile range as a measure of variation), or treating the data as interval (numeric) and summarizing the data with the mean and standard deviation. Differences in progress between Targets (Question 2) or through time can be assessed for significance using non-parametric tests for ordinal data, or parametric tests for interval data. The factors associated with progress (Question 3) can be modelled for both ordinal and interval data using regression models, though with different sets of assumptions.

The distinction in terms of treatment of data is critical. Treating the data as ordinal is statistically more rigorous, but potentially harder to interpret and less powerful. Treating the data as interval (numeric) is less statistically rigorous, but a frequently used approach across the scientific literature (e.g. for the IUCN Red List Index), and may be more powerful and have easier interpretation.

It should be noted that this is a **proof of concept** report on **methodology**; that is, it aims to assess whether answering the questions above is feasible and to test approaches with preliminary or simulated data, rather than to perform a complete and comprehensive analysis on all national-level target data. A follow-on project could have the aims of: (a) conducting the complete analysis; (b) generalising the approach and computer code to ensure that it can be applied as easily as possible on an annual basis to assess progress, and; (c) producing a peer-reviewed paper outlining the methodology and analysing the results.

1. Introduction

The Aichi Biodiversity Targets are part of the Strategic Plan for Biodiversity 2011-2020, adopted at the tenth meeting of the Conference of the Parties to the Convention on Biodiversity (CBD COP 10) held in Nagoya (Japan) in October 2010. The aim is to provide an overarching framework for biodiversity-related conventions, biodiversity management and policy development. In 2014, global progress towards the Aichi Targets was assessed and published in the Global Biodiversity Outlook 4 (GBO-4) report and an associated peer-reviewed paper (CBD 2014; Tittensor *et al.* 2014).

In addition to progress at the global scale, progress towards Aichi Targets based on national reporting has been compiled from the 5th national reports to the CBD. Progress was graded on a 5 point scale based on both national self-reporting (where possible), and on an internal CBD assessment of progress when not self-reported. The five categories of progress (**moving away from target, no progress, progress but at an insufficient rate, on track to meet the target, on track to exceed the target**) have been determined for >100 countries globally. A proportion of these data (for 60 countries) were visualised on page 131 of the GBO-4 report. However, these data were not summarized in a single metric for each Target, nor compared for statistically significant differences in progress based on the national reports. This UNEP-WCMC report presents research into the most suitable approach to summarizing and analysing these data. It shows example analyses intended to elucidate the differences between and highlight strengths and weaknesses of potential approaches. Furthermore, based on the findings herein, we make recommendations for:

(i) an appropriate and defensible approach to summarising the national progress reports into a single metric for each Target [Section 2];
(ii) how to assess the statistical significance of differences in aggregated progress between Targets, and differences in aggregated progress across multiple years for each individual Target [Section 3];
(iii) how to model the factors associated with aggregated national progress towards each Target [Section 4].

Due to differences in national targets in terms of levels of ambition and approaches to assessing progress, such as assessing outcomes versus processes, these data are not suitable for comparing progress made by one Party to another. We therefore focus on assessing differences in aggregated progress between Targets and within Targets over time. The analyses on this report are based on preliminary data provided by the CBD Secretariat to UNEP-WCMC in the Spring of 2015; these results may not reflect the final analysis to be conducted with the full data set of all compiled national reports, and should be considered in that light. The subset of Targets analysed herein (five Targets, being 1, 5, 10, 11 and 17) are those that were suggested by the CBD Secretariat, and provided as a spreadsheet with assessments for 123 countries. Data on socioeconomic variables were also provided. Targets 1 and 17 have been separately identified as having substantial gaps in terms of global indicators (Chenery *et al.* 2015).

1.1 Concepts and potential analytical approaches

Data come in many forms, from categorical (e.g. eye colour) to continuous (e.g. height or weight). In statistical terms, *measurement theory* is concerned with understanding the most appropriate ways to deal with different types of data. Stevens (1946) classified data into four types: **nominal, ordinal, interval**, and **ratio**. Nominal data are categorical but with no specific ordering (for example, eye colour). Ordinal data are also categorical, but have an explicit ordering (e.g. a survey for which the responses are 'did not like', 'neutral', 'liked'). Interval data have an explicit degree of difference and are

often (but not always) numerical (e.g. the Celsius temperature scale). Ratio data are continuous with a meaningful (unique) zero value (e.g. the Kelvin temperature scale).In his original classification, Stevens (1946) laid out the types of statistics that can be used and analyses that can be performed on each of these types of data. For example, he specified that the mean should not be calculated on ordinal data, and the mode or the median used as a measure of central tendency.

However, within the social and medical sciences (where these types of data are more common, such as from patient survey results), treating ordinal data as numeric and calculating the mean and other statistics is a common approach. Some literature surveys suggest that the majority of studies use this approach (Harwell & Gatti 2001). There is debate in the scientific literature about the appropriateness of this (e.g. Knapp 1990, Jamieson 2004, Norman 2010). Social scientists frequently use the scale of measurement known as a Likert scale. A typical Likert scale will have 5, 7, or 10 responses, such as 'strongly disagree', 'disagree', 'neutral', 'agree', and 'strongly agree'. These data are often copied to a numerical scale (e.g. 1-5) and then analysed using parametric techniques for interval data. A key issue is that there is no guarantee that the distance between two consecutive classes will be the same – i.e. that 'strongly disagree' (1) to 'disagree' (2) will be the same distance as 'disagree' (2) to 'neutral' (3). The question then becomes whether such a scale is justified, and whether statistical tests are robust to this assumption.

In the literature, the empirical evidence seems to suggest that many statistical tests are indeed robust to treating ordinal data as interval (summarised in Norman 2010). The use of parametric statistics on Likert data (i.e. converting to numeric) is consistent with the empirical literature dating back nearly 80 years, and if such approaches were rejected then a large proportion of research on education, health status, and quality of life assessment would have to be discarded (Norman 2010). However, it should be noted that the debate continues and there are still many orthodox statisticians who would view this type of approach as inappropriate.

Given these ongoing issues, a number of methodologies have been used within the scientific literature to analyse this type of data. Three key approaches are:

1. Use methods specifically designed for ordinal data (such as ordinal regression). Such an approach is statistically appropriate but may lack power compared to option (2) below.

2. Treat the data as interval (i.e. ignore that they are ordinal and assign values to each category), and proceed with the analysis. This may violate statistical assumptions, but such an approach is very commonly used (Harwell & Gatti 2001).

3. Use a modelling approach to add an extra step in the data analysis to convert the ordinal data to interval data (e.g. Item Response Theory) and then use interval data methods (Harwell & Gatti 2001).

With all of these approaches, there is a trade-off in terms of simplicity for repeated application, minimizing bias, statistical robustness, and acceptance within the literature. A useful approach in this situation is to test multiple methodologies, compare how well they work, and how frequently the results agree. Of the approaches above, (3) is a complex subject about which whole books have been written (e.g. Embretson 2000), and applying it would consist of a lengthy research project; we therefore consider it outside the scope of this analysis. In the remainder of this report, we therefore use the other two approaches: (1) analysing the data as if it were interval, which is supported in terms of robustness by Norman (2010), and (2) analysing the data as ordinal, which is a more statistically rigorous approach. The trade-off here is that option (1) opens up a wider range of statistical tools and potentially greater power to detect differences, whereas option (2) has more theoretical (though not necessarily empirical) support, but may lack power in terms of detecting significance. We assess these trade-offs below for example data.

An additional consideration is that ordinal scales can be symmetric (equal numbers of negative and positive responses) or asymmetric. The National Indicators data are asymmetric, with more positive responses; this is likely to bias the results upwards (positively). In combination with self-reporting, which also produces a positive bias, this could be an important effect, and needs to be highlighted in the caveats section of any analysis or paper.

The data on progress towards the national 'Aichi Biodiversity Targets' are **ordinal** – the measurements are in categories, but they have an order, from worst ('moving away from') to best ('on track to exceed'). The aggregated data for the preliminary analysis are summarised in Figure 1. These are provisional sample data, not final. The mean number of countries reporting results for each of these five Targets (i.e. for which there were data) was 107 (out of 123 on the spreadsheet). However, different countries missed data for each Target. In this instance, missing data may represent an issue for some statistical approaches if they assume a balanced design (i.e. the same countries reporting results across all Targets). In this report, we leave missing data out from analyses, and where necessary perform pairwise comparisons when some but not all countries have reported results.

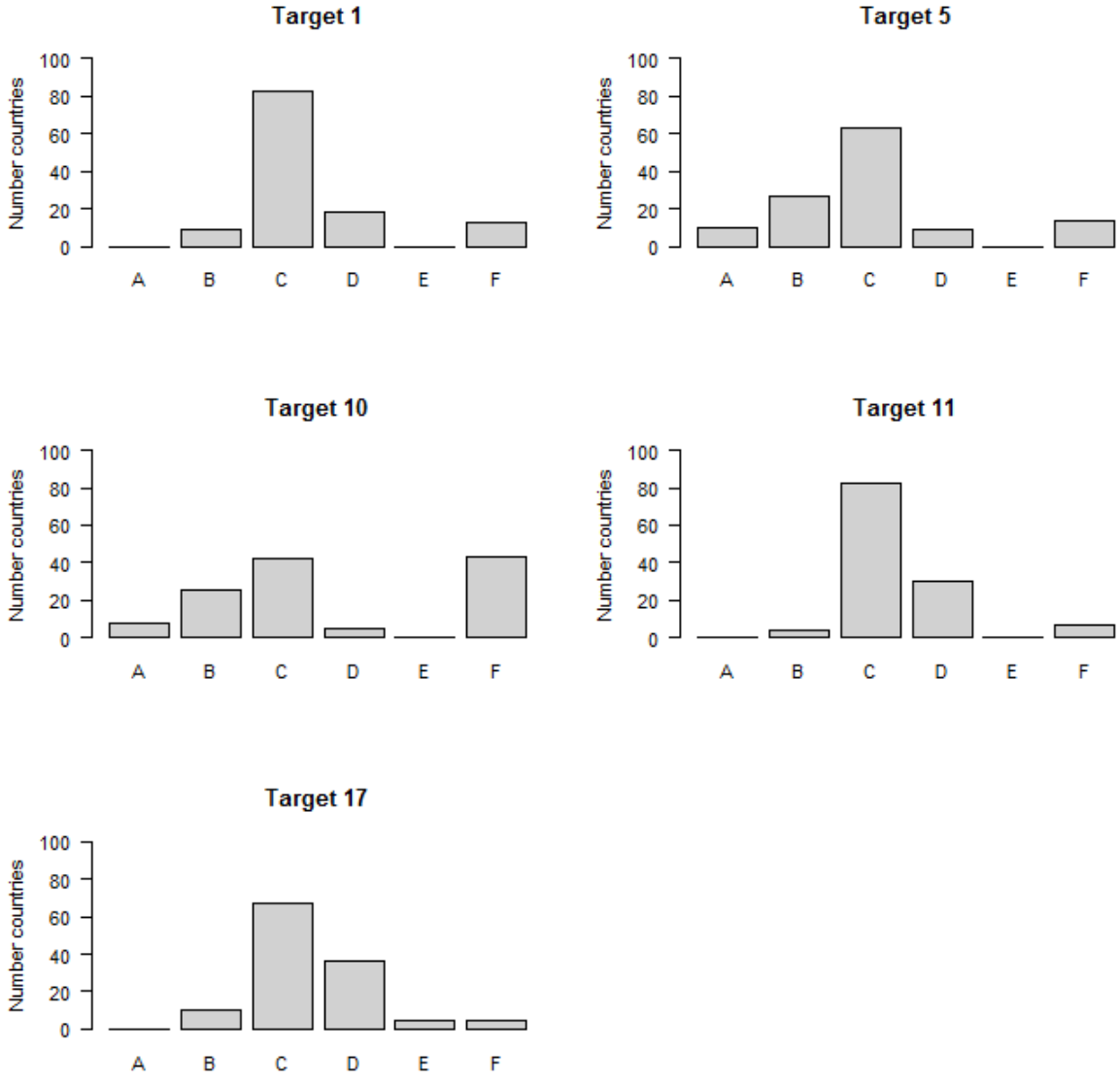

Figure 1: *Bar plot of preliminary data. The six categories under which countries can fall are: A (moving away from Target); B (no progress); C (progress but at insufficient rate; D (on track to meet the Target); E (on track to exceed the target); and F (no information).*

