Ranking for Success: a No-brainer?

RUTH DIXON and CHRISTOPHER HOOD

If you can rank Olympic sports teams and restaurants, why should there be such a fuss about doing so for university research? What's wrong with replacing the tacit or implicit qualitative knowledge about scholars, research units or journals that everyone worked on forty years ago with precise metrics based on clear criteria? After all, we rank many things other than sport teams and restaurants today. The last four decades have seen a striking growth in rankings of 'governance' by bodies like the World Bank and Transparency International, and if we can rank something as complex as that, what could be the objection to applying the same method to research quality and impact, as is due to happen in next year's Research Excellence Framework (REF)? 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, so that the whole society is better off as a result of the higher quality, higher-impact research they produce? And don't rankings go 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?

Well, it depends on whether high-pressure rankings can truly distinguish the units being ranked – whether the underlying metrics are valid and reliable. It also depends on whether the rankees choose to respond to rankings in ways that really bring benefits to society as a whole. The heretical view about rankings is that neither of those things can be taken for granted, and that there is no reason to expect research quality and impact to be exempt from some familiar problems in using rankings. Why should that be?

Even in some (highly) imaginary world where rankees didn't respond strategically to rankings, there will be some error or uncertainty in the process for at least two reasons. One, familiar to anyone who has ever marked an examination, is that different assessors vary in the score they give to an item to be ranked, and it is impossible to imagine that the research impact narrative cases which are a central feature of the REF could be exempt from such variation. Another is that (even in another imaginary world where all coders gave identical scores to every case) the item being scored may itself be an imperfect basis for measuring the quality that the ranking aims to get at. For example, research income cannot be a perfect measure of research quality, since it will tend to under-value shoestring research and over-value expensive research, and indeed if it is used as a measure of quality it will encourage scholars to pursue the most expensive possible ways of carrying out their research.

Now if we (quite conservatively) assume that such measurement errors in research ranking are not less than what has been carefully documented for school league tables in numerous studies[1], what would be the effect of such errors on the scores allocated by a hypothetical REF expert panel to 100 impact case studies? We can assume that these scores are normally distributed about a mean (but the exact shape of the distribution makes little difference), each case study is ranked according to its score, and scores are plotted against ranks, giving the black line in Figure 1. Assigning the four grades (1*, 2*, 3* and 4*) to those 100 case studies in the same proportions as in the REF pilot exercise[2], we obtain the grade boundaries shown in the figure. Now, making the 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 appears anywhere in the reports of the REF pilot exercise. When we take into account those confidence intervals, we can indeed say that the work of Brainbox University on the top left hand side of the figure is clearly distinguishable from that of the University of Dullsville on the bottom right hand side. But there is almost no genuine discrimination between 2* and 3* scores – for example between the University of Watermouth and Poppleton University here. And even Brainbox's score cannot be reliably distinguished from that of Watermouth, nor Poppleton's from Dullsville's, for that matter. Yet there are no mechanisms for correcting categorization errors in the REF exercise, given that decisions cannot be appealed, each case study is assessed by a single panel and the record of panel deliberations destroyed, as was apparently the case with the RAE.[3]

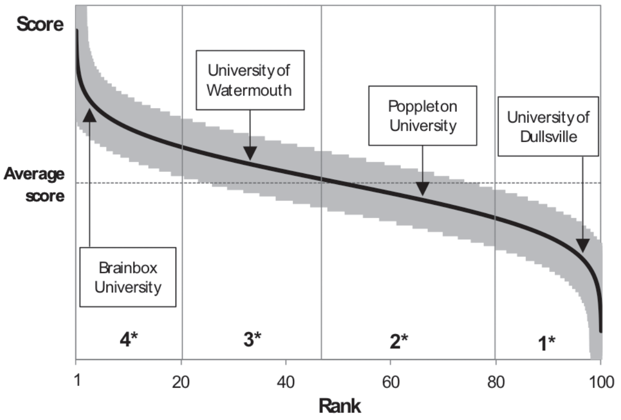

Figure 1. 100 Hypothetical REF Impact Cases Scored by an Expert Panel

Now let's introduce some strategic behaviour on the part of rankees into the mix and consider its possible effect on the eventual outcome. Like any ranking, the REF process is constructed around a set of rules and assumptions about what is to be counted as good research. Regardless of the merits or demerits of those particular choices of the criteria of good research, any set of scores has to be boiled down to a single number to make ranking possible. And that can only be done by weighting and compositing that will set off strategic responses to maximize point scores that may come to distort and undermine key values of university research that the rankings are intended to measure. Several studies of ranking and benchmarking behaviour in other contexts have found evidence of reductions in variety and innovation, resulting for instance in less emphasis on innovative teaching methods and increasing homogeneity of curricula.[4]

Is there any reason to suppose that research rankings might be immune to such effects? Probably not, considering the possible combined effect of emphasis on impact with the requirement that research only counts if it is published within a limited time-frame and achieves at least medium academic quality (normally indicated by acceptance for publication through academic peer-review systems). Suppose that in the absence of high-stakes rankings, researchers in any given field might spread themselves out in the pattern depicted in the top panel of Figure 2, which depicts two dimensions on which research styles can vary, namely the degree of risk in projects undertaken (that is, the odds against research leading 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).

