Review Essay on Pardo-Guerra's *The Quantified Scholar*

Étienne Ollion*

Department of Sociology, CNRS, French National Centre for Scientific Research (France)

Submitted: August 28, 2023 – Accepted: December 20, 2023 – Published: March 12, 2024

Abstract

What do research evaluation protocols do to research, and why should we care? In his latest book, sociologist Juan Pablo Pardo-Guerra explores this pressing question through an in-depth investigation of the REF, the research evaluation framework in the United Kingdom. The results are, to say the least, discomforting.

Keywords: Science of science; Social sciences; Research evaluation; Scoring.

* etienne.ollion@polytechnique.edu

In *The Quantified Scholar*, Juan Pablo Pardo-Guerra (2022) has written the book that many of us wanted see in print. By investigating the effects of evaluation on research and on researchers in the United Kingdom, he not only explores an object that will pique the curiosity of arguably self-obsessed academics. He also picks a central question that is relevant far beyond the precincts of science: what do the conditions of scholarly work do to the content of the work produced?

Informed by his background as a science and technology scholar, Pardo-Guerra knows that our academic productions are shaped by the "practices, communities and institutions" (p. 13) they are produced in. In this case, the institutions he wants to scrutinize are all the forms of metrics that have disseminated over the last decades and which are now routinely used by universities and by governments (writ large) to assess the quality of our research.

Such quantifications of academic performance are in no way a new phenomenon. For more than fifty years, publications have been tallied to appraise the productivity of a researcher, and citations have long been counted to gauge visibility or contribution to a field. Put otherwise, the worth of academics has been routinely and regularly scored. This tendency is nonetheless on the rise. In the last decades, this process has been at the same time extended in its scope, and systematized. In many countries, researchers applying for grants, promotions, sabbatical leaves are now assessed via this series of metrics.

The current digitalization of everyday life has made this more prevalent and more accessible at the same time. In just a few clicks, anyone can find the list of publications of any other researcher, the citations to their work, their number of followers on social media (this latter measure is increasingly interpreted as a token of visibility, and consequently valorized as "scientific dissemination"). Or again, using a search engine, they can read about the latest news coverage of a colleague's research, but also see how many times it was featured in the media.

Has this omnipresent scoring modified the way researchers go about their work, from what they write about to how they write about it, and how their careers are shaped? These two broad questions lie at the core of this investigation.

1 The REF as an Ecological Disaster

The results are discomforting. Using wide a range of data and variegated methods, Pardo-Guerra contends that the exercise of the Research Excellence Framework (REF) has made the disciplines under study more prone to conformism and more risk-adverse when it comes to new ideas. Atypical scholars have been normalized or driven out of academia, and departments with peculiar interests have been invited to participate in debates they may have had little interest in.

While the author avoids drawing strong normative conclusions about this, the message driven home throughout the book is clear: the REF caused an ecological disaster. By reducing intellectual diversity, it limited intellectual productivity (it is a well-established fact that monoculture yields diminishing returns). And by forcing institutions to care about topics they had no appetite for, it certainly destroyed some important nature preserves for ideas. The fact that the CCCS at Birmingham, home to Stuart Hall and cradle of cultural studies, was dismantled after a Research Assessment Exercise is often mentioned as evidence of this last point.

This conclusion will come as no surprise to many. It echoes many of the oft-heard criticisms about the REF, some of which are mentioned throughout the book. Yet the data and methods mobilized throughout the book make the contribution much more original.

To conduct his investigation of the impact of the REF on the social sciences, Pardo-Guerra chose to look into four disciplines (sociology, anthropology, economics, and political science) in the United Kingdom. The choice of the country is, in part, opportunistic. The author started his career in Scotland before he went to England, making him acutely aware of the evaluation processes and their impact on research. But there is another rationale for selecting the UK. Since the mid 1980s, the country has implemented a recurrent assessment exercise during which all higher education institutions are evaluated by external committees which rank their worth. Carried out at the departmental levels, the exercise is used to determine the sums allotted to the universities for the next period, creating an incentive for them to perform well under this Research Excellence Framework, or REF.