The example data also include an assessment of confidence in the reported value. The confidence values are on a three-point scale from 1 (lowest) to 3 (highest). Despite the numeric values, this is again an ordinal scale, and the same issues for conversion from ordinal to interval data also apply. Furthermore, many non-parametric tests do not specifically or easily work with weighted data. Therefore, if the data are treated as ordinal, typically either the confidence values will have to be ignored, or a threshold set (e.g. only data with a confidence level of 3 used) for incorporating data, which will reduce the power of the test as only a subset of the data are used. When treating the data as interval, however, tests can include weighted data. We demonstrate the use of weighted data in the sections 3 and 4 of the report below.

1.2 Recommendations

1. It is recommended that the full analysis makes a choice between two common approaches to optimize the trade-off between simplicity, interpretability, and robustness. These approaches are to assess and analyse the data on an ordinal scale (1 above) and on an interval scale (2 above). The final approach for the full analysis should then be chosen after considering the results from these preliminary analyses and weighing the trade-offs for the intended study.
2. It is important to note in any analyses, reports, or peer-reviewed papers that several factors, namely the asymmetric scale and self-reporting, may tend to bias the results positively ('optimistically').

2. Combining national progress indicators into a single metric

The purpose of this section is to explore options for generating a single metric of progress towards each Target based on a summary statistic of the data across all countries. We follow the approach suggested in the Introduction, which is to perform two separate sets of analyses, the first of which is to treat the data as ordinal and the second as interval.

2.1 Analyzing as ordinal data (no arbitrary scale)

If the data are treated as ordinal, then the approaches to summarising them and compressing them into a single metric are more limited. Ordinal data are often summarised in a contingency table, akin to a numeric version of Figure 1. If a single metric is required, the mean is not considered to be appropriate for ordinal data, so the mode (most frequent value) and the median (the number separating the higher half of the data from the lower half of the data; can be calculated because data are ordinal not categorical) are used instead.

In the examples shown below, the mode and the median are both the same, though this may not always be the case.

Table 1: *Mode and median values for the preliminary data. Missing data are excluded from calculations.*

	Mode	**Median**
Target 1	Progress but at an insufficient rate	Progress but at an insufficient rate
Target 5	Progress but at an insufficient rate	Progress but at an insufficient rate
Target 10	Progress but at an insufficient rate	Progress but at an insufficient rate
Target 11	Progress but at an insufficient rate	Progress but at an insufficient rate
Target 17	Progress but at an insufficient rate	Progress but at an insufficient rate

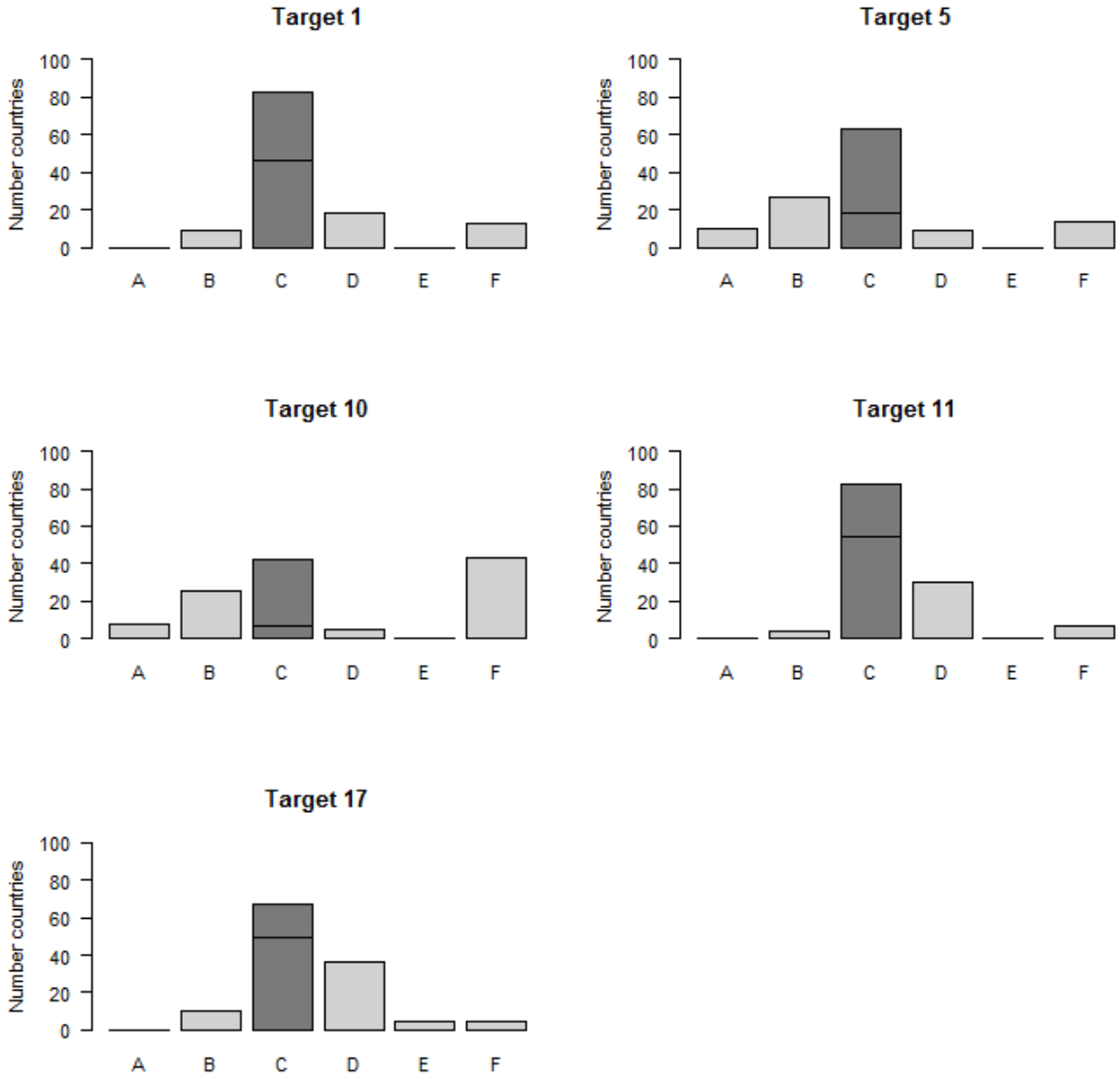

Figure 2: *Bar plot of preliminary data, showing mode (dark grey) and median (horizontal line). The six categories under which countries can fall are: A (moving away from Target); B (no progress); C (progress but at insufficient rate; D (on track to meet the Target); E (on track to exceed the target); and F (no information).*

The advantage of this approach are that there is no need to apply an arbitrary scale, and hence it may be more defensible and give reviewers less reason to take issue with the approach.

The disadvantages are (i) that it is harder to interpret, as most people are more familiar with the mean than the median or the mode; (ii) that for fairly coarse data such as these, the median and the mode may frequently be the same, as they are here, and (iii) that it is quite insensitive to small changes (i.e. numerous countries may need to change category before the median or the mode changes), and so might require substantial changes to reliably demonstrate progress.

In terms of a single metric of the variation or spread of the results, a common metric is the interquartile range. This is the distance from the 25th to the 75th percentile of the data, given in terms of number of

categories spanned (where zero means all data between the 25th and 75th percentile are within the same category). The interquartile ranges for the example data are shown in Table 2.

Table 2: *Interquartile range for all example Target data*

	Interquartile range
Target 1	0
Target 5	1
Target 10	1
Target 11	1
Target 17	1

Thus, from the table above, all countries between the 25th and the 75th percentile for Target 1 are in the same category, whereas for the other Targets they have a spread of one (i.e. are in two categories), meaning that they are less tightly clustered. In addition, if a single metric is not essential the 25th and 75th percentiles can be reported in addition to the mean and the interquartile range.

In terms of weighting data by its confidence value, while a weighted mode does not exist, technically a weighted median could be calculated (being the value at which 50% of the weight lies above and 50% of the weight lies below). However, in the example data, the confidence values are not absolute measures of variance (e.g. standard errors) but are assessments graded on an (again arbitrary) numerical scale. Thus we recommend against using this approach, as applying an arbitrary scale to confidence values to calculate a weighted median while not applying an arbitrary scale to categorical responses does not make logical sense. Instead, we recommend choosing a threshold (e.g. confidence level '3' only) and calculating median values using only these data, as a robustness check on the results when using all data. In this case, when using such an approach, the results are unchanged and hence the same as those in Table 1.

