Journal of Informetrics 11 (2017) 937–940

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Correspondence

When social scientists disagree: Comments on the Butler-van den Besselaar debate

1. Introduction

In 2003, Linda Butler found evidence that as a result of "the increased culture of evaluation faced by the [Australian higher education] sector . . . [in which] significant funds are distributed to universities, and within universities, on the basis of aggregate publication counts, with little attention paid to the impact or quality of that output . . . journal publication productivity has increased significantly in the last decade [the 1990s], but its impact has declined" (Butler, 2003a, p. 143). More recently, however, Peter van den Besselaar, Ulf Heyman and Ulf Sandström (hereafter BHS) have concluded from their bibliometric analysis that "Australia not only improved its share of research output but also increased research quality, implying that total impact was greatly increased", and hence "Butler's main conclusions are not correct" (van den Besselaar et al., 2017). How can we explain this disagreement?

In the social sciences, it is not so uncommon to find similar studies on supposedly the same issue coming to diametrically opposed conclusions. For example, during the early years of science policy research, a battle raged between two competing models of innovation, the 'science-push' linear model espoused by Vannevar Bush and senior scientists, and the 'demand-pull' model proposed by certain economists. In Project Hindsight commissioned by the US Department of Defense, an analysis of the critical research events that made possible 20 major military innovations concluded that 95% of them were motivated by a recognised defence need – i.e. they were the result of 'demand-pull' (Sherwin & Isenson, 1967). A year later, however, a rival study named TRACES commissioned by the US National Science Foundation, which adopted a broadly similar approach in examining the critical research events that made possible five major civilian innovations, found that 70% of them were 'non-mission-oriented' – i.e. 'science-push' was much more important (IIT, 1968). How could two similar studies arrive at such diametrically opposed conclusions? Closer inspection revealed one key methodological difference. Whereas Project Hindsight restricted its attention to critical research events over the preceding 20 years, TRACES traced the research origins of its innovations back 100 years. The further back one went, the more likely those critical research events were to be of the 'science-push' type.

A second example involved a bibliometric debate. During the 1980s, I and colleagues at SPRU published a number of articles reporting bibliometric analyses showing the UK's world share of scientific publications was declining (Irvine et al., 1985; Martin et al., 1984; Martin et al., 1987). In contrast, Loet Leydesdorff, from his bibliometric analysis, claimed that "British scientific output was relatively stable over the 1970s, then showed a remarkable increase from 1981 onwards" (Leydesdorff, 1988, p. 149). A joint working party was set up by Cees le Pair (Director of the Dutch Technology Foundation, STW) to identify the reasons for these very different findings. This concluded that Leydesdorff's use of 'whole counting' failed to take account of the fact that, with this particular indicator, virtually all countries' shares were increasing (because of growing international collaboration) while his use of on-line bibliometric searching (then still in a fairly primitive state, involving the use of Lockheed's DIALOG Online Information Service) introduced various errors (e.g. double counting publications co-authored by researchers in two or more of the four home countries of the UK; accidental inclusion of papers with 'New England' or 'New South Wales' in the address as part of the UK) (Anderson et al., 1988).

Differences between social scientists can arise because of using different data (part of the explanation for the differences between SPRU and Leydesdorff), not least because in the social science realm all indicators are imperfect or partial to a greater or lesser extent. Differences can also arise because of adopting different methodologies (a feature in both the examples above), since a given methodology generally contains somewhat arbitrary assumptions or choices. Differences can also arise because of the adoption of a different conceptual framework or theory.

In what follows, I explore the differences between the findings of Butler (2003a) and van den Besselaar et al., 2017 and the apparent reasons for those. Section 2 examines the original Butler paper and Section 3 the paper by BHS. Section 4 tries to clarify the main issue in contention between the two, namely exactly when the effects of performance-based funding might be expected to show up in the Australian bibliometric data, while Section 5 summarises the main conclusions.

http://dx.doi.org/10.1016/j.joi.2017.05.021
1751-1577/

2. Butler's 2003 study

It is worth re-reading this 2003 paper as it has several noteworthy features. First, it describes an explicit attempt to relate bibliometric indicators to a specific policy question – namely, the effect of a funding system based partly on publication numbers. Secondly, it shows careful awareness of the limitations of indicators – for example, that citations are at best an indication of impact, not quality. Thirdly, the research design exhibits a systematic and rigorous approach as to possible explanations of the trends exhibited by the data.

