[The present document is an extract from evidence presented to the Higher Education Funding Council for England and its Independent Review of the Role of Metrics in Research Assessment.

Details of the review and access to all the evidence submitted is available at: http://www.hefce.ac.uk/whatwedo/rsrch/howfundr/metrics/]

Responses to the Metrics Review Call for Evidence

Submitted to HEFCE by 30 June 2014

Response to the Call for Evidence

to the Independent Review of the

Role of Metrics in Research Assessment

Professor Ben Martin, SPRU (University of Sussex);
Associate Fellow, CSaP, and Senior Visiting Fellow,
CBR (both University of Cambridge)

Professor Paul Nightingale, SPRU (University of Sussex)

Dr. Ismael Rafols, Research Fellow, Ingenio (CSIC-UPV);
Visiting Fellow at SPRU (University of Sussex)

June 2014

Response to the Call for Evidence to the Independent Review of the Role of Metrics in Research Assessment

The authors

Ben Martin is Professor of Science and Technology Policy Studies at SPRU, where he served as Director from 1997 to 2004. He has carried out research for over 30 years in the field of science policy. He helped to establish techniques for evaluating scientific laboratories, research programmes and national scientific performance. He also pioneered the notion of 'technology foresight'. Since 2004, he has been Editor of *Research Policy.*

Paul Nightingale is Professor of Strategy at Sussex and Deputy Director of SPRU. He conducts research on the use of models, and their associated metrics, in both science and industry. His research explores how models and metrics inform organisational decision making in both positive and negative ways. He is currently an Editor of *Research Policy* and *Industrial and Corporate Change*. He is on the ESRC Evaluation Committee where he has taken a keen interest in how the over-enthusiastic use of metrics to manage research can negatively influence the social impact of funded research.

Ismael Rafols conducts research at the interface of scientometrics and science policy at Ingenio (CSIC-UPV, València) and SPRU. He is on the advisory board of the journal *Scientometrics,* and has given advice on the use of indicators to various organisations including the OECD, the US National Science Board (with consultancy SRI, International) and the Spanish Foundation for Science and Technology (FECyT).

Statement of evidence

1. Metrics are potentially helpful in research assessment but they need to be carefully harnessed

As a result of the increasing size and complexity of scientific communities and organisations, various measures that facilitate understanding of the properties and trends of scientific activities have been developed and these can potentially be very helpful in research assessment. In particular they have great potential to help better align the UK research system with the UK innovation system, and indeed to better align research with societal needs.

Measuring the properties of science is difficult. Metrics need to be appropriate to the property under investigation (i.e. fit for purpose), they need to be reliable (i.e. statistically stable) and they should be applicable to the whole area or system under study (i.e. robust to changing contexts such as scientific fields).[1] Some indicators that are widely used, such as the Journal Impact Factor or the Hirsch index, fail to meet these criteria in most conventional uses. Other indicators such as citation impact are only reliable above certain levels of

[1] Gingras, Y., 2014. Criteria for evaluating indicators, in: Cronin, B., Sugimoto, C. (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact.* MIT Press, Cambridge, MA and London, UK, pp. 109–126.

aggregation (usually not at the level of the individual researcher) and need careful mathematical normalisation to be used across diverse research areas.

Thus, metrics need to be properly harnessed in order to become reliable and robust sources of information for research assessment purposes.

2. **Metrics cannot substitute for judgement. Metrics are most useful in facilitating deliberation.**

Metrics should be used to inform, rather than to substitute for, expert judgement. Metrics by themselves cannot be set up to provide an algorithm to make decisions, nor can one assume that past performance (what is measured is inevitably in the past) is a reliable guide to future prospects. This point was clearly and thoroughly presented in a recent report by the Canadian Academies Council (2012)[2] based on contributions by leading science evaluation practitioners and scholars, a report which we suggest the HEFCE review consider as an important source of evidence.

Quantification of properties such as impact with a specific metric is inevitably associated with a specific perspective (e.g. citations in the clinical guidelines of a paper) at the expense of other equally legitimate perspectives (e.g. citations in patents). Given that each metric thus offers a limited or partial picture, metrics are most useful in deliberation processes in which various viewpoints are discussed. The fact that metrics are partial and thus debatable is what makes them valuable in facilitating discussion as a prelude to decision-making.[3]

Given the partial and fallible nature of metrics, a single metric is often not reliable. When various contrasting metrics aiming to capture the same (or related) concepts suggest similar insights, this "convergence of partial indicators" offers more convincing evidence about the property observed.[4] In contrast, a lack of agreement between various metrics may indicate that (subjective) perspectives taken in the deliberation are crucial in the final results.[5]

For example, Rafols et al. (2012) used different metrics to measure interdisciplinarity and research performance across a sub-sample of UK social science. The different measures of interdisciplinarity converged on a common ranking. However, the metrics of research performance failed to converge, suggesting it would be possible to rank the performance of the different research groups in almost any order one wanted by selective choice of a particular metric of performance. As such, the 'objective' ranking would simply reflect subjective choices rather than intrinsic features of the research outputs.

