Engaging Science, Technology, and Society 3 (2017), 108-118 DOI:10.17351/ests2017.115

Bridging the Evaluation Gap

PAUL WOUTERS [1]
LEIDEN UNIVERSITY

Abstract
Paul Wouters' essay is concerned with bridging the gap between what we value in our academic work and how we are assessed in formal evaluation exercises. He reflects on the recent evaluation of his own center, and reminds us that it is productive to see evaluations not as the (obviously impossible) attempt to produce a true representation of past work, but rather as the exploration and performance of "who one wants to be." Reflecting on why STS should do more than just play along to survive in the indicator game, he suggests that our field should contribute to changing its very rules. In this endeavor, the attitude and sensibilities developed in our field may be more important than any specific theoretical concepts or methodologies.

Keywords
evaluation; performativity; citation analysis; impact

Unexpected Disappointment
I had a sinking feeling. I had just left the introduction interview with the Evaluation Committee that was conducting the site visit at our institute, the Centre for Science and Technology Studies, Leiden University (CWTS), as part of the Dutch institutional research assessments,[2] and it was not the start I had hoped for. Nervously, I phoned my colleague who would be the next person to be interviewed about our bibliometric self-analysis and who had led its design. "They are rather disappointed," I told him. "They think we violated the professional standards in the field by not normalizing the indicators; standards that we—as they pointed out—helped develop ourselves.

[1] Paul Wouters, Email: p.f.wouters@cwts.leidenuniv.nl

[2] Such assessments occur every six years.

Apparently, our attempt at bibliometrics 3.0 is seen as old-fashioned bibliometrics 1.0. You will really have to pursue this further and discuss the technical fine points as well."

Clearly, playing the indicator game (Fochler and de Rijcke 2017) is risky, especially when expectations clash, as happened to me in this interview (for the assessment outcome see http://bit.ly/2druTwm). Ironically, the clash resulted from our attempt to prevent that the assessment would become a ritual that we needed to formally attend to, but that would have no real meaning for the further development of our research. We also wanted to trigger the committee's genuine interest and develop common intellectual interests, while respecting the different roles we would have to play in the game. In other words, we wanted to bridge what we have earlier called "the evaluation gap" (Wouters et al. 2014).

The Evaluation Gap

The evaluation gap is the phenomenon—experienced by many academics—that the criteria in assessments do not match the character or goals of the research under evaluation or the role that the researcher aims to play in society. Research evaluation has tended to privilege a rather narrow view of what counts as legitimate and valuable, and important aspects of scholarly performance may be ignored or deemed irrelevant in it. The resulting evaluation gap is relevant in this debate about the indicator game, because the gap is partly of our own making. In fact, popular strategies in the indicator game tend to confirm and reify the evaluation gap, each in their own way. For example, a famous physicist, who spends his Saturday mornings reviewing physics papers for the journal *Science*, once explained to me that one must produce a steady flow of publishable papers in order to survive, so that one can do "the real thing" (risky and adventurous research that has no guaranteed outcome) in the time that is left. Clearly, this keeps the evaluation machines focused on international journal articles nicely humming. While a rational strategy for the individual, it creates a shadow reality that decouples evaluation and knowledge creation and thereby sustains, legitimates and perhaps even increases the evaluation gap.

We wanted to bridge the gap by doing the opposite: not closing off our "real thing" but exposing our work in progress, including our doubts and uncertainties, to the assessment. We had hoped that this would be reciprocated by the evaluators entering our space and joining us in a common research enterprise. In other words, we wanted to turn the evaluation into what Fochler & de Rijcke (2017) call an "evaluative inquiry."