2.2 Analyzing as interval data (converting to arbitrary numeric scale)

In order to use a broader range of approaches to summarizing the data, they need to be converted from ordinal (i.e. semi-quantitative) to fully quantitative (an interval scale), by assigning a numeric value to each category. A common approach is assigning sequential integer values to each category (e.g. moving away = 1, no progress = 2, moving towards = 3, on track = 4, on track to exceed = 5), particularly for Likert scale data. However, an infinite variety of mappings exist which retain the ordered characteristic of the data, including non-integer values (e.g. 1,3,6,10,15 or -2, -1, 4, 5.5, 6). The sequential integer mapping approach appears to be used most frequently because it is (a) straightforward, (b) relatively easy to interpret, and (c) invokes the assumption that the categories are equally spaced (i.e. equal distance between 'strongly disagree' and 'disagree' as that between 'disagree' and 'neutral'). There is no reason why this need necessarily be the case; however, Norman (2010) suggests that most tests (though not explicitly summary metrics) are robust to this process. Given the range of options, from hereon in we use two example approaches to demonstrate the effect that they have on the analyses:

Approach 1: Equal spacing between categories

Assume that the spacing between each category is the same (i.e. treat it as a balanced scale). Although we could choose any reference point, if we set 'no progress' to have the value zero, for the sake of interpretability, then the numeric values assigned would be as follows:

Table 3: *Approach 1 for transforming data from ordinal to interval scale*

Progress	Moving away from	No progress	Moving towards	On track to meet	On track to exceed	No data
Numeric value assigned	-1	0	1	2	3	-

Approach 2: Unequal spacing between positive categories

It seems reasonable that the spacing between 'moving away from' and 'no change' is the same as that between 'no change' and 'moving towards'. Next make the assumption that the difference between 'moving towards' and 'on track to meet' is half that between 'no change' and 'moving towards', and similarly for 'on track to meet' and 'on track to exceed'. While this approach would be less common in the literature, and may make interpretation more difficult, it has the advantage of somewhat mitigating the positive bias, because the uppermost category is only two times the magnitude of the lowest category, rather than three times in Approach 1. The numeric values assigned would be:

Table 4: *Approach 2 for transforming data from ordinal to interval scale*

Progress	Moving away from	No change	Moving towards	On track to meet	On track to exceed	No data
Numeric value assigned	-2	0	2	3	4	-

Once the data are transformed, mean values and measures of variability can be calculated. These are shown in Table 5 for each approach.

Table 5: *Mean and (standard deviation) for the different approaches too conversion from ordinal to interval scale. Missing data are excluded from all calculations.*

	Approach 1	Approach 2
Target 1	1.08 (0.49)	2.0 (0.70)
Target 5	0.65 (0.76)	1.22 (1.40)
Target 10	0.55 (0.76)	1.03 (1.42)
Target 11	1.30 (0.60)	0.44 (0.60)
Target 17	1.31 (0.69)	0.57 (0.88)

From the table, firstly it is clear that relative progress can be compared between Targets by examining their numerical ranking in terms of their mean value. Secondly, the order of this ranking does not change between the two approaches. However, the relative position between categories can change. For example, in Approach 1, the Target 17 mean value (1.08) is between 'moving towards and 'on track to meet', whereas in Approach 2, it is exactly 'moving towards'. This demonstrates the effect of shrinking

the distance between positive categories, which acts in some extent to offset the bias of having more positive categories than negative – though the magnitude of this effect relative to the bias is unknown.

Although the effect above is small, an extreme example demonstrates an additional issue. If 'moving away from' from Approach 2 is set to -20 (instead of -2), then the mean value becomes -0.43 (instead of 1.22), and has in fact moved categories from being between 'no change' and 'moving towards' to between 'moving away from' and 'no change'.

Thus although relative orderings of means are preserved independently of the weighting scheme, the category in which the mean value resides is not. Although the previous example might seem extreme, this effect could also occur in some instances of particular distributions of countries within categories. For example, if there are many more cases of 'moving away from' the target than the positive categories, changing the value associated with 'moving away from' may change the category of the mean.

Relative to the ordinal data analysis above, the mean value appears more sensitive (i.e. no Targets have a tied rank), and the standard deviation provides a more interpretable measure of the spread of the data. The standard deviation also gives an idea of the relative differences between Targets, and the next section of this report will explore the statistical significance of such differences. The interval approach described here also opens up a wider range of statistical approaches towards testing differences between Targets, which some suggest have greater power to detect such differences (though see Knapp 1990).

Thus the mean values can be compared between Targets and time intervals, and parametric statistical approaches used to assess the significance of change. However, using different approaches for translating to an interval scale results in different values, so the interpretation is complicated. Nonetheless, in both cases, a value of less than zero indicates moving away from target, zero indicates no change, and positive indicates moving towards target, with 2 (Approach 1) and 3 (Approach 2) respectively denoting being on track to meet a particular target.

A weighted mean based on the confidence values in the spreadsheet can be calculated, but since the confidence values are again on an arbitrary scale (from 1 [lowest] to 3 [highest]), these have to be converted into appropriate weights for the data. Making the assumption that the confidence values are on a balanced, equal-interval scale, we set the weights to be [confidence value / 3], giving a range from 0.33 to 1. This will down-weight a data point with a confidence value of '1' to only contribute 33% of the weight of data with a confidence value of '3'. Such a scale makes most intuitive sense, unless the three confidence values are extremely non-linear in terms of their relationship with one another. Here, applying the weights and recalculating the mean gives results similar to those in Table 6 (all values within 0.1 of unweighted mean values). Calculating the weighted standard deviation is similarly straightforward.

2.3 Recommendations

1. The mean is more sensitive to differences, and more straightforward to understand, than the median or the mode. Thus treating the data as interval rather than ordinal provides a more easily understood metric.

2. However, as explained earlier, treating the data as interval, while common practice within the literature, has been criticized by some in the statistical community. However, the numerous, likely majority, of empirical studies in the social and medical scientific literature using this approach (e.g. Vickers 1999; Norman 2010) can be cited and invoked as support for this methodology. Furthermore, indices in the biological sciences such as the Red List Index have also used this approach (e.g. Butchart *et al.* 2004).

3. Given the two points above, we recommend using an interval scale, selecting an appropriate scale to convert from ordinal to interval data, and presenting the results of the analysis using the mean and standard deviation.

4. However, if the above approach is taken, a strong justification needs to be made for the numerical value assigned to each category. Both of the approaches used above seem reasonable, but the final choice will be for those who have good knowledge of the assessment process. Applying an equal interval scale would likely be more interpretable and is more frequently used in the literature, but using the second approach above may reduce positive bias and more accurately reflect the distances between categories in the unknown (latent) progress variable. Given the choice, if appropriate justification can be provided, we would recommend applying an equal interval scale, as is done for the IUCN Red List (Butchart *et al*. 2004).

5. We also recommend that while only a single conversion scale is presented in the main report or paper, that calculations on an ordinal scale are conducted as a test of robustness, and in case reviewers ask for the data to be treated as ordinal.

6. Notwithstanding the above, it also does not make sense to combine ordinal and interval analyses for the same report or paper: the data should be treated as one or the other. Thus if the data are considered interval for the presentation of summary metrics, they should also be considered interval for the purposes of statistical tests of differences between Targets and years within Targets, though approaches using ordinal testing methodology (outlined in the next section) can also be presented as robustness checks.

7. We recommend that, if they are considered to be roughly evenly spaced, the three confidence levels provided are treated as a linear scale, with weights equal to [confidence value / 3]. From this, it is easy to calculate a weighted mean and weighted standard deviation, which through comparison to the unweighted mean can give insight into what effect the confidence in each national report assessment is having.

3. Assessing progress within and between Targets

3.1 Concepts

A key element of an analysis of National Indicators is the quantification of the statistical significance of any differences in progress between Targets (CBD 2014), and the significance of any progress over time within a single Target – i.e., what is the strength of evidence for progress towards Targets over time?

The approach used to answer these questions again depends on whether the data are considered to be on an ordinal or interval scale. Furthermore, as explained below, it depends on the independence of the data points, both over time and between targets. Finally, it depends on whether the data are to be 'weighted' based on the relative confidence in them. Below we test analyses of differences between Targets and changes in Targets over time. For the latter, we have generated 'simulated' data as no temporal data were included in the preliminary data set: these simulated data are used to assess the applicability and sensitivity of the approaches, not to understand or predict changes over time.

When considering the difference between progress towards Targets (e.g. between Target 1 and Target 5) for ordinal data, the approach used depends on whether we consider the data to be independent or paired. Independence in this instance would mean that the value of a country's progress towards Target 1 is unrelated with its progress towards Target 5, for example. Paired data would mean that progress towards Targets is correlated within countries. It seems reasonable to assume that responses for a single country across multiple Targets are relatively independent, so we continue with this assumption in mind. However, this assumption may not hold given the synergies among Targets. Were they not considered to be independent, than the same approaches as in the section on progress within Targets (3.3) would be used.

For the whole of this section, when treating data as interval, we use Approach 1 from section 2.2 to assign numeric values to each category.

3.2 Assessing progress between Targets

Ordinal scale

To detect differences between pairs of Targets, the Mann-Whitney *U*-test can be used (McCrum-Gardner 2008). In Table 6 below, we present the results of the Mann-Whitney *U*-test for the difference between an example pair of Targets. Although in this instance the median of Target 5 is equal to the median of Target 1, there is a significant difference due to the distribution of category scores.

Table 6: *Results of Mann-Whitney U-test for differences between Targets 1 and 5.*

Comparison	**Target 1 to Target 5**
Result	There is a significant difference in the distribution of national progress categories between Targets 1 and 5 ($p < 0.0001$).

It is important to note that the Mann-Whitney test can only be used to compare differences between a pair of Targets, and that if multiple comparisons are carried out, the P-values would have to be adjusted to reflect this (i.e. to make the tests more conservative). Therefore, if comparisons are to be made between more than two Targets, the Kruskal-Wallis test should be used. If the results of the Kruskal-Wallis test show a significant difference, then post-hoc tests corrected for multiple pairwise comparisons can be used to identify which Targets are showing significant differences (Sachs 1997, Pohlert 2015). Here we use the Chi-square approach due to the number of ties present in the data. Note that it is only legitimate to proceed to pairwise comparisons with the post-hoc correction if a global Kruskal-Wallis test is significant ($p < 0.0001$, not shown). Table 7 shows the results from the Kruskal-Wallis test post-hoc comparison results.

Table 7: *Results from Kruskal-Wallis test, p-values. Significant differences are highlighted in bold. Post-hoc comparison conducted using Chi-squared correction for multiple comparisons.*

	Target 1	**Target 5**	**Target 10**	**Target 11**
Target 5	**< 0.0001**			
Target 10	**0.002**	0.96		
Target 11	0.78	**0.002**	0.069	
Target 17	0.36	**0.022**	0.27	0.96

Thus Target 1 is showing a significant difference from Target 5 and Target 10 in terms of progress (Target 1 showing lesser progress with a median category of 'No progress'), and Target 5 is showing a difference in terms of progress from Targets 11 and 17 (Target 5 showing greater progress, with a median value of 'Moving towards'). All other differences between Targets are non-significant; i.e. the evidence is insufficient to demonstrate differences in progress.

Although some work has been conducted on extending these non-parametric test to work with weighted data (Xie & Priebe 2002; Lumley & Scott 2013), and the application for a single comparison is relatively straightforward, correcting post-hoc tests for multiple comparisons is not straightforward and not yet broadly implemented (and would require some considerable time and expertise to do so). Therefore, we recommend using the same approach as for calculating the median, and re-running tests with only data of a certain confidence level or above. This will have the effect of reducing the power of the test as only a subset of the data will be included, though depending on the distribution of scores and confidence levels it may cause a result to either increase or decrease in terms of significance.

