Ranking Academic Research Performance: A Recipe for Success?

Ruth Dixon and Christopher Hood
University of Oxford

Accepted for publication in Sociologie du travail (2016)

Abstract

Using the example of a governance system that allocates public funding for research on the basis of rankings of research quality and impact (as has been developed in the UK over the past three decades), this paper explores three conditions needed for such rankings to be effective as a basis for genuine performance improvement over time. First, the underlying metrics must be capable of meaningfully distinguishing the performance of the institutions being ranked. Second, the basis of assessment must be stable enough for changes in performance over time to be identified. Third, the ranking system should avoid perverse consequences arising from strategic responses by the institutions being assessed. By means of a hypothetical example of a series of research assessment exercises, this article demonstrates the difficulty of fulfilling all three conditions at the same time, and highlights the dilemma between reliability and validity that assessors face. This analysis is relevant to governance by indicators more broadly, because any comparative assessment of institutional performance faces similar issues.

Key words (6 to 8)
Research assessment exercise
University rankings
Governance
Metrics
Unintended consequences
Validity
Reliability
Research impact

Article

1. University research rankings

If you can rank Olympic sports teams and restaurants, what could be the objection to doing so for university research? What could be wrong with replacing the tacit or implicit qualitative knowledge about the performance of scholars, research units, or academic journals that everyone worked on forty years ago with a modern system of precise performance metrics based on clear criteria? And who could object to a governance system that bases public research funding on such indicators?

After all, we rank many things other than sports teams and restaurants today. The ‘governance by indicators’ phenomenon highlighted in this issue has even extended to rankings of ‘governance’ itself over the past four decades by bodies like the World Bank and Transparency International, and if we can meaningfully rank something as complex as that, what could be the objection to applying the same method to the quality and impact of academic research? Aren't rankings a proven way to harness the power of competition to raise effort and reward success, keep everyone on their toes, and make researchers work ever harder to out-do their peers? Doesn't such competition make   the whole society better off as a result of the higher quality, higher-impact research universities produce? And don't rankings go particularly well with the grain of the hyper-competitive culture of the world of academic research, where (almost) everyone loves to rank everyone else in their field, and gossip endlessly about who's up, who's down, and who's better than whom?

That is the thinking behind the official rankings of academic research units that have been produced in recent decades in several countries, to supplement media rankings of universities based on surveys or other data. The UK has been prominent in the research rankings field ever since it was the first country to produce, three decades ago, an official ranking of the research quality of every department in every university or higher education institution (HEI) in the country, and link public research funding to research rankings on the 'best to best' principle. Over the subsequent thirty years that rankings exercise has been conducted some seven times with apparently increasing sophistication and certainly immense cost, effort, and seriousness by all concerned (Bence and Oppenheim, 2005; Stern, 2016). The most recent exercise, in 2014, involved the assessment by 39 panels of academic experts of over 50,000 academic outputs (mainly papers, books and patents) across 154 HEIs. An innovation in 2014 was the mandatory inclusion of 'impact case studies' in which the social or economic impact of individual pieces of research was required to be evidenced through an audit trail (HEFCE 2014). (The cost, largely borne by HEIs, was estimated to be £246 million (about €300 million) in the 2014 exercise, up from an estimated £66 million (about €80 million) in 2008 (Technopolis, 2015).)

But whether such high-pressure rankings can really be a recipe for fostering and rewarding improvement in research performance depends on at least three conditions. First, it depends on whether those rankings can truly distinguish the units being ranked – whether the underlying metrics used, whatever they are, are valid. Second, it depends on whether the basis of the rankings can be stable enough from one iteration to another for us to be able to reliably assess improvement (or the reverse) in research performance over time. And third, it depends on whether those being ranked choose to respond to research rankings in ways that really do bring benefits to society as a whole. The heretical view about rankings is that none of those three conditions can be taken for granted, and that there is no reason to expect HEI research quality and impact to be exempt from some familiar problems in using rankings for governance by indicators.

