THE PROBLEM OF EVIDENCE IN SSH RESEARCH EVALUATION

LAI MA

School of Information and Communication Studies

University College Dublin

SUMMARY. This paper poses questions about the epistemology of research evaluation by examining the different kinds of evidence and by explicating the epistemological assumptions of using non-intersubjective evaluation criteria in research evaluation. It is argued that the over-emphasis on the wrong kinds of evidence can lead to erroneous assumptions about people and disciplines which in turn lead to undesirable changes in research practices and knowledge production. It is also argued that some issues in SSH research evaluation are rooted in epistemological assumptions that have led to the use of methodologies deprived of intersubjective understanding.

KEYWORDS: evidence, bibliometrics, societal impacts, SSH research evaluation, epistemology of research evaluation.

INTRODUCTION

Research evaluation is a process for assessing the quality of research and scholarship. The functions, objectives, and methodologies vary at different levels, namely, government, funding agency, university, research field, and individual. Currently, 'impact' is an important criterion in research evaluation at all levels. The two kinds of impacts—research and societal—are mostly evidenced by citation-based metrics and impact case studies. However, the validity and legitimacy of citation-based metrics and impact case studies as *evidence* in research evaluation have been contested and debated, while some have shown that they can lead to academic misconduct and malpractices (see Biagioli & Lippman 2020).

For one, as many have discussed, citation counts are not necessarily a good indicator of quality, while societal impacts are not necessarily traceable, measurable, and observable. Broadly speaking, these problems are more significant in the social sciences and humanities (hereafter 'SSH') since their publication and citing practices do not lend themselves to produce high number of citations (Ma & Ladisch 2019) and their public engagement tend to generate transformative experiences rather than quantifiable impacts. As Sivertsen and Meijer (2020) state, "The quest

for *evidence* of impact seems to assume that the science-society interaction is not normal but might sometimes take place in any unexpected place and only in particular and extraordinary cases" (p. 68, emphasis added). The problem of evidence in research evaluation seems to be apparent however unresolved.

Hence, if we reconsider the validity and legitimacy—and the process and criteria—of research evaluation, particularly in SSH, we must ask some fundamental questions about the *epistemology* of research evaluation, including: What kind(s) of evidence are best for validating the quality of research and demonstrating the societal impacts of research? What kind(s) of evidence should be used and in what context? Without clear answers to these questions, there is little scope to reconfigure the process and criteria of SSH research evaluation.

In this paper, I aim to answer these questions by examining the different kinds of evidence and by explicating the epistemological assumptions of using non-intersubjective evaluation criteria in research evaluation. I will argue that the over-emphasis on the wrong kinds of evidence may lead to erroneous assumptions about people and disciplines which in turn lead to undesirable changes in research practices and knowledge production. I will also argue that some issues in SSH research evaluation are rooted in epistemological assumptions that have led to the use of methodologies deprived of intersubjective understanding.

EVIDENCE OF RESEARCH IMPACT

RESEARCH IMPACT—CONTENTS AS EVIDENCE

To evaluate the quality of research and its potential and actual impacts, the kinds of evidence can be broadly categorized as the *contents* of scholarly work and research outputs and the *data* about scholarly work and research outputs such as citation-based metrics.

Using *contents as evidence* in research evaluation, the quality is largely assessed through the peer review process. The evaluators—peer reviewers—are those who have knowledge and experience in the subject matters, methodologies, and methods. The evaluation requires the reading and understanding of the *contents* of the scholarly work and research outputs. If the evaluation process involves more than one person, agreements or compromise have to be made between peer reviewers. Non-experts are not eligible or able to judge or make decisions with *contents as evidence* because they do not have the education, expertise, or experience to support their claims as to whether the scholarly work and research outputs are good or bad, right or wrong, or worthwhile to be published or funded.

In the context of journal submission, the editor is to make the decision as to whether an article *should* be published based on reviewers' recommendations. In other words, the decision is based on an intersubjective understanding among the editor(s) and the reviewers whether the article is of sufficient quality to be published. In the context of grant reviews, the reviewers are asked to evaluate whether a research proposal *should* be funded based on criteria including, but not limited to, the quality of the research proposal and the credentials of the applicant(s). The funders then make decisions based on the recommendations and also the objectives of the funding programme which sometimes involve economic and political considerations. Nevertheless, the evaluation process involves understanding the *contents* of the research proposals.