Interval scale

Again considering the data to be independent, a *t*-test can be used to determine the difference in progress between a pair of Targets (McCrum-Gardner 2008). The results are shown in Table 8 below for an example comparison:

Table 8: *Results of t-test for difference between Targets.*

Comparison	**Target 1 to Target 5**
Result	*Mean* value of Target 5 is significantly larger than *mean* value of Target 1 ($p < 0.0001$)

It is important to note that the t- test can only be used to compare differences between a pair of Targets, and that if multiple comparisons are carried out, the p-values would have to be adjusted to reflect this (i.e. to make the tests more conservative). Therefore, if comparisons are to be made between more than two Targets, one-way ANOVA (or, equivalently, linear regression) should be used. We use linear regression because the data may be unbalanced which could present problems for ANOVA. The approach that we use to correct p-values for multiple comparisons is Tukey's Honest Significant Difference (Miller 1981, Hothorn *et al.* 2008). Table 9 shows the results of this comparison.

Table 9: *Results from linear regression, p-values. Significant differences are highlighted in bold. Post-hoc comparison conducted using Tukey's Honest Significant Difference comparison.*

	Target 1	**Target 5**	**Target 10**	**Target 11**
Target 5	**< 0.0001**			
Target 10	**< 0.0001**	0.82		
Target 11	0.46	**<0.0001**	**0.0047**	
Target 17	0.068	**0.0005**	0.066	0.87

Here it is important to note that the p-values are smaller than those using the Kruskal-Wallis test for the ordinal data, and that in fact this results in an additional comparison being significant (between Targets 10 and 11). Thus in contrast to the suggestion of Knapp (1990), at least in this instance it does indeed appear that the parametric tests used when treating the data as interval are more powerful at detecting differences than the non-parametric tests when treating the data as ordinal.

For weighting data by confidence level while treating the, as interval, the same approach as in Section 2 is used (dividing numeric values of confidence by 3 to rescale them from 0.33 to 1). A *t*-test with weighted data (Lumley 2004) shows no change from the results in Table 8. The results of the weighted linear regression (Lumley 2004) are given in Table 10.

Table 10: *Results from weighted linear regression, p-values. Significant differences are highlighted in bold. Post-hoc comparison conducted using Tukey's Honest Significant Difference comparison.*

	Target 1	**Target 5**	**Target 10**	**Target 11**
Target 5	**< 0.0001**			
Target 10	**0.0003**	0.81		
Target 11	0.38	**<0.0001**	**0.0008**	
Target 17	**0.01**	**0.0002**	0.0511	0.60

The p-values in Table 10 have changed from those in Table 9, in some cases quite dramatically (e.g. Target 17 vs Target 1). Thus it is possible in the full analysis that using weighted data based on confidence could move some results from being significant to non-significant, and vice versa.

3.3 Assessing annual progress within Targets

The preliminary example data provided to UNEP-WCMC by the Secretariat of the Convention on Biological Diversity did not include multi-year data for any single Target. Thus we generated 'simulated' example data for Target 1. We generate example values for each country in '2016' and '2017'. These data are correlated with the 2014/2015 values with a coefficient of 0.4-0.55, as we estimate that the correlation between years for countries could be substantial.

When comparing progress for a Target, the data are considered non-independent since they are repeated measures from the same countries. Thus for the purposes of the statistical tests below, the data are considered paired.

Ordinal scale

When assessing annual progress within Targets, for a comparison between just two years the Wilcoxon signed rank test should be used (McCrum-Gardner 2008). Table 11 shows the result for an example comparison

Table 11: *Results of Wilcoxon signed-rank test for difference in progress between 2014/2015 and 2016 for Target 1.*

Comparison	**Target 1 (2014/2015) to Target 1(2015)**
Result	No difference in distribution of national progress categories between 2014/2015 and 2016 data (p = 0.91)

When multiple years are to be compared within a single Target, the appropriate test to be used is a Friedman's test (McCrum-Gardner 2008). We demonstrate this analysis approach using comparisons between 2014/2015, 2016 and 2017 with the results shown in Table 12 below, and using the same Chi-square approach to correct for multiple comparisons. Note that it is only legitimate to proceed to pairwise comparisons with the post-hoc correction if a global Friedman's test is significant ($p < 0.0001$, not shown).

Table 12: *Results from Friedman's test, p-values. Significant differences are highlighted in bold. Post-hoc comparison conducted using Chi-squared correction for multiple comparisons.*

	2014/2015 progress	**2016 progress**
2016 progress	1	
2017 progress	0.25	0.32

This test shows no significant differences between years in this instance.

In the same manner as assessing progress between Targets, correcting post-hoc tests for multiple comparisons is not straightforward. Therefore, if the use of the confidence data is required, we recommend using the same approach as for calculating the median, and re-running tests with only data of a certain confidence level or above. However, note that this will reduce the power of the test (because fewer data will be included), though results may become more or less significant depending on the distribution of scores and confidence levels.

Interval scale

On the interval scale, if only two years are being compared, the appropriate approach is the paired samples *t*-test (McCrum-Gardner 2008). The results are shown below in Table 13 for an example comparison.

Table 13: *Results of paired t-test for difference in progress between 2014/2015 and 2016 for Target 1.*

Comparison	**Target 1 (2014/2015) to Target 1 (2016)**
Result	*Mean* value of Target 1 in 2016 is not greater than *mean* value of Target 1 in 2014/2015 ($p = 0.91$)

When multiple years are to be compared within a single Target, repeated measures ANOVA should be used (McCrum-Gardner 2008), or, equivalently, a linear mixed-effects model with country included as a random effect. Here we use a linear mixed-effects model as it is better able to cope with unbalanced data and more generalizable. We use Tukey's Honest Significant Difference to correct for multiple comparisons. Results are presented in Table 14 below.

Table 14: *Results from linear mixed-effect model test, p-values. Country is set to be a random effect. Significant differences are highlighted in bold. Post-hoc comparison conducted using Tukey's Honest Significant Difference correction for multiple comparisons.*

	2014 progress	**2015 progress**
2015 progress	1	
2016 progress	0.20	0.17

As with the non-parametric test (Table 12), no significant differences between years are detected, although the p-values are closer to significance in this test, suggesting greater power to detect differences.

Weighted approaches (paired *t*-tests and linear mixed-effects models) can be used to determine the significance of differences given the confidence in each estimate. In none of the examples here do any comparisons become significant (not shown), though of course this might not be the case in the full data analysis.

3.4 Recommendations

1. When assessing relative progress between Targets, treating the data as interval provides more power to detect statistical differences in progress. Therefore, we recommend treating the data as interval and using parametric tests (t-tests, ANOVA, linear regression) to estimate statistical significance.
2. Furthermore, we also recommend calculating ordinal results to test robustness and in case reviewers request them, though this may indicate that some parametric test results may not be considered as robust (if the non-parametric tests do not detect a significant difference).
3. We recommend using linear regression instead of ANOVA due to the capability to more effectively deal with unbalanced data.
4. It is important to use a correction for the *p*-values when performing multiple comparison, e.g. across multiple years or multiple targets.
5. When assessing progress across years within an individual Target, treating the data as interval again provides more power to detect statistical differences in progress. Therefore, we recommend treating the data as interval and using parametric tests (repeated measures ANOVA, mixed-effects models) to estimate whether differences are significant.
6. We recommend using mixed-effects models instead of repeated-measures ANOVA due to the capability to more effectively deal with unbalanced data.
7. It is not straightforward to calculate weighted test results for the ordinal data. When treating the data as interval, we recommend that, if they are considered to be roughly evenly spaced, the three confidence levels provided are treated as a linear scale, with weights equal to [confidence value / 3]. From this, those weights can then be applied to the parametric approaches (*t*-tests, linear regressions and mixed-effects models) to give insight into what effect the confidence in each national report assessment is having.

4. Modelling factors associated with progress

4.1 Concepts

An important aspect of analzying the national reports is to understand the reasons for varied levels of progress between different countries. This can be achieved through relating the progress category of each country to factors such as GDP, population size, whether or not it was a self-assessment, and so forth. Here we show some examples of how to approach this problem with models, while noting that these are preliminary results and could change (as could the obstacles faced) when using a full data set.

For the whole of this section, when treating data as interval, we use Approach 1 from section 2.2 to assign numeric values to each category. Furthermore, we do not do any 'model selection' in terms of removing insignificant factors, but rather choose to focus on the full models with all factors included.

Finally, in this section, we do not consider a model including progress across years (i.e. multiple years of data). At the point in time when multiple years of data exist for all Targets, if they are to be included in a model, then careful thought will need to be given as to the model structure.

4.2 Assessing factors affecting progress (ordinal data)

If the data are considered to be ordinal, then classical linear regression techniques should not be applied. Instead, ordered logistic regression (also known as the proportional odds model, the ordered logit model, (a subset of) ordinal regression, and cumulative link model) should be applied (Agresti 2010, Christensen 2011). Fitting an ordinal regression model to the data for Target 1 and the predictive variables provided in the preliminary data gives the following output (summarized):

Table 15: *Results from ordinal regression for Target 1*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Log (GDP)	0.27	0.15	1.81	0.07
Log (Population)	-0.34	0.25	-1.33	0.18
Log (Area)	0.10	0.17	0.60	0.55
Self Assessment (Yes)	-0.38	0.47	-0.80	0.43
Global Benefits Index for biodiversity	0.00	0.01	0.19	0.85

The key thing to notice here is that none of the predictor variables are significantly associated with progress category (i.e. Pr(>|z|)>0.05 for all predictors), although GDP is close to significance. Thus we can conclude that none of the predictors appear to have a relationship with national progress on Target 1. It is also worth noting the sign of the estimate: GDP and area are positively correlated with progress, while population is negatively correlated, all of which seem reasonable. When countries self-assess, their average progress category is lower (negative parameter estimate), though not significantly so. Thus we can report that self-assessment does not appear to inflate progress estimation, which is helpful as it indicates that we may not need to be concerned with one of the potential positive biases of the study.

Interpretation of the coefficient estimates is not as straightforward as in a linear regression model. The easiest way to interpret them is to convert them to 'odds ratios' by taking their exponent. Thus the odds ratio for Log(GDP) is exp(0.27) = 1.31. This can be interpreted as follows: for countries to be in a class higher than *n* versus in classes less than or equal to *n* the odds are 1.31 times greater if the other variables are held constant. So, for example, when Log(GDP) increases by one, the odds of being in 'on track to meet the Target' or higher are 1.31 times greater than being in 'progress but at an insufficient rate' or lower. For negative parameter estimates, such as self-assessment, the interpretation remains the same, but the odds ratios are less than one. That is, for self-assessment, the odds ratio is exp(-0.38) = 0.69. So for self-assessed countries, the odds of being in a class higher than *n* are 0.69 versus classes *n* or lower (i.e. they are more likely to be in lower categories when self-assessed).

