4

The Mismeasurement of Quality and Impact

Paul Wouters

In this chapter, I would like to discuss the implication of the problematic relationship between "gaming the system" and "properly operating in the system" for research evaluation. As Biagioli and Lippman (Introduction, this volume) argue, the current audit culture in academia has created new dimensions of academic misconduct. The performance indicators that are supposed to objectively measure the value of scientific production have become the objects of new forms of manipulation. Because research evaluations have become reliant on proxy indicators that measure only indirectly what they are supposed to represent (either quality, impact, or societal relevance), it is increasingly hard to see the difference between indicator manipulation and authentic high-value research. According to Biagioli and Lippman (Introduction, this volume), "the traditional locus of evaluation—the publication's claims—has become technically irrelevant to metrics regimes based on impact factors." Reading has stopped being necessary in institutional evaluations, they argue, while the role of metadata has shifted from descriptive to evaluative. While I think the practice of evaluation is still quite messy (in many assessments, reading might still happen), and the role of indicators is not so straightforward (the Journal Impact Factor, though the most popular indicator, is not the only game in town and neither representative for all), I do agree that Biagioli and Lippman very perceptively sketch a portrait of the dominant tendencies in the academic audit culture and of the market enabled by the infrastructure of citation (Wouters, 2014). Indeed, if these trends remain dominant, high performance scores will be *made* identical to high-value work. Those aspects of academia that are not presentable in indicators may then no longer be supported (by researchers themselves as well as by their employers) and may have to migrate to the area of works of love (amateurs)—or become extinct. Scientific research may in such a scenario still lead to exciting innovations, but the notion of knowledge itself as

an important aspect of human culture would no longer be an important *leitmotiv* of universities. Scientific research will then have been instrumentalized in its totality.

If we wish to somehow preserve, or restore in new forms, this aspect of academic life, it seems useful to imagine alternative scenarios. This should also involve alternative forms of research assessments and performance evaluation of individual researchers, research groups, and institutions. Imagining these will require a bolder redesigning of current evaluation protocols at universities than is currently the case. In particular, I will argue, evaluation experts and scientometrians need to go beyond the popular concept of "informed peer review" (Moed, 2005). It is not sufficient to claim that peer review and indicators need to be combined in intelligent ways because the very basis of what counts as an intelligent combination is at stake. What is needed is a more radical *recontextualization* of indicators as well as qualitative evidence in assessments.

Before discussing a possible alternative to the market-oriented, indicator-driven forms of evaluation, it makes sense to briefly outline how research assessments have become influential in shaping academic life.

First of all, the conduct of research has become so strategic that it is vital for researchers to be visible at both the national and international level. Second, partly as a result of the success of scientific and technological research in many fields, and partly as the consequence of independent policy developments, the traditional academic autonomy no longer exists. Apart from some exceptions, the scholarly community on which Robert Merton built his sociology of scientific norms (Merton, 1973) has become pervaded by extra-academic interests and communications (Leydesdorff, 2000; Shapin, 2008). Third, research groups and individual researchers in the public research system in all disciplines are subjected to recurring institutional assessments in which performance must be made visible in the terms of that specific institution. Usually, scientific quality (or excellence) and societal impact are the main pillars. Other criteria such as the quality of teaching and PhD training, viability and feasibility of the research plans, and, last but not least, earning power are drawn in as well. The key issue is that the results of these evaluations produce the symbolic capital with which the researchers can—or cannot—participate in the next cycle of research.

Two different forms of research evaluation are usually distinguished in opposition to each other in the interdisciplinary debate about the best way to assess academic research and universities. The first is the qualitative judgment called peer review, based on the assessment by scientific

experts usually working in the same or a related field as the research group. The second is assessment by indicators often based on simple or complex forms of citation analysis in which high numbers of citations are seen as proxy for influence or quality. Academia has a lively debate about the weight each form of evaluation should have and about the extent to which it is possible to combine one with the other (Hicks et al., 2015; Wilsdon et al., 2015).