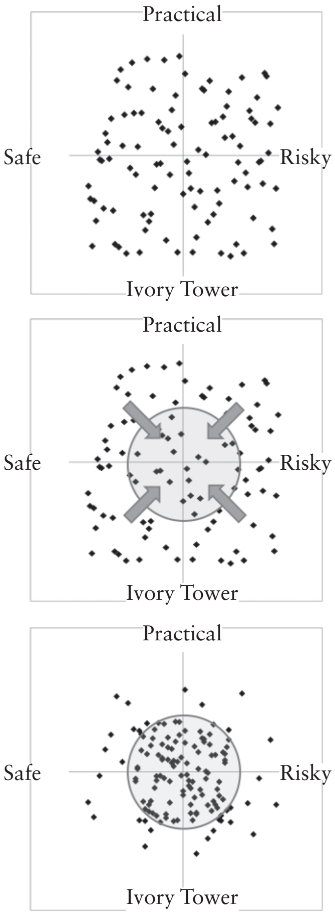

Figure 2. Homogenization Effect of Rankings: (a hypothetical example)

Now consider the combined effects of the REF rules noted above on that pattern. If Sir Terry Spreadsheet and his fellow Vice-Chancellors calculate that the combined effects of those rules will tend to penalize the extremes on both dimensions, we can expect to see a move towards the pattern shown in the lower panel of Figure 2. On the 'application' dimension, the Better Mousetraps Unit will be cajoled into putting more effort into the 'academic' peer-reviewed publications it needs to hit the 2* publication score for 'impact', while the Abstract Puzzles Unit will be put under pressure to make its work more 'relevant' to raise its impact score. 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. We end up in the much more homogenous research world of the bottom of Figure 2. And that is not just a hypothetical possibility, given the strong encouragement given to placing research into high impact factor journals, which in turn forces scholars to do the kind of research those journals favour, and affects recruitment and promotion patterns.

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 departments tend to cluster around whatever profile will optimize their REF score (the circled area in Figure 2) would represent a worrying loss of research 'biodiversity,' to the detriment of innovation and variety.

So perhaps research rankings are not quite the 'no-brainer' for public benefit they might at first sight appear, and they raise several tricky issues other than the ones mentioned here. But even from this very limited analysis we can draw two conclusions. One is that basing high-stakes financial consequences on statistically insignificant differences in scores can turn the funding process into a lottery – just what rankings purport to avoid. The other is that a ranking system that cannot satisfactorily capture all the relevant dimensions, including those depicted in Figure 2, may come to threaten variety and innovation itself.

[1] E.g. Goldstein, H., and S. Thomas (1996) 'Using Examination Results as Indicators of School and College Performance.' *Journal of the Royal Statistical Society, Series A*, 159(1): 149-163.
Wilson, D., and A. Piebalga (2008) 'Performance measures, ranking and parental choice: an analysis of the English school league tables.' *International Public Management Journal*, 11(3): 344-366.

[2] Research Excellence Framework impact pilot exercise: Findings of the expert panels (2010) ***http://www.ref.ac.uk/pubs/refimpactpilotexercisefindingsoftheexpertpanels/***. 500 case studies were assessed by 5 subject panels, and graded for impact as follows: 93 4*, 124 3*, 152 2*, 96 1* and 35 unclassified. Each case study was assessed by one panel, and there is no indication in the report that the panels made any estimate of grade uncertainty.

[3] Corbin, Z. (2008) 'Panels ordered to shred all RAE records.' *Times Higher Education*. ***http://www.timeshighereducation.co.uk/story.asp?storycode=401501***

[4] E.g. Sauder, M. and W.N. Espeland (2009) 'The Discipline of Rankings: Tight Coupling and Organizational Change.' *American Sociological Review* 74(1): 63-82;
Ellwood, J.W. (2008) 'Challenges to Public Policy and Public Management Education.' *Journal of Policy Analysis and Management* 27(1):172-187;
Frederickson, H.G., and E.C. Stazyk. (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*, 63-80. Oxford: OUP.