3. **The unwanted consequences of metrics**

When incentives in science become associated with certain metrics, it is highly likely that researchers will strategically shift their attention towards activities that offer them the largest

[2] See http://www.scienceadvice.ca/en/assessments/completed/science-performance.aspx

[3] Barré, R., 2010. Towards socially robust ST indicators: indicators as debatable devices, enabling collective learning. Research Evaluation 19, 227–231.

[4] This was demonstrated in several SPRU studies – e.g. Martin, B.R., Irvine, J., 1983. Assessing basic research: Some partial indicators of scientific progress in radio astronomy. Research Policy 12, 61–90; Martin, B.R., 1996. The use of multiple indicators in the assessment of basic research. Scientometrics 36, 343–362.

[5] Rafols, I., Ciarli, T., Van Zwanenberg, P., Stirling, A., 2012. Towards indicators for opening up S&T policy. STI Indicators Conference. http://2012.sticonference.org/Proceedings/vol2/Rafols_Towards_675.pdf

rewards (e.g. publishing on fashionable topics or in highly cited journals). This may result in behavioural changes that are sometimes desirable, but which often also generate unintended or even undesirable consequences.[6] For example, the increase of number of publications in Australia fostered by incentives to publish more articles resulted in a relative decline in the standing of the journals of publication.[7] We provide other evidence of more serious problems with gaming in our answer to the questions on gaming (see below).

While the use of metrics is often predicated on the assumption that they provide an objective way to assess science activities, the diversity of practices means that subjective judgement is needed to weight their relative importance in a given context. For example, the relative propensity to exhibit a certain metric (a citation, a tweet on twitter, etc.) varies widely across scientific areas and any comparison must therefore involve some form of normalisation. In the lack of appropriate weighting, one likely outcome of directly associating incentives with rewards is that this will foster certain types of research or research behaviour over others. In general, those fields with strong institutional structures (such as prestigious journals or conferences) are likely to benefit at the expense of less well integrated fields. This means that fields like oncology are more likely to "perform well" in metrics than smaller ones such as epidemiology, which in turn are likely to perform better than interdisciplinary fields. Such differences may be transformed into financial or human resources – in effect the bias in the metrics generates biases in the resource distribution, which in turn ends up generating more bias in science. Such biases have been found, for example, with respect to language[8], gender[9], interdisciplinary research[10], and clinical research[11].

Given the specific conditions of each research assessment regarding field, national or local context, level of aggregation and so on, different types of biases and unwanted consequences may almost inevitably result. The available evidence suggests that in general peripheral countries, disciplines or topics will be disadvantaged and will receive less credit or resources than they deserve. Since these peripheral activities enhance scientific diversity, which is an invaluable source of creativity and societal relevance, *a likely consequence of the inappropriate use of metrics is the suppression of diversity and creativity and hence also of socio-economic impact of research.*

[6] Weingart, P., 2005. Impact of bibliometrics upon the science system: Inadvertent consequences? Scientometrics 62, 117–131.

[7] Butler, L., 2003. Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. Research Policy 32, 143–155. doi:10.1016/S0048-7333(02)00007-0

[8] Van Leeuwen, T., Moed, H., Tijssen, R.W., Visser, M., Van Raan, A.J., 2001. Language biases in the coverage of the Science Citation Index and its consequences for international comparisons of national research performance. Scientometrics 51, 335–346.

[9] Leahey, E., 2007. Not by Productivity Alone: How Visibility and Specialization Contribute to Academic Earnings. American Sociological Review 72.

[10] Rafols, I., Hopkins, M.M., Hoekman, J., Siepel, J., O'Hare, A., Perianes-Rodríguez, A., Nightingale, P., 2012. Big Pharma, Little Science? A bibliometric perspective on big pharma's R&D decline. Technological Forecasting & Social Change In press.

[11] Van Eck, N.J., Waltman, L., van Raan, A.F.J., Klautz, R.J.M., Peul, W.C., 2013. Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. PLoS ONE 8, e62395.

4. Beyond measures of performance – multi-dimensional metrics that illuminate hidden dynamics in the science.

Most use of, and debates about, metrics in research assessment has been dominated by discussion of performance concepts such as production or excellence, which aim to examine whether "more" or "better" science is being produced. Sometimes, other dimensions such as collaboration, internationalisation or impact are also discussed – but most often in scalar (i.e. one-dimensional) terms: i.e. whether more or less of that particular metric is achieved.