When running an ordinal regression, the 'proportional odds assumption' – the assumption that the relationships between pairs of outcome categories are the same - must be tested to ensure that it is an appropriate model. If this assumption is violated, a more complex modelling framework (generalized ordered logit model) should be used. In the current model, this assumption is not violated (chi-squared test p>0.05).

A more complex model is one that includes all Targets. Target number then becomes a predictive variable in the model. We also switch to a mixed-effects ordinal regression model, with country as a random effect, as the responses between countries are assumed to be correlated (note: this is a different assumption to that in Section 3). The results from this model are shown below:

Table 16: *Results from mixed-effects ordered logistic model (with country as a random effect) for factors affecting progress towards Targets. Significant results are in bold.*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Log(GDP)	0.23	0.09	2.60	**0.01**
Log(Population)	-0.17	0.15	-1.12	0.26
Log(Area)	-0.03	0.10	-0.30	0.76
Self Assessment (Yes)	-0.01	0.28	-0.05	0.96
Global Biodiversity Index	0.01	0.01	1.41	0.16
TargetNumber5	2.28	0.32	7.16	**0.00**
TargetNumber10	1.78	0.33	5.32	**0.00**
TargetNumber11	0.60	0.29	2.08	**0.04**
TargetNumber17	0.80	0.29	2.74	**0.01**

This time, there are some factors that have a significant effect on the assessed category for countries. GDP has a positive effect: a higher GDP means that a country is significantly more likely to be in a higher category. Furthermore, Targets 5, 10, 11 and 17 are significantly different from Target 1 (the baseline comparison), suggesting a different distribution of scores within Targets. We are not actually interested in this factor, because we have already conducted the assessment of progress across Targets, but it is important to retain it to account for its effects on the other predictor variables.

In terms of the proportional odds assumption, it becomes challenging to test once moving to a mixed-effects framework, and visual assessments of conformity to the assumption may have to be used. Again, if the proportional odds assumption is breached, a more complex random-effects framework may have to be used.

Another choice that needs to be made is whether 'Region' should be fitted as a random effect (with country nested inside, or as a fixed-effect. This depends upon whether it is considered to be a variable of interest. Here we treat it as a random effect, with country nested within (i.e. it is not directly of interest). In this instance, adding region as a random effect adds little over and above including country (results not shown).

In terms of weighting the data by confidence, again we recommend against treating the confidence level as an (arbitrary) numeric scale while using an ordinal modelling framework. Instead, it is more consistent to drop data below a specific confidence level. As an example, the below shows the same random effects model as fitted in Table 16 but restricted to only data with a confidence level of three (about half the data points):

Table 17: *Results from mixed-effects ordered logistic model (with country as a random effect) for factors affecting progress towards Targets, data only with confidence level 3 included. Significant results are in bold.*

	Estimate	**Std. Error**	**z value**	**Pr(>|z|)**
Log(GDP)	0.26	0.11	2.27	**0.02**
Log(Population)	-0.10	0.18	-0.56	0.58
Log(Area)	-0.01	0.12	-0.07	0.94
Self Assessment (Yes)	-0.42	0.34	-1.24	0.21
Global Biodiversity Index	0.01	0.01	0.75	0.45
TargetNumber5	3.21	0.54	5.95	**0.00**
TargetNumber10	2.40	0.55	4.37	**0.00**
TargetNumber11	0.68	0.46	1.50	0.13
TargetNumber17	1.39	0.46	3.06	**0.00**

In this instance, the estimates, standard errors, and p-values have all changed somewhat, with the consequence that Target 11 is no longer significantly different from Target 1.

4.3 Assessing factors affecting progress (interval data)

When the data are interval, modelling them is more straightforward, as standard linear regression and mixed-effects models can be used. Table 18 gives the results of a regression model treating the data as interval for Target 1 (compare to Table 15):

Table 18: *Results from linear regression for Target 1*

	Estimate	**Std. Error**	**t value**	**Pr(>|t|)**
(Intercept)	1.178267	0.569947	2.067	**0.0413**
Log(GDP)	0.053448	0.030599	1.747	0.0838
Log(Population)	-0.06531	0.052051	-1.255	0.2126
Log(Area)	0.018281	0.034895	0.524	0.6015
Self Assessment (Yes)	-0.07754	0.096903	-0.8	0.4255
Global Biodiversity Index	0.000786	0.003039	0.259	0.7964

The results are very similar to those from the ordinal regression (the intercept can be ignored in this case as it is not a factor of interest). The interpretation of the coefficients is somewhat more straightforward, however. For example, if LogGDP increases by 1, the numerical category value is estimated to increase by 0.05.

Fitting the mixed-effects model for comparison across all Targets gives the following result (compare to Table 16):

Table 19: *Results from mixed-effects model (with country as a random effect) for factors affecting progress towards Targets. Significant results are in bold.*

	Value	**Std.Error**	**DF**	**t-value**	**p-value**
(Intercept)	1.02	0.44	385.00	2.31	**0.02**
Log(GDP)	0.06	0.02	110.00	2.57	**0.01**
Log(Population)	-0.04	0.04	110.00	-0.96	0.34
Log(Area)	-0.01	0.03	110.00	-0.37	0.71
Self Assessment (Yes)	0.01	0.08	110.00	0.20	0.85
Global Biodiversity Index	0.00	0.00	110.00	1.40	0.16
TargetNumber5	0.57	0.08	385.00	7.30	**0.00**
TargetNumber10	0.45	0.08	385.00	5.37	**0.00**
TargetNumber11	0.15	0.08	385.00	1.96	0.05
TargetNumber17	0.21	0.08	385.00	2.78	**0.01**

For the full data set, GDP is significant, with a positive effect on progress, and Targets 5, 10, and 17 are significantly different in terms of progress from Target 1 (very similar to the first column of Table 9). When including country as a random effect nested within region, GDP becomes non-significant, probably because it is highly correlated with region.

When treating the data as interval, an arbitrary scale can be applied to the confidence levels; we use the same approach as in Sections 2 and 3 divide numeric values of confidence by 3 to rescale them from 0.33 to 1, where 1 is high confidence.

Table 20: *Results from mixed-effects model (with country as a random effect) for factors affecting progress towards Targets, data weighted by confidence level. Significant results are in bold.*

	Value	**Std.Error**	**DF**	**t-value**	**p-value**
(Intercept)	0.99	0.43	385.00	2.32	**0.02**
Log(GDP)	0.06	0.02	109.00	2.60	**0.01**
Log(Population)	-0.03	0.04	109.00	-0.87	0.39
Log(Area)	-0.01	0.03	109.00	-0.34	0.74
Self Assessment (Yes)	-0.01	0.07	109.00	-0.17	0.87
Global Biodiversity Index	0.00	0.00	109.00	1.27	0.21
TargetNumber5	0.61	0.08	385.00	7.90	**0.00**
TargetNumber10	0.50	0.09	385.00	5.83	**0.00**
TargetNumber11	0.16	0.08	385.00	2.11	**0.04**
TargetNumber17	0.25	0.07	385.00	3.39	**0.00**

Weighting the data by confidence level changes the estimates and p-values slightly, and the difference between Target 11 and Target 1 becomes significant at $p<0.05$.

4.4 Recommendations

1. When modelling the preliminary data, the results from treating the data as ordinal and treating it as interval are fairly similar (compare Tables 16 and 19). Assuming this remains the case with the full data, we therefore recommend choosing to treat the data based on consistency with choosing a single metric (Section 2) and assessing relative progress across Targets and across years (Section 3). That is, if an ordinal approach is used in those sections, an ordinal approach should also be used to test factors associated with progress, and vice versa. The main difference between ordinal and interval regression in terms of the preliminary data is in the interpretation of the regression coefficients, which are somewhat easier to understand when treating the data as interval
2. We recommend using a mixed-effects model and finding factors associated with progress across all Targets. Consider including Region as a random effect. Also consider including Target as a random effect, to increase the degrees of freedom and reduce overlap with assessing progress across Targets.
3. If including data across multiple years, consider carefully whether Year should be a random effect or an auto-correlated fixed effect (i.e. with a temporal autocorrelation structure).
4. It is important to use a correction for the *p*-values when performing multiple comparison, e.g. across multiple years or multiple targets.
5. When assessing progress across years within an individual Target, treating the data as interval again provides more power to detect statistical differences in progress. Therefore, we recommend treating the data as interval and using parametric tests (repeated measures ANOVA, mixed-effects models) to estimate whether differences are significant.
6. We recommend using mixed-effects models instead of repeated-measures ANOVA due to the capability to more effectively deal with unbalanced data.
7. It is not straightforward to calculate weighted test results for the ordinal data. When treating the data as interval, we recommend that, if they are considered to be roughly evenly spaced, the three confidence levels provided are treated as a linear scale, with weights equal to [confidence value / 3]. From this, those weights can then be applied to the parametric approaches (*t*-tests, linear regressions and mixed-effects models) to give insight into what effect the confidence in each national report assessment is having.

5. Final recommendations

The data associated with the national reports provide a lot of opportunities for interesting analysis. It is a relatively large data set, which provides reasonable power at determining differences. Through this report, we have explored two different approaches: treating the data as ordinal (categorical) data, which is more statistically appropriate but harder to interpret and potentially less powerful, and treating them as interval (numeric) data, which is technically not true, but a frequently used approach in the literature nonetheless, and may be more powerful and easier to interpret. Table 21 below summarises our findings from the previous three sections for the two approaches.

Table 21: Approach and results summary.

Approach	**Ordinal (categorical)**	**Interval (numeric)**
Single metric	Mode and median used as the single metric. Interquartile range used as the measure of variability. All Targets have same median and mode national progress for all Targets.	Mean used as the single metric, and standard deviation used as the measure of variability. All Targets show different mean values, though not all will be significantly different from one another.
Assessing relative progress across Targets	Mann-Whitney / Kruskal-Wallis test used. Target 1 is significantly different from Targets 5 and 10, and Targets 11 and 17 are also significantly different from Target 5.	t-test / linear regression used. Target 1 is significantly different from Targets 5 and 10, Targets 11 and 17 are also significantly different from Target 5, and Targets 10 and 11 are significantly different.
Assessing progress across time for a single Target	Wilcoxon signed-rank / Friedman's test used. No evidence for significant difference in progress between years (using simulated 2016 & 2017 data).	Paired t-test / linear mixed effects model used. No evidence for significant difference in progress between years (using simulated 2016 & 2017 data).
Assessing factors associated with progress	Ordinal regression / mixed-effects ordered logistic model used. GDP significantly associated with progress (and Targets show differences).	Linear regression and mixed-effects model used. GDP significantly associated with progress (and Targets show differences).