When *contents as evidence* are used in research evaluation, the best way to achieve good results is to produce good contents. Although there are known problems of peer review as an evaluation system/mechanism (Lamont 2010; Langfeldt 2001; Luukonen 2012), it is generally trusted and regarded as the most important mechanism to uphold the quality and trustworthiness of scholarly work and research outputs. The mechanism of peer review has not been reported as a factor contributing to gaming strategies or misconduct and malpractices in research and publication.

RESEARCH IMPACT—METRICS AS EVIDENCE

Metrics are the *data about* or the *metadata* of scholarly work and research outputs (Biagioli 2018). The data about scholarly work and research outputs are usually described as *evaluative bibliometrics*, or simply, *metrics*, that are largely based on the number of citations. The most commonly known and used evaluative bibliometrics are Journal Impact Factor (JIF) and h-index, while there are many others created and provided by commercial and non-commercial entities. These data are used by both experts in a subject area and non-experts, with the assumption that metrics represent the quality and impact of scholarly work and research outputs and sometimes also the productivity of scholars and researchers. In other words, when *metrics* are used as evidence in research evaluation, the evaluators do not have to understand the *contents* of the work. Metrics are used as evidence because they are seemingly 'objective' with no need of interpretation.[1] The validity and limitations of evaluative bibliometrics are seemingly not questioned in the design and process of research evaluation.

Metrics are often used by scholars and researchers to maintain a track record and to benchmark with others, as well as to make decisions about hiring, promotion,

---

[1]   See Espeland & Stevens (2008) for an excellent discussion of the process of commensuration with regard to quantitative indicators/measures.

and tenure. At the same time, metrics are also used by university administration and management, funding agencies as well as governmental bodies as evidence of excellence in research, considered in strategic planning, priorities setting, resource allocation, and so on. The use of evaluative bibliometrics is also propagated by university rankings, many of which use data provided by commercial indexing companies such as Scopus.

When *evaluative metrics as evidence* are used in research evaluation, the best way to achieve good results should be the same as using contents as evidence with the assumption that the better the quality of one's work, the higher the metrics will be. Unfortunately, evaluative bibliometrics can be poor indicators of quality because there are underlying problems with using citations as evidence of research impact; not to mention that these indicators can be manipulated by altering research and publication practices. Gaming strategies are well documented as many attempt to score higher metrics as a way to collect evidence for promotion, tenure, awards, and so on (Biagioli, Lippman 2020).

EVIDENCE OF SOCIETAL IMPACTS

Most agree that REF 2014 was the first to introduce the impact case studies method, which has since been adapted and adopted in many European countries (Sivertsen, Meijer 2020). As an *ex post* impact assessment, impact case studies can include, but are not limited to, cultural, economic, environmental, health, political, and societal impacts. In this article, we use 'societal impacts' as an umbrella term to include all impacts beyond academia.

Impact case studies are read, interpreted, and assessed using peer review process, although the expertise of reviewers is often unknown and their roles in society can be diverse. Impact case studies are populated with narratives as well as quantitative evidence such as the number of downloads, views, visitors, and so on. Some argue that the evaluation can be subjective because societal impact is a fluid concept and is not necessarily easy to compare and score. For example, consider the impact of these three groups: (a) a spin-off company developing a vaccine for Covid-19, (b) a group of social scientists who publish reports about how Covid-19 has affected lives of people in vulnerable conditions, and (c) a group of artists, musicians, and poets hosting public events to bring solidarity to people and to console the bereaved and the suffering. Which group has the highest impact? Is one kind of impact more important than the other?

Impact case studies are evaluated based on what are deemed important to a society—be they cultural, economic, or political. What are *deemed* important is

based on a certain level of intersubjective—normative—agreements among members of the society. We can postulate that the reviewers make their judgements about societal impact based on their own perceptions of social good. Bozeman and Boardman (2009), for instance, maintain that a notion of 'social good' is required to evaluate potential societal impacts.