Given that science is a complex system (or an "ecosystem"), such one-dimensional descriptions are of limited use. In order to make decisions – in both assessment and also in strategic science policy that seeks to better align the science system with society's need – it is usually more important to understand the specific axes (or direction) in which growth occurs – for example, towards which disciplines, which countries, or which topics, rather than simply knowing about the aggregate growth or decrease. Science metrics, appropriately used, can offer a much richer palette to characterise science, and such a multi-dimensional portrayal can be extremely helpful in informing (rather than substituting for) decision-making.[12]

While adding more dimensions has the obvious disadvantage of making the metrics more complex,[13] new interactive visualisation tools offer great potential in providing the means to facilitate a deeper intuitive understanding. Network and science-mapping visualisations have considerably enhanced the capacity to convey complex information to users. These tools are now sufficiently mature to be used not only available in academia[14] but also in consultancy and funding organisations. The visualisation of NIH-funded grants offers an example of how metric-based approaches can provide relevant and very detailed yet accessible information for research assessment purposes (http://nihmaps.org).[15] These tools not only facilitate visualisation, but specific metrics can be associated with them, for example regarding knowledge flows[16] or disciplinary/topic diversity[17].

5. The challenge of use when metrics become an established part of the "infrastructure"

Metrics are becoming increasingly included by default in data research infrastructure, such as the Current Research Information System (CRIS) (used by university managers) or in

[12] Rafols, I., Ciarli, T., Van Zwanenberg, P., Stirling, A., 2012. Towards indicators for opening up S&T policy. STI Indicators Conference. Obtainable: http://2012.sticonference.org/Proceedings/vol2/Rafols_Towards_675.pdf (accessed June 30, 2014).

[13] Stirling, A., 2010. Keep it complex. Nature 468, 1029–1031.

[14] See for example, Börner, K., Chen, C., Boyack, K.W., 2003. Visualizing Knowledge Domains. Annual Review of Information Science & Technology 37, 179–255. Rafols, I., Porter, A.L., Leydesdorff, L., 2010. Science overlay maps: a new tool for research policy and library management. Journal of the American Society for information Science and Technology 61, 1871–1887.

[15] Talley, E.M., Newman, D., Mimno, D., Herr, B.W., Wallach, H.M., Burns, G.A.P.C., Leenders, A.G.M., McCallum, A., 2011. Database of NIH grants using machine-learned categories and graphical clustering. Nat Meth 8, 443–444.

[16] Boyack, K.W., Börner, K., Klavans, R., 2009. Mapping the structure and evolution of chemistry research. Scientometrics 79, 45–60.

[17] Rafols, I., Meyer, M., 2010. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. Scientometrics 82, 263–287.

electronic journals (e.g. in PLoS).[18] Such developments may be positive in terms of enhancing the transparency and accountability of science, but they do assume that users have the necessary capabilities to make an appropriate interpretation of the raw data or metrics. Some officers in funding agencies and specialised evaluation units may have the understanding to do so. However, many lower-level managers (e.g. heads of department) and individual researchers do not have such training and have often been using metrics without sufficient understanding of the limitations. The recent San Francisco Declaration on Research Assessment (DORA)[19] points to the evident dangers here.

Ease of access to certain indicators means that the users of those indicators are becoming "locked in" and relying in those forms of metrics that are more readily available (e.g. Google Scholar – sometimes termed 'desktop bibliometrics' or even 'the poor man's bibliometrics') rather than those which are more appropriate and rigorous.

6. New metrics are worth investigating but are as yet untested for assessment purposes

The wider uptake of research is often highly dependent on framing the research and the accompanying paper to maximise its chances of gaining attention. This might involve coming up with a catchy title or suggesting a link with a topical or attention-grabbing issue. This connection between attention and linking to attention-grabbing issues is even more pronounced when social media are involved, where links (however remote) with sex, pets, crime or celebrities can be associated with research gaining greater attention. New metrics attempting to capture the wider impact of research as reflected through social media may therefore end up capturing little more than the ability of the researcher concerned to come up with an eye-catching aspect of the research. In short, some of these indicators are likely to further encourage work aimed at attention grabbing.

Moreover, there is extensive evidence from research on the sociology of networks that attention in networks is heavily influenced by traditional sociological variables. Powerful groups in the centre of networks with large numbers of connections get more attention than marginalised groups. Metrics that are intended to inform understanding of quality may simply end up capturing network centrality and people in key positions as 'bridgers and brokers' between networks.

Lastly, certain kinds of research, particularly on social problems that are complex, require sensitive, long-term research engagement that can be damaged by publicity. Engagement with research users and subjects is likely to be discouraged if they are concerned about publicity, particularly in a format restricted to 140 characters.

Alternative metrics may in the future provide useful insights, but currently they are at a very early stage.[20] They should not be used in composite indicators such as the European

[18] Wouters, P., 2014. The citation: from culture to infrastructure, in: Cronin, B., Sugimoto, C. (Eds.), Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Impact. MIT Press, Cambridge, MA and London, UK, pp. 47–66.

