

## Annex J: Stability of indicators

1. When we have calculated the bibliometric indicators for each submission to the REF pilot, we have (implicitly) assumed that the indicators are the true values for the submission. In practice, this is unlikely to be the case; for example, some outputs that were submitted will not have been matched to the bibliometric database being used (or even be included in it). The aim of the work in this annex is to get a feel for whether this is likely to have a significant effect on the outcomes of the assessment process, and whether some indicators are less susceptible to such perturbations than others.
2. We can make an estimate of the importance of capturing ‘everything’<sup>1</sup> for the model by randomly removing some of the outputs included in a submission, and looking at the effect this has on the outcomes. By repeating this process many times, we can estimate the effect of an incomplete data set on the outcomes.
3. The work may also help to inform decisions about panel interpretation of small sample sizes in the assessment process<sup>2</sup>.
4. Note that this process is not what we would do in an ideal world. We know that all of the actual submissions are incomplete in the sense that not every output is included in the bibliometric database, or can be matched to it. These uncovered or unmatched outputs are unlikely to be representative of the outputs included in the model as a whole. For example, an unusual page or article numbering scheme of a journal could have made it very difficult to successfully match outputs published in it to a submission.
5. We use this randomisation process on the baseline model. In this model, outputs in each submission are those that are linked to one or more RAE-submitted members of staff associated with the UOA. In future, we plan to look at the ‘top six’ model on a similar basis; however this is a more complex process because we need to recalculate each author’s top six outputs having removed a proportion of outputs from the model.
6. We proceed as follows:
  - we take all of the submissions in a UOA (we use UOA 19 (Physics) as an example in this section)

---

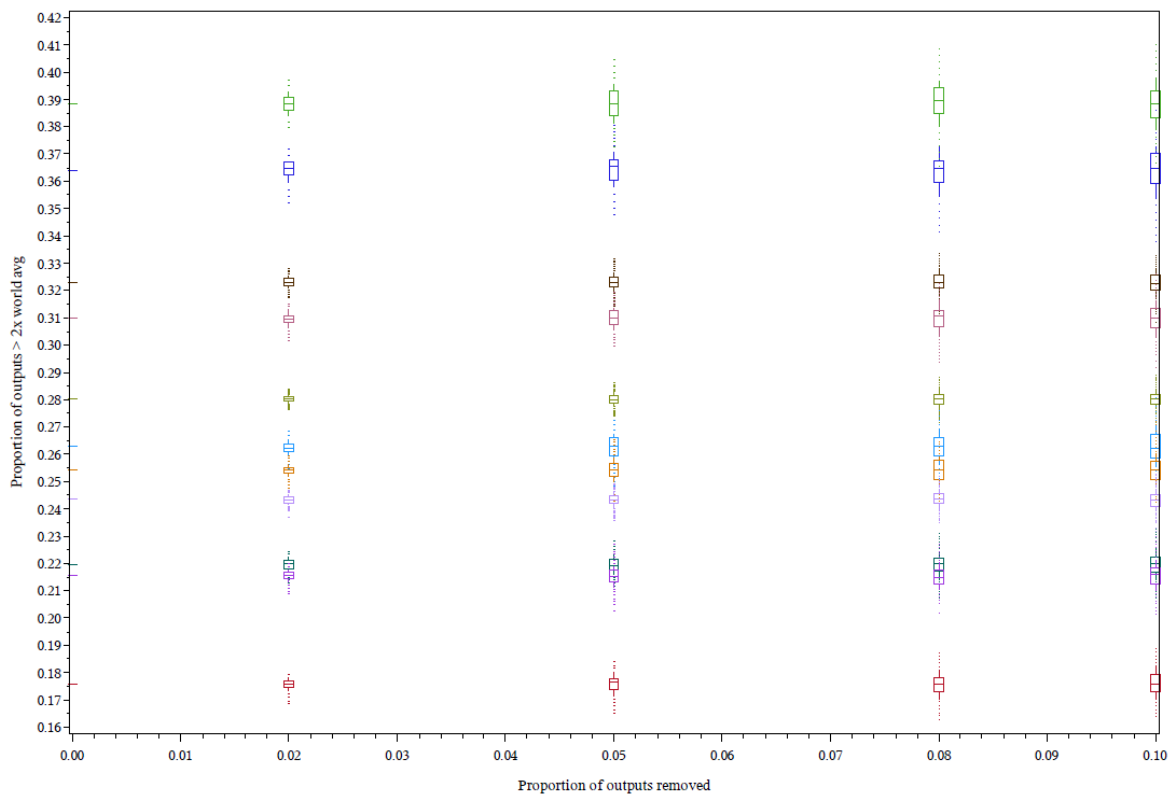
<sup>1</sup> ‘Everything’ in this context could be every output from the department, or a subset of outputs limited by, for example, staff status. The key thing is that every item we want to include (based on our eligibility criteria) is included.