If the case is relevant, given that metrics are probably more prevalent in the UK than in any other Western country when it comes to assessing research, what makes the book stand out is the set of methods used. To reach his conclusions, Pardo-Guerra assembled a vast trove of data. He looked at publications, at citations, which he collected via a commercial website (the *Social Sciences Citation Index*). But parting way with the gargantuan literature in scientometrics that takes them as primary material, he also collected the text of these articles. And to analyze them, he resorted to natural language processing algorithms to measure the evolution of each of the four disciplines according to a series of dimensions.

2 Score Me, Maybe?

The mechanism laid out by the author is pretty simple. Faced with a daunting and highly consequential review process, departments will try to attract valued scholars (said to be highly "REF-able") to increase their chances to get a high score in the next exercise. In the process, they will try to hire a more varied crew of scholars than they normally would have — at the detriment of some topical, methodological, or theoretical consistency they used to have. One could rightfully argue that this could have no effect besides an increased rotation between universities: after all, social scientist could pursue their own agenda in different places. This is where the interviews, carried out with dozen of scholars, convincingly buttress Pardo-Guerra's argument. Except for a few individuals, many interviewees recognized that the REF made them change where and how often they planned to publish, forcing them to confront other literatures and unamicable reviewers.

And while this can be, in theory, a productive request, it can also be powerful tool for intellectual normalization. In fact, the few interviewees who said they did not change their research strategy were those who were already perfectly attuned to the demands of the exercise criteria. This is perfectly exemplified by the interview with "Carl", the institution made man, on page 136. Because the REF adopted the standards that his institution already had in place, he never had to worry about them: he was already publishing in top journals or internationally, had a policy impact, was widely cited, etc.

But why did the REF have such an impact? The sums allocated are undeniably important, but so are the human and financial costs sustained by the institutions to maximize their chances in a REF. Every five years, some hire external consulting companies and they ask some of their most senior scholars to prep for the review. And between two exercises, they often run mock evaluations. Pardo-Guerra tells us that the reason for this is that the expected returns are not only material, but also symbolic. Success in the exercise is seen as a token of quality, something universities can brag about, and that makes REF-able scholars feel good about themselves. Academia, a work milieu filled with competitive, self-branding individuals anxious about their scholarly worth, proves a fertile ground for the development of theses metrics. Put otherwise, if it were not for the acceptance of these metrics by scholars, the REF could not have such an impact. The quantified scholar is, at the same time, an active quantifying scholar obsessed with their own scores.

The claim could be regarded as tautological. How, in effect, could an instrument have an impact if it is not used or even considered? It is, rather, an intervention in the public and scholarly debates about the roles of scores in our everyday lives. From his Science and Technology Studies background, Pardo-Guerra has inherited more than an interest for the conditions of production of knowledge. He also knows that scoring devices differently affect individuals and institutions, depending on the use they make of them.

Chapter 5, "Hierarchies of quantification", further develops this point as it investigates the factors that favored an investment in the assessment exercise. He points to an unexpected tendency: the REF, he writes, mostly affects departments and scholars placed in intermediary positions. Those on top of the rankings don't need to pay attention to it, while those at the bottom cannot really hope to gain much from it. The "culture of audit" favored by the exercise does not have monotonic nor univocal effects.

3 The Computational Scholar

There is a lot to like in Pardo-Guerra's latest book, not least the ambition that fuels this endeavor. The REF and, more generally, evaluation exercises, have been the object of countless studies, though few are effectively cited. But rather than a disregard for this abundant literature, it seems that this concealment is caused by the goal that the author tried to achieve. The idea is to capture some systemic changes caused by the introduction of this generalized system of evaluation. To do so, and to analyze not only the numbers (of publications, of citations), Pardo-Guerra turned to the content. He collected and investigated more than 70,000 abstracts published in four disciplines, over a period of more than three decades. To carry out this daunting task, he dug into the developing toolbox of Computational Social Science. In many ways, the book itself is a demonstration of what CSS can bring to the human and social sciences. Taking cues from the literature on scoring practices, building upon on a rich qualitative study, the author crafted hypotheses he was later able to test on this massive corpus. Put otherwise, computational methods allowed him to evaluate the gigantic material — in this case, text — at scale.