[19] Obtainable from http://am.ascb.org/dora/files/sfdeclarationfinal.pdf (accessed on 30 June 2014).

[20] Wouters , P., & Costas , R. (2012). Users, narcissism and control — tracking the impact of scholarly publications in the 21st century. Amsterdam: SURF foundation (downloaded from

Innovation Scoreboard (EIS) since Mendeley, Twitter, Facebook and the like differ considerably and hence may capture (as the EIS does) very different aspects of societal impact.

7. Summary: Between Scylla and Charybdis

New developments in the use of metrics have a major potential to improve the governance of the science system and to better align the structure of the science system with societal needs. However, in achieving this potential, the science policy community needs to avoid two dangers.

The first danger is that the clear and inherent problems with the use of metrics are used to discredit the entire enterprise and hence to throw the baby out with the bathwater. Metrics such as citations capture academic impact, rather than quality, and capture that impact only partially. They are not appropriate for capturing quality, particularly at the individual and small group levels, and many metrics are used in very inappropriate ways. The temptation is to dismiss the use of all metrics for all purposes and return to an un-strategic and largely unaccountable pattern of governance based around disciplinary peer-review.

A system of governance based only in peer-review is problematic because it overlooks how choices need to be made between disciplines. As such it often assumes that any area of science is just as likely to generate benefits as any other, and therefore simply increasing the amount of science increases the public good. Both assumptions are false. The links between science and impact may be uncertain, but impacts are generally in the areas where research takes place – pharmaceutical firms provide funding for molecular biology rather than astronomy. Moreover, many societal problems are caused by technology that science often directly contributes towards. The key science policy issue is to ensure the greatest compatibility between the outputs of the research system and the ability of society to utilise those outputs and generate long term, sustainable benefits.[21] In this respect, metrics can be hugely valuable.

In particular, a range of partial metrics can play an important supporting role in managing research to improve this societal alignment. Their role is to open up and encourage informed decisions in evaluation, not to displace judgement. They provide a useful counter to the potential problems of having powerful parts of the research community dominating decision making (e.g. via peer review in appointment or funding committees). Research with diffused social benefits and hence a bigger 'free rider' problem, tends to be politically vulnerable under such circumstances. As Sarewitz (1996, p.49) notes the effect of allowing these powerful groups to dominate decision making is to 'discourage the creation of new knowledge in neglected areas of science, concentrating new expenditure on areas that are

http://www.surf.nl/binaries/content/assets/surf/en/knowledgebase/2011/Users+narcissism+and+control.pdf on 30 June 2014).

[21] Sarewitz D. (1996), *Frontiers of Illusion*. Temple University Press.

already generously supported, stifling democratic discourse over research priorities and insulating the basic research system from social and political accountability'.[22]

The second danger the science policy community needs to avoid is allowing metrics to be used in inappropriate ways. As this introduction section has shown, there are major problems with the application of metrics in inappropriate ways. The lessons of the financial crisis (see appendix) suggest that the misapplication of metrics to govern science could have very negative implications. Metrics should not be used to manage science and displace expert judgement. However, they do have important roles in opening up democratic deliberations over research priorities, improving accountability and better aligning the research system with the social and economic needs of society. When used in this way, metrics have the potential to be hugely beneficial, and we strongly support their further development and use.

[22] Ibid.

Specific responses to questions

Identifying useful metrics for research assessment

- *What empirical evidence (qualitative or quantitative) is needed for the evaluation of research, research outputs and career decisions?*

The empirical evidence needed will be highly dependent on the type of research and the goals pursued. There are no 'magic bullets', either in terms of qualitative or quantitative methods. Different levels of analysis (national, organisational, individual) require different types of evidence. Different fields (biology, medicine, history, art) require different evidence. Different stages in a researcher's career are often associated with different contributions (focused scholarship, PhD training, community services such as editorship). If we also include evidence of societal impact, the forms of evidence are even more diverse.[23] Given this **extreme diversity**, there are significant dangers of adopting overly rigid standards with regard to metrics.

Given this diversity of evidence, and the basic incommensurability between different types of contributions, the difficulty of using quantitative indicators is all too apparent. They can certainly be useful, but they are never more than an incomplete indicator of the characteristic being sought (e.g. 'output', 'productivity', 'impact', 'value') and often a very biased one at that. Therefore, **while metrics or indicators can help evaluators in making a judgment, they should not be considered as providing judgement on their own**. Metrics on citations for example may suggest impact, but only judgement will distinguish between this impact being driven by genuine research contributions or by other factors, such as fashion or fad, that do not reflect quality but nevertheless drive up that particular metric.