The Quantitative and the Qualitative

An important theme in our little experiment is the distinction between quantitative and qualitative research assessment. In many countries in Europe, national research assessment exercises are either mainly metrics or mainly peer review based, and in some countries they are strongly linked to research funding (e.g. in the UK) whereas in others they are not (e.g. in the Netherlands) (Hicks 2012; Aagaard, Bloch, and Schneider 2015; VSNU, NWO, and KNAW 2015; Wilsdon et al. 2015; Sivertsen 2016; Whitley and Gläser 2007). This policy context has strongly shaped the discourse in the form of a dichotomy between indicator-driven or peer review based research evaluation, also with respect to other forms of evaluation (e.g. in human resource management and project funding). The result is a highly politicized debate in which every critique of performance indicators (or their underlying data) results in strengthening the case for peer review, and vice versa. It has also led to the pragmatic concept of "informed peer review," which combines peer judgment with quantitative indicators and data in an attempt at triangulation (Moed 2005; Butler 2007; Moed 2007). Most bibliometricians saw this as an orderly process in which professionally produced high quality bibliometric data "inform" the peer reviewers who subsequently exercise their professional judgment.

However, in the practice of knowledge creation and evaluation, we observe a much more diversified interaction and mixing of peer based evaluations and metrics informed assessments (Wouters 2016b; Rushforth and de Rijcke 2015; Wouters et al. 2015; Sauder and Espeland 2009; Hicks et al. 2015; Wouters 2016a). As a result, "metrics" and "peer review" are mutually interpenetrating each other. With the exception of new areas of interdisciplinary research, peer review and citation communities (if we can define them as such for the moment) share many members and as a result citation patterns and peer based recognition interact strongly. The decision to cite a particular work is in the end based on a qualitative value judgment. Moreover, since the institutionalization of the citation in university management, researchers are aware of the potential value of a reference to the cited authors and institutions and may take this into account in their publishing and citing behavior. Hence, citation scores will be influenced by perception of reputation, whereas peer judgment may be informed by the value of performance metrics. In other words, most peer review may already be "informed" by metrics, albeit perhaps not in the systematic and expert led way the proponents of informed peer review would have wished for (Moed 2005; Leydesdorff, Wouters, and Bornmann 2016; Wouters et al. 2013; Moed, Glänzel, and Schmoch 2005).

The Performative Nature of Assessment

A second theme is the performative nature of research evaluations. By purporting to measure reality, they effectively transform reality. And this is not only a matter of representation and symbolic interaction, but of deeply material forms of exchanging and acting. Reality is enacted (Law 2004; Mol 2002). Peer review and performance metrics do not so much *record* pre-existing quality, reputation or excellence, but literally *enact* them. Outside of quality assessment practices, "quality" does not actually exist. Given the strong interactions between quantitative indicators and qualitative judgment, then, "quality" in all its versions (excellence, impact, top, etc.) is a hybrid result of the research evaluation exercises that have blossomed at all levels in academia since the 1980s. The way we configure research assessments shapes the criteria for quality in the next generation of researchers and influences their academic identity. Thus, there is much more at stake in the indicator game than the choice between qualitative judgment and quantitative indicators.

Bringing these two themes together, we conceptualized the research assessment of our center as a form of "situated intervention" (Zuiderent 2015), which would help us to bridge the evaluation gap (rather than mind the gap). This should not be seen as an attempt to get a "real representation" of our research in the evaluation. Rather, we wanted to enact how we saw ourselves both in the recent past and in the near future. In other words, we hoped to shape the research evaluation according to criteria not only of what we already *were* but especially of what wanted *to become*. This can be seen as a specific form of future-oriented evaluative inquiry.

To realize this, we had to overcome existential anxieties in the preparation of the assessment, as well as in the course of it. To mention one: on what basis could we decide what criteria should prevail in the assessment? Our criteria might very well not square with those of the auditors. Our wish to make the research assessment an enjoyable learning exercise was also problematic. Didn't this mean that we were naively ignoring the power games in and around assessments? Why did we expect that the auditors would play along? And how could we ascertain that we were playing the *same* game? Come to think of it, how could we actually know that all members of our *own* staff were playing the same game? On top of this, some of us were concerned about the costs of the game. It is all very well to learn from the exercise, but what if the assessment were negative and the center was threatened with closure as a result? Wouldn't it be much wiser to take the instrumental view on the indicator game, as Bal discusses as "adjusting to the system" (Bal 2017), and define the game as closely as possible to the rules of the book of the national evaluation protocol (the "Standard Evaluation Protocol")? As said, we decided to take the risk and did the opposite.