The impact case studies method, however, has its shortcomings and limitations: attribution, the counterfactual argument, the time lag between research outputs and impact, and so on (European Science Foundation 2012; Luukkonen 1998; Penfield, et al. 2014; Spaapen, van Drooge 2014). Recently, Sivertsen and Meijer (2020) critique that the impact case studies method celebrates extraordinary impact and undermines normal impact. All of the above issues can be attributed to the need for traceable, measurable, observable, and tangible impacts as evidence.

Further, quantitative indicators such as the number of attendees, downloads, views, and sometimes even the number of tourists and hotel nights are used to demonstrate economic impact of a cultural event. Similar to evaluative bibliometrics, these quantitative indicators are seen as 'short-cuts' to estimate societal impact, that is, without the need to read and understand the complexity of dissemination and influences of scholarly work and research in the society. They are presumably easy to understand, interpret, and manipulate in the decision-making process.

These numbers, of course, are not bad in themselves and they do indicate societal impact to a certain extent. However, the over-reliance on these numbers can be problematic. Consider our previous example about Covid-19: Do more downloads or likes on a social media platform about one group indicate higher impact than another? Can these numbers replace the intersubjective agreements as to what are important and influential in our community and the society at large? Can these numbers be boosted using marketing strategies to win the prize of the best impact case studies?

THE USE OF EVIDENCE IN RESEARCH EVALUATION

Do scholars and researchers use *metrics as evidence* of quality work? Possibly both yes and no. We often hear that scholars and researchers browse public profiles such as ResearchGate and Google Scholar, including the metrics, when they do not know a person—whether they are potential collaborators or candidates or competitors for an academic/research position. But most would agree that that the most important indicator of quality is the work itself which they get to know from attending presentations and discussions at conferences and not least reading the publications and other research outputs. In other words, they depend on *contents*

as evidence to make judgments about quality. The wide use of metrics by funding agencies, universities, publishers, and indexing services, however, has heightened the importance of metrics over contents as evidence.

The regime of new public management has been discussed as one of the culprits of the audit culture in academia where scholars and researchers must prove their worth in terms of research impact as well as societal impact (see, for example, Burrows 2012; Power 1997). Since most stakeholders cannot evaluate research in-and-of-itself, they depend on metrics as *de facto* indicators of research quality. Biagioli (2018) has suggested that evaluative bibliometrics are a 'double alien' form of knowledge: both produced and used by people who are not practitioners of the field to which the publications belong" (p. 252). Similarly, Gingras (2014) observes that the uses of evaluative bibliometrics shift qualitative peer review to quantitative expert review, "which subtly shifts the ground from 'peers' (that is, real experts in the field) to a more ill-defined 'expert', who may in fact be no 'peer' and no 'expert' either, and thus needs crutches like journal rankings to evaluate researchers since they cannot themselves evaluate the content of the research being evaluated" (p. 56).

Impact case studies have also been discussed in light of the audit culture. It is perceived that they are used as evidence of value-for-money. The critiques of the impact case studies method (for example, Sivertsen and Meijer 2020; Smith, Ward and House 2011; Watermeyer2014, 2016) explicate the problems of the method, the evidence it requires and, all in all, the purposes of the evaluation exercises. These critiques point to the implications of the impact case studies method in knowledge production and research integrity. We must make clear that, however, they are not questioning the responsibility of scholars and researchers to produce 'impact'. Rather, they critique the emphasis on *tangible evidence* required, that is, evidence that is traceable, measurable, and observable, as well as the neglect of the fact that 'impacts' are collectively and collaboratively produced and are often not attributed to a single individual or project.

The different kinds of evidence are required in research evaluation as a matter of objectivity because they are often used by non-experts in a research field, including administrators, managers, and marketers. Nevertheless, the kinds of evidence used can affect our perspectives and thinking about research, scholarship, and research evaluation. So we must ask: Why do we need evidence? What kinds of evidence should we use? In what follows, I take a small step in thinking about these questions by considering the epistemological assumptions of using quantitative measures in *decision-making processes* such as tenure and promotion, grant funding, and university rankings.

THE PROBLEM OF EVIDENCE IN RESEARCH EVALUATION—AN EPISTEMOLOGICAL EXPLORATION

Using *contents as evidence* requires intersubjective understanding of texts, contents, and the contexts in which the scholarly work and research outputs are created. This kind of evidence does not provide a 'short-cut' to know about a scholar or a researcher, a piece or a corpus of work. It is because reading and understanding take time, not to mention the required education, training, and experience in a research field and the knowledge domain. Therefore, peer review is irreplaceable as a means of quality assurance despite some of the known problems and issues, and despite that it is time-consuming.

