Journal of Informetrics 11 (2017) 923–926

Contents lists available at ScienceDirect

Journal of Informetrics

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

Correspondence

Some considerations about causes and effects in studies of performance-based research funding systems

1. Introduction

We thank the Editor for providing us with the opportunity to reflect and comment upon the critique of Linda Butler's analyses of the potential effects of the Australian performance-based funding system (PRFS) in the article by Van den Besselaar, Heyman, and Sandström (2017; hereafter van den Besselaar et al.) in this special section of Journal of Informetrics.

In 2010 Butler wrote that "*[a]ssessing the impact of performance-based research funding systems (PRFS) is a fraught exercise, which perhaps explains the paucity of broad authoritative texts on the subject*" (Butler, 2010, p. 128). As it is well known to most readers of this journal, Butler herself has conducted one of the few authoritative studies on this subject (documented e.g. in Butler, 2002, 2003a, 2003b, 2004). Her study has since then not only been an almost mandatory reference point for other articles on this subject. It has also influenced both policy discussions and designs of PRFS around the world. In this light the present article by van den Besselaar et al. should be welcomed as it gives us an occasion to revisit and discuss Butler's seminal work. However, as Butler also highlights in her quote above, examining effects of PRFS is an extremely challenging task. These challenges will be our main focus in this commentary, where some of the problems associated with attributing behavioral causality to PRFS are discussed. First, we outline some basic conditions which should be met in order to attribute causality to PRFS. Secondly, we compare the approaches of Butler and van den Besselaar et al. with regard to these issues. Finally, we round off with a few concluding remarks

2. Attributing causality to PRFS

In several of her publications, Butler uses causal language suggesting that the Australian "funding formula" and its implementation *caused* pivotal changes in publication behavior among university researchers. The result was higher publication activity in general, but noticeably, the increase was relatively strongest in journals with lower Journal Impact Factors (e.g., Butler, 2003a, 2004). According to Butler the direct consequence was a corresponding decline in overall Australian citation impact. In their article, van den Besselaar et al. also make causal claims, but these are partly opposite to Butlers. The authors claim that the Australian funding model did indeed *cause* higher productivity, but that the productivity increase eventually led to higher national citation impact instead of a decline. Both parties acknowledge the problems of attributing causality in a strict sense, but the causal claims are nevertheless prominent throughout the contributions.

This is however not particular for these two studies. Causal language and corresponding claims are rife in the social sciences, giving the impression that many causal relationships are well established and seemingly straightforward to document using various designs and modelling techniques. But in reality the examination of causal effects from non-experimental data poses immense challenges and requires a lot of ingenious work. In the following we will for both the study of Butler and the study of van den Besselaar et al. discuss some of the minimum conditions for causal claims to be valid, including the presence of precedence, correlation, and non-spuriousness. In relation to PRFS there are in most cases fundamental challenges to all of the three conditions outlined above.

2.1. Precedence

The first condition concerns the simple fact that cause needs to precede effect. But even such a basic condition can sometimes be surprisingly difficult to substantiate. The question of when exactly a specific PRFS *de facto* took effect is often ambiguous due to the way in which the long and complex multi-level implementation processes play out. Also the question of time lags is much more complicated than often assumed. While it may be reasonable to operate with a substantial time lag in relation to project funding, the situation is less straightforward with regard to PRFS. Institutions and individuals can in this situation change their behavior more or less from day to day. Outlet strategies can be altered until the last day before

http://dx.doi.org/10.1016/j.joi.2017.05.018
1751-1577/© 2017 Elsevier Ltd. All rights reserved.

submission, and old drafts which were forgotten in the drawer can suddenly be transformed into submissions if the situation requires it. Behavior can even change before the cause has been implemented based on expectations.