Engaging the Auditors

To address these risks and to overcome our anxieties we decided to engage the auditors in an open intellectual debate about what the future of our STS center should be and how its societal role should be defined; this was possible thanks to the institutional design of the Dutch evaluation protocol with its emphasis on communication and site visits. So, formally the assessment was retrospective; but we hoped it could also become a process of envisioning the future, spelling out the criteria of quality and impact that were essential to realizing this future, and then on that basis judging how far we already were on that road and what would lie ahead of us. This turned the assessment into an exercise in collective future making, rather than a game of trying to score as high as possible on a set of indicators that were more or less relevant to our work. In other words, we invited the evaluators into our research space, regarded them as better "versions of ourselves" (Strathern 1997, 319), and sent them as much information as possible.

They clearly welcomed this since they asked for even more information, including the full text of all our publications. We wrote a self-evaluation on the basis of our assessment in 2008 and the midterm review in 2012. We performed a SWOT analysis (strengths, weaknesses, opportunities, and threats) of our institute, guided by a professional outsider. Separately, we had a SWOT analysis performed about our contract research and service work. We wrote a short vision about the direction our new research program might take. We worked especially hard on two case studies in which we repositioned several research projects and themes. One focused on university rankings in which we brought together our technical bibliometric work with the critical analysis of the social effects of rankings on academia. The other case study concerned the evaluation of biomedical research in which we synthesized our evaluation studies, ethnographies in medical centers, studies on PhD students and postdoc careers, and (reflexively) our studies of how societal impact of biomedical research is realized as well as imagined. And we performed a bibliometric self-analysis. Given the fact that we are seen as one of the leading centers in bibliometrics, we were quite confident that we could do this ourselves. We frankly had not expected at all that this would disappoint the evaluation committee.

No Role Playing

The massive amount of information and data we provided the evaluation committee was meant not only to enable them to make an informed judgment about our performance, but also and primarily to level the playing field and create the conditions for this shared journey into the future. A related measure was our decision not to fake anything. We took more than a year to

prepare the evaluation but we deliberately abstained from any role-playing exercises or simulations to prepare what each of us was supposed to say to the committee. To drive this point home, as the director I told everyone that if they had any complaints about how things were run, this was the committee to file these complaints to in order to change things. The PhD students and the postdocs were interviewed independently and I hope they used the occasion to the full. (One of the recommendations was that we should improve the structure of our PhD supervision, so I think they did speak up!) The idea of not scripting our interaction with the committee was that scripting would destroy the basis for a shared "evaluative inquiry."

We performed a citation analysis that was at the same time a critical inquiry of what it means to do a citation analysis. An exercise in method as much as in substance. This was partly motivated by the simple and rather old-fashioned criterion of consistency: if we are learning things in our research that have bearing on the accuracy and/or legitimacy of the evaluation process, we should bring that knowledge to the evaluation itself rather than pretend that the customary/existing evaluation method is sufficient to realizing the objectives that the evaluators themselves bring to the process. For example, our bibliometric research had shown that citation densities differ considerably within disciplines and the usual normalization of indicators does not correct for those differences (van Eck et al. 2013). Because of this, we felt we couldn't really make the claim that the relatively high citation scores of the bibliometric research at our center was in itself already evidence of its high quality. In other words, our bibliometric research had undermined a simple, instrumentally advantageous interpretation of bibliometric indicators. Our research had resulted in a possible solution, which we tried out for the first time upon ourselves in this evaluation: "contextualised scientometrics" (Waltman and Eck 2016). Whereas the emphasis in the standard approach is on normalized indicators, contextualised scientometrics is based on simplicity, diversity of indicators and contextual information. It introduces a qualitative element in the core of the bibliometric analysis by linking information about publications to information about their intellectual and social context. The latter must be judged qualitatively. And it contradicts received wisdom in scientometrics about the need for normalized indicators. It probably was this dissonance that puzzled the evaluation committee. Yet, in the end, they gave us the benefit of the doubt on this exercise and their assessment of our bibliometric work as a whole was the highest possible: "excellent." Apparently, our position was strong enough to try out experimental approaches and question some of the evaluation criteria.