<sup>2</sup> This is discussed more fully by Evidence in Annex L. However, as Evidence notes, this work is likely to provide a relatively poor proxy for such an analysis because we are making submissions smaller by removing items at random. In practice, a small department’s submission is likely to look rather different to this. For example, it may consist of a single research group, rather than the several groups one would typically expect to see in a larger departmental submission. We cannot capture this structure in the data that we hold, making such an analysis impossible.

- for each submission, we remove either a fixed proportion (say, 5 per cent) of outputs, or a fixed number of outputs from the submission, at random, and note the values of the bibliometric indicators that result from this amended submission
- we repeat this process 1,000 times, recording the values of the indicators after each of the iterations.

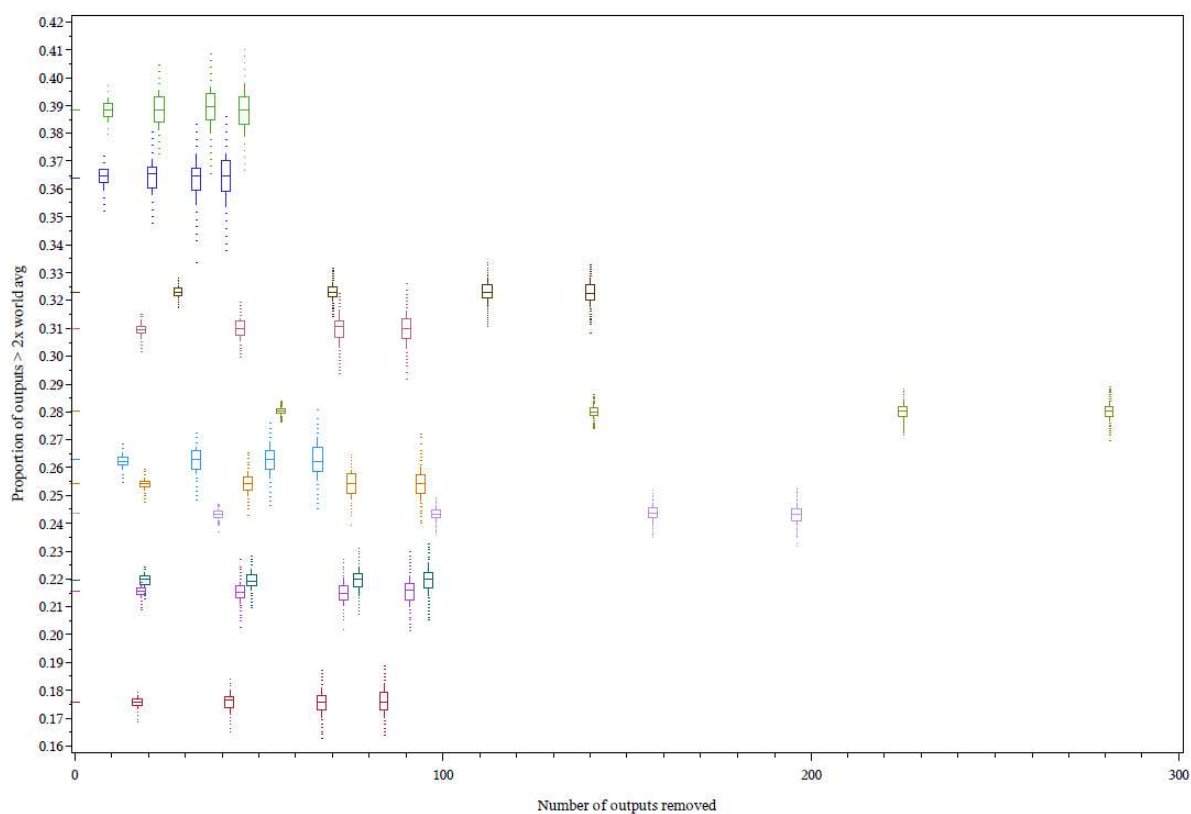
7. Figure J1 shows the effect of removing a fixed proportion of outputs from each submission in UOA 19 on the value of the indicator ‘proportion of work greater than twice the world average’. As we remove an increasingly large proportion of the outputs from the submission, we see that the range of indicator scores grows. Importantly, we see that this does not appear to systematically increase (or decrease) any institution’s score. We do, however, see that some institutions’ range of indicator scores is larger than others’.