More specifically, the argument relies on topic models, a Bayesian statistical technique aimed at uncovering the latent structure of texts. This statistical procedure is designed to extract a number of "themes" or "topics" that compose the backbone of all documents at hand. Starting from this first operation of dimensionality reduction, Pardo-Guerra and his team of research assistants devised three clever metrics: one that looks at the *similarity* of a scholar with respect to their colleagues, one that looks at the *typicality* of the department (how is it different from others) and one that, once again using the same topic models, measures the overall *distance* between departments. This latter measure (Pardo-Guerra, 2022, p. 111) is central to the demonstration of increased conformism: since the 1990s and the first exercises of the REF, the distance between departments has drastically decreased.

Invented in the early 2000s, topic models have been immensely popular in the human and social sciences, owing greatly to their inductive character and their ease of use. Rightfully so, given that the methods allow a researcher to take a guided tour in a giant corpus, to extract information from it, and to visualize changes and stability. But can they be used to model the thematic structure of a given field of research, and to measure it accurately (i.e., with the same rate of error) for four disciplines, over a relatively long period of time? More fundamentally, this begs the question: can a set of inductive topics generated by an unsupervised algorithm aptly capture the intellectual diversity of scientific fields? This is an open question, one that can partly be answered by painstakingly looking at the results of the analyses, and by confronting them with exterior sources of knowledge. There is no doubt that the author and his team of assistants did so. In fact, in the book, the author points to a few signs of external validation, as well as to various tests and specifications that were carried out.

Yet because of the onus placed on the method — virtually all statistical results in the book rely at some point on this analysis — it seems only fair to discuss these choices. For instance, the author decided after a series of attempts to set the number of topics to a fixed number ($k = 40$). This is certainly possible, but it begs a question: is forty the right number for *all* disciplines, including those (like anthropology) with a much lower number of articles? And were all the themes easily interpretable, or — as should be expected — were some of them more ambivalent? Did they map the field according to categories that make sense for their practitioners, or did the algorithm sort the abstracts according to some distinct terms — for instance by area of study, which may be a wise choice in certain disciplines, but not in others? Finally, did the topic model have the same consistency over the period? The use of word embeddings later in the book reveals that concepts drastically change meaning over time. Is this not a problem for the topic models used elsewhere?

As a practitioner of the method, I know full well that there is no obvious measure of quality for these tools. But at times when the reliability of topic models is questioned (for an overview, see Shadrova, 2021), a few robustness checks would go a long way to strengthen the argument of the book. In turn, this begs a question for the entire Computational Social Science community: how could one measure conformism and diversity in science in ways that are both accurate and refined? This is certainly a task worth of interest, one that the current developments in NLP (Natural Language Processing) help respond to.

4 What’s in a REF?

A central argument of the book is that the REF played a central role in the reshaping of university departments and, more generally, of academic production, in the United Kingdom. For UK scholars, who are regularly confronted with the daunting perspective of a novel “exercise” and who talk about this more than every five years, there is little doubt this is the case. *The Quantified Scholar* only provides further qualitative and quantitative evidence of this.

But readers from outside of the United Kingdom may be left wondering: what happens in countries where evaluation is not so prevalent? To this, Pardo-Guerra has a response: the UK is just a good testing ground of the impact of evaluation, which in one form or another happens everywhere. The country is, in the words of one of his esteemed predecessors, a “site of strategic research” to investigate what the scoring of academic performance does to academics (Merton, 1973), irrespective of its local implementations. And the criteria evoked in the book will look familiar to scholars from many countries: publications in top journals, internationalization, and policy impact are often on the top of the list of what matters for university administrators.

Two questions remain, one about the general cause of this shift, and one about the specific mechanisms. Let us start with the former. As a French sociologist, I cannot but read Pardo-Guerra’s book thinking about the evolution of the field in my own country. There too, important transformations have happened, some of them in line with what the book describes. Looking at French sociology from a historical perspective, Heilbron (2015) showed for instance that the long-established division between schools of thought receded at the turn of the 2000s.