Unexpected Enthusiasm

Another challenge was how we would be able to present our work in qualitative science and technology studies. Since 2010 we have broadened the research agenda of the center and started up a number of new research themes: ethnographic research of evaluation practices; survey work about the development of academic careers; and inquiries into the uses and societal impacts of research. Part of this work resulted not in journal articles but in chapters in edited books. These outputs are notoriously less visible in bibliometric indicators. An interesting extra twist is that our qualitative research is *about* research assessments and questions the very role of evaluation and assessment as well as the criteria on which they are based. And we had no idea, really, about how the auditors would read this. Would they be tempted to question their own assumptions? And, given the necessary broadness of their disciplinary span, would they understand one another and work together creatively?

I had expected enthusiasm from the committee about our bibliometric experiment and got puzzlement instead. Also somewhat surprisingly, we received full and enthusiastic support for the new directions our center had taken, including a strong emphasis on the development of our new research lines in STS and policy studies. Not only did they endorse the need to further strengthen this research, they upped the game by advising us to make the synergy between quantitative and qualitative research approaches the central motif for our new research program. We had not yet done this ourselves in such explicit terms and our plans were still coached in terms of strengthening our qualitative STS and policy themes while maintaining and developing our bibliometric strengths. So in a way our assessors perhaps saw more clearly what we were up to than we did ourselves at the time! I have since gone through the notes of my discussions with the committee members (we were not privy to their internal conversations) and my conclusion is that the authenticity and sincerity of our junior and senior researchers must have been pretty convincing. It was not that the committee had become uncritical fans: their report clearly enough points out that some research lines seemed to stagnate or falter, that supervision of junior researchers needs further improvement, and that internal communication requires strengthening. But the key point is that the committee gave these critiques from the point of view of a joint partner in an evaluative inquiry.

STS Informed Attitude was Key

Looking back, I think the key for the success of this little adventure was not the amount of information, the well-crafted nature of our writing, or the sophistication of our indicators. They

key was much more an STS-informed *attitude* that enabled the research assessment to become a self-critical inquiry of the possibilities in the future of our center (given our current abilities and resources) and of the very criteria by which these future activities should be valued. Interestingly, this also created the space for disagreement among us (between and among CWTS researchers as well as committee members) without this jeopardizing the exercise.

Of course, there are a variety of possible solutions to perform this. Our case is just one example and each field and research context would require a different specification. But I hope that it may inspire more experiments in research evaluation. In this respect, the advice from Vikkelsø (2007) seems particularly relevant: "it is neither fruitful nor necessary to make a choice between a politicized, a sterile, an anecdotal or an action-oriented science. Any scientific endeavor runs the risk or has the opportunity—depending on temper—to be all of the above. The key question is how the researcher relates to the object of study and to the socio-material collective in which he or she operates" (Vikkelsø 2007, 298).

Indeed.

Acknowledgements

I would like to thank all CWTS staff as well as the members of the CWTS Evaluation Committee for their role in the CWTS assessment. I thank Sarah de Rijcke, Ludo Waltman, Thomas Franssen, Alex Rushforth, Thed van Leeuwen, Björn Hammarfelt, Wolfgang Kaltenbrunner, and Tjitske Holtrop for the discussions and their comments on earlier versions of this argument. Last, I would like to thank the anonymous reviewers for their valuable comments, and the editors of this volume Max Fochler and Sarah de Rijcke as well as the journal editors Katie Vann and Daniel Kleinman for their guidance that made this contribution possible.

