Research Policy 44 (2015) 451–466

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

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

Article history:
Received 29 October 2013
Received in revised form 28 July 2014
Accepted 18 August 2014
Available online 18 September 2014

Keywords:
Research assessment
Informed peer review
Bibliometric evaluation
VQR

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),

☆ The authors have been, respectively, president of the panel evaluating Italian research in the area "Economics and Statistics" (Tullio Jappelli), coordinators of the sub-panels for Economics (Graziella Bertocchi), Management (Alfonso Gambardella) and Statistics (Franco Peracchi), and assistant to the panel (Carmela A. Nappi). We acknowledge helpful suggestions and comments from the members of the panel and from Sergio Benedetto, national coordinator of the VQR. We are grateful to Dimitris Christelis for implementation of the multiple imputation model and comparison with the baseline model. We also thank the editor of this journal and three anonymous referees for their valuable advice.

* Corresponding author.
E-mail addresses: graziella.bertocchi@unimore.it (G. Bertocchi), alfonso.gambardella@unibocconi.it (A. Gambardella), tullio.jappelli@unina.it, jappelli@unina.it (T. Jappelli), carmelaanna.nappi@anvur.org (C.A. Nappi), franco.peracchi@uniroma2.it (F. Peracchi).

http://dx.doi.org/10.1016/j.respol.2014.08.004
0048-7333/© 2014 Elsevier B.V. All rights reserved.

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

[1] The area is denominated by ANVUR "Economics and Statistics" but also includes Management. From now on, we call it Economics and Statistics to be consistent with the official label by ANVUR.

[2] On the comparison between expert assessment and bibliometric indicators, see for instance Allen et al. (2009) and Eyre-Walker and Stoletzki (2013). Waltman and Costas (2013) analyze the correlation between recommendations and citations.

[3] On post-publication peer review see Allen et al. (2009), Eyre-Walker and Stoletzki (2013), and Waltman and Costas (2013). Hicks and Wang (2011) discuss the issue of assessing the scholarliness of a journal within potentially fragmented scientific communities.

reviewers are likely to be influenced by the perceived quality of the publication outlet, their perceptions are highly correlated with other indicators of the quality of an article and are not the leading factor in their assessment.

The remainder of this paper is organized as follows. Section 2 illustrates the characteristics of the assessment exercise implemented by the VQR. Section 3 describes the construction of the journal list database and presents descriptive statistics. Section 4 deals with the imputation of missing values of the bibliometric indicators, describing a simple two-step procedure and a more elaborate procedure based on multiple imputations. Section 5 presents the ranking and summary statistics for the distribution of journals in the different merit classes, by research sub-areas (Economics, Economic History, Management, Statistics). Section 6 describes the comparison of informed peer review and bibliometric evaluation for the random sample and also reports various robustness checks. Section 7 concludes. Two appendices provide information on the technical aspects of the multiple imputation procedure and the referees' evaluation forms.

2. The Italian evaluation of research quality 2004–2010 (VQR)

The VQR was conducted by ANVUR, on behalf of the Italian Ministry of Education, University and Research (MIUR), to evaluate the scientific production of Italian public research institutions, namely public universities, private universities awarding officially recognized academic degrees, and public research institutions.[4] Admissible research papers included journal articles, books, book chapters, conference proceedings, etc., published between 2004 and 2010. The evaluation was at two levels: the individual research institutions and their departments. University researchers had to submit three research papers each, whereas researchers from institutions with no teaching duties had to submit six.[5] The VQR has been the first massive university evaluation exercise in Italy covering all research fields.[6] The exercise is meant to be repeated in the future on a regular basis. Starting from 2013, the results of the VQR have contributed to determining the share of the MIUR University fund allocated to each university.

For each of 14 broadly defined research areas,[7] ANVUR selected a panel of experts based on their scientific qualifications and previous experience with evaluation procedures. The number of panelists varied with the number of researchers in each area. Further, specific sub-panels were set up for areas with a strong disciplinary heterogeneity and a large number of papers submitted for evaluation. For Area 13 (Economics and Statistics), the number of panelists was 36 and three sub-panels were set up to cover Economics (including Economic History), Management and Statistics.

Each panel defined with ANVUR the general principles to be adopted in the evaluation and had the responsibility to assign a rating to each papers and compile a ranking of institutions. In particular, each panel could select one of two evaluation methods: bibliometric analysis, based on the citations of a paper and the impact factor (or other bibliometric indicator) of the journal where it was published, and informed peer-review evaluation, carried out by external experts chosen by the panel (normally two for each paper, independently selected by two different panelists).[8] Each panel could also choose to adopt both methodologies. The panel for Area 13 decided to evaluate all journal articles by bibliometric analysis (after imputing missing values of bibliometric indicators through a procedure described in more detail in Section 4), and all other papers by informed peer review. It also decided to send to the informed peer-review process a random sample of journal articles. Therefore, for journal articles from Area 13, we are able to compare the results of the two evaluation methods.

