Bibliometric evaluation vs. informed peer review: Evidence from Italy☆

Graziella Bertocchi [a], Alfonso Gambardella [b], Tullio Jappelli [c,*], Carmela A. Nappi [d], Franco Peracchi [e]

[a] Department of Economics "Marco Biagi", University of Modena and Reggio Emilia, Viale Berengario, 51, 41121 Modena, Italy
[b] Department of Management & Technology and CRIOS, Bocconi University, Via Roentgen, 1, 20136 Milan, Italy
[c] Department of Economics and Statistics and CSEF, University of Naples Federico II, Via Cinthia, 21, 80126 Napoli, Italy
[d] ANVUR, Piazza Kennedy, 20, 00144 Rome, Italy
[e] Department of Economics and Finance, University of Rome Tor Vergata, Via Columbia, 2, 00133 Rome, Italy

ARTICLE INFO

ABSTRACT

A relevant question for the organization of large-scale research assessments is whether bibliometric evaluation and informed peer review yield similar results. In this paper, we draw on the experience of the panel that evaluated Italian research in Economics, Management and Statistics during the national assessment exercise (VQR) relative to the period 2004–2010. We exploit the unique opportunity of studying a sample of 590 journal articles randomly drawn from a population of 5681 journal articles (out of nearly 12,000 journal and non-journal publications), which the panel evaluated both by bibliometric analysis and by informed peer review. In the total sample we find fair to good agreement between informed peer review and bibliometric analysis and absence of statistical bias between the two. We then discuss the nature, implications, and limitations of this correlation.

1. Introduction

Measuring research quality is a topic of growing interest to universities and research institutions. It has become a central issue in relation to the efficient allocation of public resources which, in many countries and especially in Europe, represent the main component of university funding. In the recent past, a number of countries – Australia, France, Italy, Netherlands, Scandinavian countries, UK – have introduced national assessment exercises to gauge the quality of academic research. We have also seen a new trend in the way funds are being allocated to higher education in Europe, on the basis not only of actual costs but also, to promote excellence, academic performance. Examples of performance-based university research funding systems (OECD, 2010; Hicks, 2012; Rebora and Turri, 2013) include the British Research Excellent Framework (REF) and the Italian Evaluation of Research Quality. Performance-based funding, however, comes with substantial costs in terms of time and resources, and such costs may differ considerably across evaluation methods (Geuna and Martin, 2003; Martin, 2011).

The main criteria for evaluating research performance combine, in various ways, bibliometric indicators (Moed, 2005; Nicolaisen, 2007) and peer review (Bornmann, 2011). Bibliometric indicators typically use the number of citations that a paper receives, which in turn represents a measure of its impact and international visibility (Burger et al., 1985). Perhaps their simplest application is to the ranking of scientific journals. Although journal rankings have been introduced in various countries, such as Australia, France and Italy, the fact that bibliometric indicators come from different databases (ISI Thompson Reuters, Scimago, Google Scholar, etc.) raises the problem of how to combine the information that they contain (Bartolucci et al., 2013). An additional problem is that journal rankings are only an imperfect proxy for the quality of a paper. We refer to Seglen (1997), Oswald (2007), Bornmann and Daniel (2008),

Fagerberg et al. (2012) and Rafols et al. (2012) for further discussion on the limits of bibliometric analysis as a tool for evaluating research.

Peer review is in principle a better way of evaluating the quality of a paper because it relies on the judgment of experts. However, it is not without its problems. First, there are issues of feasibility and, perhaps, reliability. As all journal editors know, it is not easy to find qualified referees and to provide the right incentives for them to devote adequate effort to the evaluation of a paper. This issue becomes even more serious in the context of a large-scale research assessment. In addition, peer review may be subject to conflicts of interest, and the assessments may not be uniform across research papers, disciplines or research topics. Moreover, what specific criteria reviewers should take into account in their evaluation is subject to extensive discussion (Rinia et al., 1998; Martin and Whitley, 2010). Finally, peer review is much more costly and time demanding than bibliometric evaluation.

Since no evaluation method appears to dominate, it is important to understand whether one can effectively combine bibliometric indices and peer review in order to assess research quality (Butler, 2007; Moed, 2007). This requires the selection of bibliometric indices and an analysis of the correlation between bibliometric and peer review evaluations. This article explores these issues in the context of the Italian Evaluation of Research Quality 2004–2010, hereafter VQR.