For the preliminary data, there is not much difference between the ordinal and interval approaches in terms of the significant of results, with the difference that more Targets are shown to be significantly different from one another using the interval approach. On the whole, however, the approaches give fairly similar results, though this may not necessarily be the case with all of the data. However, the interval data are easier to interpret (e.g. mean instead of median, easier interpretation of regression coefficients), and make it easier to display differences in terms of progress when compressed to a single metric.

Thus for the full, final analysis, the choice about which approach to use comes down to a trade-off between statistical rigour, which would favour treating the data as ordinal, and interpretability and ease of understanding of results, which would favour treating the data as interval.

In this case, given the importance of communicating the results, we recommend treating the data as interval. There are many examples in the literature of this approach being taken, and of the robustness of statistical tests to this apparent inconsistency. However, a very strong justification must be used for the numerical scale that is used to convert the data from categories to values, and possible biases or issues with the approach must be presented up front. Ultimately, if it can be supported, we would recommend using an equal interval scale, as per the IUCN Red List (Butchart *et al*. 2004).

The risk associated with doing this is that reviewers may take issue with the approach and ask for the data to be treated as ordinal, or a model such as item-response theory (IRT) to be applied when converting from ordinal to interval – a not inconsiderable analysis. In response to this, we have three recommendations. Firstly, cite literature showing the robustness of statistical tests to this assumption. Secondly, highlight the importance of interpretability for these results in terms of international policy processes. Thirdly, and optionally, also perform an analysis treating the data as ordinal and present this in the supplementary material. With these recommendations, and the others given throughout this report, we feel that a robust and comprehensive analysis can be conducted on the national reports of progress towards the Aichi targets, and an insightful and interpretable product produced.

6. References

Agresti, A. (2010) *Analysis of ordinal categorical data* (2nd Ed.) Wiley, New York.

Butchart, S. H. M., Stattersfield, A. J., Bennun, L. A., Shutes, S. M., Akcakaya, H. R., Baillie, J. E. M., Stuart, S. N., Hilton-Taylor, C. & Macr, G. M. (2004). Measuring global trends in the status of biodiversity: Red List Indices for birds. *PLoS Biology*, 2: e383.

CBD (Secretariat of the Convention on Biological Diversity) (2014). Global Biodiversity Outlook 4. Montreal, Canada.

Chenery, A., McOwen, C., Dixon, M., Ivory, S., Alison, H., Walpole, M., and Regan. E. (2015) Review of the global indicator suite, key global gaps and indicator options for future assessment of the Strategic Plan for Biodiversity 2011-2020. UNEP-WCMC: Cambridge, UK.

Christensen, R. H. B. (2015). Regression models for ordinal data. R-package version 2015-1-21. http://www.cran.r-project.org/package=ordinal/.

Harwell, M. R. & Gatti, G. G. (2001). Rescaling ordinal data to interval data in educational research. *Review of Educational Research*, 71: 105-131.

Hothorn, T., Bretz, F. & Westfall, P. (2008). Simultaneous inference in general parametric models. *Biometrial Journal*, 50: 346-363.

Jamieson, S. (2004). Likert scales: how to (ab)use them. *Medical Education*, 38: 1212-1218.

Knapp, T. R. (1990). Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing Research*, 39: 121-123.

Lumley, T. (2004). Analysis of complex survey samples. *Journal of Statistical Software*, 9: 1-19.

Lumley, T. & Scott, A. J. (2013). Two-sample rank tests under complex sampling. *Biometrika*, 100: 831-842.

McCrum-Gardner, E. (2008). Which is the correct statistical test to use? *British Journal of Oral and Maxillofacial Surgery*, 46: 38-41.

Miller, R. H. (1981). *Simultaneous Statistical Inference*. Springer, Berlin.

Norman, G. (2010). Likert scales, levels of measurement and the "laws" of statistics. *Advances in Health Science Education*, 15: 625-632.

Pohlert, T. (2015). Calculate pairwise multiple comparisons of mean rank sums. R Package v 1.1. http://CRAN.R-project.org/package=PMCMR

Sachs L (1997). *Angewandte Statistik.* 8th Edition. Springer, Berlin.

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, 103: 667-680.

Tittensor, D. P. *et al.* (2014). A mid-term analysis of progress toward international biodiversity targets. *Science*, 346: 241-244.

Vickers, A. J. (1999). Comparison of an ordinal and a continuous outcome measure of muscle soreness. *International Journal of Technology Assessment in Health Care* , 15: 709-716.

Xie, J. & Priebe, C. E. (2002). A weighted generalization of the Mann-Whitney-Wilcoxon statistic. *Journal of Statistical Planning and Inference*, 102: 441-466.