2. Three problems with ranking research: a hypothetical example

To explore these conditions, consider the case of an imaginary country ('Metricia') whose policy-makers are seized with zeal for ranking the performance of academic research units according to the scientific quality and social and economic impact of their work. For assessment of 'impact,' Metricia chooses to follow the UK's current definition of this prized but elusive quality, namely 'an effect on, change or benefit to the economy, society, culture, public policy or services, health, the environment or quality of life, beyond academia' (HEFCE, 2016). Those conducting or presiding over Metricia's research quality and impact rankings will have to face at least three major issues.

2.1 Measurement error

First, there will be some unavoidable error or uncertainty in the measurement or categorization that underlies the rankings. Why? Even if we assume Metricia's imaginary rankees are so honest (or naïve) that they do not respond strategically to the rankings exercise, simple perplexity is likely to produce less than perfect concordance among different raters' assessments of quality and impact. Different assessors will vary in the score they give to an item to be ranked – a phenomenon familiar to anyone who has ever graded an examination together with other assessors. If Metricia's impact assessors have to arrive at their ranking scores by coding what they read in narrative accounts of research impact, it is impossible to imagine that they could be exempt from such variation, any more than judges of musical or artistic performance.

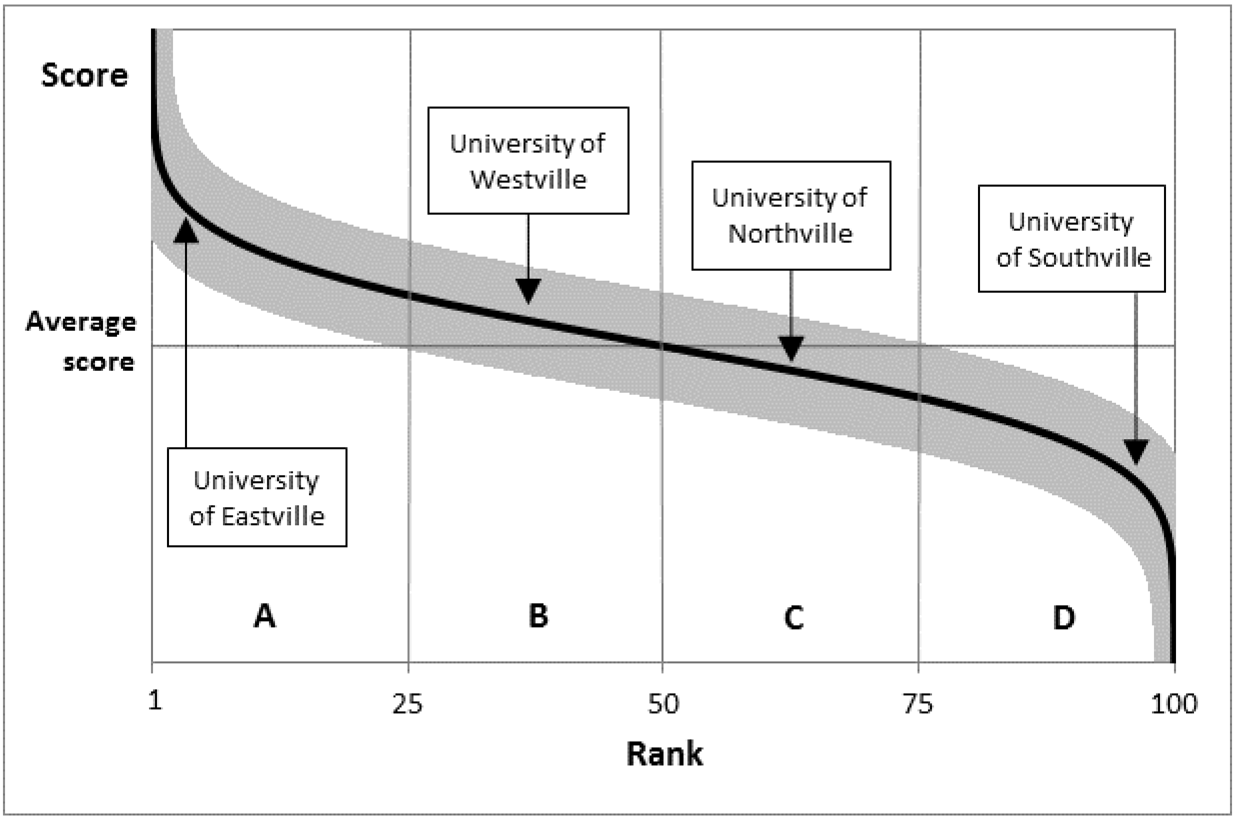
But even if we assume that major problem away and imagine that all Metricia's coders give identical scores to every case, as is more likely to happen if they work from summary statistics or numéraires rather than qualitative narratives, the item being scored may itself be an imperfect basis for measuring the quality that the ranking aims to get at. Indeed, that problem is more likely to occur if the coders work from summary administrative statistics rather than qualitative accounts. For example, the quantity of research funding obtained cannot be a perfect measure of research quality, since focusing on inputs rather than outputs will tend to under-value low-cost 'shoestring research' and over-value high-cost research. Indeed if the level of funding secured is used as a measure of quality, that will encourage HEI researchers to pursue the most expensive possible ways of researching their ideas – hardly an outcome likely to be beneficial to Metricia's society.