The VQR rules required classifying each paper into one of the following six categories or "merit classes"[9]:

A. Excellent: the publication is in the highest 20% of the quality ranking shared by the international scientific community (weight 1);
B. Good: the publication is in the 60–80% segment (weight 0.8);
C. Acceptable: the publication is in the 50–60% segment (weight 0.5);
D. Limited: the publication is in the lowest 50% (weight 0);
E. Not assessable: the publication does not belong to the typologies included in the evaluation exercise; or it includes attachments or files that are not adequate for evaluation; or it has been published in previous or subsequent years (weight −1);
F. Fraud or plagiarism (weight −2).

For Area 13, the categorization of journal articles was based on the journal classification produced internally (see Section 5 for details) and the number of citations that each article received. For papers sent to informed peer review, the evaluation relied on the general criteria defined by ANVUR, namely relevance, originality/innovation, and internationalization/international standing. Analytic assessments from each reviewer had to be converted into numerical scores and then assigned to the six categories described above. The panel then produced a final evaluation of each paper subject to informed peer review through a Consensus Group consisting of the two panelists in charge of the paper, plus a third when needed. Notice that, since external experts received €30 for each review, the cost of sending a paper to informed peer review was €60, to be added to the fixed costs of the research assessment.

3. The journal list

The panel for Area 13 compiled an initial journal list based on ISI-Thomson Reuters Web of Science (WoS). This list included all journals in the ISI-JCR Social Science Edition[10] belonging to the subject categories relevant to the area, plus other journals in the ISI-JCR Science Edition.[11] This initial list was then expanded, using the

[4] Other public and private research institutions were allowed to participate in the evaluation upon request.

[5] Exceptions to this rule were possible depending on the year of recruitment or the periods of maternity or sickness leave of a researcher.

[6] An early evaluation exercise was carried out in Italy in 2006 with reference to papers published in 2001–2003. Each university was required to submit a relatively small number of papers (half the number of its researchers), with no restrictions across areas. For additional details, see Rebora and Turri (2013) and the references therein.

[7] The 14 research areas are: Mathematics and Computer Sciences (Area 1); Physics (Area 2); Chemistry (Area 3); Earth Sciences (Area 4); Biology (Area 5); Medicine (Area 6); Agricultural and Veterinary Sciences (Area 7); Civil Engineering and Architecture (Area 8); Industrial and Information Engineering (Area 9); Ancient History, Philology, Literature and Art History (Area 10); History, Philosophy, Pedagogy and Psychology (Area 11); Law (Area 12); Economics and Statistics (Area 13); Political and Social Sciences (Area 14).

[8] By the VQR rules, at least half of all papers over all areas had to be evaluated through peer review.

[9] See http://www.anvur.org/attachments/article/122/bando_vqr_def_07_11.pdf.

[10] The subject categories included are: DI (Business), DK (Business, Finance), FU (Demography), GY (Economics), NM (Industrial Relations and Labour), PS (Social Sciences, Mathematical Methods), PE (Operations Research and Management Science), XY (Statistics and Probability).

[11] Other journals have been included from the following ISI Science subject categories: AF (Agricultural Economics), JB (Environmental studies), KU (Geography), NE (Public, Environmental and Occupational Health), PO (Mathematics, Interdisciplinary Applications), WY (Social Work), YQ (Transportation).

U-GOV dataset[12], to include all journals containing articles published by Italian researchers from Area 13 during the period 2004–2010.

To account for differences in publication style, the panel defined four sub-areas: Business, Management and Finance (hereafter Management); Economics; Economic History and History of Economic Thought (hereafter History); Statistics and Applied Mathematics (hereafter Statistics); plus an additional sub-area comprising three general-interest journals, namely *Science*, *Nature*, and *Proceedings of the National Academy of Sciences*. To avoid ranking the same journal differently in different sub-areas, each journal was assigned to one and only one sub-area.

As a summary of the differences across these sub-areas, it is enough to notice that journal articles represent 77% of all research papers submitted to the VQR for Economics, 76% for Statistics, 46% for Management, and only 32% for History. These differences reflect the distinct publication styles prevailing within the various sub-areas of the Italian research community; for instance, publishing monographs is more common in accounting or economic history than in econometrics, economics or statistics.

The next step was to integrate the available bibliometric information for the journals in the WoS with the additional information from Google Scholar available for all journals. Thus, in April 2012, the panel collected *h*-index data from Google Scholar for all journals in the preliminary list for Area 13.[13] Several studies within the social sciences have concluded that the degree of agreement between the bibliometric indicators from WoS and Google Scholar is high[14], and that the rankings of journals for which both sets of indicators are available tend to be similar. This is especially true when the objective is classification into broad categories – as in the VQR – rather than comparison across individual journals.

At the end of April 2012, the panel published the preliminary list of journals and solicited comments and suggestions from the scientific community. The final list, published at the end of July 2012, contains several changes based on the comments received.[15] Table 1 shows that the list includes 1903 journals, of which 767 (40%) belong to Management, 643 (34%) to Economics, 445 (23%) to Statistics, and 48 (2%) to History. ISI journals represent 49% of the total, but their fraction varies by sub-area, ranging from 40% for History to 42% for Management, 53% for Economics, and 56% for Statistics.