Appendix A: Computer code to run example analyses

The code below is for the 'R' statistical programming language (freely available from www.r-project.org)

```
require(foreign)
require(ggplot2)
require(MASS)
require(Hmisc)
require(reshape2)
require(PMCMR)
require(lsmeans)
require(cwhmisc)
require(survey)
require(weights)
require(lmerTest)
require(ordinal)
require(VGAM)
require(nlme)

# Directory in which data file is located. This needs to be changed to the
location of the CBD provided data.
data_directory = "c:/users/derekt/work/research/data/aichi national
targets/"
data_file = "cbd_national_indicator_data.csv"

#-------------------------------------------------------------------------
----
# Read in the data and transform it to appropriate formats and structures
# Load the data
natind = read.csv(paste(data_directory, data_file, sep = ""))
natind[natind == "NI"] = NA

# Create data table. Note NO NAs
targetdata = factor(c(natind$Target.1,natind$Target.5),levels =
c('1','2','3','4','5'))
targetnumber = c(rep("Target 1", length(natind$Target.1)),
rep("Target5",length(natind$Target.5)))
TargetTable = table(targetnumber,targetdata)

# Create a data frame
targetcountry = rep(natind$Country,2)
TargetDataFrame = data.frame(targetdata,targetnumber,targetcountry)
colnames(TargetDataFrame) = c("Score","TargetNumber","Country")

# Aggregate data for the purpose of statistics
Target1 = as.numeric(levels(natind$Target.1))[natind$Target.1]
Target1Sum =  c(sum(Target1 == 1, na.rm = T), sum(Target1 == 2, na.rm = T),
sum(Target1 == 3, na.rm = T), sum(Target1 == 4, na.rm = T), sum(Target1 ==
5, na.rm = T), sum(is.na(Target1)))
Target1Pct = 100 * Target1Sum / length(Target1)

Target5 = as.numeric(levels(natind$Target.5))[natind$Target.5]
```

```
Target5Sum = c(sum(Target5 == 1, na.rm = T), sum(Target5 == 2, na.rm = T),
sum(Target5 == 3, na.rm = T), sum(Target5 == 4, na.rm = T), sum(Target5 ==
5, na.rm = T), sum(is.na(Target5)))
Target5Pct = 100 * Target5Sum / length(Target5)

Target10 = as.numeric(levels(natind$Target.10))[natind$Target.10]
Target10Sum = c(sum(Target10 == 1, na.rm = T), sum(Target10 == 2, na.rm =
T), sum(Target10 == 3, na.rm = T), sum(Target10 == 4, na.rm = T),
sum(Target10 == 5, na.rm = T), sum(is.na(Target10)))
Target10Pct = 100 * Target10Sum / length(Target10)

Target11 = as.numeric(levels(natind$Target.11))[natind$Target.11]
Target11Sum = c(sum(Target11 == 1, na.rm = T), sum(Target11 == 2, na.rm =
T), sum(Target11 == 3, na.rm = T), sum(Target11 == 4, na.rm = T),
sum(Target11 == 5, na.rm = T), sum(is.na(Target11)))
Target11Pct = 100 * Target11Sum / length(Target11)

Target17 = as.numeric(levels(natind$Target.17))[natind$Target.17]
Target17Sum = c(sum(Target17 == 1, na.rm = T), sum(Target17 == 2, na.rm =
T), sum(Target17 == 3, na.rm = T), sum(Target17 == 4, na.rm = T),
sum(Target17 == 5, na.rm = T), sum(is.na(Target17)))
Target17Pct = 100 * Target17Sum / length(Target17)

# Create a new data frame with a dummy target variable to test for
differences between targets
TwoTargetComparison = data.frame(Score = c(natind$Target.1,
natind$Target.5), TargetNumber =
as.factor(c(rep(1,length(natind$Target.1)), rep(5,
length(natind$Target.5)))))

# Create a new data frame with a dummy target variable to test for
differences between targets
AllTargetComparison = data.frame(Score = c(natind$Target.1,
natind$Target.5, natind$Target.10, natind$Target.11, natind$Target.17),
TargetNumber = as.factor(c(rep(1,length(natind$Target.1)), rep(5,
length(natind$Target.5)), rep(10, length(natind$Target.10)), rep(11,
length(natind$Target.11)), rep(17, length(natind$Target.17)))))


#-------------------------------------------------------------------------
-----
# Rescale using the ordinal to numeric conversion approaches
t1RescaledApp1 = c(rep(-1,Target1Sum[1]),rep(0, Target1Sum[2]),
rep(1,Target1Sum[3]), rep(2,Target1Sum[4]), rep(3,Target1Sum[5]))
t1RescaledApp2 = c( rep(-2,Target1Sum[1]),rep(0, Target1Sum[2]),
rep(2,Target1Sum[3]), rep(3,Target1Sum[4]), rep(4, Target1Sum[5]))
t5RescaledApp1 = c(rep(-1,Target5Sum[1]),rep(0, Target5Sum[2]),
rep(1,Target5Sum[3]), rep(2,Target5Sum[4]), rep(3, Target5Sum[5]))
t5RescaledApp2 = c( rep(-2,Target5Sum[1]),rep(0, Target5Sum[2]),
rep(2,Target5Sum[3]), rep(3,Target5Sum[4]), rep(4, Target5Sum[5]))
t10RescaledApp1 = c(rep(-1,Target10Sum[1]),rep(0, Target10Sum[2]),
rep(1,Target10Sum[3]), rep(2,Target10Sum[4]), rep(3, Target10Sum[5]))
t10RescaledApp2 = c( rep(-2,Target10Sum[1]),rep(0, Target10Sum[2]),
rep(2,Target10Sum[3]), rep(3,Target10Sum[4]), rep(4, Target10Sum[5]))
t11RescaledApp1 = c(rep(-1,Target11Sum[1]),rep(0, Target11Sum[2]),
rep(1,Target11Sum[3]), rep(2,Target11Sum[4]), rep(3, Target17Sum[5]))
t11RescaledApp2 = c( rep(-2,Target11Sum[1]),rep(0, Target11Sum[2]),
rep(2,Target11Sum[3]), rep(3,Target11Sum[4]), rep(4, Target10Sum[5]))
t17RescaledApp1 = c(rep(-1,Target17Sum[1]),rep(0, Target17Sum[2]),
rep(1,Target17Sum[3]), rep(2,Target17Sum[4]), rep(3, Target17Sum[5]))
```

```
t17RescaledApp2 = c( rep(-20,Target17Sum[1]),rep(0, Target17Sum[2]),
rep(2,Target17Sum[3]), rep(3,Target17Sum[4]), rep(4, Target17Sum[5]))

# Calculate medians and modes (though using rescaled data for convenience)
Mode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}

# Using approach 1 rescaled data
median(t1RescaledApp1)
Mode(t1RescaledApp1)
median(t5RescaledApp1)
Mode(t5RescaledApp1)
median(t10RescaledApp1)
Mode(t10RescaledApp1)
median(t11RescaledApp1)
Mode(t11RescaledApp1)
median(t17RescaledApp1)
Mode(t17RescaledApp1)

# Calculate interquartile range using approach 1 rescaled data
print(paste("T1 IQR: ", quantile(t1RescaledApp1, 0.75) -
quantile(t1RescaledApp1, 0.25), sep = ""))
print(paste("T5 IQR: ", quantile(t5RescaledApp1, 0.75) -
quantile(t5RescaledApp1, 0.25), sep = ""))
print(paste("T10 IQR: ", quantile(t10RescaledApp1, 0.75) -
quantile(t10RescaledApp1, 0.25), sep = ""))
print(paste("T11 IQR: ", quantile(t11RescaledApp1, 0.75) -
quantile(t11RescaledApp1, 0.25), sep = ""))
print(paste("T17 IQR: ", quantile(t17RescaledApp1, 0.75) -
quantile(t17RescaledApp1, 0.25), sep = ""))

# Calculate means and SDs
mean(t1RescaledApp1)
sd(t1RescaledApp1)
mean(t1RescaledApp2)
sd(t1RescaledApp2)
mean(t5RescaledApp1)
sd(t5RescaledApp1)
mean(t5RescaledApp2)
sd(t5RescaledApp2)
mean(t10RescaledApp1)
sd(t10RescaledApp1)
mean(t10RescaledApp2)
sd(t10RescaledApp2)
mean(t11RescaledApp1)
sd(t11RescaledApp1)
mean(t11RescaledApp2)
sd(t11RescaledApp2)
mean(t17RescaledApp1)
sd(t17RescaledApp1)
mean(t17RescaledApp2)
sd(t17RescaledApp2)

# Calculate medians for only data with confidence >= 3
median(t1RescaledApp1[natind$Confidence >= 3], na.rm = T)
median(t5RescaledApp1[natind$Confidence.1 >= 3], na.rm = T)
median(t10RescaledApp1[natind$Confidence.2 >= 3], na.rm = T)
median(t11RescaledApp1[natind$Confidence.3 >= 3], na.rm = T)
```

```
median(t17RescaledApp1[natind$Confidence.4 >= 3], na.rm = T)

# Calculate weighted means. Remember that we need to convert to the
Approach 1 numeric values
weighted.mean(c(-1,0,1,2,3)[match(Target1,c(1,2,3,4,5))], natind$Confidence
/ 3, na.rm = T)
weighted.mean(c(-1,0,1,2,3)[match(Target5,c(1,2,3,4,5))],
natind$Confidence.1 / 3, na.rm = T)
weighted.mean(c(-1,0,1,2,3)[match(Target10,c(1,2,3,4,5))],
natind$Confidence.2 / 3, na.rm = T)
weighted.mean(c(-1,0,1,2,3)[match(Target11,c(1,2,3,4,5))],
natind$Confidence.3 / 3, na.rm = T)
weighted.mean(c(-
1,0,1,2,3)[match(Target17,c(1,2,3,4,5))][!is.na(natind$Confidence.4)],
natind$Confidence.4[!is.na(natind$Confidence.4)] / 3, na.rm = T)

# To calculate weighted standard deviations, use wtd.var from package HMisc
#-------------------------------------------------------------------------
-----


#-------------------------------------------------------------------------
-----
# PLOTTING DATA
# Plot graphs
par(mfrow = c(3,2))
xlabels = c("A", "B", "C", "D", "E", "F")

barplot(Target1Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey82",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 1", ylab = "Number countries")
axis(2, las = 1)
barplot(Target5Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey82",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 5", ylab = "Number countries")
axis(2, las = 1)
barplot(Target10Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey82",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 10", ylab = "Number countries")
axis(2, las = 1)
barplot(Target11Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey82",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 11", ylab = "Number countries")
axis(2, las = 1)
barplot(Target17Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey82",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 17", ylab = "Number countries")
axis(2, las = 1)

# Plots highlighting the median and the mode
halves = c(55,55,40,58,59)
dev.new()
par(mfrow = c(3,2))
barplot(Target1Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey48",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 1", ylab = "Number countries")
axis(2, las = 1)
lines(x = c(2.6,3.6), y = c(46,46))
```

```
barplot(Target5Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey48",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 5", ylab = "Number countries")
axis(2, las = 1)
lines(x = c(2.6,3.6), y = c(18,18))
barplot(Target10Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey48",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 10", ylab = "Number countries")
axis(2, las = 1)
lines(x = c(2.6,3.6), y = c(7,7))
barplot(Target11Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey48",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 11", ylab = "Number countries")
axis(2, las = 1)
lines(x = c(2.6,3.6), y = c(54,54))
barplot(Target17Sum, ylim = c(0,100), col = c("grey82", "grey82", "grey48",
"grey82", "grey82", "grey82", "grey82"), yaxt= "n", names.arg = xlabels,
main = "Target 17", ylab = "Number countries")
axis(2, las = 1)
lines(x = c(2.6,3.6), y = c(49,49))
#-------------------------------------------------------------------------
-----



#-------------------------------------------------------------------------
-----
# ASSESSING DIFFERENCES BETWEEN TARGETS AND BETWEEN YEARS
# Perform Mann-Whitney test
wilcox.test(Target1, Target5, conf.int = TRUE)

# Now for weighted data
design <- svydesign(ids = ~0, data = TwoTargetComparison)
svyranktest(formula = Score ~ TargetNumber, design = design)

# Kruskal-Wallis test for multiple Targets
kruskal.test(list(Target1 = Target1, Target5 = Target5, Target10 =
Target10, Target11 = Target11, Target17 = Target17))

# Kruskal-Wallis test for multiple Targets with posthoc nemenyi test. With
most post-hoc tests it is safer to remove complete rows of NAs. For
example, Friedman's test results can change if rows of NAs are included.
AllTargetComparison2 =
AllTargetComparison[!is.na(AllTargetComparison[,1]),]
attach(AllTargetComparison2)
posthoc.kruskal.nemenyi.test(x = Score, g = TargetNumber, dist =
"Chisquare")
detach(AllTargetComparison2)

# Weighted comparison. Note that it is not straightforward to do a
corrected post-hoc comparison, and it does not appear that there is any
currently implemented way of doing so in R, for example using Chi-squared
or Tukey's as per the# PCMCR package.
design <- svydesign(ids = ~0, data = AllTargetComparison)
svyranktest(formula = Score ~ TargetNumber, design = design)

# Perform t-test
t.test(Target1, Target5)
```

```
# Perform the weighted t-test. Note that for both this and the GLM, the
survey package seems to produce slightly lower p-values
tempvals = c(as.numeric(natind$Confidence),as.numeric(natind$Confidence.1))
/ 3
tempvals[is.na(tempvals)] = 0
design <- svydesign(ids = ~0, data = TwoTargetComparison, weights =
tempvals)
svyttest(formula = Score ~ TargetNumber, design = design, weights =
weights(design))

# Perform an ANOVA for multiple comparisons. Left here for comparison
purposes
#TukeyHSD(aov(Score ~ TargetNumber, data = AllTargetComparison))

# Use a linear regression in case data are unbalanced
detach("package:lmerTest")
pairs(lsmeans(lm(Score ~ TargetNumber, data = AllTargetComparison), ~
TargetNumber))