Now if we (quite conservatively) assume that the likely measurement errors in Metricia's research ranking are not less than what has been carefully documented for school league tables in numerous studies (e.g. Goldstein and Thomas 1996, Wilson and Piebalga 2008), what would be the effect of such errors on the scores allocated by Metricia's expert assessors to 100 research quality submissions, one from each of Metricia's universities? We assume (i) that these scores are normally distributed about a mean (but the exact shape of the distribution makes little difference); (ii) each submission is ranked according to its score, and scores are plotted against ranks, giving the black line in Figure 1. Each quartile is assigned a grade (A, B, C or D), giving the boundaries shown in the figure, and public research funding is then allocated on the basis of those grades. Now, making that conservative assumption that the uncertainty on each score is no less than has been established for school league tables, the grey area either side of the black line in the figure represents the ~95 % confidence intervals for each score – a vital piece of information for assessing the meaningfulness of the ranking, though not one that commonly appears in official or media ranking reports.

When we take into account those confidence intervals, we can indeed say that the quality of the submission from Metricia's University of Eastville on the top left hand side of the figure is clearly distinguishable from that of its University of Southville on the bottom right hand side. But there is almost no genuine discrimination between B and C scores – for example between the University of Westville and the University of Northville here. And even Eastville's score cannot be reliably distinguished from that of Westville, or Northville's from Southville's, for that matter.

That outcome presents those running Metricia's research rankings with an awkward dilemma. Do they try to improve the validity of the scores by setting up provisions for correcting categorization errors, for example by allowing appeals for reconsideration of any given unit's scores by the same or different assessors, in ways that could lead to years of wrangling and litigation by disappointed HEIs? Or do they choose to live with the validity problems and follow the 'rough justice' approach of the UK's Research Assessment Exercise in forbidding appeals and destroying all the assessment records as soon as the research rankings appear (Corbin, 2008)?

Figure 1. 100 Hypothetical Metrician University Research Submissions, Scored by an Expert Panel



2.2 Unstable metrics

Nor do the likely headaches of Metricia's research rankers end there. If any given iteration of the rankings exercise produces less than perfect validity for the reasons we have described, how should subsequent iterations of the rankings be designed? If Metricia's rankers repeat the exercise with exactly the same metrics as before in a later round, failing to change what or how they measure in the light of validity problems that came to light in the previous round, they will be sacrificing validity for reliability. But if they decide to change the metrics (in response to criticisms of the previous iteration or to changing social or political agendas over what sort of things research is supposed to achieve), they will be sacrificing reliability for validity. How should they handle that trade-off?

Figure 2. Effect of Methodological Choices on University Rankings over Three Hypothetical Rankings Exercises. Exercise 1 represents unweighted scores for *outputs* (below the white bar) and *impacts* (above the bar) for the four universities. Exercises 2 and 3 show the effects of increasing the weighting of impacts and decreasing the weighting of outputs in successive rounds of the assessment process.

