If you do not deign to quantify, someone else will do it for you: In support of a balanced approach to the evaluation of science

Mathieu Lizotte
University of Ottawa, Canada

Abstract
This is a commentary in support of Olof Hallonsten’s historical-sociological argument for countering the growing distrust and governance of science. From this starting point, the problem of quantification in the evaluation of science is addressed and several examples of the unintended consequences of the currently available metrics are discussed. In particular, the issue of quantification is discussed in regard to the modality of scientific research, power and research and the peer relationship. Although in approval with Hallonsten’s argument for reversing the burden of proof, reasonable skepticism is expressed regarding the persuasiveness that this counter-rhetoric will have on members of parliament, public servants and university administrators. If this long-term goal is to be accomplished, it is argued that concrete actions must be pursued in the short and medium term. In this spirit, several suggestions are formulated to further this agenda, most notably greater support for intellectual diversity, greater participation and readership in science studies by science practitioners and the promotion of the comparative approach for understanding the different ways that metrics are actually used in practice. Finally, I argue that the refusal of participating in the quantification of science is bound to hinder applied critical thinking and will most likely and regrettably exacerbate its current perverse effects.

Keywords
evaluation of science, governance, intellectual diversity, quantification, unintended consequences

Corresponding author:
Mathieu Lizotte, University of Ottawa – Sociology, 120 University Private Ottawa, ON K1N 6N5, Canada.
Email: mlizott2@uottawa.ca
Résumé
Cet article vient appuyer la proposition historico-sociologique d’Olof Hallonsten destinée à contrer la méfiance grandissante envers la science, ainsi que la gouvernance que cette dernière subit. En partant de ce constat, j’interroge le problème de la quantification dans l’évaluation des sciences et reviens sur différents exemples où les instruments de mesure actuels ont eu des conséquences inattendues. Le problème de la quantification est plus particulièrement analysé à la lumière des modalités de la recherche scientifique, du lien entre pouvoir et recherche et des relations entre les pairs. Bien qu’étant d’accord avec Hallonsten concernant le renversement du fardeau de la preuve, je pense qu’il est néanmoins nécessaire de faire monter d’un scepticisme raisonnable envers le pouvoir de persuasion que cette contre-rhétorique pourrait avoir sur les parlementaires, les fonctionnaires et les personnes travaillant dans l’administration au sein des universités. Pour atteindre cet objectif à long terme, des actions concrètes devront être prises à court et à moyen terme. Dans cet esprit, j’énumère plusieurs suggestions pour faire advenir cet agenda, parmi lesquelles on peut notamment compter : un appui plus important en faveur de la diversité intellectuelle, une participation plus importante des chercheurs aux études sur les sciences et leur prise en compte par ces derniers, ou encore la promotion d’une approche comparative qui permettrait de comprendre les différentes façons dont ces instruments de mesure sont utilisés en pratique. Pour finir, j’avance l’idée que le refus de participer à la quantification de la science nous conduira forcément à entraver la pensée critique appliquée et conduira très probablement – et ce de façon regrettable – à exacerber ses effets pervers.

Mots-clés
conséquences imprévues, diversité intellectuelle, évaluation de la science, gouvernance, quantification

Olof Hallonsten (2021) raised some very thought-provoking issues regarding the current evaluation regime in science. Like many of us, I share his concerns regarding the immense pressure the university is currently facing to demonstrate its utility (Mirowski, 2011). Much of this pressure is under the guise of economic and budgetary concerns, but there are obviously deeper ideological reasons at play (Bourdieu, 2004; Moore et al., 2011). Utility is after all an abstract concept that can be given any substantive definition (Kymlicka, 2002). In light of constant economic pressure and now the ecological crisis, the dominant discourse defines utility in terms of innovation and sustainable growth. This urgency has led some to advocate for a more intimate relationship between the university and the market, much to the dismay of others who see this analogously as merchants desecrating the temple. The dominant discourse is contingent, however, as neoliberal ideology would recommend a closer relationship between the university and the market regardless of global warming or any other crisis (Slaughter et al., 2004; Mirowski, 2013). Setting those crises aside, independent of all of this, technological development in computer science has created new possibilities for
tracking and evaluating knowledge (Van Noorden, 2010). All of these factors converge towards essentially one issue: should science be further rationalized and if so, to what extent and how? Hallonsten (2021) criticizes what appears to be the growing consensus, or perhaps implicit consent, in society that science should be further rationalized through exogenous interference.