The use of other kinds of evidence, which we can describe as 'data', are deconstextualized, for the very nature of data is to reduce complexities. We use data (rather than *contents*) in research evaluation partly because, by not having to read and understand the contents, the use of data saves time (Ma, *forthcoming*), and partly because the data can presumably be used and understood by non-experts. Therefore, by using data as evidence, research evaluation adheres to the epistemological assumptions of empiricism that neglect the understanding of cultural and social affordances and intersubjective understanding of intentions and actions. In the following I will use the DIKW (Data-Information-Knowledge-Wisdom) model to illustrate the way of knowing based on empiricist epistemologies.

The DIKW model is most commonly known in the field of knowledge management. Many regard R. L. Ackoff's presidential address to International Society for General Systems Research (ISGSR) in 1998 as the origin of the model (see, for example, Frické 2009; Rowley 2006). The DIKW model is usually illustrated in a triangular shape with data at the bottom, then progress to information, knowledge, with wisdom at the top. The model is regarded as a representation of cognition by presenting human thought and learning as data-processing as in a computational model. It has been argued that, however, "for humans, information and knowledge are not processed data; rather, we learn by being situated within and understanding complex webs of relations of persons, events, social and political structures, and many other things" (Ma 2012: 720).

Methodologically, the DIKW model adheres to data processing in computational thinking, in contrast to the phenomenological and hermeneutic approaches most commonly used in the SSH. While it is undeniable that computational methods can assist the understanding of human actions and societal interactions on a large scale, they tend to give answers about the *what* rather than the *how* and the *why*. Data alone are not sufficient to understand human and societal phenomena—their intentions and actions—because they require the understanding of

contexts: cultural, educational, economic, socio-political, and so on that constitute the complex web of human and social relationships. As Albert Einstein once put it: "Understanding physics is child's play compared to understanding child's play."

We can consider the problems of evidence, specifically, the preference for empirical evidence in research evaluation using the DIKW model. Since most administrators and managers do not have the expertise to evaluate research in-and-of-itself, that is, *contents as evidence*, they rely on 'data'—empirical evidence to be informed, notwithstanding that their understanding is out of context because these data, either in the form of evaluative bibliometrics or impact case studies, are decontextualized. To make a decision based on data as if they *represented* the quality and impact of scholarly work and research outputs implies an empiricist understanding of data, information, and knowledge that undermines interpretation and understanding.

Research evaluation requires some forms of evidence—without which one cannot assess a piece of work, or the performance of an individual. Problems arise, however, when the kinds of data as evidence do not fully represent the quality, productivity, and/or performance for which the evaluators are *informed* and make important decisions about persons and their work, particularly when many evaluators are not subject experts; in other words, when decisions are made without the acts of knowing. It is dangerous to believe in the simple DIKW model that data automatically turn into information, knowledge, and wisdom.

Therefore, we need to further consider the epistemology of research evaluation. In other words, how we *know* and *understand* evidence in the acts of research evaluation? How do we *know* if a piece of work is of good quality? How do we *know* if a scholar or researcher is producing and/or has the potential to produce works of excellent quality? Can we rely on (meta)data to tell us about persons and their work? What kinds of evidence should we use in research evaluation? All in all, we need to consider prioritizing understanding and interpretation by subject experts and perhaps limiting the use of weak indicators of quality and performance by non-experts to avoid mechanical thinking such as in the DIKW model.

CONCLUSION: VALUES AND (E)VALUATION OF SSH

In the midst of the Covid-19 crisis, our newspapers and social media are awash with new findings of the coronavirus. We are reminded of the significance of scientific research, as well as the danger of defunding scientific research and institutions such as the US Centres of Disease Control and Prevention and World Health Organisation. There is no question that it will be with scientific breakthrough(s) that we overcome this crisis in the long term. The value of the sciences

is apparent—scientific research devises new technologies, including vaccines, to improve the living conditions and health of humanity. The evidence is clear.