Table 1
Distribution of journals by sub-area and ISI code.

	Research sub-areas				
	Economics	History	Management	Statistics	Total
Non ISI %	305	29	446	195	975
	47.43	60.42	58.15	43.82	51.23
ISI %	338	19	321	250	928
	52.57	39.58	41.85	56.18	48.77
Total %	643	48	767	445	1903
	100.00	100.00	100.00	100.00	100.00

Note. The table reports the distribution of the journals included in the list by research sub-area and presence in the database ISI–Thomson Reuters.

Table 2 presents the basic statistics for our four bibliometric indicators: the Impact Factor (IF), the 5-year Impact Factor (IF5) and the Article Influence Score (AIS), all computed by ISI, and the *h*-index obtained from Google Scholar. The IF is computed using the same methodology as for IF5, but over a two-year period. The AIS excludes journal self-citations and gives more weight to citations received from higher ranked journals. Data on the journals' IF, IF5 and AIS refer to December 2011, the latest data available before the start of the evaluation process.

The IF, available for all 912 ISI journals, has a mean of 1.19 and a standard deviation of 0.97. Its mean value varies by sub-area, and is highest for Management (1.47) and lowest for History (0.49). The IF5 and the AIS are available only for a subset of 684 ISI journals. Means and percentiles for these two indices show important differences in citation patterns by sub-area, with the lowest values for the History journals. Apart from History, the distribution of the AIS appears to be more similar across sub-areas than the distribution of the IF5.

Table 3 shows the fraction of missing values on the three ISI indicators, separately by sub-area. Column (1) shows the total number of journals, columns (2) and (3), respectively, show the number and percentage of journals with missing values for the IF, and columns (4) and (5) give the same information for the IF5 and the AIS. The last two indicators have identical patterns of missingness, as the

Table 2
Statistics for impact factor (IF), 5-year impact factor (IF5), article influence score (AIS) and *h*-index (h) by research sub-area.

Research sub-area	Mean	sd	p10	p25	p50	p75	p90	iqr
Impact factor (IF)								
Economics	1.05	0.92	0.22	0.41	0.84	1.40	1.99	0.99
History	0.49	0.34	0.11	0.24	0.39	0.68	1.04	0.44
Management	1.47	1.16	0.32	0.65	1.11	2.01	2.94	1.36
Statistics	1.06	0.65	0.37	0.58	0.95	1.38	1.93	0.80
Total	1.19	0.97	0.27	0.53	0.94	1.58	2.36	1.05
5-year impact factor (IF5)								
Economics	1.55	1.18	0.42	0.79	1.33	2.00	2.89	1.22
History	0.73	0.36	0.34	0.44	0.63	1.12	1.24	0.67
Management	2.44	1.90	0.76	1.17	1.94	3.02	4.92	1.85
Statistics	1.47	0.87	0.59	0.84	1.28	1.87	2.51	1.03
Total	1.80	1.44	0.56	0.88	1.42	2.25	3.41	1.37
Article influence score (AIS)								
Economics	1.09	1.54	0.17	0.35	0.64	1.06	2.34	0.71
History	0.45	0.33	0.14	0.15	0.41	0.80	0.94	0.65
Management	0.93	1.10	0.19	0.34	0.60	0.99	2.08	0.65
Statistics	0.95	0.69	0.31	0.51	0.72	1.23	1.89	0.72
Total	0.98	1.18	0.22	0.39	0.68	1.06	2.00	0.67
h-Index (*h*)								
Economics	21.51	18.92	4.00	7.00	16.00	30.00	47.00	23.00
History	9.31	6.34	4.00	4.00	7.00	11.50	21.00	7.50
Management	22.77	20.78	4.00	8.00	17.00	31.00	47.00	23.00
Statistics	19.77	16.38	4.00	7.00	14.00	28.00	43.00	21.00
Total	21.30	19.06	4.00	7.00	15.00	29.00	45.00	22.00

Note: The table reports statistics of the four bibliometric indicators considered (impact factor, 5-year impact factor, article influence score and *h*-index). The statistics reported are: mean; standard deviation (sd); 10th, 25th, 50th, 75th and 90th percentiles (respectively, p10, p25, p50, p75, p90); inter-quantile range (iqr).

[12] U-GOV is a dataset containing the papers of all researchers in Italian public universities. From the U-GOV dataset we excluded the following publication outlets: journals clearly outside Area 13; working paper series and collections/reports of Departments/Research Institutions; journals with Google Scholar's *h*-index missing for the period 2004–2010 (or shorter periods for new journals); journals with *h*-index less than three for the period 2004–2010; and journals too recent for the *h*-index to be reliable, such as the *American Economic Journals* (*Macroeconomics*, *Microeconomics*; *Applied Economics*; *Economic Policy*) and the *Annual Review of Economics*. Publications in these journals were therefore sent to peer review.