Here, a little background context is needed. In the late 1980s and early 1990s, Australian university research faced funding cuts and policy changes, with the situation being portrayed as a state of 'crisis'. One might therefore have expected to see this crisis reflected in a fall in publication output. In fact, the data revealed the opposite. After declining in the 1980s, there was a rise in published output in the 1990s. The paper set out to explore this surprising finding. Besides analysing Australia's share of world publication and citation totals, Butler also looked at relative citation impact (RCI) and the changes in the impact of journals in which Australian academics were publishing. In terms of RCI, Australia fell from sixth to tenth position out of 11 OECD countries over the period from 1981 to 1999. And in a more detailed analysis in which journals were divided into four quartiles based on their citation impact, Butler (2002, 2003b) showed that, although Australia's share of the top two quartiles increased by around 28% and 15% respectively over the 1990s, the share of the third quartile jumped by 55% and that for the lowest quartile doubled.

Butler explored several possible explanations for this increase in publications but decline in relative citation impact, including increased rates of international collaboration, the entry of new universities (former colleges), and an increase in the number of university researchers, but these could at best only explain a small fraction of the effect. The most plausible cause instead seemed to be the increased emphasis on evaluation and publications in Australian universities. This began with the Linke report in 1991, which proposed the use of performance indicators including publications. From 1992 all Australian universities were required to supply publication data to the Australian Vice Chancellors' Committee as it became clear that publication data would shortly become incorporated in calculations of the Research Quantum, the mechanism used for distributing core funds across universities. Butler therefore suggested that these changes might have encouraged the publication of more articles during the 1990s, but with a growing percentage appearing in lower impact journals.

To explore this hypothesis further, she carried out case-studies of two universities, Queensland and Western Australia. These had pursued very different management strategies, with the former focussing especially on publication output. More detailed bibliometric analysis for the two institutions provided support for the suggestion that an emphasis on publication output would tend to increase this but at a cost of reduced relative citation impact.

Butler therefore concluded that the increased use of evaluation and bibliometric indicators "appears to be altering researchers' publication habits. . . . What is of concern for Australia is that while journal output has grown rapidly, it is increasingly appearing in lower impact journals . . . [which] raises important questions on the wisdom of a policy that rewards quantity, with scant regard to quality" (Butler, 2003a, p. 154). As BHS note, Butler's paper has subsequently been widely cited as providing empirical evidence of how evaluation and bibliometric indicators may affect the behaviour of researchers, not always as intended.

3. The BHS 2017 study

BHS set out to "redo and extend the [Butler] analysis" (van den Besselaar et al., 2017). Some of the indicators they use are the same; for example, Figure 1 shows very similar results to Butler – with an increase in Australia's publication and citation shares in the 1990s. In addition, BHS look at highly cited papers, finding that Australia produced more of these during the 1990s. They note that "despite the strongly increasing share of low impact journals in the total Australian output [i.e. confirming Butler's main finding], the average impact still increased, as did the number of highly cited papers" (van den Besselaar et al., 2017).

However, the paper's main criticism of Butler is that the effects of increasing numbers of publications and increasing share of lower impact journals came too soon for this to have been caused by changes in the funding system. According to van den Besselaar et al. (2017), performance-based funding of Australian universities only came into effect in 1995 so "the full impact cannot be expected before around 1998". I return to consider this in more detail in Section 4.

Several other observations can be made about the BHS paper. First, given that the authors are experienced bibliometricians well aware that citation data relate to impact not quality, it is surprising to find references being made to "less good papers" and Australian science "losing quality". Secondly, in several places BHS overstate Butler's position, for example saying she claimed that "Australian science policy in the early 1990s made a mistake by introducing output based funding" (ibid.), when her more nuanced conclusion was that her study "raised important questions on the wisdom of a policy that rewards quantity, with scant regard to quality" (Butler, 2003a, p. 154).

Thirdly, in two places, BHS refer to a book by Simonton 2004, citing this as the source for "creativity theory", in which (according to BHS) higher publication output is associated with "higher quality". There are two points to note here. One is that Simonton's research on creativity deals with the work of individuals rather than that of large collectives like countries. The other (and more important) point is that Simonton is actually associated with 'the equal-odds rule', reflecting the fact that "the ratio of high-impact publications to total output – the hit rate – is uncorrelated with total output" (Simonton, 2004, p. 22); in other words, the exact opposite of what BHS suggest.

Lastly, BHS carry out a 'test' of salami-publishing. Amongst other things, they assume that salami publishing will show up as a decrease in the average word length of papers (one might equally well assume that authors may try to obscure the fact they are engaging in salami-publishing by padding out their papers), and that a small sample of just 12 journals is sufficient to test this. Consequently, the test is far from convincing.

In short, there is a certain lack of care that is worrying to find in a replication study claiming to prove an earlier study was incorrect.