Once organized around leading scholars like Bourdieu, Touraine, Boudon, and Crozier, the intellectual landscape was rapidly reorganized after that. Likewise, after having been a rallying flag for several decades in various corners of the social sciences, interdisciplinarity seems to be on the wane, an analysis that would confirm the redisciplinarization mentioned in the book.

Could we say that tighter academic evaluation paved the way for these transformations, or are other factors at play? In France, the measuring of academic performance remained relatively limited in comparison to the UK. To account for this change, and for some that took place in other countries, alternative explanations come to mind. Austerity and the ensuing dearth of jobs is an obvious candidate. Did intellectual diversity plummet because of the growing pressures of evaluations, or because the pressure on hiring was not as high in the 1970s, and universities could afford to recruit atypical scholars for a workforce direly needed to teach a quickly growing student body? The rising academic conformism diagnosed by Pardo-Guerra would thus be a consequence of hiring practices from the past more than of contemporary evaluations. And we would just be witnessing a phenomenon diagnosed a long time ago by Max Planck: change, in academia, happens one funeral (or at least one retirement) at a time.

Other plausible causes could be mentioned, like the pressure universities may feel to offer a wide choice of topics in order to attract students in times of rising competition between them for tuition revenue. The variable at play would, this time, have to do with the marketization of college education, a phenomenon that certainly happened in the UK. Or again, what about the omnipresence of digital tools like *Google Scholar* or REPEC for economists? Because they allow competition-driven scholars to tally themselves up against one another in a quasi-continuous fashion, this "ordinalization" of academic worth, to paraphrase sociologist Marion Fourcade (2016), could play a central role. This is certainly quantification, but is it caused by external evaluations?

This could in turn — and this is my second question — help us disentangle what is lumped together under the term "research evaluation." What, in the REF, matters to the observed phenomena? Is it the pressure to publish? To publish articles rather than books? To publish internationally? To get grants? To be visible in the news? To have the ear of politicians or policy specialists? Is it something else? Because of the scope and the ambition of the book, all of these items are easily aggregated under the same banner, but they could gain from being distinguished as they certainly impact research in different ways. Economics, for instance, became more applied as it was called to advise policy-makers. This significant change, which largely predates evaluation, does not impact scholarship in the same way as a pressing invitation to publish internationally does.

As should be clear by now, the book is an important contribution to these debates. It asks an important question and provides new material and evidence, while also offering a blueprint for Computational Social Science research. It does so with wit, accuracy, and in a beautiful style. The questions and doubts raised in this review should thus be seen as an invitation to prolong the conversation. As Bourdieu used to say, *The Quantified Scholar* is definitely a book that is "good to think with".

References

Fourcade, M. (2016). Ordinalization: Lewis A. Coser Memorial Award for Theoretical Agenda Setting 2014. *Sociological Theory*, *34*(3), 175–195. https://doi.org/10.1177/0735275116665876

Heilbron, J. (2015). *French Sociology*. Ithaca, NY: Cornell University Press. https://doi.org/10.7591/9781501701177

Merton, R.K. (1973). *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago, IL: The University of Chicago Press.

Pardo-Guerra, J.P. (2022). *The Quantified Scholar. How Research Evaluations Transformed the British Social Sciences*. New York City, NY: Columbia University Press. https://doi.org/10.7312/pard19780

Shadrova, A. (2021). Topic Models Do Not Model Topics: Epistemological Remarks and Steps towards Best Practices. *Journal of Data Mining and Digital Humanities*, 1–28. https://doi.org/10.46298/jdmdh.7595

Étienne Ollion – Department of Sociology, CNRS, French National Centre for Scientific Research (France)
https://orcid.org/0000-0003-3099-5240 | ✉ etienne.ollion@polytechnique.edu
https://ollion.cnrs.fr/english/

Etienne Ollion is a political sociologist working as a senior fellow at the Centre National de la Recherche Scientifique in France. He is also a Professor of Sociology at École Polytechnique (France). His work focuses on political sociology and on artificial intelligence.