[13] A journal has index *h* if *h* of its *N* published articles have at least *h* citations each, and the other *N*–*h* have no more than *h* citations each. We computed the *h*-index in Google Scholar in the 2004–2010 period. Data were collected in April 2012 and checked throughout May 2012.

[14] See, e.g., Mingers et al. (2012) for Management, Linnemer and Combes (2010) for Economics, and Jacobs (2011) for Sociology. For a comparison between WoS and Google Scholar see Harzing and van der Wal (2008).

[15] The panel received comments regarding the following aspects: misclassification of journals across the four sub-areas; misreported presence of journals in the WoS; misreported values of the *h*-index; inclusion of journals that meet the panel classification requirements; exclusion of journals published after 2008 or pertaining to other disciplines; errors in the name or ISSN of journals.

Table 3
Prevalence of missing values for all three bibliometric indicators.

Research sub-area	(1)	(2)	(3)	(4)	(5)
	Total number of journals	2-year Impact Factor (IF)		5-year Impact Factor (IF5) and Article Influence Score (AIS)	
		Number of journals with a missing value	Percentage of journals with a missing value (%)	Number of journals with a missing value	Percentage of journals with a missing value (%)
Economics	643	319	49.61	399	62.05
History	48	30	62.50	37	77.08
Management	767	447	58.28	549	71.58
Statistics	445	195	43.82	234	52.58

Note: The table reports the total number of journals in the list by research sub-area and the number and percentage of journals with missing values for the three bibliometric indicators in ISI—Thomson Reuters (impact factor -IF-, 5-year impact factor -IF5-, article influence score -AIS-). IF5 and AIS have identical patterns of missingness, as the AIS can be defined only when IF5 is also defined.

Table 4
Skewness and kurtosis of the levels and logarithms of IF5 and AIS.

Research sub-area	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Levels				Logarithms			
	5-year Impact factor (IF5)		Article Influence Score (AIS)		5-year Impact factor (IF5)		Article Influence Score (AIS)	
	Skewnewss	Kurtosis	Skewnewss	Kurtosis	Skewnewss	Kurtosis	Skewnewss	Kurtosis
Economics	2.320	11.515	4.038	22.691	−0.674	4.179	−0.284	4.253
History	0.283	15.009	0.539	1.735	−0.351	4.384	0.054	1.433
Management	2.158	9.458	3.167	15.009	−0.483	4.450	−0.303	4.384
Statistics	1.526	6.500	1.938	8.397	−0.702	5.696	−1.006	7.273

Note. The table reports the indices of skewness and kurtosis for 5-year impact factor (IF5) and article influence score (AIS) in levels (columns (1)–(4)) and in logarithms (columns (5)–(8)).

AIS is defined only when the IF5 is also defined. The fraction of missing values is notable for all three indicators, but especially for IF5 and AIS. Looking by sub-area, the journals in History and Management are the most affected by missingness, while the journals in Statistics are the least affected.

It is useful to inspect the distribution of non-missing values of the various indicators, as this is relevant for the choice of imputation model described in Section 4. Nonparametric kernel estimates of the density of the IF5 and the AIS (not reported for brevity) reveal right-skewness and long right tails. This is true for all four sub-areas, but especially for Economics and Management. The indices of skewness and kurtosis shown in columns (1)–(4) of Table 4 confirm these findings. Skewness and long right tails are a well-known feature of bibliometric indicators in science, particularly for individual scientists or articles (Seglen, 1992). Our findings confirm existing evidence of this phenomenon across journals as well (Stern, 2013).

The substantial skewness and kurtosis in the distribution of the bibliometric indicators make estimation of regression models in levels problematic, as the outliers in the long right tail of the distribution are likely to be very influential. To reduce their impact, we chose to estimate our models in logarithms rather than levels. The logarithmic transformation is strictly increasing, so it does not change the ranking of journals, but makes the distribution of all indicators much more symmetric and closer to the normal (Gaussian) distribution. This can be seen by the much reduced values of skewness and kurtosis shown in columns (5)–(8) of Table 4. In Economics, Management and Statistics, and for both the IF5 and the AIS, the logarithmic transformation brings skewness closer to zero and kurtosis closer to three, which are the values for a normal (Gaussian) distribution.

Table 5 shows the correlations between the various indicators after the logarithmic transformation. The correlation between the three ISI indicators is very high: for instance, the correlation between log(IF) and log(IF5) is always higher than 0.9, while the correlation between log(IF5) and log(AIS) is always higher than 0.8.

