USING ELSEVIER'S FIELD WEIGHTED CITATION INDEX SCORE FOR REDUNDANCY SELECTION AT THE UNIVERSITY OF LIVERPOOL

Executive summary

Senior management at the University of Liverpool intend to make 47 academic staff members in the Faculty of Health and Life Sciences redundant as part of a restructure titled 'Project SHAPE'. One of the two criteria it has used to select staff for redundancy is an Elsevier metric called the 'Field Weighted Citation Impact' (FWCI). The University of Liverpool has defined an FWCI score of < 2 as the threshold for redundancy selection.

The UCU University of Liverpool branch has consistently warned, in public and in negotiations with the University, that the FWCI contains very significant errors in its methodology and corruptions in the algorithms it employs. We have also cited peer assessments and warnings by Elsevier's own data scientists that the FWCI is meaningless when it is applied to individual research profiles. Those warnings have been consistently ignored.

Intitial evidence shows no correlation between the University of Liverpool's minimum individual FWCI score and academic excellence.

- We ran all of the 127 government SAGE advisers who are affiliated to universities through SciVal to generate FWCI scores for the period 2015-2020 (the same period the University of Liverpool used). We found that more than half of this group had individual scores of < 2.
- We ran all Nobel Prize Winners between 2018 and 2020 (a total of 25) through SciVal to generate FWCI scores for the period 2015-2020. We found that 10 of those 25 Nobel Prize winners (or 40%) had SciVal scores of < 2.

Analysis of empirical evidence from a sample of researchers at the University of Liverpool reveals serious unexplained errors in the algorithms.

- FWCI is fundamentally unstable over time. Some researchers scores can be exponentially increased or decreased with very small variations in the time period used for analysis.
- Coding errors in SciVal tend exclude a large number of high-quality publications, and, at the same time, include low-quality publications. This significantly distorts the FWCI score of the

majority of researchers.

- Unexplained coding errors produce hugely different scores for researchers with identical profiles.
- In our sample of 90 academics, randomly selected from the Faculty of Health and Life Sciences, we found that FWCI has precisely the same mathematical forecasting ability for research performance as rolling a dice.

The selection of staff for redundancy at the University of Liverpool has been based on a deliberately random process. It is a process has allowed management to unfairly select almost any individual from a redundancy pool that was designed to capture the vast majority of teaching and research staff in the Faculty.

Introduction

Senior management at the University of Liverpool intend to make 47 academic staff members in the Faculty of Health and Life Sciences redundant as part of a restructure titled 'Project SHAPE'. The 47 academic staff members have been selected using the following criteria:

- research grant income;
- a 'Field Weighted Citation Impact' (FWCI) score of < 2.

FWCI is a metric produced by Elsevier, and freely available for use on Elsevier's SciVal platform. It draws data from the Scopus database, which is an Elsevier platform. The FWCI metric is designed to indicate how the number of citations received by a particular entity (a University or a Department within a University) compares with the average number of citations received by other entities indexed in the Scopus database.

FWCI compares the mean citation rate of an entity (for example a University or a Department) in comparison to the global mean. The novelty of FWCI is that it claims that its algorithms weight for different 'Subject' citation rates and therefore that entities can be compared across disciplines.

We deal with the research grant income metric in another paper. In this paper we analyse the FWCI.

The intended purpose of FWCI

FWCI tries to account for citation distribution across large aggregations of research profiles. In this respect, Elsevier guidance for using FWCI notes that when "entities are small.... the metric may fluctuate significantly and appear unstable over time, even when there is complete Scopus coverage." Individual researchers are, by definition, the smallest of entities.

Indeed, Elsevier has said regularly in its training and presentations to users that the FWCI should not be used for "small publication sets (such as a researcher's output)"." This is why it offers no guidance on the use of FWCI to assess individuals. Elsevier only offers guidance on its use to measure: "An institution and departments (Groups of Researchers) within that institution; A country and small research institutes within that country; A geographical region and countries within that region."

Those very significant limitations mean that the use of FWCI should never be used to assess the performance of individuals. This point was confirmed as recently as 3rd March 2021 during a social media discussion between research metrics professionals on FWCI. In this discussion, one of Elsevier's senior data specialists with responsibility for SciVal training said publicly that "FWCI should never be used at an individual level. The number of publications are too small, this will mean that the FWCI value will be meaningless." In other words, the University of Liverpool has generated a series of "meaningless" values and used those meaningless values to assess the worth of its researchers.