In this debate, it is often overlooked that the two forms of assessment are intrinsically and intensely linked to each other. This was not yet the case when the Science Citation Index was invented by Eugene Garfield (Garfield, 1955; Wouters, 1999, 2014; Csiszar, this volume, chapter 1). But as a result of the rise of the use of citation-based performance indicators since the early 1980s, first in national science policies and later in the management of universities and research institutes and in global university rankings, both methods are no longer purely quantitative versus qualitative but have intertwined and interpenetrated each other. This is relevant to the debate about research integrity and norms for proper scientific behavior because, as we will see, comparable forms of mixing are making the identification of improper behavior less evident. So it is worthwhile to spell out the connections and mutual pollution between peer review and assessment by indicators in more detail.

Citing relationships are, in the end, based on the decisions by authors of scholarly papers to include formal references in their bibliography to scientific work that they deem relevant (implicit references do not end up in citation indexes, although they may be very important intellectually). These citing relationships are usually concentrated within the same research area combined with additional interdisciplinary connections to other literature. The decision to cite a specific reference, and not an alternative piece of work, is shaped by a complex mixture of intellectual and strategic motives. The guidelines of scholarly journals or book publishers regarding number and type of references form the template for these decisions, but there is still a lot of leeway for the authors to express their preferences. It is difficult, if not impossible, to clearly separate intellectual and strategic motives. In addition, it should be stressed that both motives are of a social nature.

In addition to this basic connection at the individual level, there is also a group connection between peer review and citation-based indicators: they draw upon the same scientific or scholarly community. This may vary by type of document and by discipline, but, very often, the researchers in the citation network are also involved in the regular peer-review

work of journal and book publications. An interesting question, not yet frequently studied, is to what extent these communities are also the basis for post-publication peer review such as the national research assessment exercises. The latter forms of peer review may draw upon more interdisciplinary networks and hence be more removed from the core peer networks in particular research areas.

A third connection between citation and peer-based evaluation is the reflexive loop that has been created by the emergence of citation-based performance indicators. Because researchers have become aware, on a large scale, that their bibliographies may influence the careers of the researchers they cite, their "citing behavior" will be affected by this knowledge. The strategic motives may therefore have become more important because the competition among researchers has been extended to the domain of metadata such as numbers of citation. Ethnographic research of publication and evaluation practices has shown that researchers tend to reason quite strategically about both their publication outlets and their bibliographies, although it is also clear that this varies strongly by field and type of research (Rushforth and de Rijcke, 2015). This does not mean, *inter alia*, that bibliographies have become completely dishonest and unreliable as a source for intellectual queries, but it has certainly made the sociological interpretation of citation frequencies and networks more complex. It is an example of the basic reflexive nature of communication behavior and networks (Leydesdorff, 1995).

This third connection may have important consequences for the conduct of post-publication evaluation of research performance. Most formal evaluation protocols do not include instructions to use citation indicators and some even discourage the use of indicators such as the Journal Impact Factor or the Hirsch Index. However, this does not mean that these indicators do not play a role in, for example, the preparation of a committee session by individual evaluators. Because the citation indicators have basically become an easily available resource and can indeed be seen as a citation infrastructure (Wouters, 2014), consulting Web of Science–, Scopus-, or Google Scholar–based citation scores may be matter of course without much conscious deliberation. They have even led to "folk citation theories" upon which researchers draw in their interpretation of these data (Aksnes and Rip, 2009; Rushforth and de Rijcke, 2015; Wilsdon et al., 2015). For example, biomedical researchers are usually quite aware of the technical limitations of the impact factor and use these as components of a field-specific interpretation of the journals. High impact factors may then be interpreted as indicating journals with a large

number of submissions and a high rejection rate. Hence, publishing in these journals is a sign of success in the fierce competition for recognition. Rushforth and de Rijcke (2015) show how the impact factor is used in comparable ways as a judgment device that is already deeply engrained in collaboration and publication strategies. This does indeed indicate that the impact factor, as Biagioli and Lippman (Introduction, this volume) argue, functions as a measure of value in a market, in other words as a currency.