Despite the provocative tone, Hallonsten (2021) ultimately supports a middle of the road approach, as he repeats Linda Butler’s plea for sanity. As quantitative performance measures were being introduced to Australia and the United Kingdom, Butler (2007) promoted a balanced approach, arguing that although metrics had their place, for instance by making the process more efficient and cost effective, qualitative peer review should remain the keystone in research evaluation. This implies that both extremes should be avoided: i) the categorical refusal of any metric for any purpose; and ii) the dystopia where peer-review is abolished in favor of an algorithm. Whether we like them or not, metrics are here to stay, but it is up to us how we use them in our practice and how we tolerate others to use them without our blessings. I think the balanced approach is ultimately the one to follow, but this is easier said than done, as it necessitates a concerted effort to simultaneously use, develop and criticize them. Opinions are bound to vary in both the choice of metrics and how they are used for scientific and administrative purposes. Supporting Butler (2007), Hallonsten (2021) contributes to this discussion by making his own plea that historical evidence should be taken into consideration when evaluating the productivity of science. Basically, he is arguing that science has been immensely productive well before performance benchmarks were ever conceived – so productive in fact that he promotes shifting the burden of proof over to those who dare claim otherwise.

The perverse effects of the current benchmarks

Hallonsten (2021) argues that much of the current practice in quantitative performance evaluation is pointless and counterproductive. While the former predicate adjective is needlessly provocative, there is already a growing literature criticizing the unbalanced usage of metrics in the evaluation of science (Gendron, 2008; Adler and Harzing, 2009; Espeland and Sauder, 2009, 2016). Many of these negative effects may be labeled unintended consequences or perverse effects (see Merton, 1936; Boudon, 2016). For the sake of brevity, I shall limit myself to highlight only a few of these, categorized in three broad themes: i) the modality of scientific research; ii) power and research; iii) the peer relationship.

Firstly, the current evaluation regime can actually foster conformity with benchmarks that are counter-productive to originality, scientific rigour and ethical conduct. For instance, it can incentivize researchers to ignore worthwhile fields or issues that demonstrate less citing and reference potential. Science is somewhat of a market of ideas where practitioners are guided by a combination of curiosity, utility, passion and popularity. These can rarely be followed without making compromises. Do we want to institutionalize science in such a way where people curtail their passions and intuitions in favor of topics that demonstrate more publishing or hiring potential? This is most certainly counterproductive to true innovation. By the latter, I am reclaiming a definition of innovation,
which is not synonymous with applicability and marketability. Furthermore, it has been suggested that there is an overproduction of academic papers which is leading to a decline in standards and straining the peer-review system (Harley and Acord, 2011). Some argue that journal rankings can suppress interdisciplinary studies (Rafols et al., 2012). Harsher criticisms have been directed towards the social sciences, regarding which some argue outright that there has been a ‘proliferation of meaningless research’ (Alvesson et al., 2017: 4) which adds no value to society and modest value to its authors apart for career purposes. While the latter view is too cynical, I believe that most researchers would agree that the pendulum needs to shift back towards quality at the expense of quantity. Most distressing of all, evidence has been found, at least in the pure sciences, that journal rank is a significant predictor of the incidence of scientific fraud and retractions (see Brembs et al., 2013).

Secondly, the fetishization of the citation index (or other similar metrics) can reinforce practices that are more akin to power and prestige-seeking than actual scientific progress (Hazelkorn, 2015). While ideally scientists should adhere to reason in their quest for truth (as epistemologically difficult as it sounds), they are also human beings who are members and representatives of networks, institutions, schools of thought and even ideological orientations. Less euphemistically, an overemphasis on these benchmarks can promote cronyism where people make explicit or implicit pacts to cite some and not others. Christoph Bartneck and Servaas Kokkelmans (2011) have shown how an author can considerably inflate and bias their h-index through self-citation. More to the point, Crawford Spence (2019) argues that metrics ultimately undermines ‘nobler, socially minded visions of what a university should be’ (2019: 761). He further argues that universities should privilege a collegial ethos of judgment over a managerialist ethos of measurement. He argues that the metric-centered competition in the world rankings has the perverse effects of sacrificing deep scientific inquiry and pedagogy. Some critics have even used the term ‘Potemkin village’ to express concern about how focusing only on the evaluated criteria will ultimately overshadow the core missions of learning and research (Gioia and Corley, 2002; Lund Dean et al., 2020). Overall, it is not surprising that many critics have used Michel Foucault’s governmentality to highlight how metrics can be used as technologies of power (Sauder and Espeland, 2009). No serious scientists should be against accountability. But it would be naïve not to ask the deeper questions of accountable to whom, by which means and according to which values?