The *h*-index from Google Scholar, which is available for all journals in our list, also reveals differences in citation patterns across sub-areas: the lowest mean value is again for History, the highest for Management. The *h*-index correlates strongly and positively with the three ISI indicators. In particular, for Economics and Management the correlation between log(*h*) and log(IF5) and log(AIS) exceeds 0.7, for History it ranges from 0.61 for the IF to 0.72 for the IF5, for Statistics it ranges from 0.65 for the AIS to 0.73 for the IF5 (Table 5). These values made the panel confident that the *h*-index

Table 5
Correlation matrix of log bibliometric indicators by research sub-area.

	log (IF)	log(IF5)	log(AIS)	log(h)
Economics				
log(IF)	1.0000			
log(IF5)	0.9592	1.0000		
log(AIS)	0.8277	0.8887	1.0000	
log(h)	0.7173	0.7753	0.7936	1.0000
History				
log(IF)	1.0000			
log(IF5)	0.9323	1.0000		
log(AIS)	0.9384	0.9367	1.0000	
log(h)	0.6058	0.7164	0.6741	1.0000
Management				
log(IF)	1.0000			
log(IF5)	0.9192	1.0000		
log(AIS)	0.7432	0.8288	1.0000	
log(h)	0.7148	0.7636	0.7256	1.0000
Statistics				
log(IF)	1.0000			
log(IF5)	0.9272	1.0000		
log(AIS)	0.7478	0.8179	1.0000	
log(h)	0.6904	0.7290	0.6540	1.0000

Note. The table reports the correlation between the logarithms of the four bibliometric indicators considered (impact factor -IF-, 5-year impact factor -IF5-, article influence score -AIS- and *h*-index -*h*-) by research sub-area.

Table 6
Differences in journal rankings between the baseline and the multiple imputation methods.

Rank difference across imputation methods	(1)	(2)	(3)	(4)
	5-year impact factor (IF5)		Article influence score (AIS)	
	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals
Economics				
Rank difference = −3	7	1.09%	7	1.09%
Rank difference = −2	18	2.80%	20	3.11%
Rank difference = −1	52	8.09%	40	6.22%
Rank difference = 0	485	75.43%	494	76.83%
Rank difference = +1	66	10.26%	71	11.04%
Rank difference = +2	15	2.33%	10	1.56%
Rank difference = +3	0	0.00%	1	0.16%
Percentage of journals for which the rank difference is between −1 and +1	93.78%		94.09%	
Management				
Rank difference = −3	5	0.65%	10	1.30%
Rank difference = −2	10	1.30%	25	3.26%
Rank difference = −1	66	8.61%	74	9.65%
Rank difference = 0	607	79.14%	543	70.80%
Rank difference = +1	63	8.21%	86	11.21%
Rank difference = +2	16	2.09%	28	3.65%
Rank difference = +3	0	0.00%	1	0.13%
Percentage of journals for which the rank difference is between −1 and +1	95.96%		91.66%	
Statistics				
Rank difference = −3	3	0.67%	6	1.35%
Rank difference = −2	8	1.80%	15	3.37%
Rank difference = −1	23	5.17%	28	6.29%
Rank difference = 0	380	85.39%	338	75.96%
Rank difference = +1	28	6.29%	49	11.01%
Rank difference = +2	3	0.67%	9	2.02%
Rank difference = +3	0	0.00%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	96.85%		93.26%	

Note. The table reports the differences in the journal rankings obtained with the two imputation methods (the baseline imputation method -BIM- and multiple imputation method -MIM-) by research sub-area. Note that the table does not report the results for the research sub-area History since the multiple imputation model was not used for the above mentioned sub-area because of the small number of observations.

was a strong predictor to use for imputing missing values of IF, IF5 and AIS.

4. Imputation of bibliometric indicators

We now describe the procedure adopted by the panel to impute missing values for the three ISI indicators (IF, IF5 and AIS). After taking logarithms of all three indicators, the imputation methods considered are:

(i) A baseline imputation method (BIM) which regresses the logarithm of each of the three ISI indicators on a constant and the logarithm of the *h*-index. We use the *h*-index as a predictor because it is always available. Regressions are carried out separately by sub-area and, for each indicator/sub-area combination, the estimation sample consists of the observations with non-missing values for the indicator of interest. We then fill-in the missing values with the values predicted by the regressions.

(ii) A more elaborate multiple imputation method (MIM) which produces multiple imputed values for each missing observation. The principle of multiple imputations, introduced by Rubin (1987), is widely used in micro-data surveys.

Unlike BIM, which produces a single imputed value for each missing observation, MIM recognizes that imputation is subject to uncertainty and produces multiple imputed values. This allows one to estimate not only the expectation of the missing value but also the extra variance due the imputation process. This is important because ignoring this additional uncertainty, as BIM does, may result in biased standard errors.

In our version of MIM, each indicator to be imputed is regressed not only on a constant term and the logarithm of the *h*-index, but also on the observed or imputed values of the other indicators. For example, to impute the IF we use as predictors the IF5 and the AIS, which can have imputed values in the sample of non-missing observations for IF. Given the high correlation of the IF with the IF5 and the AIS, including these two indicators should increase the predictive power of the regression model.[16] In addition to the level of each indicator, we include its square to allow for possible nonlinearities. We also include a binary indicator equal to one for a journal published in English because this affects the probability that the journal is included in the WoS. To reduce the influence of outliers, the MIM estimation sample only retains observations with values of the dependent variable above the 1st percentile and below the 99th percentile. As a result, the estimation samples for MIM are slightly smaller than for BIM.