The VQR, which formally started at the end of 2011 and was completed in July 2013, was coordinated by the Italian National Agency for the Evaluation of University and Research Institutes (hereinafter ANVUR). The evaluation process was conducted by 14 panels, each corresponding to a broadly defined research area, and combined bibliometric analysis and informed peer review, in proportions that varied across research areas. Our study focuses on the evidence available for one of the 14 areas covered by the VQR, namely Economics and Statistics (Area 13).[1] We present evidence based on our direct involvement in the evaluation process.

The area that we consider is particularly interesting because, at least in Italy, it lies in between the "hard" sciences on the one hand and the humanities and social sciences on the other hand. While in the former most research is disseminated through academic journals and is therefore covered by bibliometric databases, the latter are characterized by a more fragmented literature and more frequent publishing in books and other outlets (Hicks, 1999), so that bibliometric databases are incomplete or almost entirely missing. While for the economic and statistical sciences we do have bibliometric databases covering journal articles, as our analysis will document these databases tend to be incomplete since many journals (published in Italy and elsewhere) are not indexed. Thus, in order to perform the bibliometric evaluation, our first task was to compile a list of all the academic journals – inclusive of non-indexed ones – in which Italian researchers published during the 2004–2010 period covered by the VQR.

We describe the construction of this list and the statistical procedures used to impute bibliometric indicators when missing in order to produce a uniform classification. We then compare the results of the two evaluation methods – bibliometric evaluation and informed peer review – using a random sample of journal articles assessed using both methods. Since comparison is based on a genuine randomized control trial, it represents a significant

contribution to current knowledge, and the results could be useful for other research areas.[2]

Our main finding is that there is adequate agreement between bibliometric evaluation and informed peer review. Although bibliometric evaluation tends to be more generous than informed peer review – it assigns more papers to the top class than informed peer review – in the total sample we find no systematic differences between the two evaluation tools.

We would like to stress that the VQR relies on informed peer review, not just peer review. There are important differences between these two methods. While uninformed peer review is anonymous and double-blind, informed peer review is anonymous, but the referees know the identity of the authors of the item. Further, in the type of informed peer review adopted by the VQR, the evaluation refers to published journal articles, not unpublished manuscripts (as is the case when referees review submitted papers). Since referees know which journal published the paper, this information may influence their evaluation.[3] This is an important issue that we need to clarify at the outset of our analysis. First, it means that we do not seek to assess the intrinsic correlation between peer review and bibliometric evaluation, let alone the intrinsic validity of the latter. The very nature of informed peer review, as opposed to blind peer review, or peer review for short, implies that the reviewer is influenced by both the intrinsic quality of the paper and information about the publication outlet. Second, the structure of the evaluation process, which is fixed and given to us by the VQR, constrains our analysis: the VQR evaluates published material, and the reviewers are informed about the sources of the publications. As a result, we cannot compare bibliometric outcomes with those of uninformed peer review to establish the intrinsic consistency between the two processes. In other words, we cannot disentangle whether the correlation that we observe depends on an intrinsic relation or on the influence of the information on publication outlets on the reviewers.

As a consequence of this caveat, we need to be clear about the policy implications that we can draw from our analysis. Particularly, as noted, we cannot make any claim about the validity of bibliometrics as a substitute for peer review, let alone advocating the substitution of the informed peer review process with bibliometric assessments. However, the correlation between informed peer review and bibliometrics suggests that in any large-scale evaluation exercise, like the one that we carried out, informed peers will produce assessments broadly consistent with the bibliometric indicators. This may be because of an intrinsic correlation or because reviewers update their assessments from their information about the source.

While we cannot distinguish between the two sources, our finding is informative. For example, large scale assessment exercises, which combine bibliometric analyses and informed peer review, are costly, especially because they mobilize several reviewers, so they are usually carried out infrequently. Our result suggests that bibliometric analyses, possibly between two large-scale assessment exercises, may provide a more continuous monitoring consistent with informed peer review. In addition, we check whether the perceived quality of a journal carries a disproportionate weight in the evaluations by employing background information about the refereeing process. We find that even when

---