Finally, I believe that the current evaluation regime cultivates animosity, resentment and disdain between colleagues (i.e., beyond the unavoidable minimum). It fuels this between those who play the game, those who refuse, and perhaps to some extent those who cannot or choose not to. Like any other human endeavor, academia is a field; it has its games, its rules, its material and symbolic rewards (Bourdieu, 2004). The unhealthy cocktail of hypercompetition, perverse incentives and a lack of funding can cultivate unethical behavior (Edwards and Roy, 2017). Perhaps we should use this discussion on metrics to address a deeper issue (Trank and Rynes, 2003): does the current reward structure actually recognize the whole gamut of intellectual contributions to its just value? I am not convinced that it does. Quite to the contrary. There are different ways to be a scholar. Some professors do not publish as much but are esteemed by their peers and are sought after for their expertise. Some are recognized internationally, while others go
unnoticed within their own department. Some refrain from grant hunting, while others colloquially play ball. Some are worshiped by their students; others are scorned as if they had leprosy. Some scholars seek out the limelight and metaphorically die if they are ignored. Others are content to do quality work regardless of whether or not they will gain any lifetime or posthumous recognition, etc. What do all of these colleagues have in common? They are exercising their profession according to their competence and vision and they have arguably earned the right to do so. Academia would perhaps generate less depression and burnout (see Gill, 2016) if it recognized contributions more holistically.

Concrete actions towards a more diverse and just evaluation regime

Regarding accusations that the university suffers from a want of productivity, Hallonsten’s (2021) recommendation to reverse the burden of proof is bold. That being said, brilliant as it is as a counter-rhetoric, public officials and university administrators are not likely to be particularly impressed given the immense pressure they are under to rationalize public services for the sake of a narrow and ideological kind of ‘efficiency’ (e.g. new public management, see Chandler et al., 2002; Lorenz, 2012). Reversing the burden of proof is certainly a worthy long-time goal, but it cannot be accomplished overnight. More concrete actions are needed in the short-term. Let us discuss what this might entail.

Firstly, I believe that a plea for intellectual diversity is in order. There is much discussion at the moment towards inclusion and diversity in the social sense – discussions which are long overdue. However, there is also another kind of diversity that requires attention: the recognition that there exist different types of scholars and worthwhile contributions. The administration can evaluate us all they want, they can create hundreds of metrics, etc., but we need to stand firm in the recognition of different career paths. If you regularly apply for funding – good. If you only apply from time to time when a project is ready and well-developed – just as good. If you never apply and just theorize – fine. If you want to concentrate on supervision and teaching, why not? In some ways, many are forced to assume responsibilities that others neglect – whether or not they excel in research. Many departments would be greatly helped if a few colleagues would concentrate on teaching and administration. To some extent, this is what happens in practice anyway, but those that concentrate on these tasks get little to no recognition. If we stand firm on the legitimacy of the different ways of being a scholar, the evaluative threat is countered to a large degree. The first step is perhaps to convince ourselves. This struggle for recognition will be easier in universities where the faculty is unionized (Kezar et al., 2019). This also means to value those professors who have lost their taste for research and who contribute more to administration. I believe that much political progress has been sacrificed by disdaining those colleagues who have both the competence and the standing to contest unbalanced evaluative practices. Our collective disdain does probably not inspire them to improve our work conditions – nor should it. I am not saying that each type of contribution should get the same merit or esteem, but rather that citations and funding are not – nor should be – the only criteria by which researchers esteem one another.
Secondly, I believe that researchers should be more involved in the field of science studies (Latour, 1999; Latour and Woolgar, 2013) and other key disciplinary fields whose work is of general scientific interest, such as the sociology of valuation (Cetina-Knorr, 2009; Lamont, 2012; Helgesson and Muniesa, 2013). While most of us are not actively involved in these scientific fields, we are all concerned given our scientific practice. Such fields are more necessary than ever to address the profusion of performance metrics that are already here (Van Noorden, 2010) and those that are to come. We should keep in mind that new measures are regularly created, and they will inevitably be used in unintended ways. This was the case for the Journal Impact Factor which was originally developed to guide libraries in their purchasing and indexing decisions (see Garfield, 1963; McKiernan et al., 2019). They are now often used to some capacity in the review, promotion and tenure (RPT) process (Abbott et al., 2010). In many respects, most researchers have no idea that Pandora’s box has already been opened. Our only hope is to play a greater part in the conversation and to apply critical thinking in how we choose to use them in our individual and collective practice. The balanced approach means encouraging attempt to develop ‘responsible’ metrics and practices (see Wilsdon et al, 2017). While much work remains to be done, I would argue that most are also currently unaware of the considerable effort to criticize and modify the current metrics (Sauder and Espeland, 2006). For instance, Emilio Ferrara and Alfonso Romero (2013) have suggested ways to mitigate the self-citation bias in the $h$-index.