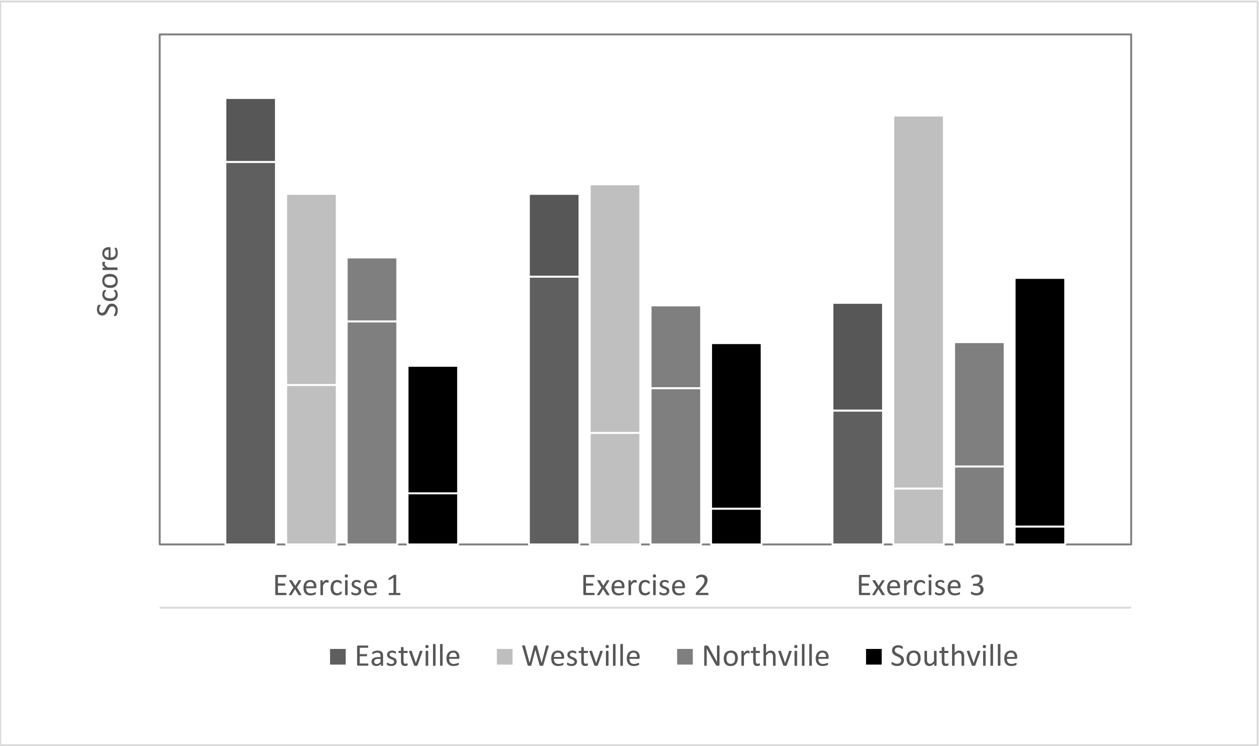


Figure 2 illustrates the problem by depicting three hypothetical iterations of Metricia's research ranking system. It shows how the four universities mentioned in Figure 1 rank relative to one another in each of those three iterations.

For the purpose of this illustration, we assume Metricia uses two main criteria for assigning a score for research quality on each of the three iterations: the academic quality of the outputs (the score below the white bars in Figure 2), and the social and economic impact of the research (the score above the white bars). Further, we assume that for one reason or another Metricia's policy-makers choose to alter the relative weights attached to those two criteria from one ranking episode to the next.

In the first round, Exercise 1, let us suppose that Eastville and Northville gain most of their scores from the quality of their outputs but have relatively low scores for impact. By contrast, Southville gains most of its score from impact, while Westville is more evenly balanced. In the next round of Metrician research rankings, let us assume that impact is weighted more heavily than in the first. *Assuming no change in performance* for any of the four universities, those whose strengths lie in 'impact' benefit most from this change. As shown in Figure 2, Southville's and Westville's scores go up – an apparent improvement in performance due solely to the change in indicator weighting – while the other universities show an apparent decline.