MIM runs iteratively until convergence, which occurs when predicted values hardly change from one iteration to the next. We set a maximum of 100 iterations and, after checking for convergence, we used the predictions from the last iteration as our final imputations.

[16] The particular implementation of MIM that we used is from van Buuren et al. (2006). Details are given in Appendix A.

Table 7
Differences in journal rankings across bibliometric indicators, baseline imputation method.

Rank difference across bibliometric indicators	(1)	(2)	(3)	(4)	(5)	(6)
	IF5 versus AIS		IF5 versus *h*-index		AIS versus *h*-index	
	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals	Number of journals	Percentage of all journals
Economics						
Rank difference = −3	0	0.00%	0	0.00%	0	0.00%
Rank difference = −2	4	0.62%	13	2.02%	9	1.40%
Rank difference = −1	43	6.69%	43	6.69%	27	4.20%
Rank difference = 0	554	86.16%	508	79.01%	542	84.29%
Rank difference = +1	39	6.07%	71	11.04%	61	9.49%
Rank difference = +2	2	0.31%	7	1.09%	4	0.62%
Rank difference = +3	1	0.16%	1	0.16%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	98.91%		96.73%		97.98%	
History						
Rank difference = −3	0	0.00%	0	0.00%	0	0.00%
Rank difference = −2	0	0.00%	1	2.08%	0	0.00%
Rank difference = −1	2	4.17%	2	4.17%	5	10.42%
Rank difference = 0	44	91.67%	41	85.42%	38	79.17%
Rank difference = +1	2	4.17%	4	8.33%	5	10.42%
Rank difference = +2	0	0.00%	0	0.00%	0	0.00%
Rank difference = +3	0	0.00%	0	0.00%	0	0.00%
Percentage of journals for which the rank difference is between −1 and +1	100.00%		97.92%		100.00%	
Management						
Rank difference = −3	1	0.13%	0	0.00%	0	0.00%
Rank difference = −2	5	0.65%	13	1.70%	11	1.43%
Rank difference = −1	25	3.26%	31	4.04%	41	5.35%
Rank difference = 0	701	91.40%	662	86.31%	652	85.01%
Rank difference = +1	31	4.04%	54	7.04%	56	7.30%
Rank difference = +2	2	0.26%	5	0.65%	4	0.52%
Rank difference = +3	2	0.26%	2	0.26%	3	0.39%
Percentage of journals for which the rank difference is between −1 and +1	98.70%		97.39%		97.65%	
Statistics						
Rank difference = −3	1	0.23%	0	0.00%	2	0.45%
Rank difference = −2	7	1.57%	12	2.70%	9	2.02%
Rank difference = −1	39	8.76%	40	8.99%	46	10.34%
Rank difference = 0	356	80.00%	342	76.85%	332	74.61%
Rank difference = +1	33	7.42%	44	9.89%	44	9.89%
Rank difference = +2	9	2.02%	6	1.35%	10	2.25%
Rank difference = +3	0	0.00%	1	0.23%	2	0.45%
Percentage of journals for which the rank difference is between −1 and +1	96.18%		95.73%		94.83%	

Note: The table reports the differences in the journal rankings from the baseline imputation method (BIM) comparing (by pair) the results obtained using impact factor (IF), 5-year impact factor (IF5) and article influence score (AIS).

For each missing observation, we produced 500 imputations. Following Rubin (1987), the missing value of the logarithm of an indicator for a particular observation was filled in using the average over the 500 imputations for that observation. Because the sample available for History is very small, we did not use the MIM method in this case.

The estimation results show that for both the AIS and the IF5 the adjusted R^2 of BIM is always high (between 0.5 and 0.6, depending on the research sub-area), indicating good predictive power despite this method using only the logarithm of the *h*-index as a predictor. As already discussed, MIM includes a richer set of predictors. In fact, the adjusted R^2 for MIM is higher than for BIM (between 0.6 and 0.8).

5. Classification of journals

After producing imputations using both BIM and MIM, we compare the two methods in a more formal way by examining the differences in the implied journal classification. To classify journals, we first create deciles of the distribution of the logarithm of the IF5, the AIS and the *h*-index for each sub-area, using both the non-imputed and the imputed values. Then, following the VQR rules, we classify journals into four classes using the following criteria: journals in the lowest five deciles are assigned to class D, those in the sixth decile to class C, those in the seventh and eighth deciles to class B, and those in the top two deciles to class A. After creating these four classes, we compare how the classification of journals differs across both imputation methods and bibliometric indicators.

Table 6 shows substantial agreement between BIM and MIM, but also reveals some differences in journal ranking between these two imputation methods. For example, for the AIS there are 40 journals in Economics with a rank difference of minus one, i.e., they rank one level lower under BIM compared to MIM. On the other hand, for the IF5 there are 28 journals in Statistics with a rank difference of one, i.e., they rank one level higher under BIM compared to MIM.