Author Biography

Paul Wouters is Professor of Scientometrics at the Centre for Science and Technology Studies, Leiden University. He has published on the history of the Science Citation Index, on and in scientometrics, and on the way the criteria of scientific quality and relevance have been changed by the use of performance indicators. His PhD thesis "The Citation Culture" (1999) is available here. He has also studied the role of information and information technologies in the creation of new scientific and scholarly knowledge. In this area, he was appointed as leader of two research programs by the Royal Netherlands Academy of Arts and Sciences: Networked Research and

Digital Information (Nerdi) (2000 - 2005) and The Virtual Knowledge Studio for the Humanities and Social Sciences (VKS) (2005 - 2010). The experiences and insights gained in the VKS were condensed in Virtual Knowledge. Experimenting in the Humanities and Social Sciences, a collection edited in collaboration with Anne Beaulieu, Andrea Scharnhorst and Sally Wyatt (MIT Press 2013). He was Principal Investigator of several European research consortia, among others ACUMEN on research careers and evaluation of individual researchers. Paul was coordinator of the Dutch STS Graduate School Science, Technology, and Modern Culture (WTMC) together with Annemiek Nelis (2001-2005). Currently he is chair of the WTMC board. In 1999, he helped create Onderzoek Nederland, a leading professional journal on Dutch science policy (part of Research Professional) and has since published in the journal. He is a member of the editorial board of Social Studies of Science, Journal of the Association of Information Science and Technology, and Cybermetrics, was member of the Council of the Society for the Social Studies of Science from 2006 to 2008, and sits on various advisory boards of international programs and projects. Currently, he is involved in, among others, the PRINTEGER project on integrity in science, KNOWSCIENCE, the Center for Research Quality and Policy Impact Studies at NIFU in Oslo, and he is member of the program board of the ZonMW program to promote responsible research behavior.

References

Aagaard, K., C. Bloch, and J. W. Schneider. 2015. "Impacts of Performance-Based Research Funding Systems: The Case of the Norwegian Publication Indicator." *Research Evaluation* 24 (2): 106–17. doi:10.1093/reseval/rvv003.

Bal, R. 2017. "Playing the Indicator Game: Reflections on Strategies to Position a Group in a Multidisciplinary Environment." *Engaging Science, Technology, and Society* 3: 41-52. DOI: 10.17351/ests2016.111.

Butler, L. 2007. "Assessing University Research: A Plea for a Balanced Approach." *Science and Public Policy* 34 (8): 565–74. doi:10.3152/030234207X254404.

Fochler, M. and S. de Rijcke. 2017. "Implicated in the Indicator Game? An Experimental Debate." *Engaging Science, Technology, and Society* 3: 21-40. DOI:10.17351/ests2017.108.

Hicks, D. 2012. "Performance-Based University Research Funding Systems." *Research Policy* 41 (2). Elsevier B.V.: 251–61. doi:10.1016/j.respol.2011.09.007.

Hicks, D., P. Wouters, L. Waltman, S. de Rijcke, and I. Rafols. 2015. "The Leiden Manifesto for Research Metrics." *Nature* 520: 429–31. doi:10.1038/520429a.

Law, J. 2004. *After Method: Mess in Social Science Research*. London: Routledge.

Leydesdorff, L., P. Wouters, and L. Bornmann. 2016. "Professional and Citizen Bibliometrics: Complementarities and Ambivalences in the Development and Use of Indicators—a State-of-the-Art Report." *Scientometrics*, October, forthcoming. doi:10.1007/s11192-016-2150-8.