Also the second condition concerning **correlation** can be challenging when we use bibliometric indicators as the dependent variable. There are a number of issues which should be considered here. One has to do with the fact that we are working with dynamic databases characterized by strong growth over time in number of journals, number of articles within journals, and number of references and citations attributed to each article. These developments severely challenge the interpretation of time series. Another factor has to do with the effects of increased internationalization over time in terms of co-authorships, which results in auto correlation between the developments of individual countries. Finally, a third factor relates to the vast degrees of freedom with regard to selection of indicators, time-periods, length of citation windows, counting methods, choice of database etc. When effects are big, these choices may not alter the overall picture much, but when effects are more modest, these effects may in reality be artefacts of methodological choices rather than effects of the PRFS under examination. In some cases there may not be much choice for the individual researcher (when InCites® or other applications are used), but the effects of the choices made by the companies behind the databases may still be equally important.

The third and final factor, **non-spuriousness**, relates to the question of how to isolate the effects of a specific PRFS in a non-experimental setting, when everything else is in constant flux. The long time-frames necessary to be able to detect any noteworthy patterns make it almost impossible to hold anything constant, neither by design or statistics. Deep contextual knowledge is therefore necessary to interpret the bibliometric macro data as a number of potential confounding variables always need to be considered; although the number of these essentially is unknown. This is in particular the case when assessments of PRFS have the focus on macro-trends only; i.e. where causality is claimed from the macro-intervention to macro-effects. The actual behavior we aim to explain, is however taking place at the *meso*- and micro-levels: at different institutions where the intervention may have been translated into local policies in various ways; within different scientific fields and different researcher cohorts with different publication cultures and incentive structures; and under very different contextual conditions. To assume that a macro-intervention will influence all actors uniformly is thus highly unrealistic.

2.2. Implications for studies of PRFS

Studies of PRFS accordingly have to be extremely carefully designed in order to minimize the effects of these and other challenges, and considerable caution has to be shown when conclusions are drawn based on imperfect data, under-developed concepts, weak causal designs and uncertain assumptions. How robust the results of a particular study are, thus depends on the design, the data and the methodological choices made along the way. The next section discusses the way in which Butler and van den Besselaar et al. have dealt with these issues.

3. A comparison of the approaches of Butler and van den Besselaar et al.

Based on these general challenges to the assessments of PRFS, it is no surprise that the studies of both Butler and van den Besselaar et al. must be considered weak from a strict research design perspective with regard to their ability to attribute causality. A "strong" design-based approach would systematically model conditions of causality such as correlation, precedence and non-spuriousness in a "controlled" setting with proper comparison groups enabling counterfactual analyses between them.

While weak compared to an ideal situation, Butler's study is, however, relatively strong compared to other studies of PRFS in general – and to the study of van den Besselaar et al. in particular. This is especially the case with regard to the question of **non-spuriousness**. Butler examines her presented developments in a (more or less) orderly way. She examines some potential spurious relations (e.g., influx of research staff to the system), disaggregates outcomes to main fields and presents and discusses suitable controls or "counterfactual-like" units of analysis (e.g., sector research institutions not included in the PRFS, and universities with different policies towards publication rewards and hiring). By disaggregating and isolating her data Butler thus makes a fairly convincing comparative analysis. Another strong feature of Butler's study is her extensive contextual and historical knowledge. This is imperative for a solid analysis.

The causal design in the study by van den Besselaar et al., on the other hand, is clearly inferior to Butler's with regard to this factor. The claim about higher productivity eventually leading to higher impact therefore merely ends up as a postulate. Linking the claim to Simonton's "theory" of creativity constitutes no causal evidence whatsoever (Simonton, 2004). In comparison to Butler's design there are two major differences: First, the study by van den Besselaar et al. is a pure macro–macro design, in contrast to Butler who uses the *meso*-level to substantiate her claims. The inclusion of the *meso*-level is essentially what enables the counterfactual analysis in her case. Secondly, the contextual knowledge in the study by van den Besselaar et al. appears to be largely missing. This also weakens their claims substantially.