- *What metric indicators are currently useful for the assessment of research outputs, research impacts and research environments?*

There is an extensive body of literature describing the responsible use of bibliometric indicators such as number of publications, citations, etc. This is well established in the scientometrics community, which has arrived at a certain consensus on the following:

  - Numbers of publications and citations may be useful, but they need to be compared within the same field. Comparison between fields is however, controversial and in the case of interdisciplinary or applied research it may be necessary to investigate the effects of various forms of normalisation.[24]
  - The H-index is a problematic indicator since it conflates production, scientific impact and age of the researcher(s) involved.[25]
  - The Impact Factor of journals should not be used as an estimate of article quality or impact, given the very highly skewed distribution of article quality within a journal.[26]

[23] Molas-Gallart, J., Salter, A., Patel, P., Scott, A., Duran, X., 2002. Measuring third stream activities. Final report to the Russell Group of Universities. Brighton: SPRU, University of Sussex.

[24] Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., Stirling, A., 2012. How journal rankings can suppress interdisciplinarity. The case of innovation studies and business and management. Research Policy 41, 1262–1282. Van Eck, N.J., Waltman, L., van Raan, A.F.J., Klautz, R.J.M., Peul, W.C., 2013. Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research. PLoS ONE 8, e62395.

[25] Waltman, L., van Eck, N.J., 2012. The inconsistency of the h-index. J. Am. Soc. Inf. Sci. 63, 406–415.

- Bibliometric indicators are only reliable above a certain statistical threshold size and hence they are generally not reliable for single individuals.

With regard to research impact assessment (and indeed research environment), the problems in attempting to use indicators are even greater than those found in bibliometrics. Research impact, whether economic or societal, takes an enormous variety of forms,[27] varying not only with research field (impact for an engineer is very different from that for a biomedical researcher, a social scientist or a historian) but also with the nature of the research, the orientation and mission of the research performing institution and the wider regional or national environment and culture. Assessment systems based on the use of only a limited number of impact indicators risk skewing the activities of researchers towards forms of research that score highly on those particular indicators, to the detriment of other, less easily measurable forms of impact that may ultimately be more valuable. Moreover, any national-level assessment system must find some way of normalising across fields, a task that is inherently more difficult than in the case of bibliometric indicators.

- *What new metrics, not readily available currently, might be useful in the future?*

With the growth of the internet and the more recent emergence of 'big data', various web-based or other digital indicators have been proposed, in particular for trying to capture the wider impact of research. However, while such alternative indicators (or alt-metrics) can be devised, it is much less evident what they are in fact capturing or measuring. Moreover, it is far from clear whether such data will ever be rigorous and systematic enough to derive reliable and robust indicators for assessment or other policy purposes.[28]

- *What are the implications of the disciplinary differences in practices and norms of research culture for the use of metrics?*

As noted above, in any large-scale use of indicators to assess research, there is a natural tendency to keep the number of indicators to a reasonable number. Those chosen will inevitably favour the research practices of certain fields (e.g. engineering, biomedical research) but fail to capture adequately the research outputs and impact of researchers in other fields. The latter researchers will then be tempted to change their research topic and practices to a form better suited to scoring highly in the assessment system, regardless of whether this is appropriate for improving the quality or impact of their research.

- *What are the best sources for bibliometric data? What evidence supports the reliability of these sources?*

Currently the Web of Science (WoS produced by Thomson-Reuters) and Scopus (Elsevier) offer the largest coverage of research publications and are probably the most useful for analysis in the UK and Europe (with Scopus arguably being better in the Social Sciences). Readily available is data from Google Scholar. However, Google has been very secretive

[26] Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. British Medical Journal 314, 498–502.

[27] Molas-Golart, J. et al., 2002. *Measuring Third Stream Activities* (downloaded from http://ict-industry-reports.com/wp-content/uploads/sites/4/2013/10/2002-Measuring-University-3rd-Stream-Activities-UK-Russell-Report.pdf on 30 June 2014).

[28] Wouters , P., & Costas , R. (2012). Users, narcissism and control — tracking the impact of scholarly publications in the 21st century. Amsterdam: SURF foundation.

with regard to what sources it includes or excludes, so little credibility can be placed on such data or indicators based on them.

- *What evidence supports the use of metrics as good indicators of research quality?*

Research 'quality' is an ideal that is notoriously difficult to define, let alone opêrationalise or 'capture' with some form of indicators. One of the earliest attempts to grapple with the conceptual issues involved can be found in Martin and Irvine (1983)[29], which distinguishes between quality, influence and impact, and argues that citations provide at best a partial indicator of impact. However, such conceptual distinctions are all too often overlooked, with many users of bibliometric data assuming that citations measure research quality.

Moreover, most bibliometric indicators capture impact on other researchers rather than on society. Given the points made about the need to align the structure of the science system with societal need, it is not necessarily the case that what counts as research quality at the local disciplinary level will coincide with what counts as good research for society.