To compare better the different rankings obtained under the two methods, for each sub-area/indicator combination we compute the percentage of journals for which the two imputation methods

Table 8
Final classification of journals.

	Research sub-area				
	Economics	History	Management	Statistics	Total
A %	152	10	172	112	446
	23.64	20.83	22.43	25.17	23.44
B %	118	9	144	81	352
	18.35	18.75	18.77	18.20	18.50
C %	61	5	76	37	179
	9.49	10.42	9.91	8.31	9.41
D %	312	24	375	215	926
	48.52	50.00	48.89	48.31	48.66
Total	643	48	767	445	1903
	100.00	100.00	100.00	100.00	100.00

Note: The table reports the final journal classification by research sub-area and merit classes.

Table 9
Distribution of journal articles in the population and in the sample.

	Population	Sample	%
Economics	2361	235	10
History	147	37	25
Management	1750	175	10
Statistics	1423	143	10
Total	5681	590	

Note. The table reports the distribution of journal articles by research sub-area in the population of articles submitted and in the random sample.

Table 10
Distribution of bibliometric rankings in the population and in the sample.

	N population	% population	*N* sample	% sample
Economics				
A	923	39.09	95	40.43
B	337	14.27	29	12.34
C	434	18.38	49	20.85
D	667	28.25	62	26.38
History				
A	35	23.81	9	24.32
B	43	29.25	12	32.43
C	25	17.01	7	18.92
D	44	29.93	9	24.32
Management				
A	465	26.57	44	25.14
B	238	13.60	22	12.57
C	231	13.20	31	17.71
D	816	46.63	78	44.57
Statistics				
A	507	35.63	51	34.92
B	382	26.84	38	27.76
C	166	11.67	16	11.27
D	368	25.86	37	26.06

Note. The table reports the number and percentage of journal articles by research sub-area and by merit class in the population and in the random sample.

produce rankings that are “not too dissimilar”, in the sense that their rank difference is between minus one and one. It turns out that 95% of the journals belong to this category, the lowest percentage being 92% for the AIS in Management. In fact, most journals rank the same.

Thus, while BIM and MIM may sometime give different results for individual journals, for the purposes of classifying journals according to the VQR rules both methods give essentially equivalent results. Therefore, for our final journal classification we use the ranking produced by BIM, which is simpler and more easily implementable.

Having chosen BIM, the panel then looked at the differences in journal rankings between pairs of indicators. Again, most journals rank the same, no matter which indicator is used. This emerges clearly in Table 7, which shows the distribution of the differences in rank between pairs of indicators. Most journals rank very similarly under all three indicators. The differences are largest for the AIS and the *h*-index for the Statistics sub-area. However, even in this case, the percentage of journals with a rank difference of at most one in absolute value is 94.8%, while the percentage of journals that rank the same is 74.6%. This is not surprising as all indicators are strongly positively correlated and the *h*-index is a crucial predictor when imputing the IF, the IF5 and the AIS.

The strong correlation between the various indicators means that, in principle, one could employ any of them for classification purposes. Given these considerations, the panel decided to base the final classification of journals on the maximum between their AIS and IF5 rank. It also decided to make the final classification of each journal article dependent on the individual citations it received in the WoS. Specifically, the panel upgraded articles published in ISI journals by one level if they received at least five citations per year in 2004–2010. No upgrading was made for articles not published in ISI journals because of lack of reliable citation data.[17]

Table 8 shows the final journal classification by sub-area. Overall, 48.7% of the journals are in class D (“limited”), 9.4% in class C (“acceptable”), 18.5% in class B (“good”), and 23.4% in class A (“excellent”). These proportions are slightly different from those recommended by the VQR guidelines (namely 50%, 10%, 20% and 20%). This mainly reflects three factors: the rule of the maximum between AIS and IF5 ranks, the presence of ties in the imputed values of the AIS and the IF5, and the panel decision to upgrade some Italian journals to class C.[18] The fraction of papers in class D is similar for all sub-areas. The fraction in class A is slightly above average for Statistics (25.2%) and slightly below average for Management and History (22.4% and 20.8%, respectively). In terms of absolute numbers, Management has the largest number of journals in class A (172), followed by Economics (152), Statistics (112), and History (10).

6. Comparison between informed peer review and bibliometric evaluation

The set of articles submitted to the VQR and published in one of the journals in the list for Area 13 consists of 5681 articles. From this population, a stratified sample of 590 articles was randomly drawn, corresponding to 10% of the journal articles for Economics, Management and Statistics, and 25% for History.[19] Oversampling of History was necessary due to the small size of its population of articles (147 articles). Articles in this sample were then sent out to informed peer review, with the goal of comparing the results with bibliometric evaluation.

Table 9 shows the distribution of both the population and the sample of journal articles by sub-area. Table 10 shows the same distribution by merit class (A, B, C or D). The population and the

[17] The panel made this decision because of concerns about the fact that journal citations may not reflect well the citations received by a single journal article. However, the practical effect of the upgrading was negligible, since the few articles that received a sufficient number of citations already appeared in A-class journals, with the exception of only six papers. Only one of these six papers turned out to be included in the random sample described in the following section.