There is however one point in the critique of Butler which deserves further scrutiny. Van den Besselaar et al. raise a central question in relation to **precedence**: When did the effect(s) actually set in? As argued above, precedence is crucial to causal analyses, i.e. the time order between the cause and its potential effect. Both pre-implementation expectations and/or lag of effects can however make it very difficult to disentangle and isolate the "real" effects of a PRFS. Indeed, both Butler and van den Besselaar et al. are very aware of this issue, but they nevertheless disagree substantially in their assumptions. Butler argues for the effects of an expectation period previous to the actual implementation in 1996, while van den Besselaar et al. argue for the exact opposite: that the effects of the PRFS would be unlikely to materialize before a lag period of 1–2 years from

the formal implementation in 1996 – in other words around 1998. This pivotal question in the interpretation of the Australian case can however only be settled by persons with deep contextual knowledge of the Australian system. To our knowledge Butlers assumption has not been questioned prior to this, which in addition to her extensive contextual knowledge weighs in favor of her interpretation. But only other persons with the same type of contextual insight will ultimately be able to judge the soundness of this assumption.

However, even if we accept Butler's argument about the effects starting to materialize prior to the formal implementation, the case is not clear cut. The third issue regarding **correlation** raises some important questions for both studies. With regard to the effects it is important to notice that both Butler and van den Besselaar et al. operate with first and second-order effects of the PRFS. The first order effects concern productivity and here both parties seem to agree that the intervention led to higher productivity. They disagree however with regard to the second order effects. Butler concludes that the higher productivity in particular materializes in lower impact journals which then lead on to a general drop in Australian impact. Van den Besselaar et al. on the other hand conclude that the higher productivity eventually leads to an increase in Australian impact in the beginning of the 2000s. The interpretation of such causal claims is obviously highly dependent upon exact time of treatment and the assumption of no confounding factors as discussed above. The interpretation is however equally dependent on the actual effects sizes of both the changes in publication output and eventual citation impact in the bibliometric data. These sizes are again dependent upon the way in which the actual time series are constructed.

As argued above the construction of such time series is susceptible to considerable researcher degrees-of-freedom. Both Butler and van den Besselaar et al. discuss differences and defend their actual choices. In this case it is however important to notice that the relative effect sizes, upon which the conclusions are drawn, appear modest. In other words, the data do not speak clearly for themselves. Furthermore, when we scrutinize the time series produced by Butler and van den Besselaar et al. (and some reproduced by ourselves in Schneider, Aagaard, and Bloch (2016), it is possible to point at some longer trends that do not seem to corroborate to the narratives of the "sudden" impact on behavior as the result of the intervention, neither in Butler's version nor in the version of van den Besselaar et al.

Eventually this accordingly boils down to interpretations of time series which by no means are the products of an exact science (due to arbitrary counting methods, database differences, the dynamic character of these, language biases etc.). If we look at our own time series in Schneider et al. (2016) based on individual publication data (which also are the product of a number of methodological choices) some clear trends are visible. There appears to be an increase in publication activity in all journal classes, but it apparently sets in before Butler's treatment demarcation. Furthermore, the general journal publication behavior from 1980 onwards measured by the Mean Normalized Journal Score (MNJS) Waltman, van Eck, van Leeuwen, Visser, and van Raan (2011) shows a decline during the 1980s, a stabilizing period in the 1990s and then a rise in the 2000s. This trend correlates very well with the actual mean normalized citation score (MNCS) for the Australian publications shown in our article. Our time series thus opens up for a third interpretation: That the introduction of the Australian PRFS in reality did not matter much in either direction. The observed development may have been well under way even before Butler's claimed effects could have set in.