- *Is there evidence for the move to more open access to the research literature to enable new metrics to be used or enhance the usefulness of existing metrics?*

The shift to more open access for research literature will doubtless open up possibilities for developing new metrics. However, such metrics will face essentially the same problems as those based on the traditional research literature, including both conceptual issues (what the indicators actually capture and what are they missing?) and methodological problems (e.g. how to normalise for field, especially for more interdisciplinary research).[30]

How should metrics be used in research assessment?

- *What examples are there of the use of metrics in research assessment?*

Metrics are increasingly used for research assessment purposes at a variety of levels or units of analysis. They are used at the individual level in decisions involving hiring, promoting, granting tenure and so on. This is despite the fact that most evaluation experts (including bibliometricians) are extremely wary about applying such indicators at the individual level, where too many other factors are at work influencing the degree to which publication and citation data do, or do not, adequately capture whether the research is excellent or not. At most, metrics should be only one of the many inputs discussed by those making such individual-level decisions.

Metrics are used at the departmental level, for example by university managers in trying to decide which are doing better research and therefore deserving of more resources. All too often, issues to do with ensuring adequate normalisation across fields are ignored, resulting in resources being allocated preferentially to areas that happen to score well under the chosen indicators to the detriment of those that do not. It is important to note that there is no single objective normalisation method, and hence the subjective choice of the method inevitably has an influence on outcomes. Such framing choices are rarely made explicit,

[29] Martin, B.R., Irvine, J., 1983. Assessing basic research: Some partial indicators of scientific progress in radio astronomy. Research Policy 12, 61–90.

[30] Bornmann, L., 2014. Validity of altmetrics data for measuring societal impact: A study using data from Altmetric and F1000Prime. ArXiv e-prints: http://arxiv.org/abs/1406.7611 (accessed June 30 2014).

despite their influence on outcomes. As noted, this problem can be partially addressed by the use of multiple partial indicators to establish whether convergence can be achieved or not.

Metrics are used nationally, for example by research funding agencies to compare universities or laboratories. Here, the problems associated with individuals and small numbers are less pronounced, and the users of such indicators tend to be more knowledgeable, aware of the dangers of trying to compare across fields without adequate normalisation, as well as the need to ensure that there is full and proper peer review, with the metrics merely being used to raise questions and focus the discussion rather than as a mechanical tool for arriving at decisions.

With regard to national-level research assessment exercises, the way was pioneered by the UK, starting in 1986. However, up to now, these assessments have been based primarily on peer review rather than on metrics. In the 1990s, one RAE asked for bibliometric data to be provided but the conclusion at that stage was that it added little to the peer review process. In the most recent REF, bibliometric data was used by some panels to help arrive at assessments of research excellence in their fields, while in other areas it was widely believed that panels would not read most of the submissions and would instead make judgements based on Journal Impact Factors alone.

- *To what extent is it possible to use metrics to capture the quality and significance of research?*

Bibliometric indicators such as citations **do not** capture the 'quality' of research – at best, they provide a partial indicator of the impact of the research on other researchers as reflected in subsequent publications. High quality research is not always particularly highly cited, while some low quality research may for various reasons end up being relatively highly cited. Even if it was the case that high quality research was always highly cited, this would not imply that highly cited research was always high quality.

- *Are there disciplines in which metrics could usefully play a greater or lesser role? What evidence is there to support or refute this?*

Scientometric indicators (and indeed most research indicators) were originally devised for use in relation to science, engineering and biomedical research. They are less appropriate for social sciences and currently of little use with regard to arts and humanities. However, as some social sciences have become more 'science-like', both with regard to methodological approach and in terms of relying increasingly on articles in international journals rather than books and other forms of output, so there are now greater opportunities to develop appropriate indicators for social sciences and even to some extent the humanities. Some proposals for how this might be achieved can be found in a SPRU report to the European Science Foundation and four national research councils (including ESRC).[31]

- *How does the level at which metrics are calculated (nation, institution, research unit, journal, individual) impact on their usefulness and robustness?*

[31] B.R. Martin et al., 2010, *Towards a Bibliometric Database for the Social Sciences and Humanities – A European Scoping Project*, A report produced for ESF, ANR, ESRC, DFG and NOW, Brighton: SPRU (downloaded from https://globalhighered.files.wordpress.com/2010/07/esf_report_final_100309.pdf on 30 June 2014).

As stressed above, most indicators become less reliable and less robust when applied at lower levels of aggregation, in particular at the level of individuals, where there are just too many other factors at work influencing the various indicators besides the aspects that the evaluator is interested in. However, at higher levels of aggregation (national, institutional, journal), many of those 'other factors' tend to cancel out so the indicators can provide a more reliable means for comparing output and impact.