The mis-use of FWCI for individual performance

We were first alerted to problems of applying of the FWCI to individual performance in February 2021, after the University of Liverpool informed us of the redundancy selection criteria they had used. We started by testing well-known researchers who could reasonably be expected to be very highly cited in their disciplines, and in global terms. We were surprised by some very low FWCI scores. The majority of the most widely cited researchers we ran through SciVal to test their FWCI score, achieved scores well below the University Liverpool redundancy threshold of 2. They included: Paul Gilroy (cultural studies), Wendy Brown (political science), Slavoj Zizek (Marxism and psychotherapy), Helen King (classics), Noam Chomsky (political theory and linguistics), Judith Butler (sociology), David Harvey and Ron Johnson (both geography). The majority we looked at, including all of those named here had FWCI scores well below 2, and often below 1. At the same time, we tested some very obscure researchers with very low numbers of publications, who are very poorly cited. Some of them had FWCI scores at two or three times the University Liverpool redundancy threshold of 2. This didn't make sense. It looked like the measure didn't work in any predictable way, and often generated the lowest scores for the most widely cited scholars.

We tested this anecdotal evidence with two, more systematic, tests.

- We ran all of the 127 government SAGE advisers who are affiliated to universities through SciVal to generate FWCI scores for the period 2015-2020 (the same period Liverpool used). We found that more than half had scores < 2.

- We ran all Nobel Prize Winners between 2018 and 2020 (a total of 25). We found that 10 (or 40%) had 2015-2020 SciVal scores < 2.

Our evidence to the University of Liverpool has sought answers to those spectacular anomalies.

This rest of this paper uses empirical evidence drawn from the SciVal profiles of researchers at the University of Liverpool to illustrate three major problems: first the fundamental instability of the FWCI as a metric; second a series of visible and 'explainable' coding errors; and third a range of more fundamental and unexplained coding errors.

The problem of instability of the data

When we conducted the exercise noted above, we found that the FWCI was extremely unstable over time. We observed that if we selected different time periods, some profiles generated wildly different scores across those time periods. Here are two examples:

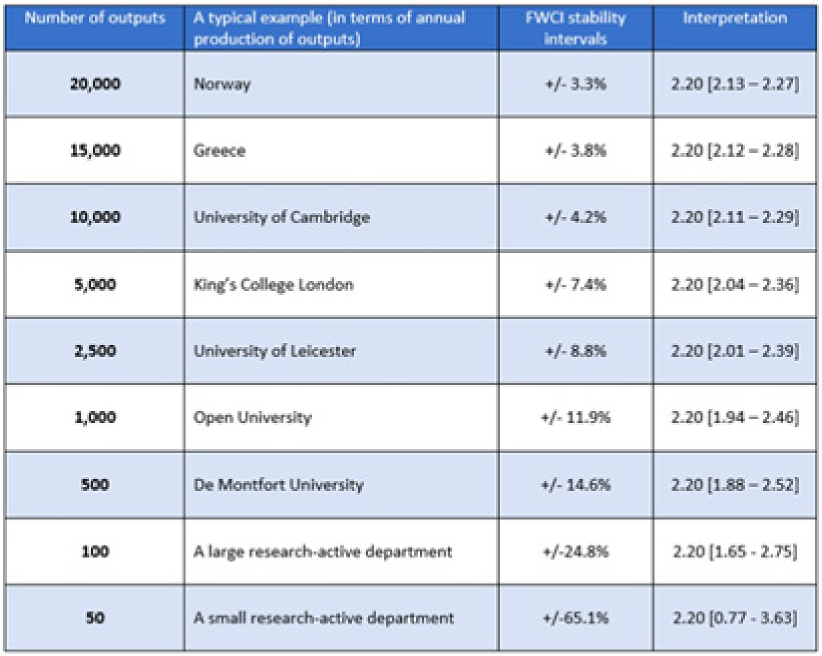
- Professor Stephen Hawking's FWCI score is 0.57 if calculated 2010-2015, but is multiplied 9 times to 5.45 if calculated 2015-2020.
- One Nobel prize winner, Professor Greg Winter of Cambridge, has an FWCI score of 0.70 when calculated 2015-2019; his score is multiplied 13 times to 9.30 when it is calculated 2015-2020.