In the midst of the Covid-19 crisis, our movements are restricted and some of us are experiencing anxiety and stress, and despair and sadness. We turn to music and poems, novels and films, literary works and history books to find solace and meaning. We turn to digital library and museum collections although their physical doors are closed. Artists and writers offer free classes to keep minds young and old creative and active. The value of the arts and humanities is apparent—they are our channels for feelings and thoughts and hope and fears in our time and throughout history. The evidence is clear.

In the midst of the Covid-19 crisis, our society is undergoing unprecedented changes. We are preparing for the most severe economic downfall since the Great Depression of the 1930s at a time when income inequality is a grave concern in many countries. Educators, economists, and political scientists are weighing in their expertise to understand the current societal challenges and to plan for the post-Covid-19 new normal. The value of the social sciences is apparent—the knowledge of societal complexities is vital to build and re-build a fair and prosperous future. The evidence is clear.

The value of knowledge is acknowledged and recognized in our society, particularly in times of crisis. The evidence seems to be clear—we are making use of our knowledge to solve scientific problems and to resolve social issues and to find meanings. We accept that the value of knowledge can be justified by measurable or immeasurable, tangible or intangible evidence in public discourse. It is apparent that there is an intersubjective understanding of the value of knowledge, research, and scholarship. It is also apparent that data alone cannot tell a complete story.

In this paper, I argue that the use of certain kinds of evidence in research evaluation, particularly when used by non-subject experts, adheres to the cognition model represented by the DIKW model, which can lead to uninformed decision-making due to a lack of contextual understanding. I also argue that the epistemological assumptions can undermine the value of the SSH in the academe and the society at large.

Furthermore, I argue that the over-reliance on certain kinds of data as evidence can also alter research and publication practices, resulting in the change of perception of education, knowledge, and universities by the public and other stakeholders. If the kinds of evidence in research evaluation is to hold research institutions and researchers accountable, it might well work backwards when the various measures lead to goal displacements, not to mention misconduct by some. Hence, it is pertinent to consider the epistemologies of research evaluation.

And hence, before asking for evidence, we must ask what are the values that we must show. The values should be recognised by all stakeholders—the public, the government, the funding agencies, the universities, the scholars, and the researchers themselves—such that we can find the right kinds of evidence in research evaluation that is not designed to censor or to audit, but to motivate and to nurture the production of knowledge as a social good. Consequently, then, not only can we improve the validity and the legitimacy of research evaluation, but we can also (re) gain public trust in experts and in expertise.

REFERENCES

Ackoff, Russell. From data to wisdom: Presidential address to ISGSR, June 1988. *Journal of Applied Systems Analysis*, 1989, no. 16, 3–9.

Adams, Jonathan, Karen Gurney, Tamar Loach and Martin Szomszor. Evolving Document Patterns in UK Research Assessment Cycles. *Frontiers in Research Metrics and Analytics*, 2020, no. 5, 2; https://doi.org/10.3389/frma.2020.00002.

Bod, Rens. *A New History of the Humanities: The Search for Principles and Patterns from Antiquity to the Present*. Trans. L. Richards. Oxford University Press, 2013.

Biagioli, Mario. Quality to impact, text to metadata, publication and evaluation in the age of metrics. *KNOW*, 2018, no. 2 (2), 249–275.

Biagioli, Mario, and Alexandra Lippman. *Gaming the Metrics: Misconduct and Manipulation in Academic Research* (pp. 1–23). Cambridge, Massachusetts: The MIT Press, 2020.

Burrows, Roger. Living with the h-index? Metric assemblages in the contemporary academy. *The Sociological Review*, 2012, no. 60 (2), 355–372.

Collini, Stefan. *What Are Universities For?* London: Penguin, 2012.

Espeland, Wendy. N., and Mitchell L. Stevens. A sociology of quantification. *European Journal of Sociology*, 2008, no. 49 (3), 401–436.

European Science Foundation. *The Challenges of Impact Assessment.* Internet access: http://archives.esf.org/coordinating-research/mo-fora/evaluation-of-publicly-funded-research.html [retrieved 2019 10 09].

Frické, Martin. The knowledge pyramid: A critique of the DIKW hierarchy. *Journal of Information Science*, 2009, no. 35 (2), 131–142.