'Gaming' and strategic use of metrics

- *What evidence exists around the strategic behaviour of researchers, research managers and publishers responding to specific metrics?*

Whenever one attempts to 'measure' certain aspects of a system, one inevitably changes that system.[32] Researchers are intelligent rational actors who respond to the incentives linked to the application of certain indicators, modifying their behaviour to maximise their 'score' in terms of those indicators. In some cases, those changes in behaviour can be argued to correspond to improvements in the quality or impact of the research, but in other cases the relationship is far less obvious. Indeed, some of the changes of behaviour encouraged by the use of metrics are certainly detrimental. In earlier forms of the Australian research assessment system, researchers were rewarded for the number of articles they published in journals. Many responded by splitting their published outputs into smaller components ('least publishable units') and seeking publication in lesser status journals.[33] This enabled them to maximise the financial returns under that resource distribution system but was almost certainly detrimental to the longer-term health of Australian research. When the effects of this game-playing became apparent, the assessment system was changed.

One metric that has gained much attention in recent years is the Journal Impact Factor (JIF). Researchers or groups of researchers are often now judged on their ability to get their articles published in journals with the highest impact factors. Those journals inevitably tend to be in the mainstream of established disciplines, which means that those pursuing more interdisciplinary research are at a considerable disadvantage. Researchers (particularly younger ones) are therefore encouraged to carry out research that fits within the top disciplinary journals, even though such research may be less creative and important in the longer term.

Journal editors are now increasingly judged by publishers and others by their success in raising the impact factor of their journal. This has encouraged various forms of game-playing. Editorials have appeared, for example supposedly to inform young or less experienced researchers what sorts of papers the journal in question publishes, which then just happen to cite all the publications in that journal from the last two years (thereby single-handedly raising the JIF by 30% or more). Many authors, on being informed by a journal editor that their paper is now close to being accepted, are 'requested' to add in a number of articles in that journal *from the last two years.* This process of 'coercive citation' was systematically

[32] There are obvious parallels here with Heisenberg's uncertainty principle and with
the Hawthorne effect in management (Martin, B.R., 2011. The Research Excellence Framework and the 'impact agenda': are we creating a Frankenstein monster?, *Research Evaluation*, 20, 247-254 – see p.250).

[33] Butler, L., 2003. Explaining Australia's increased share of ISI publications—the effects of a funding formula based on publication counts. Research Policy 32, 143–155.

investigated in 2013 and found to be now all too prevalent.[34] More recently, evidence has emerged of cross-journal citation circles systematically adding in references to another journal (e.g. in 'Special Issues') to inflate that other journal's JIF artificially. As a result of these various practices, the JIF metric has now lost virtually all its credibility.

- *Has strategic behaviour invalidated the use of metrics and/or led to unacceptable effects?*

Yes (see example above with regard to JIF). There is also growing evidence that the increasingly widespread use of research performance indicators has been one of the factors encouraging certain researchers to 'cut corners' and engage in various forms of research misconduct (e.g. fabricating data, plagiarism, self-plagiarism) or at least in research behaviour that most would regard as inappropriate (e.g. salami publishing).[35]

- *What are the risks that some groups within the academic community might be disproportionately disadvantaged by the use of metrics for research assessment and management?*

Those whose research falls outside the disciplinary mainstream (perhaps because they have been responding to government policy to engage more closely with 'users' and their research needs) will inevitably be penalised by the use of research metrics, as will those whose research outputs are such that they are not suitable for publication in 'top' journals. Over time, this is likely to discourage more risky, longer-term research, to the detriment of the health of the nation's science.

There is a danger that inappropriate use of such metrics to inform appointments could lead to a significant sorting effect in the academic labour market. Moreover, given that the citation patterns within a portfolio of research projects will be highly skewed and subject to significant survivor bias, there is a danger that resources are inappropriately diverted. Under such conditions, focusing resources on research strategies with high variance in outcomes may well have a detrimental impact on overall outcomes if high-risk high-reward strategies are associated with below-average performance. Similarly, more innovative approaches with lower private returns but higher social returns will be discouraged.

For example, the experiences of many years of evaluations of ESRC projects has highlighted the value of long-term, high-trust, interdisciplinary engagement with research users in generating societal impact. Many researchers perceive, correctly or not, that this style of research is less well supported in research evaluations and they are actively discouraged from undertaking it by research managers as it conflicts with the generation of outputs likely to score highly on assessment metrics.

- *What can be done to minimise 'gaming' and ensure the use of metrics is as objective and fit-for-purpose as possible?*

Researchers will always devise ways to maximise their 'score' on the chosen indicators. As a result, once an indicator is introduced as a performance measure, it rapidly loses its ability

[34] Wilhite, A.W., Fong, E.A., 2012. Coercive Citation in Academic Publishing. Science 335, 542–543.