An analysis published by data specialist Ian Rowlands succinctly identifies a fundamental problem of instability in FWCI data that worsens as sample size declines:

"The trouble is that, typically, the distribution of citations to outputs is highly skewed, with most outputs achieving minimal impact at one end and a small number of extreme statistical outliers at the other... This effect is likely to be more marked the smaller the sample size..."

Rowlands measured stability intervals, using an automated a re-sampling procedure, to test the integrity of the FWCI. The results are set out in figure 1.

Figure 1: Stability intervals in FWCI distribution of citations to outputs



Number of outputs	A typical example (in terms of annual production of outputs)	FWCI stability intervals	Interpretation
20,000	Norway	+/- 3.3%	2.20 [2.13 – 2.27]
15,000	Greece	+/- 3.8%	2.20 [2.12 – 2.28]
10,000	University of Cambridge	+/- 4.2%	2.20 [2.11 – 2.29]
5,000	King's College London	+/- 7.4%	2.20 [2.04 – 2.36]
2,500	University of Leicester	+/- 8.8%	2.20 [2.01 – 2.39]
1,000	Open University	+/- 11.9%	2.20 [1.94 – 2.46]
500	De Montfort University	+/- 14.6%	2.20 [1.88 – 2.52]
100	A large research-active department	+/-24.8%	2.20 [1.65 - 2.75]
50	A small research-active department	+/-65.1%	2.20 [0.77 - 3.63]

It is clear from this analysis that the FWCI becomes highly unstable, to the point that its usability is of little value, even when the unit of analysis or 'entity' is a small research-active Department. For entities smaller that this, FWCI has even less statistical stability.

It is important to reiterate that at no point in its guidance does Elsevier anticipate FWCI being used at the level of individuals, or anything close to this scale of entity. In its guidance, Elsevier further warns against significant fluctuations in the metric, especially in small data sets.

This instability is illustrated when we drill down and look at a single department at the University of Liverpool, as we do in figure 2 below.

Figure 2. FWCI for 90 randomly selected T&R staff in HLS plotted for period 2015-2017 (three years) vs 2018-2020 (three years).

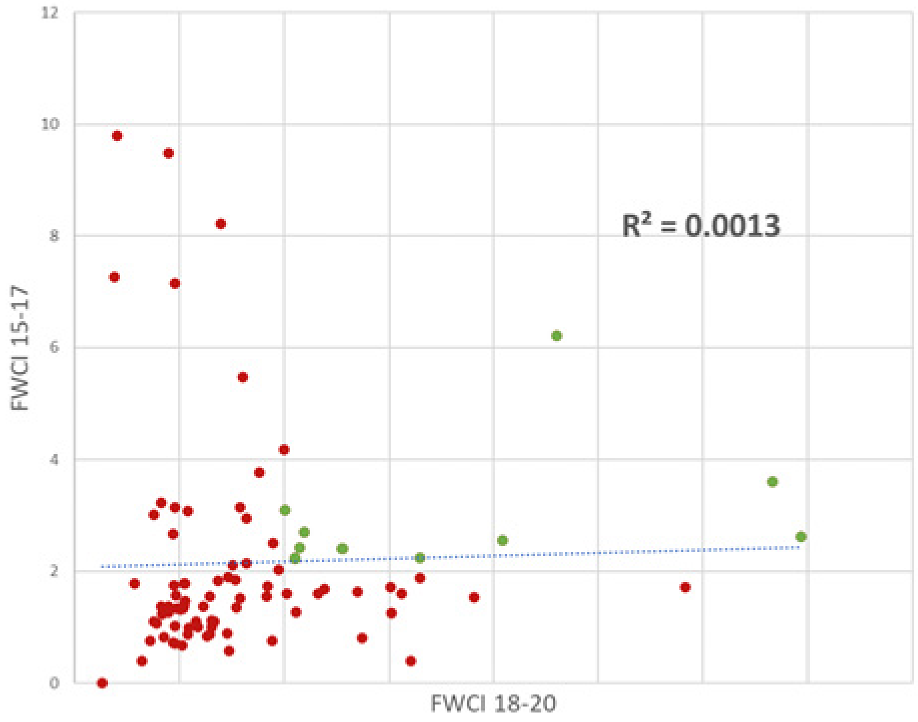


Figure 2 explores the lack of stability in FWCI values for 90 randomly selected HLS staff for two time periods 2015-2017 vs 2018-2020. The plot demonstrates some key points. First, only 11out of 90 staff achieve FWCI > 2 in both periods, further indicating the inappropriate nature of this threshold. Second, there is no correlation at all between the two time points, showing a lack of stability in the measure. If the time points are extended to a four by four year (or longer time period), correlations are still absent. This demonstrates an unfairness in the system – depending on the time period chosen, there is a random effect as to who falls below the line.