In Exercise 3 (again assuming no change in actual performance) another shift in weighting towards impact pushes Southville's score still further up, putting Southville in second place in this league table, up from a poor fourth in the first round of the exercise. Eastville's and Northville's reliance on high-quality outputs has done them no favours and they end up in third and fourth places respectively, while Westville ends up top of the league table. Since the rankings would have remained unchanged across all iterations of the exercise if the weightings had stayed constant, Figure 2 shows how much both the ranking of each unit and its direction of change vary according to those key choices in the method of arriving at a composite score. This illustration is necessarily simplified but reflects observed patterns in some real-world ranking exercises. For instance, a detailed simulation study a decade or so ago showed how minor changes in indicator distribution and weightings could produce major changes in the performance grades given to hospitals and local authorities on the ranking systems then applying in England (Jacobs and Goddard, 2007).

That validity/reliability dilemma lies at the heart of governance by indicators. Other scholars have pointed to trade-offs between validity and reliability as an ineluctable feature of inter-organizational comparisons (Gormley and Weimer, 1998) and in our own earlier work we have shown an inverse relationship between validity and reliability in fifteen international rankings of governance and public administration (Hood et al. 2008). If you do not measure performance in a consistent way in successive iterations, it is impossible to know whether the system as a whole, or any of the individual units within it, are improving, deteriorating or staying the same (Hood and Dixon, 2015).

Further, successive attempts to pursue higher validity by frequent changes in the basis on which rankings are constructed may have the unintended social effect of undermining their credibility by making their basis incomprehensible to all but a small group of experts (Tsoukas, 1997). All we can assess from rankings whose basis changes significantly from one iteration to another (subject to validity limitations, of course) is how well the system or the units within it relate to whatever desiderata are measured on each occasion.

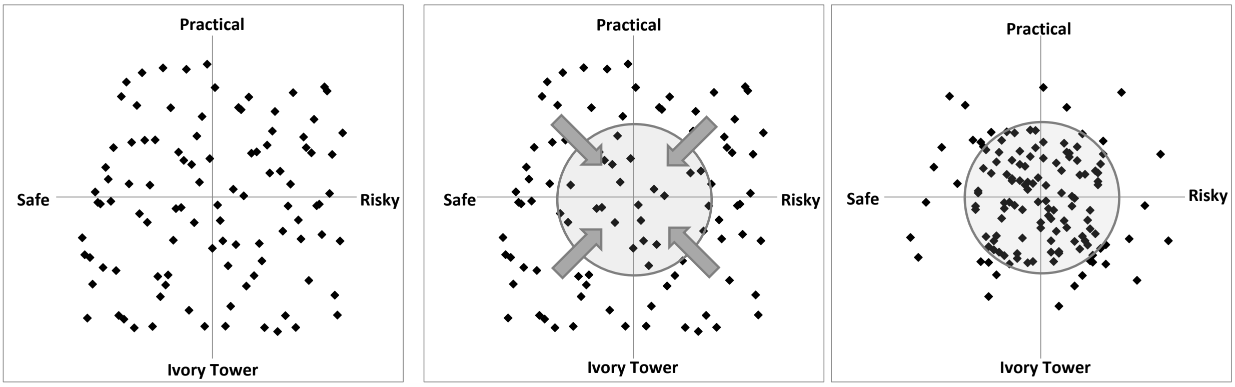
2.3 Homogenizing effects

But now let us suppose that those operating Metricia's research ranking exercise respond to that validity/reliability dilemma in quite the opposite way than that we have just explored, by instead choosing to keep the measurement system reasonably consistent between ranking episodes. What strategic behaviour might we expect of Metricia's rankees in that case and what possible effects might that behaviour have on the eventual outcome for Metrician society?

Any ranking has to be constructed around a set of rules and assumptions about what is to be valued, in this case high-quality, high-impact research. Regardless of the merits or demerits of any criteria of such polyvalent qualities that might be chosen, any set of scores has to be boiled down to a single number to make ranking possible. That can only be done by a process of weighting and compositing that can be expected to set off strategic responses to maximize point scores. And over time those strategic responses may come to distort and undermine the values the rankings are intended to measure. Several studies of high-consequence ranking and benchmarking behaviour in other contexts, such as teaching performance, have found evidence of reductions in variety and innovation, for instance in less emphasis on innovative teaching methods, more emphasis on 'teaching to the test' and increasing homogeneity of curricula (see e.g. Elwood, 2008; Sauder and Espeland, 2009; Frederickson and Stazyk, 2010). And there is no reason for supposing that research rankings can be immune to such effects.