Gingras, Yves. *Bibliometrics and Research Evaluation: Uses and Abuses.* Cambridge, Massachusetts: The MIT Press, 2014.

Lamont, Michèle. *How Professors Think: Inside the Curious World of Academic Judgment.* Harvard University Press, 2010.

Langfeldt, Liv. The decision-making constraints and processes of grant peer review, and their effects on the review outcome. *Social Studies of Science*, 2001, no. 31 (6), 820–841.

Luukkonen, Terttu. The difficulties in assessing the impact of EU framework programmes. *Research Policy*, 1998, no. 27 (6), 599–610; https://doi.org/10.1016/S0048-7333(98)00058-4.

Ma, Lai. Meanings of information: The assumptions and research consequences of three foundational LIS theories. *Journal of the American Society for Information Science and Technology*, 2012, no. 63 (4), 716–723.

Ma, Lai, and Michael Ladisch. Evaluation complacency or evaluation inertia? A study of evaluative metrics and research practices in Irish universities. *Research Evaluation*, 2019, 28 (3), 209–217; https://doi.org/10.1093/reseval/rvz008.

Ma, Lai. Metrics as time-saving devices. In *Inquiring into Academic Timescape*. Ed. Filip Vostal. Emerald. *Forthcoming*.

Müller, Ruth, and Sarah de Rijcke. Exploring the epistemic impacts of academic performance indicators in the life sciences'. *Research Evaluation*, 2018, no. 26 (3), 157–168.

Penfield, Teresa, et al. Assessment, evaluations, and definitions of research impact: A review. *Research Evaluation*, 2014, no. 23 (1), 21–32.

Power, Michael. *The Audit Society: Rituals of Verification*. Oxford: Oxford University Press. 1997.

Readings, Bill. *The University in Ruins*. Cambridge, Massachusetts: Harvard University Press. 1996.

Rowley, Jennifer. The wisdom hierarchy: representations of the DIKW hierarchy. *Journal of Information Science*, 2006, no. 33 (2), 163–180.

Sivertsen, Gunnar, and Ingeborg Meijer. Normal versus extraordinary societal impact: how to understand, evaluate, and improve research activities in their relations to society? *Research Evaluation*, 2020, no. 29 (1), 66–70; https://doi.org/10.1093/reseval/rvz032.

Small, Helen. *The Value of the Humanities.* Oxford University Press, 2013.

Smith, Simon, Vicky Ward and Allan House. 'Impact' in the proposals for the UK's Research Excellence Framework: Shifting the boundaries of academic autonomy. *Research Policy*, 2011, no. 40 (10), 1369–1379; https://doi.org/10.1016/J.RESPOL.2011.05.026.

Spaapen, Jack, and Leonie van Drooge. Introducing 'productive interactions' in social impact assessment. *Research Evaluation*, 2011, no. 20 (3), 211–218; doi:10.3152/095820211X12941371876742.

Watermeyer, Richard. Issues in the articulation of 'impact': the responses of UK academics to 'impact' as a new measure of research assessment. *Studies of Higher Education*, 2014, no. 39 (2), 259–377.

Watermeyer, Richard. Impact in the REF: issues and obstacles. *Studies in Higher Education*, 2016, 41 (2), 199–214; https://doi.org/10.1080/03075079.2014.915303.

Lai Ma

ĮRODYMŲ PROBLEMA, KYLANTI VERTINANT HSM TYRIMUS

SANTRAUKA. Šiame straipsnyje keliami klausimai apie mokslinių tyrimų vertinimo epistemologiją analizuojant įvairius įrodymus ir įvardijant epistemologines prielaidas, kurios daromos taikant neintersubjektyvius vertinimo kriterijus. Čia teigiama, kad netinkamų įrodymų pernelyg didelis sureikšminimas gali lemti klaidingą mokslininkų ar disciplinų vertinimą, kuris, savo ruožtu, veda link nepageidaujamų mokslinės veiklos pokyčių. Taip pat tvirtinama, kad kai kurios HSM vertinimo problemos kyla iš epistemologinių nuostatų ir taikomų metodologijų, kurioms nebūdingas intersubjektyvusis supratimas.

RAKTAŽODŽIAI: įrodymai, bibliometrija, socialinis poveikis, HSM vertinimas, mokslinio vertinimo epistemologija.