4. Timeline for policy changes in Australia

In 1990, Australia introduced the Research Quantum mechanism for distributing institutional core funding to universities. Initially this was based on research grant data. However, the following year a report was published reviewing performance indicators and proposing the use of publication and other performance-based indicators (Linke, 1991). In 1992, Australian universities began supplying research output data to the Australian Vice Chancellors' Committee (AVCC). In 1993, ACCC set up a working party to develop new indicators, and that same year AVCC and universities began lobbying for the publication component of the Research Quantum to be increased to 50%. Moreover, by 1995 concerns were already emerging about universities giving excessive weight to publication data.

It is Butler's contention that the response of universities and their faculty to a funding mechanism based partly on publication numbers "started as soon as the universities knew what was coming – by 1993", whereas van den Besselaar et al. (2017) argue that "the full impact cannot be expected before around 1998". In the light of the above timeline, the former date would seem much more plausible.[1] The imminent arrival of a publication-based funding system seems likely to have influenced the publication behaviour of Australian academics from around 1993, contributing to an increase in published output over the 1990s but with a corresponding increase in the share of articles appearing in lower impact journals. Nevertheless, a small puzzle still remains, since both the Butler paper and the BHS paper (see Figure 1 in each) show Australia's rise in publications seems to have begun around 1990, while the former shows that Australia's RCI apparently began to fall after 1988 (Butler, 2003a, Figure 4). This would suggest that some other force may have been at work in the years immediately before 1993.

5. Conclusions

This debate illustrates the importance of replication in research, especially for studies with a significant policy impact. However, as we have seen, researchers can arrive at very different conclusions. Replication requires careful attention to methodological detail.

The debate also points to the need for deep knowledge of the prevailing policy context in order to draw valid policy conclusions. In this case, close examination of the policy context leaves one more inclined to accept Butler's posited date for the likely effect to begin to appear (i.e. 1993) than that of BHS (i.e. 1998). However, the challenge posed by BHS highlights that the rise in Australian publications and decline in relative citation impact each seemingly began a little before 1993, suggesting that, during this period at least, forces other than a funding formula based partly on publication counts may have been at work.

Declaration of interest

I worked with Linda Butler and her colleague, Professor Paul Bourke, during the 1980s and 1990s. I am also an Editor of *Research Policy*, the journal in which her 2003 paper was published (although I was not an editor at that time and did not handle the paper). In addition, I worked with Professor Peter van den Besselaar on a European Science Foundation project during 2009–10, and co-authored a report with him in 2010.

References

Anderson, J., Collins, P. M. D., Irvine, J., Isard, P. A., Martin, B. R., Narin, F., et al. (1988). On-line approaches to measuring national scientific output – A cautionary tale. *Science and Public Policy*, *15*, 153–161.
Butler, L. (2002). A list of published papers is no measure of value. *Nature*, *419*, 878.
Butler, L. (2003a). Explaining Australia's increased share of ISI publications – The effects of a funding formula based on publication counts. *Research Policy*, *32*, 143–155.
Butler, L. (2003b). Modifying publication practices in response to funding formulas. *Research Evaluation*, *12*, 39–46.
IIT. (1968). *Technology in retrospect and critical events in science (Project TRACES)*. Chicago: Illinois Institute of Technology (IIT) Research Institute.
Irvine, J., Martin, B. R., Peacock, T., & Turner, R. (1985). Charting the decline in British science. *Nature*, *316*, 587–590.
Leydesdorff, L. (1988). Problems with the 'measurement' of national scientific performance. *Science and Public Policy*, *15*, 149–152.
Linke, R. D. (1991). *Performance indicators in higher education*. Canberra: Australian Government Publishing Service.
Martin, B. R., Irvine, J., & Turner, R. (1984). The writing on the wall for British science'. *New Scientist*, *104*(1429), 25–29.

[1] It is also significant that Linda Butler and her former colleague, Professor Paul Bourke, were very much at the centre of Australian science policy in the late 1980s and early 1990s, and therefore likely to be better informed about the above policy changes and the timing of their effects.

Martin, B. R., Irvine, J., Narin, F., & Sterritt, C. (1987). The continuing decline of British science'. *Nature*, *330*, 123–126.
Sherwin, C. W., Isenson, R. S., Sherwin, C. W., & Isenson, R. S. (1967). Project Hindsight: A defense department study of the utility of research. *Science*, *156*, 1571–1577.
Simonton, D. K. (2004). *Creativity in Science: Chance, logic, genius, and zeitgeist*. Cambridge: Cambridge University Press.
van den Besselaar, P., Heyman, U., & Sandström, U. (2017). Perverse effects of output-based research funding? Butler's Australian case revisited. *Journal of Informetrics*.

Ben R. Martin
SPRU, University of Sussex, United Kingdom
E-mail address: B.Martin@sussex.ac.uk
Available online 21 June 2017