Suppose that before Metricia introduced its research quality and impact rankings, its researchers in the four imaginary universities were spread out in the pattern depicted on the left-hand-side of Figure 3, which shows two readily-observable dimensions on which research styles can vary, namely the degree of risk in projects undertaken (that is, the likelihood that research will lead to significant academic publications in a limited time-frame) and the extent to which research work is 'applied' (that is, the extent to which research has an obvious and practical application at the time when it is undertaken).

Figure 3: Homogenization Effect of Rankings on Metrician Research: A Hypothetical Example



Now consider the combined effects on that pattern of putting together measures of academic research quality and impact into a single composite index needed for arriving at an overall ranking. If Metricia's university managers calculate that combining those two elements into a single ranking will tend to penalize the extremes on both dimensions, we can expect to see a move towards the pattern shown on the right-hand-side of Figure 2.

For example, let us suppose that at the outset of the research ranking system, each of Metricia's four universities has something equivalent to a 'Better Mousetraps Unit,' focusing on severely applied and practical work, and an 'Abstract Puzzles Unit,' focusing on solving intellectual problems that have no immediate practical applications. The ranking system we have described makes it likely that Metricia's university managers will cajole the equivalents of the Better Mousetraps Unit into putting more effort into the 'academic' peer-reviewed publications they need to reach the appropriate quality standard, while they will put the equivalents of the Abstract Puzzles Unit under pressure to make their work more socially or economically 'relevant' to raise their impact scores. Similarly on the 'risk' dimension, those with high risk appetites will tend to be reined in, while those with risk appetites that are too low (for example in only doing almost exact replications of other studies) will be encouraged to play it rather less safe.

The result of such strategic responses to governance by indicators leads Metricia to end up in the much more homogenous research world of the right-hand side of Figure 3. And that is not just a hypothetical possibility, given the strong encouragement in most research assessment systems to reporting results in high impact factor journals (that is, journals whose papers are most frequently cited in other academic publications), which in turn forces scholars to do the kind of research those journals favour, and shapes the recruitment and promotion patterns of the whole research community.

What could be wrong with that? Nothing, if what you value is greater uniformity. But if you believe society can be better served by a high degree of diversity in research styles rather than by everyone converging on a similar level of risk and application, an outcome in which research units tend to cluster around whatever profile will optimize their composite score (the circled area in Figure 3) would represent a worrying loss of research 'biodiversity' within Metricia, to the possible long-term detriment of innovation and variety in its research community.

3. Implications

So perhaps research rankings as a particular application of governance by indicators are not quite the 'no-brainer' (unambiguously beneficial choice) for improving public service performance that they might at first sight appear. And indeed they raise several tricky issues other than the ones mentioned here. That is not to deny that rankings of organizations on items such as research quality can have socially positive effects in some circumstances. But examples of dysfunctional monocultures and persistence of failing institutional structures or processes are plentiful in the literature on social and economic development. Indeed, despite the uncertainties and perverse effects discussed above, performance ranking systems such as that for UK research quality tend to persist, and indeed to expand, once established. Such persistence and expansion are seen in the worldwide growth of ranking systems in many areas of governance (see e.g. Besancon, 2003; Johnsen, 2005; Arndt, 2008; Hood et al., 2008). A full account of reasons for such persistence and expansion is well beyond the scope of this paper, but would be likely to include some familiar ingredients of institutional path-dependence— given that statistical systems can themselves be understood as institutions (Desrosières, 1998). For. For instance, organizations which benefit from a ranking system are unlikely to wish to rock the boat, and as those organizations become better at conforming more closely to the metrics of the rankings, those tasked with assessing the outcomes can make claims about improvement in the system as a whole over time (see for instance Stern, 2016). But that does not mean that ranking systems as institutions can be immune to the familiar mechanisms that produce crises or punctuated equilibria.

However that may be, even from this very limited analysis we can draw three conclusions whose significance for the effect of governance by indicators goes well beyond the specific case of research assessment.