Moed, H. F. 2007. "The Future of Research Evaluation Rests with an Intelligent Combination of Advanced Metrics and Transparent Peer Review." *Science and Public Policy* 34 (8): 575–83. doi:10.3152/030234207X255179.

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

Moed, H. F., W. Glänzel, and Ulrich Schmoch. 2005. *Handbook of Quantitative Science and Technology Research : The Use of Publication and Patent Statistics in Studies of S&T Systems*. Dordrecht etc.: Kluwer Academic Publishers.

Mol, A. 2002. *The Body Multiple : Ontology in Medical Practice*. Durham: 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). Springer Netherlands: 117–39. doi:10.1007/s11024-015-9274-5.

Sauder, M. and W. N. Espeland. 2009. "The Discipline of Rankings: Tight Coupling and Organizational Change." *American Sociological Review* 74 (1): 63–82. doi: 10.1177/000312240907400104.

Sivertsen, G. 2016. "Publication-Based Funding: The Norwegian Model." In *Research Assessment in the Humanities: Towards Criteria and Procedures*, edited by Ochsner M, Hug SE, and Daniel HD, 79–90. Zurich: Springer International Publishing. doi:10.1007/978-3-319-29016-4_7.

Strathern, M. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305. doi:10.1017/S1062798700002660.

van Eck, N. Jan, L. Waltman, A. F. J. van Raan, R. J. M. Klautz, and W. C. Peul. 2013. "Citation Analysis May Severely Underestimate the Impact of Clinical Research as Compared to Basic Research." *PLoS ONE* 8 (4). doi:10.1371/journal.pone.0062395.

Vikkelsø, S. 2007. "Description as Intervention: Engagement and Resistance in Actor-Network Analyses." *Science as Culture* 16 (3): 297. http://www.informaworld.com/10.1080/09505430701568701.

VSNU, NWO, and KNAW. 2015. "Standard Evaluation Protocol 2015 - 2021. Protocol for Research Assessments in the Netherlands."

Waltman, L. and N. Jan Van Eck. 2016. "The Need for Contextualized Scientometric Analysis: An Opinion Paper." In *STI Conference 2016. Peripheries, Frontiers and beyond*, 1–9. Valencia: STI.

Whitley, R. and Jochen Gläser. 2007. *The Changing Governance of the Sciences: The Advent of Research*

Evaluation Systems. Springer.

Wilsdon, J., L. Allen, E. Belfiore, P. Campbell, S. Curry, S. Hill, R. Jones, et al. 2015. *The Metric Tide : Report of the Independent Review of the Role of Metrics in Research Assessment and Management*. doi:10.13140/RG.2.1.4929.1363.

Wouters, P. 2016a. "Semiotics and Citations." In *Theories of Informetrics and Scholarly Communication*, 72–92. Berlin/Boston: de Gruyter.

Wouters, P. 2016b. "The Mismeasurement of Quality and Impact." In *Gaming Metrics*, edited by Alexandra Lipmann and Mario Biagioli, forthcoming. MIT Press.

Wouters, P., J. Bar-Ilan, M. Thelwall, I. F. Aguillo, Ü. Must, F. Havemann, H. Kretschmer, et al. 2014. "Acumen Final Report." Brussels. http://cordis.europa.eu/result/rcn/157423_en.html.

Wouters, P., W. Glänzel, J. Gläser, and I. Rafols. 2013. "The Dilemmas of Performance Indicators of Individual Researchers: An Urgent Debate in Bibliometrics." *ISSI Newsletter* 9 (3): 48–53.

Wouters, P., M. Thelwall, K. Kousha, L. Waltman, S. De Rijcke, A. Rushforth, and T. Franssen. 2015. "The Metric Tide: Literature Review (Supplementary Report I to the Independent Review of the Role of Metrics in Research Assessment and Management)," 188. doi:10.13140/RG.2.1.5066.3520.

Zuiderent, T. 2015. *Situated Intervention: Sociological Experiments in Health Care*. Cambridge Mass.: The MIT Press.