**Figure J1 The effect of removing a fixed proportion of outputs from a submission**



8. We can look at whether this is simply due to the different submission sizes from the institutions by plotting the same data against the number of outputs that were removed from each submission. These data are shown in Figure J2. If we take a ‘slice’ through the data at an (approximately) fixed number of outputs, we see that the level of variability between institutions’ submissions is not constant. For example, if we remove about 50 papers from a submission, some institutions’ submissions will have a wider range of possible indicator scores than others.

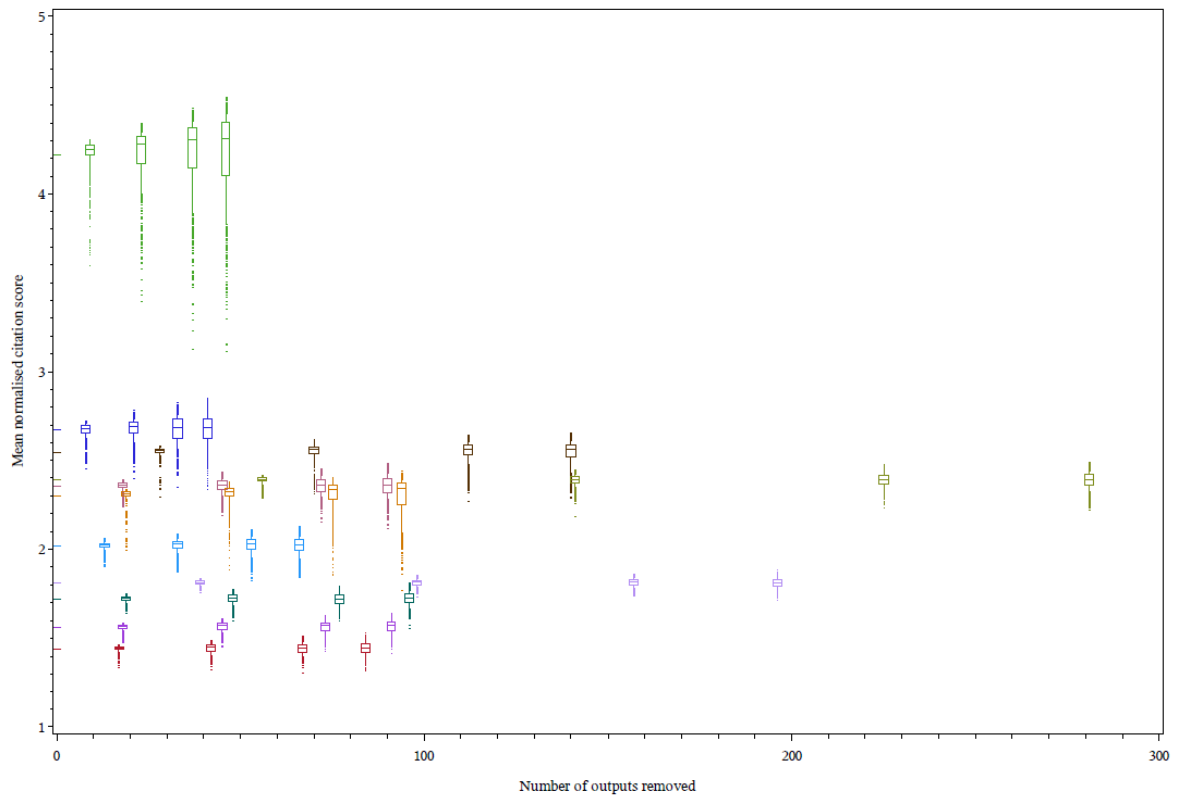
**Figure J2** The data in figure J1 shown by number of (rather than proportion of) outputs removed.



9. Figures J1 and J2 show that the amount of variation we see for an institution's indicator, when we remove some outputs, cannot be fully explained by the size of the institution. Importantly, they show that the exemplar indicator that we have used does not appear to systematically advantage or disadvantage any of the institutions when some outputs are removed from the submission.

10. Figure J3 shows the effect of repeating the same removal process on the mean normalised citation score for each submission. We see that the range of scores for some institutions is much more strongly affected than for others.

**Figure J3 The effect of removing a number of outputs from a submission on mean citation score.**



11. We know that the distribution of normalised citation scores is highly skewed. Most outputs have normalised citation scores relatively close to one; however, a few outputs have very large normalised citation scores. The mean is sensitive to outlying data points. If the papers with large citation scores happen to be removed from the submission, this will have a large effect on the value of the indicator. If we use an indicator such as the proportion of work above a certain threshold, or the median, then the level of fluctuation caused by the outlying data points is reduced. This suggests that such indicators are likely to be more suitable for our purposes.