Figure 2 shows that the FWCI for individuals has zero forecasting ability. Even if FWCI values reported on contemporary research performance, which as noted above they do not, the lack of forecasting ability is again extremely problematic. If you wish to predict those HLS teaching and research staff with FWCI > 2 for the next period, consider the following. If you take a five- sided dice, rolled it for every staff member and only passed those getting a 'five', you would have the same success rate for predicting future "research performance" as using the current FWCI scores for HLS staff.

Explainable coding errors

There are some coding errors that arise in relation to the correct identification of individuals and publications. This has caused problems, but they are minor compared to the systematic coding problems that we identify in this paper. We are not ignoring coding errors that arise in relation to the identify of individuals and publications, and we know they can skew FWCI scores very significantly. However, we regard those as problems that can be adjusted and fixed relatively easily. We are concerned with a much more significant order of coding error; errors that expose the very significant flaws in the the University of Liverpool's use of FWCI..

i. Missing publications

SciVal's source of raw citation data is the Elsevier database, Scopus. What is crucial here is that Elsevier explicitly acknowledges in its manuals that Scopus coverage is NOT complete. It is also important to note that SciVal official manuals do warn that there can be significant skewing if one publication is missing from a small entity.

"Care is advised, however, when comparing small entities, from which a single missing publication may have a significant negative impact; for example, an academic's performance may suffer due to gaps in database coverage of their portfolio, as well as gaps in publications citing those items of the portfolio that are indexed."

We sampled nine researchers, across three faculties at the University of Liverpool to explore the accuracy of Scopus profiles, and the process by which SciVal uses data from Scopus.

In our analysis, we found that all nine reported significant publications missing from their profiles. There was a total of 15 refereed journal articles missing from those nine profiles; two thirds (n=6) had refereed journal articles omitted from their profiles; and two had a sole-authored book omitted from their profiles.

There appears to be no pattern in the missing data here. The missing work includes papers published in the most established and leading journals in those researchers' disciplines.

In other words, there are coding errors in FWCI that mean that some of the most important academic work is likely to be missing from any given researchers' profile.

ii. Low quality publications

The algorithms that scrape the data for SciVal profiles also apparently include low quality publications that can significantly skew citation scores.

The issue here is that the pre-built filters in SciVal do not have the capacity to filter in high quality publications. The pre-built filters, for example, cannot produce a profile of refereed journal articles. Neither is there an option in pre-built filters to 'filter out' publications like magazine articles and opinion pieces. The later types of publication remain in the pre-built 'Article' filter, buried in the same category as refereed journal articles.

This is important because publications not intended for an exclusively academic audience such as magazine articles, opinion pieces, pamphlets, etc. are not written to be cited, and are produced with the understanding that they are not likely to be cited at all (often in the deliberate expectation they are likely to generate zero citations). They therefore may drag scores down, regardless of field-weighting. In our sample of nine researchers, we found that four had publications that significantly pulled their citation score down, including one which was a two-page advertisement for an academic conference in a professional journal.