[35] B.R. Martin, 2013, 'Whither Research Integrity? Plagiarism, Self-plagiarism and Coercive Citation in an Age of Research Assessment', *Research Policy*, 42, 1005-1014.

to capture and measure the original intended characteristic of research.[36] To stay one step ahead of such gaming, those responsible for the assessment system would need to keep changing the chosen indicators. However, this is costly in terms of effort as well as sending the research community continuously changing messages about what sort of research they are trying to encourage. At best, one will end up with a 'Red Queen effect', where researchers are running ever faster in order to stay in the same place. In short, there is no 'technological fix' (or 'indicator fix') to the problem of gaming by those being assessed. To imagine otherwise is a snare and a delusion.

[36] In economics, this is known as Goodhart's Law (see Goodhart, C.A.E., 1975. Monetary relationships: a view from Threadneedle Street. In *Papers in Monetary Economics*, Vol. 1. Reserve Bank of Australia.)

Appendix: A Warning from the Financial Crisis

The use of metrics to measure and assess institutions, people, processes and outputs is not unique to science policy. While sometimes such metrics have been extremely useful in helping to co-ordinate and improve the performance of complex socio-technical systems, in other instances their impact has been less positive.[37]

Where metrics are useful, they are often applied to help co-ordinate relatively simple systems where the representational accuracy of a single measure can be used to co-ordinate a system.[38] However, in other instances, where the thing being co-ordinated is difficult to measure and define, the application of metrics is more difficult and if used inappropriately can lead to dysfunctional behaviour. This is particularly a problem when metrics imperfectly capture what they are intended to measure, and feedback loops emerge through inappropriate incentive systems that influence behaviour in ways that generate qualitatively different outcomes.

The risk to the research system would be that eventually metrics on quality would be used to *define* quality, generating a system in which the point of academic research is to produce citations and get published in journals with high impact factors. If there are easier ways of generating citations than publishing high-quality work, eventually they will tend to dominate, and confidence in the entire system runs the risk of collapsing if funders, which in most countries are governments, realise that research has become an irrelevant game, disconnected from society and its problems.

This can be seen with the recent financial crisis where the inappropriate use of a particular metric to indirectly evaluate risk was combined with an incentive system that encouraged gaming. The problems emerged with the use of collateralized debt obligation (CDOs) where large numbers of assets would be pooled, and then sliced into tranches (equity, mezzanine, senior, etc) based on the order in which they would subject to default.[39] This was an innovative way of managing risks, and allowed financial institutions to generate AAA-graded products from 'junk'. The value of the resulting products was clearly influenced by the risk of default, which can be calculated (imperfectly) by analysing previous patterns of default and assuming the future will be like the past.

However, rather than calculating those risks, a new set of metrics were introduced that indicated the risks of default from the cost of insurance (using Credit Default Swaps, insurance contracts that pay out in the event of default) rather than direct calculation. New methods to work out the correlations in defaults from the cost of insuring a particular tranche allowed bespoke products to be valued in relation to standardized indices (with second order index tranching used to hedge against shifts in correlation). Banks could then create and sell products and buy insurance against default, seemingly enabling them to increase the value of 'junk' while keeping the risks off the books. This led to an increase in the value of 'junk' assets, and the creation of a very large market for CDOs and CDSs that was in part driven by people realising the risk was becoming massively underpriced. Eventually the market for CDSs collapsed and the CDOs could not be properly priced, leaving huge uncertainties

[37] See for example Power, M (1997) *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press.

[38] See Beniger M. (1986) *The Control Revolution*, Harvard University Press

[39] A similar process is at work with firm financing, where debt investors get paid back first when a firm goes into bankruptcy, and therefore take less risk than equity investors who get paid last.

about the liabilities and hence the value of banks' balance sheets. This led to a collapse in inter-bank lending and the financial crisis.

Using metrics to evaluate research runs a similar potential risk. The quality of research, like the quality of bundles of loans, can be either evaluated through expert analysis, using peer review for research or relationship banking for loans. Or, we can use an indirect indicator like citations (or insurance costs). In both cases a logical jump is made. With science we move from assuming high quality science will have larger numbers of citations, to assuming larger numbers of citations implies high quality science. With loans we move from assuming higher quality loans will have lower insurance costs, to assuming lower insurance costs imply a higher quality loan. In both cases, even if we accept the assumed link between quality and the metric, "all y are x" does not imply "all x are y". If we then introduce a reward system, we create a feedback loop that can be dysfunctional, with researchers focused on creating citations (which can be through high quality work, but also through producing more disciplinary, incremental, fad-driven work, or by gaming the system), and bankers focused on profiting from selling mispriced assets. In both instances one would expect professional norms to become eroded, given their lack of influence on outcomes and lack of alignment with reward mechanisms.

The financial crisis therefore offers a valuable lesson about the inappropriate use of metrics linked to reward systems. The widespread belief that metrics are used in this way in academia, and evidence of the corrosive effect this belief has had on professional norms is therefore concerning. Metrics have enormous potential to improve the governance of science when used in an appropriate way. But the lessons of the financial crisis suggests regulators need to tread very carefully.