Finally, as a corollary to the previous point, much is to be gained by applying the comparative approach to the meta-evaluation inherent in science studies. This approach would provide the reflexive insights needed to develop a truly informed and balanced approach. It would highlight how evaluation practices differ between disciplines, universities and countries, and it would help identify the implicit values behind all evaluative methods and their consequences. In some ways, I am less concerned by the practices themselves as much as the intentions and values behind them. As Baruch Spinoza would say, nothing is inherently bad in and of itself. It rather depends on how it is used. In short, while we do not have the power to stop the ‘evaluation frenzy’ (Hallonsten, 2021: 14), we do possess the capacity of evaluating their evaluative practices in return. Accountability goes both ways: evaluate not, lest ye be evaluated. This comparative meta-evaluation might even inspire new measures. Why not create a whole series of metrics for important issues in science which are being swept under the rug? Measures could be created to track and compare the level of government support for science. Other potential benchmarks: i) the percentage of GDP spent on higher education; ii) the ratio of tenure track positions to enrolled students; iii) an index of academic freedom. There will be purists who refuse to play this game, arguing that creating alternative measures serves to legitimize the quantification of performance. But it seems to me that a pragmatic approach should be adopted. One is almost tempted to formulate a law for this modern age: if you do not quantify what is important to you, rest assured that others will quantify what is important to them. The consequences of inaction can be devastating in the long run. As professors, perhaps we should do to governments one of the things we do best: grade them. We taught those in power once, so why should we refrain from evaluating them now?
Funding
The author received no financial support for the research, authorship, and/or publication of this article.

ORCID iD
Mathieu Lizotte https://orcid.org/0000-0002-6381-8112

References
Abbott A, Cyranoski D, Jones N, et al. (2010) Metrics: Do metrics matter? Nature News 465: 860–862.
Adler NJ, Harzing AW (2009) When knowledge wins: Transcending the sense and nonsense of academic rankings. Academy of Management Learning & Education 8(1): 72–95.
Alvesson M, Gabriel Y, Paulsen R (2017) Return to Meaning: A Social Science with Something to Say. Oxford: Oxford University Press.
Bartneck C, Kokkelmans S (2011) Detecting $h$-index manipulation through self-citation analysis. Scientometrics 87(1): 85–98.
Boudon R (2016) The Unintended Consequences of Social Action. Berlin: Springer.
Bourdieu P (2004) Science of Science and Reflexivity. Cambridge: Polity.
Brembs B, Button K, Munafo M (2013) Deep impact: Unintended consequences of journal rank. Frontiers in Human Neuroscience 7: 1–12.
Butler L (2007) Assessing university research: A plea for a balanced approach. Science and Public Policy 34(8): 565–574.
Cetina-Knorr K (2009) Epistemic Cultures: How the Sciences Make Knowledge. Cambridge: Harvard University Press.
Chandler J, Barry J, Clark H (2002) Stressing academe: The wear and tear of the new public management. Human Relations 55(9): 1051–1069.
Edwards MA, Roy S (2016) Academic research in the 21st century: Maintaining scientific integrity in a climate of perverse incentives and hypercompetition. Environmental Engineering Science 34(1): 51–61.
Espeland WN, Sauder M (2009) Rating the rankings. Contexts 8(2): 16–21.
Espeland WN, Sauder M (2016) Engines of Anxiety: Academic Rankings, Reputation, and Accountability. New York: Russell Sage Foundation.
Ferrara E, Romero AE (2013) Scientific impact evaluation and the effect of self-citations: Mitigating the bias by discounting the $h$-index. Journal of the American Society for Information Science and Technology 64(11): 2332–2339.
Garfield E (1963) Citation indexes in sociological and historical research. American Documentation 14(4): 289–291.
Gendron Y (2008) Constituting the academic performer: The spectre of superficiality and stagnation in academia. European Accounting Review 17(1): 97–127.
Gill R (2016) Breaking the silence: The hidden injuries of neo-liberal academia. Feministische Studien 34(1): 39–55.
Gioia DA, Corley KG (2002) Being good versus looking good: Business school rankings and the Circean transformation from substance to image. Academy of Management Learning & Education 1(1): 107–120.
Hallonsten O (2021) Stop evaluating science: A historical-sociological argument. Social Science Information 60(1): 7–26.
Harley D, Acord SK (2011) Peer Review in Academic Promotion and Publishing: Its Meaning, Locus, and Future. Berkeley: Center for Studies in Higher Education.