In the recent Higher Education Funding Council for England report *The Metric Tide* (Wilsdon et al., 2015), a number of norms for proper evaluation have been proposed. The unifying concept here is "responsible research metrics," which makes an important reference to the European policies on "responsible research and innovation." Responsible research metrics should be seen as those practices in using quantitative performance indicators that are attuned to five core principles:

- **Robustness**: basing metrics on the best possible data in terms of accuracy and scope
- **Humility**: recognizing that quantitative evaluation should support—but not supplant—qualitative, expert assessment
- **Transparency**: keeping data collection and analytical processes open and transparent, so that those being evaluated can test and verify the results
- **Diversity**: accounting for variation by field, using a variety of indicators to reflect and support a plurality of research and researcher career paths
- **Reflexivity**: recognizing the potential and systemic effects of indicators and updating them in response

These principles have been formulated on the basis of the recognition of the complex interplay between peer judgment and citation-based indicators. If national and institutional research assessments and the building of the databases used in them would consistently adhere to these principles, it would surely represent important progress in evaluation practices. But how should we interpret the second principle, "humility"? It can easily be read as the incorporation of quantitative indicators within a framework that is dominated by peer-review and expert judgment. But is this really a form of humility?

The recent Leiden Manifesto formulated a comparable norm for the use of quantitative performance indicators (Hicks et al., 2015): "Quantitative evaluation should support expert assessment." The manifesto is

a warning against exaggerated forms of performance-based indicators and pleads for a judicial combination of quantitative and qualitative evidence in research evaluation. This basically builds forth on the concept of "informed peer review," which also aims to combine citation analysis with peer review, both in the context of peer review itself (where indicators would simply be a form of evidence next to other forms of qualitative or quantitative evidence) and as a check on the integrity of the peer-review process itself (Moed, 2005; Butler, 2007).

Informed peer review can be seen as an attempt at triangulation: if we can collect more evidence and this evidence points in the same direction, surely we have a more robust foundation for our conclusions? However, this is based on the assumption that the two forms of evidence are independent of each other. As we have seen, this is only partially the case. In addition, a practical problem arises when citation analysis and peer judgment are in conflict with each other about a particular research performance. On what basis should the evidence be weighted? The idea that expert judgment is always better ignores the gatekeeper role of scientific reviewers and is naïve with respect to the strategic motives operating in the process of peer review. Conflicting outcomes of peer review and citation analysis may for example be a signal that the reputational mechanisms in the academic system are not keeping up with novel developments in research. Peer review may also play in the hands of "old boys" networks and discriminate against women and ethnic and intellectual minorities in science. Relying mainly on peer review may delay interdisciplinary innovation because this often entails not only a reconfiguration of substantive or methodological research areas, but may also mean a redefinition of the very criteria of what counts as high quality in research. In other words, peer review in a particular discipline may simply filter out radical innovation because it is not recognized as high quality.

So we cannot always rely on peer review as the ultimate arbiter. But the same holds for quantitative performance indicators. Neither the number of publications nor their number of citations, normalized for field, document type, and age or not, can simply be interpreted as proxies for quality or impact. Researchers with exceptional publication numbers may be opening up an exciting new field or they may be very good in gaming the performance system. A high number of citations may indicate great research with a huge influence or they may come from humdrum me-too research or even citation cartels. And a low number of citations may result both from less interesting research and from path-breaking studies that are not

yet recognized. So we need some form of judgment to assess the value of publication or citation performance criteria (the same holds for indicators of earning power). But this brings us back to the experts involved in the peer-review system. We seem to be caught between a rock and a hard place.

And yet, this does not hinder assessments to take place. In fact, the international research system is a buzzing evaluation machine (Dahler-Larsen, 2012). The construction of peer review and metrics-based assessment as two opposites, common in generalized debates about research evaluation and in most research policy discussions, is a false one in the sense that it is not what happens in the varied practices of research evaluation. Rather than the dichotomy of qualitative versus quantitative, or peer review versus measurement, we should focus on the context of evaluation. The tendency to speak of research evaluation as such tends to ignore the wide variety of practices that are bundled in this container concept. The weight of impact indicators, for example, varies strongly between the assessment of the results of a PhD student and the ranking of a university in comparison to its international peers. The tendency to speak of research evaluation as a somehow integrated institution is a form of purification that tends to make invisible precisely that which should be foregrounded in an alternative discourse. In this light, the proposals to formulate principles of responsible metrics and responsible evaluation are only a first step. These could still easily be incorporated in the framework of a market-oriented audit culture. To enable a true alternative to the dominant trends in academic research evaluation, we need to complement the concept of responsible metrics with the recognition that valuing is principally an act of judgment in context. Decontextualized information, whether peer based or indicator based, needs to be put back into context if we wish to create a strong barrier in assessment practices to academic misconduct in all its novel forms.