Obviously, we would never claim this interpretation to be valid based on such a superficial analysis. Neither would we claim that our time series are more "correct" than the others (universities are for instance not separated from other sectors in our data). But the example serves to show that even minor differences in technical, methodological choices open up for quite different interpretations when effect sizes are modest. Hence, even if there were agreement on the issue of when real effects could be expected, there could be disagreement based on methodology. Ultimately, one interpretation may be valid at the macro level while others are more fitting at selected lower levels. The central point is thus that we are inferring between levels here. At the macro-level the effect may be marginal and influenced by many other factors. Yet, at *meso*- and micro-levels the PRFS can have important consequences, albeit only for selected individuals or groups, and this may not necessarily be something that is immediately readable from the macro statistics. We document such varying treatment effects in our analysis of the Norwegian PRFS (e.g. Aagaard, 2015; Schneider et al., 2016).

4. Concluding remarks

Among the two studies under examination here, Butler's stands out with the strongest design and the most convincing in-depth contextual knowledge. The study by van den Besselaar et al. on the other hand is lacking in both respects. Design and contextual knowledge do not, however, do the job alone. The data also need to show a clear effect, and here we are not fully convinced by any of the causal claims of the two studies. The observable effects are at best modest in both. Furthermore, if we look at different time series, it appears to become more uncertain whether there is a particular effect right after Butler's demarcation line. Finally, in both studies there is hardly any substantiation of the actual causal mechanisms at the micro level. We fully acknowledge however that many of the same points of critique rightfully can be applied to our own study of the Norwegian PRFS (Schneider et al., 2016). It just underlines the initial quote from Butler that to examine effects of PRFS is a fraught exercise.

But while we will never see a perfect study of these issues, it is still highly relevant to study them at the best of our ability – and equally important to discuss the results as done in this special section of *Journal of Informetrics*. When we do so, we should keep renowned statistician David A. Freedman's advice in mind:

> "Causal inferences can be drawn from nonexperimental data. However, no mechanical rules can be laid down for the activity. (. . .) Instead, causal inference seems to require an enormous investment of skill, intelligence, and hard work. Many convergent lines of evidence must be developed. Natural variation needs to be identified and exploited. Data must be collected. Confounders need to be considered. Alternative explanations have to be exhaustively tested" (Freedman, 2005, p. 1071).

References

Aagaard, K. (2015). How incentives trickle down: local use of a national bibliometric indicator system. *Science and Public Policy*, *42*(5), 725–737.

Butler, L. (2002). 'A list of published papers is no measure of Value—The present system rewards quantity, not quality – but hasty changes could be as bad'. *Nature*, *419*(6910), 877.

Butler, L. (2003a). Explaining Australia's increased share of ISI Publications—The effects of a funding formula based on publication counts'. *Research Policy*, *32*(1), 143–155.

Butler, L. (2003b). 'Modifying publication practices in response to funding formulas'. *Research Evaluation*, *12*(1), 39–46.

Butler, L. (2004). 'What happens when funding is linked to publication counts?'. In H. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research*. Netherlands: Kluwer Academic Publishers [pp. 389–405].

Butler, L. (2010). 'Impacts of performance-based research funding systems: a review of the concerns and the evidence'. In *Performance-based funding for public research in tertiary education institutions* [pp. 127-65].

Freedman, David A. (2005). Linear statistical models for causation: a critical review. In B. S. Everitt, & D. C. Howell (Eds.), *Encyclopedia of statistics in behavioral sciences* (Vol. 2) (pp. 1061–1073). Chichester, UK: John Wiley & Sons.

Schneider, J. W., Aagaard, K., & Bloch, C. W. (2016). What happens when national research funding is linked to differentiated publication counts? A comparison of the Australian and Norwegian publication-based funding models. *Research Evaluation*, *25*(3), 244–256.

Simonton, D. K. (2004). *Creativity in science: Chance, logic, genius, and zeitgeist*. Cambridge, UK: 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*.

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.

Kaare Aagaard *
Jesper W. Schneider
Danish Centre for Studies in Research and Research Policy (CFA), Department of Political Science, Aarhus University, Bartholins Allé 7, 8000 C, Aarhus, Denmark

* Corresponding author.
E-mail address: jws@ps.au.dk (K. Aagaard)

Available online 29 June 2017