Hazelkorn E (2015) Rankings and the Reshaping of Higher Education: The Battle for World-Class Excellence. Berlin: Springer.

Helgesson CF, Muniesa F (2013) For what it’s worth: An introduction to valuation studies. Valuation Studies 1(1): 1–10.

Kezar A, DePaola T, Scott DT (2019) The Gig Academy: Mapping Labor in the Neoliberal University. Baltimore: Johns Hopkins University Press.

Kymlicka W (2002) Contemporary Political Philosophy: An Introduction. Oxford: Oxford University Press.

Lamont M (2012) Toward a comparative sociology of valuation and evaluation. Annual Review of Sociology 38: 201–221.

Latour B (1999) Pandora’s Hope: Essays on the Reality of Science Studies. Cambridge: Harvard University Press.

Latour B, Woolgar S (2013) Laboratory Life: The Construction of Scientific Facts. Princeton: Princeton University Press.

Lorenz C (2012) If you’re so smart, why are you under surveillance? Universities, neoliberalism, and new public management. Critical Inquiry 38(3): 599–629.

Lund Dean K, Fornaciari C, Bento R, et al. (2020) Premises, promises, and perils of the Academic Potemkin Village. Management Learning 51(4): 491–510.

McKiernan EC, Schimanski LA, Nieves CM, et al. (2019) Meta-research: Use of the journal impact factor in academic review, promotion, and tenure evaluations. eLife 8: e47338.

Merton RK (1936) The unanticipated consequences of purposive social action. American Sociological Review 1(6): 894–904.

Mirowski P (2011) Science-Mart. Cambridge: Harvard University Press.

Mirowski P (2013) Never Let a Serious Crisis Go to Waste: How Neoliberalism Survived the Financial Meltdown. New York: Verso.

Moore K, Kleinman DL, Hess D, et al. (2011) Science and neoliberal globalization: A political sociological approach. Theory and Society 40(5): 505–532.

Rafols I, Leydesdorff L, O’Hare A, et al. (2012) How journal rankings can suppress interdisciplinary research: A comparison between innovation studies and business & management. Research Policy 41(7): 1262–1282.

Sauder M, Espeland W (2006) Strength in numbers? The advantages of multiple rankings. Indiana Law Journal 81(1): 205–227.

Sauder M, Espeland WN (2009) The discipline of rankings: Tight coupling and organizational change. American Sociological Review 74(1): 63–82.

Slaughter S, Slaughter SA, Rhoades G (2004) Academic Capitalism and the New Economy: Markets, State, and Higher Education. Baltimore: Johns Hopkins University Press.

Spence C (2019) ‘Judgement’ versus ‘metrics’ in higher education management. Higher Education 77(5): 761–775.

Trank CQ, Rynes SL (2003) Who moved our cheese? Reclaiming professionalism in business education. Academy of Management Learning & Education 2(2): 189–205.

Van Noorden R (2010) Metrics: A profusion of measures. Nature News 465: 864–866.

Wilsdon J, Bar-Ilan J, Frodeman R, et al. (2017) Next-generation metrics: Responsible metrics and evaluation for open science. Brussels: Report of the European Commission Expert Group on Altmetrics.
Author biography

Mathieu Lizotte is a sociologist of inequality and justice. He teaches sociology at the School of Sociological and Anthropological Studies of the University of Ottawa, Canada. His main topics of research are inequality, perceptions of inequality and distributive justice. His current projects focus on wealth inequality, the stigma of poverty, household debt and bankruptcy.