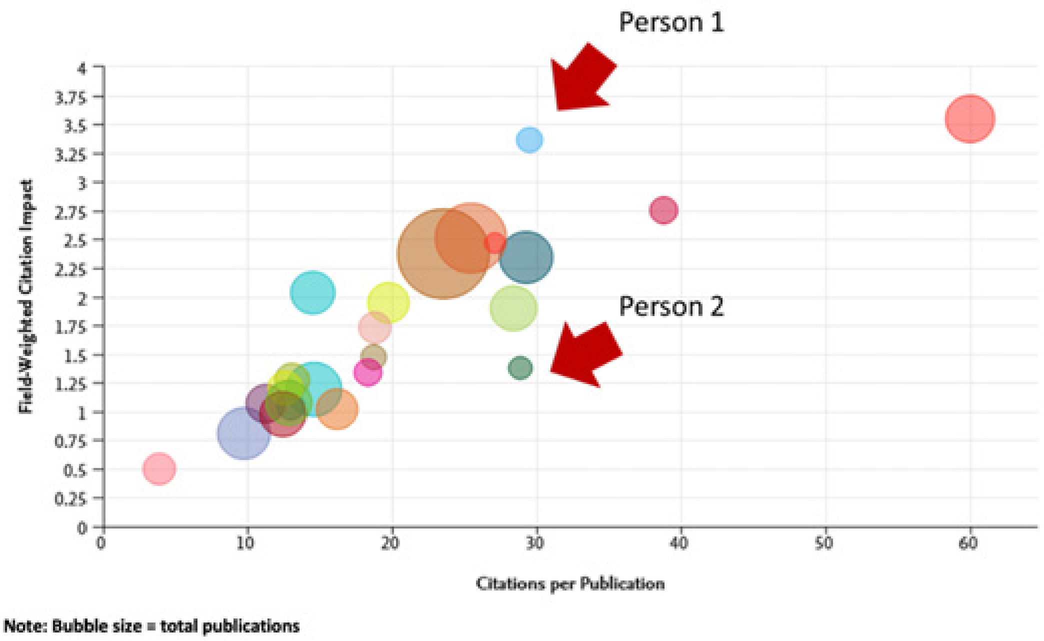
This has a profound impact on FWCI scores. Our evidence showed that prior to cleaning out magazine articles, one researcher at the University of Liverpool had an FWCI score of 1.20, and after the cleaning process, had an FWCI score of 2.00.

In other words, there are coding flaws relating to the inclusion of publications which can significantly impact on a given researcher's score.

Unexplained coding errors

Figure 3 identifies a major difference in FWCI scores between two researchers with almost identical profiles.

Figure 3: Two researchers working in exactly same subject, plotting FWCI 2015-2020 vs citations per publication.



Both person 1 and person 2 work in the same field and publish in the same journals. They have almost identical citations per publication in their profiles and the strongest paper of each is published in the same journal. Person 1 has 29.5 citations per paper and has a FWCI score of 3.4, yet Person 2 has 28.8 citations per paper and has an FWCI score of 1.4.

On more detailed inspection it appears that the difference between Person 1 and Person 2 can largely be put down to a variation in FWCI scores for two papers published by each of those researchers. Those papers are the top cited papers in each of those researchers' profiles, and happen to be published in the same academic journal. Person 1's paper has 102 citations and an FWCI score of 19.1. Person 2's paper has 129 citations and an FWCI score of 3.34. It is the difference across those papers – more than any other difference – that accounts for the difference in their 'overall' score profiles. This difference (which ensures that one can reach the University of Liverpool threshold, whilst the other can fall well below) is as yet unexplained.

Yet those differences are significant enough to put one researcher clearly inside the University of Liverpool redundancy zone, whilst the other is well outside the redundancy redundancy zone.

We stress that this exemplar is not an anomaly or 'one-off'; similarly distorted FWCI pairings can be found for most researchers in SciVal.

Conclusion

The empirical evidence presented in this paper has been given in further detail to the University of Liverpool during the negotiations over the redundancy selection process. This evidence shows very precisely that they have generated a series of meaningless values and used these to assess the worth of its researchers. The University of Liverpool has so far failed to admit that it has used FWCI in a way that contravenes manufacturers' instructions and has caused scandal in the data science community.

It is a process that would be comical if these redundancy proposals had not been so traumatic and stressful for so many. There are many jobs and researcher's livelihoods at stake, not to mention the livelihoods of their families. In the absence of answers that explain the skewing and the corruption in the data, they need to know why the University is still applying a metric that is palpably meaningless when it is applied to measuring the research performance of individuals.

University of Liverpool managers must withdraw their metric-based redundancy proposal immediately. Whilst we have presented a deep analysis of FWCI here, fundamental flaws can be found in any metric. Indeed, our evidence on the use of research income as a metric, set out in another paper, reveals a similar order of mathematical flaws.

No metric could ever capture the richness of the contribution made by a teacher and a researcher to the life of the University and the community.

UCU University of Liverpool Branch, April 2021