Technically, my proposal is to replace the notion of "informed peer review" as the supposedly most nuanced approach in research assessments by "contextualized judgment," which 1) puts context central and 2) does not create a false dichotomy between peer review and indicator-based assessment. It takes into account the flexible, and often quite ingenious, ways in which researchers attach meaning to constructs like peer opinion and citation indicators (Rushforth and de Rijcke, 2015).

Politically, following this approach would put two questions central in the construction of new evaluation protocols and procedures:

1. How will this evaluation design influence the creative process of knowledge creation?
2. Who is in control of the agenda setting and the research process?

The first question is about coping with matters of perverse effects, the alignment of the criteria in the evaluation and the mission of the specific research group or program, and effects on the texture of power in the field. The second question addresses the way quality is managed in the field and its connections with stakeholders, nonacademic partners, and society at large. Both questions point to the political aspects of norms for evaluation. This is as it should be since matters of evaluation are deeply political matters and deal with the question of how we wish to live and what kind of society we are creating (Mol, 2002; Thurtle and Mitchell, 2002).

Notes

Acknowledgement: I would like to thank Thomas Franssen, Sarah de Rijcke, and the referees and editors for their comments on an earlier version. This work was partly funded by the Norwegian Research Council through the Center for Research Quality and Policy Impact Studies (R-QUEST) (https://www.r-quest.no/) and by the Riksbankens Jubileumsfond (Sweden) through the KNOWSCIENCE project (https://www.fek.lu.se/en/research/research-groups/knowscience).

References

Aksnes, D. W., and A. Rip. 2009. "Researchers' Perceptions of Citations." *Research Policy* 38(6):895–905. http://doi.org/10.1016/j.respol.2009.02.001.

Butler, L. 2007. Assessing University Research: A Plea for a Balanced Approach." *Science and Public Policy* 34(8):565–574. http://doi.org/10.3152/030234207X254404.

Dahler-Larsen, P. 2012. *The Evaluation Society*. Stanford, CA: Stanford University Press. Retrieved from http://www.amazon.com/The-Evaluation-Society-Peter-Dahler-Larsen/dp/080477692X.

Garfield, E. 1955. "Citation Indexes for Science Through Association of Ideas." *Science* 122(3159):108–111.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. 2015. "The Leiden Manifesto for Research Metrics. *Nature* 520:429–431. http://doi.org/10.1038/520429a.

Leydesdorff, L. 1995. *The Challenge of Scientometrics: The Development, Measurement, and Self-Organization of Scientific Communications*. Leiden, The Netherlands: DSWO Press.

Leydesdorff, L. 2000. "A Triple Helix of University-Industry-Government Relations." *The Journal of Science and Health Policy* 1:43–48.

Merton, R. K. 1973. *The Normative Structure of Science*. Chicago: University of Chicago Press.

Moed, H. F. 2005. *Citation Analysis in Research Evaluation* (Vol. 9). Dordrecht, The Netherlands: Springer.

Mol, A. 2002. *The Body Multiple: Ontology in Medical Practice*. Durham, NC: Duke University Press.

Rushforth, A., and S. de Rijcke. 2015. "Accounting for Impact? The Journal Impact Factor and the Making of Biomedical Research in the Netherlands." *Minerva* 53(2):117–139. http://doi.org/10.1007/s11024-015-9274-5.

Shapin, S. 2008. *The Scientific Life: A Moral History of a Late Modern Vocation*. Chicago: University of Chicago Press.

Thurtle, P., and R. Mitchell. 2002. *Semiotic Flesh: Information and the Human Body*. Seattle: University of Washington Press.

Wilsdon, J., L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, R. Kain, S. Kerridge, M. Thelwall, J. Tinkler, I. Viney, P. Wouters, J. Hill, and B. Johnson. 2015. *The Metric Tide: Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. http://doi.org/10.13140/RG.2.1.4929.1363.

Wouters, P. 1999. *The Citation Culture*. Amsterdam: University of Amsterdam. Retrieved from http://garfield.library.upenn.edu/wouters/wouters.pdf.

Wouters, P. 2014. "The Citation: From Culture to Infrastructure." In B. Cronin and C. R. Sugimoto (Eds.), *Beyond Bibliometrics: Harnessing Multidimensional Indicators of Scholarly Performance* (pp. 47–66). Cambridge, MA: MIT Press. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10650330.