Response to the Submission "Citations and Sub-Area Bias in the UK Research Assessment Process" by Alan Dix

Robert Jäschke
L3S Research Center
Appelstraße 4, 30167 Hannover, Germany
jaeschke@l3s.de

1. INTRODUCTION

The paper that is targeted by this response [1] analyzes REF2014 – the UK research assessment process – by comparing the REF score distribution of sub-areas of Computer Science with scores derived from citation counts. The results uncover a bias towards more theoretical sub-areas whose citation count is lower than one would expect given their high REF scores. The results also suggest that even on the institutional level a bias exists towards pre-1992 universities.

I consider the paper an important contribution that can and hopefully does spark a discussion about consequences for research evaluation in general and the procedure in the periodic UK research assessment in particular. In the following discussion I want to address three aspects of the paper that I think are worth to discuss and investigate: the different *observation levels* at which the assessment results can be compared, potential *explanations* for the findings, and *questions* raised by the paper.

2. DISCUSSION

Observation Levels. A remarkable observation is that there are differences in the correlation between citations and REF scores depending on the observation level. In the context of this study we can identify three levels of granularity:

1. The individual institutions represent the *macro* level.
2. The individual outputs represent the *micro* level.
3. Between those two, at the *meso* level we can consider the "Units of Assessment" (UoA) or the subject areas of a discipline, where different subject areas typically belong to the same UoA.

Both on the macro and micro level prior works have reported correlations between citation counts and REF ratings. A "strong correlation between citations and ratings" was observed by earlier works (references 1, 3, 4, 5, 12 in the paper) on the institutional (macro) level and Sloman's analysis (reference 11) showed a "correlation between citation and REF ratings on an output-by-output basis" (i.e., the micro level). However, this study found a "large divergence [...] at the level of subject areas", i.e., the meso level.

On the one hand, this shows the importance of analyzing the results at different levels of granularity and it would be interesting to perform a similar comparison with the Units of Assessment and with other disciplines. On the other hand, it raises the question whether the observed divergence could be attributed to some statistical artifact, e.g., Simpson's paradox [2], which can occur when different partitions of the data are analyzed.

Explanations. Given the page restrictions for the submission it is difficult to discuss in detail or even further investigate potential explanations for the observed discrepancies. Differences in the citation behavior between sub-areas are an obvious first explanation but they are accounted for by the normalization of the Scopus citation data. The explanations given in the paper ("halo effects when assessing papers from 'good' institutions" or "inter-area bias") seem reasonable and it would be interesting to analyze whether similar patterns can be observed in other disciplines.

One could also pick up the well-known argument that pure citation counts are not a good measure to assess research and that maybe the REF scores better reflect the excellence of the universities. Since I am not familiar with the process underlying the scoring of research outputs in REF supporting or denying such arguments is difficult. Overall, I regard the findings as quite remarkable and suggest that similar comparisons are performed for other disciplines.

Questions. The questions that the paper immediately raises are: Can we trust the results from REF2014? What do the results reflect and how can the differences to citations be explained? Is human judgment more accurate than numerical metrics like citation counts? What can we do and what shall we do? Or more fundamentally: Do we want or need such kind of research assessment? Answering them is beyond the scope of this response but I would like to connect the challenges raised by the paper with the 'W' in the workshop's acronym: scholarly communication on the Web is growing

and researchers are working on methods to assess the quality and impact of such activities. Probably it is not the question *if* such activities will become a part of future research assessment processes but *when*. The resulting questions and challenges are similar and could be discussed at the workshop. Which differences between institutions, disciplines, and sub-areas exist in the usage of the Web for scholarly communication? And which biases could be introduced by considering such outputs in research assessment?

Considering a potential bias between sub-areas of Computer Science I would argue that more applied areas could benefit from the consideration of alternative metrics, since their results are easier to communicate via the Web to the general public which is very likely also more interested in applied research. I wonder whether this is true and whether it is different in other disciplines like Physics, where large basic research projects (e.g., the LHC at CERN) often get much publicity.

In any case, assessing scholarly communication on the Web and incorporating it into research assessment is quite challenging and a fruitful direction for further research.

3. REFERENCES

[1] Alan Dix. Citations and sub-area bias in the UK research assessment process. In *Proceedings of the Quantifying and Analysing Scholarly Communication on the Web Workshop*, 2015.

[2] Edward H. Simpson. The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, 13:238–241, 1951.