# Now do the weighted linear regression. Note that it is better practice to
use as.  as.numeric(levels(natind$Population))[natind$Population]
#tempvals =
c(as.numeric(levels(natind$Confidence))[natind$Confidence],as.numeric(level
s(natind$Confidence.1))[natind$Confidence.1],
as.numeric(levels(natind$Confidence.2))[natind$Confidence.2],
as.numeric(levels(natind$Confidence.3))[natind$Confidence.3],
as.numeric(levels(natind$Confidence.4))[natind$Confidence.4]) / 3
tempvals = c(natind$Confidence,natind$Confidence.1,natind$Confidence.2,
natind$Confidence.3, natind$Confidence.4) / 3
tempvals[is.na(tempvals)] = 0
lm(Score ~ TargetNumber, weights = tempvals, data = AllTargetComparison)
pairs(lsmeans(lm(Score ~ TargetNumber, weights = tempvals, data =
AllTargetComparison), ~ TargetNumber))

library(lmerTest)

# Equivalent using svyglm
#design <- svydesign(ids = ~0, data = AllTargetComparison, weights =
tempvals)
#svyglm(formula = Score ~ TargetNumber, design = design, weights =
weights(design))

# Equivalent using multcomp
#glht(lm(Score ~ TargetNumber, data = AllTargetComparison),
mcp(TargetNumber = "Tukey"))

# Generate fake data
set.seed(34)
Target1Fake = Target1 + sample(c(-1,0,1), size = length(Target1), replace =
TRUE)
Target1Fake2 = Target1 + sample(c(-1,0,1), size = length(Target1), replace
= TRUE)

# Create a matrix with temporal data in
DataForComparisonAcrossYears = matrix(cbind(Target1, Target1Fake,
Target1Fake2), ncol = 3, dimnames = list(1:length(Target1),
c("y2014","y2015","y2016")))

# Perform Wilcoxon signed rank test
wilcox.test(Target1, Target1Fake, paired = TRUE, conf.int = TRUE)
```

```
# Now perform Friedman's test for all the different years of data. Note
that NAs must be removed before conducting this comparison - VERY IMPORTAT
DataForComparisonAcrossYears =
DataForComparisonAcrossYears[!is.na(DataForComparisonAcrossYears[,1]),]
posthoc.friedman.nemenyi.test(DataForComparisonAcrossYears)

# Perform paired t-test
t.test(Target1, Target1Fake, paired = TRUE)

# Paired and weighted t-test. In fact, use paired and weighted LM.
set.seed(21)
ConfidenceFake1 = as.numeric(natind$Confidence) + sample(c(-1,0,1), size =
length(Target1), replace = TRUE)
ConfidenceFake2 = as.numeric(natind$Confidence) + sample(c(-1,0,1), size =
length(Target1), replace = TRUE)
tempvals = c(as.numeric(natind$Confidence), ConfidenceFake1) / 3
tempvals[is.na(tempvals)] = 0
TempDat = data.frame(Score = c(Target1,Target1Fake), Year =
as.factor(c(rep(2015,123), rep(2016,123))), Country = rep(natind$Country,
2))
pairs(lsmeans(lmer(Score ~ Year + (1|Country), weights = tempvals, data =
TempDat), ~Year))

# Put data in form for LME or repeated measures ANOVA model
DataForComparisonAcrossYearsLM = data.frame(Score = c(Target1, Target1Fake,
Target1Fake2), Year = as.factor(c(rep(2014, length = length(Target1)),
rep(2015, length = length(Target1Fake)), rep(2016, length =
length(Target1Fake2)))), Country = rep(natind$Country, 3))

# AOV version. May struggle with unbalanced data
#pairs(lsmeans(aov(Score ~ as.factor(Year) + Error(Country), data =
DataForComparisonAcrossYearsLM), ~as.factor(Year)))

# Run LME and calculate pairwise comparison with posthoc test
pairs(lsmeans(lmer(Score ~ Year + (1|Country), data =
DataForComparisonAcrossYearsLM), ~Year))

# Multcomp version
#summary(glht(lmer(Score ~ Year + (1|Country), data =
DataForComparisonAcrossYearsLM), mcp(Year = "Tukey")))

# Run LME with weighted data and comparison with posthoc test
tempvals = c(as.numeric(natind$Confidence), ConfidenceFake1,
ConfidenceFake2) / 5
tempvals[is.na(tempvals)] = 0
pairs(lsmeans(lmer(Score ~ Year + (1|Country), weights = tempvals, data =
DataForComparisonAcrossYearsLM), ~Year))

#-------------------------------------------------------------------------
-----


#-------------------------------------------------------------------------
-----
# MODELLING FACTORS AFFECTING PROGRESS TOWARDS NATIONAL TARGETS

# Extract data into appropriate form & remove NAs
newdat = data.frame(Target1 = as.factor(natind$Target.1), Country =
as.factor(natind$Country), Region = as.factor(as.numeric(natind$Region)),
```

```
LogGDP = log(as.numeric(levels(natind$GDP))[natind$GDP]), LogPopulation =
log(as.numeric(levels(natind$Population))[natind$Population]), LogArea =
log(natind$Total.Area), GBI =
as.numeric(levels(natind$Global.Benefits.Index.for.Biodviersity))[natind$Gl
obal.Benefits.Index.for.Biodviersity], SelfAssess =
as.factor(natind$Self.Assessment
))
newdat = newdat[!is.na(newdat[,1]),]
newdat = newdat[!is.na(newdat[,4]),]
newdat = newdat[!is.na(newdat[,5]),]
newdat = newdat[!is.na(newdat[,6]),]
newdat = newdat[!is.na(newdat[,7]),]

# NEED TO SET IT TO ONLY HAVE THE LEVELS IN THE DATA
levels(newdat$Target1) = c(1,2,3,4,5)

# Ordinal logistic regression
m1 <- polr(Target1 ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI,
data = newdat, Hess = TRUE)

# Check significance
(ctable <- coef(summary(m1)))
p <- pnorm(abs(ctable[, "t value"]), lower.tail = FALSE) * 2
(ctable <- cbind(ctable, "p value" = p))

# Likelihood profile based confidence intervals for parameters
(ci <- confint(m1))

# Odds ratios
#exp(coef(m))
## Odds ratios and CI
#exp(cbind(OR = coef(m), ci))

# Check to see whether the proportional odds assumption is violated
# If it is, how robuest is the test?
sf <- function(y) {
  c('Y>=1' = qlogis(mean(y >= 1)),
    'Y>=2' = qlogis(mean(y >= 2)),
    'Y>=3' = qlogis(mean(y >= 3)))
}

(s <- with(newdat, summary(as.numeric(Target1) ~ LogGDP + LogPopulation +
LogArea + SelfAssess + GBI, fun=sf)))

# Plot to visualise
par(mfrow=c(1,1))
plot(s, which=1:3, pch=1:3, xlab='logit', main=' ', xlim=c(-3,3))

# Print as a table. Should be spaced roughly equally (distance between)
s[,3] = s[,3] - s[,2]
s[,2] = s[,2] - s[,2]
s

# Now fit the same model using the ordinal package
m1 <- clm(Target1 ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI,
data = newdat)
summary(m1)
# Check the proportional odds assumption formally
m2 <- clm(Target1 ~ 1, nominal = ~LogGDP + LogPopulation + LogArea +
SelfAssess + GBI, data = newdat)
```

```
summary(m2)
anova(m1,m2)

# Alternative approach. Note that this checks all variables independently,
as opposed to the above and below which check them all simultaneously.
nominal_test(m1)

# Now the same model using the VGAM package
fit1 <- vglm(as.ordered(Target1) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI, data = newdat,family=cumulative(reverse = T, parallel=T))
summary(fit1)

# Check the proportional odds assumption formally
fit2 <- vglm(as.ordered(Target1) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI, data = newdat,family=cumulative(reverse = T, parallel=F))
summary(fit2)
pchisq(deviance(fit1)-deviance(fit2),
df=df.residual(fit1)-df.residual(fit2),lower.tail=FALSE)

# Alternative approach for the VGAM package
fitx <- vglm(as.ordered(Target1) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI, propodds, data = newdat)
summary(fitx)

# OK, proportional odds assumption holds for the relatively simple model
above, and all three approaches give the same result. Now start to add
complexity.
newdatfull = data.frame(Score = as.factor(c(natind$Target.1,
natind$Target.5, natind$Target.10, natind$Target.11, natind$Target.17)),
Country =  as.factor(rep(natind$Country,5)), Region =
as.factor(rep(natind$Region,5)), LogGDP =
log(as.numeric(levels(natind$GDP))[natind$GDP]), LogPopulation =
log(as.numeric(levels(natind$Population))[natind$Population]), LogArea =
log(natind$Total.Area), GBI =
as.numeric(levels(natind$Global.Benefits.Index.for.Biodviersity))[natind$Gl
obal.Benefits.Index.for.Biodviersity], SelfAssess =
as.factor(rep(natind$Self.Assessment,5)), TargetNumber =
as.factor(c(rep(1,length(natind$Target.1)), rep(5,length(natind$Target.5)),
rep(10,length(natind$Target.10)), rep(11,length(natind$Target.11)),
rep(17,length(natind$Target.17)))))

newdatfull <- newdatfull[!is.na(newdatfull[,1]),]
newdatfull <- newdatfull[!is.na(newdatfull$LogArea),]


# Fit mixed-effects ordinal models
m1 <- clmm(Score ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI +
TargetNumber + (1|Country), data = newdatfull)
summary(m1)

# Nested random effects
m1 <- clmm(Score ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI +
TargetNumber + (1|Region/Country), data = newdatfull)
summary(m1)

# Now with region included as a fixed effect
m1 <- clmm(Score ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI +
TargetNumber + Region + (1|Country), data = newdatfull)
summary(m1)
```

```
# Need to use clmm2 to get proportional odds assumption test with a random
effect. Note that it doesn't work in this instance due to convergence
issues; need alternative approach
#m1 <- clmm2(Score ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI +
TargetNumber + Region, random = Country, Hess = TRUE, data = newdatfull)
#summary(m1)
#m2 <- clmm2(Score ~ 1, nominal = ~LogGDP + LogPopulation + LogArea +
SelfAssess + GBI + TargetNumber + Region, random = Country, Hess = TRUE,
data = newdatfull)
#summary(m2)
#anova(m1,m2)

summary(m3)

# OK, now do the same thing but with data weighted by confidence (!)
newdatfullincconf = data.frame(Score = as.factor(c(natind$Target.1,
natind$Target.5, natind$Target.10, natind$Target.11, natind$Target.17)),
Country =  as.factor(rep(natind$Country,5)), Region =
as.factor(rep(natind$Region,5)), LogGDP =
log(as.numeric(levels(natind$GDP))[natind$GDP]), LogPopulation =
log(as.numeric(levels(natind$Population))[natind$Population]), LogArea =
log(natind$Total.Area), GBI =
as.numeric(levels(natind$Global.Benefits.Index.for.Biodviersity))[natind$Gl
obal.Benefits.Index.for.Biodviersity], SelfAssess =
as.factor(rep(natind$Self.Assessment,5)), TargetNumber =
as.factor(c(rep(1,length(natind$Target.1)), rep(5,length(natind$Target.5)),
rep(10,length(natind$Target.10)), rep(11,length(natind$Target.11)),
rep(17,length(natind$Target.17)))), Confidence =
as.factor(c(natind$Confidence, natind$Confidence.1, natind$Confidence.2,
natind$Confidence.3, natind$Confidence.4)))
newdatfullincconf <- newdatfullincconf[!is.na(newdatfullincconf$Score),]
newdatfullincconf <- newdatfullincconf[!is.na(newdatfullincconf$LogArea),]
newdatfullincconf <-
newdatfullincconf[!is.na(newdatfullincconf$LogPopulation),]
newdatfullincconf <- newdatfullincconf[!is.na(newdatfullincconf$LogGDP),]
newdatfullincconf <- newdatfullincconf[!is.na(newdatfullincconf$GBI),]
newdatfullincconf <-
newdatfullincconf[!is.na(newdatfullincconf$Confidence),]
ndf3 =  newdatfullincconf[newdatfullincconf$Confidence %in% c("3") ,]

# Fit mixed-effects ordinal model on confidence >=3 data
m1 <- clmm(Score ~ LogGDP + LogPopulation + LogArea + SelfAssess + GBI +
TargetNumber + (1|Country), data = ndf3)
summary(m1)




# Now assume that these are interval data and fit a linear regression
m1 <- lm(as.numeric(Target1) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI, data = newdat)
summary(m1)

# Remove NA rows
newdatfull <- newdatfull[!is.na(newdatfull[,4]),]
newdatfull <- newdatfull[!is.na(newdatfull[,5]),]
newdatfull <- newdatfull[!is.na(newdatfull[,7]),]

# Now fit a mixed-effects model with country as a random effect
```

```
m1 <- lme(as.numeric(Score) ~ LogGDP + LogPopulation + LogArea + SelfAssess
+ GBI + TargetNumber, random = ~1|Country, data = newdatfull)
summary(m1)

# A hierarchical mixed-effects model with country nested within region as
random effects
#m2 <- lme(as.numeric(Score) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI + TargetNumber, random = ~1|Region/Country, data =
newdatfull)
#summary(m2)

# A mixed-effects model with weighted data
 m1 <- lme(as.numeric(Score) ~ LogGDP + LogPopulation + LogArea +
SelfAssess + GBI + TargetNumber, random = ~1|Country, weights
=~I(1/(as.numeric(levels(Confidence))[Confidence])), data =
newdatfullincconf)
summary(m1)

# For multiple years of data: need to decide whether year should be a
random effect. Is it a crossed effect? Yes. Could also model using temporal
autocorrelation structure
# For the same but with multiple years of data use
cor=corAR1(n,form=~1|Year), where n should be set to a 0-1 value; see
reference for details

#------------------------------------------------------------------------
```