One is that basing high-stakes financial consequences (in this case, levels of public funding allotted on the basis of rankings) on statistically insignificant differences in scores can turn the funding process into a lottery – just what rankings purport to avoid. A second is that – as in other types of governance – there are unavoidable dilemmas and tradeoffs between competing values, in this case between validity and reliability, and too many adjustments intended to improve validity can make it impossible to assess performance over time on consistent criteria. The third is that a ranking system that cannot satisfactorily capture all the relevant dimensions, including those depicted in Figure 3, may come to threaten variety and innovation itself.

4. References

Arndt, C., 2008. The Politics of Governance Ratings. International Public Management Journal 11 (3) pp. 275-297.

Bence, V., and Oppenheim, C., 2005. The Evolution of the UK's Research Assessment Exercise: Publications, Performance and Perceptions. Journal of Educational Administration and History 37 (2) pp. 137-155.

Besançon, M., 2003. Good Governance Rankings: The Art of Measurement. WPF Reports No. 36. Cambridge, MA: World Peace Foundation.

Corbin, Z., 2008. Panels Ordered to Shred all RAE Records. Times Higher Education. Online: http://www.timeshighereducation.co.uk/story.asp?storycode=401501

Desrosières, A., 1998. The Politics of Large Numbers: A History of Statistical Reasoning. Harvard University Press,Cambridge, MA.

Ellwood, J.W, 2008. Challenges to Public Policy and Public Management Education. Journal of Policy Analysis and Management 27(1) pp. 172-187.

Frederickson, H.G., and Stazyk, E.C., 2010. Ranking US Public Affairs Educational Programmes: Searching for Quality, Finding Equilibrium? In H. Margetts, P. 6, and C. Hood (Eds.), Paradoxes of Modernization: Unintended Consequences of Public Policy Reform. Oxford University Press, Oxford. pp. 63-80.

Goldstein, H., and Thomas S., 1996. Using Examination Results as Indicators of School and College Performance. Journal of the Royal Statistical Society, Series A, 159 (1) pp. 149-163

Gormley, W., and Weimer, D., 1999. Organizational Report Cards. Harvard University Press, Cambridge, MA.

Higher Education Funding Council for England (HEFCE), 2014. REF 2014: The Results. Online: http://www.ref.ac.uk/pubs/201401/

Higher Education Funding Council for England (HEFCE), 2016. Policy Guide: Research Impact. Online: http://www.hefce.ac.uk/rsrch/REFimpact

Hood, C., Dixon R., and Beeston, C., 2008. Rating the Rankings: Assessing International Rankings of Public Service Performance. International Public Management Journal 11 (3), pp. 298-358.

Hood, C., and Dixon, R., 2015. A Government that Worked Better and Cost Less? Evaluating Three Decades of Reform and Change in UK Central Government. Oxford University Press, Oxford. Chapter 3, pp.44-64.

Jacobs, R., and Goddard, M., 2007. How Do Performance Indicators Add Up? An Examination of Composite Indicators in Public Services. Public Money & Management 27 (2), pp. 103-110.

Johnsen, Å., 2005. What Does 25 Years of Experience Tell Us About the State of Performance Measurement in Public Policy and Management?, Public Money & Management, 25 (1) pp. 9-17.

Sauder, M. and Espeland, W.N., 2009. The Discipline of Rankings: Tight Coupling and Organizational Change. American Sociological Review 74 (1) pp. 63-82.

Stern, N., 2016. Building on Success and Learning from Experience: An Independent Review of the Research Excellence Framework. Department for Business, Energy and Industrial Strategy, London.

Technopolis, 2015. REF Accountability Review: Costs, Benefits and Burden. Technopolis, Brighton. Online: http://www.technopolis-group.com/?report=ref-accountability-review-costs-benefits-and-burden

Tsoukas, H., 1997. The Tyranny of Light: The Temptations and the Paradoxes of the Information Society. Futures 29 (9), pp. 827-843.

Wilson, D., and Piebalga A., 2008. Performance Measures, Ranking and Parental Choice: an Analysis of the English School League Tables. International Public Management Journal, 11(3) pp. 344-366.