[18] The panel decided to upgrade 20 Italian journals (5 in each sub-area) from class D to class C based on the value of their *h*-index.

[19] The sample was drawn before starting the peer review process using a random number generator.

sample distributions are very similar for each sub-area. We conclude that our sample is representative of the population of journal articles, both overall and within each sub-area.

The informed peer review process was managed as for a scientific journal with two independent editors. Each article was first assigned to two panelists with expertise in the article's specific field of research. Each of them assigned the article to an independently chosen informed peer reviewer.[20] Overall 610 referees were selected on the basis of their academic curricula and research interests.[21] Informed peer reviewers were instructed to evaluate the article according to three criteria: relevance, originality/innovation, and internationalization/international standing. Referees expressed their evaluation on a predefined form containing three broad questions referring to the three above-mentioned dimensions of the quality of the papers and an open field.[22] As already mentioned in Section 2, based on the informed peer reviews the panel produced a final evaluation through a Consensus Group consisting of the two panelists in charge of the article, plus a third when needed.

For each article included in our sample, the following variables are therefore available: the bibliometric indicator (*F*) based on the number of citations the article received and the classification of the journal in which it was published, the evaluation of the first referee (P1), the evaluation of the second referee (P2)[23], and the final evaluation of the Consensus Group (*P*). Each of these variables is mapped into one of four merit classes, corresponding, respectively, to the top 20% of the quality distribution of published articles (class A), the next 20% (class B), the next 10% (class C), and the bottom 50% (class D). More precisely, variables P1 and P2, originally measured on a numerical scale between 3 and 27 (with scores from 1 to 9 assigned to the three different criteria) are converted into one of the four merit classes using a conversion grid[24]; the other two (*F* and *P*) are directly expressed in the four-class format. Assignment of numerical scores to the four merit classes follows the VQR rules, namely 1 for class A, 0.8 for class B, 0.5 for class C, and 0 for class D.

To compare informed peer review and bibliometric analysis, we can compare the *F* and *P* evaluations. Other comparisons could also be informative. In particular, comparison between P1 and P2 allows us to study the degree of agreement between the referees.

6.1. The *F* and *P* distribution

Table 11 presents the distribution of the *F* and *P* indicators, while Table 12 presents the distribution of P1 and P2. The elements on the main diagonal in Table 11 correspond to cases where informed peer review and bibliometric evaluations coincide. The off-diagonal elements correspond to cases of disagreement between the two evaluations, either because *F* provides a higher evaluation (elements above the main diagonal) or because *P* provides a higher evaluation (elements below the main diagonal).

Table 11
Comparison between *F* and *P*.

Bibliometric (*F*)	Peer (*P*)				
	A	B	C	D	Total
A	98 49.49	72 36.36	19 9.60	9 4.55	198 100.00
B	11 10.78	56 54.90	26 25.49	9 8.82	102 100.00
C	4 3.88	25 24.27	39 37.86	35 33.98	103 100.00
D	3 1.60	21 11.23	45 24.06	118 63.10	187 100.00
Total	116 19.66	174 29.49	129 21.86	171 28.98	590 100.00

Note: The table tabulates the distribution of the journal articles in the sample by informed peer review and bibliometric evaluations, expressed through the merit classes. The elements on the main diagonal correspond to cases for which informed peer review and bibliometric evaluation coincide. The off-diagonal elements correspond to cases of disagreement between informed peer review and bibliometric evaluation.

Table 12
Comparison between P1 and P2.

Peer no. 1	Peer no. 2				
	A	B	C	D	Total
A	53 46.49	43 37.72	7 6.14	11 9.65	114 100.00
B	36 21.56	73 43.71	29 17.37	29 17.37	167 100.00
C	8 8.70	34 36.96	21 22.83	29 31.52	92 100.00
D	4 1.84	46 21.20	50 23.04	117 53.92	217 100.00
Total	101 17.12	196 33.22	107 18.14	186 31.53	590 100.00

Note: The table tabulates the evaluations of the two external referees, expressed through the merit classes. The elements on the main diagonal correspond to cases for which informed peer reviewers agree on the evaluation. The off-diagonal elements correspond to cases of disagreement between the two informed peer reviewers. Note that labeling the two evaluations by the two informed peer reviewers as Peer no. 1 and Peer no. 2 is purely a convention, reflecting only the order in which the referees accepted to review the paper.

Table 11 shows that the main source of disagreement between *F* and *P* is that informed peer review classifies as "A" only 116 (58.6%) of the 198 papers classified as "A" by bibliometric analysis.[25] Table 11 shows also that informed peer review classifies as "B" a larger number of papers (174 papers) than bibliometric analysis (102 papers). On the other hand, the assignment of papers to the "C" and "D" classes is similar for the two methods. Overall, bibliometric analysis (*F*) and informed peer review (*P*) give the same classifications in 53% of the cases (311 cases are on the main diagonal of Table 11), and in 89% of the cases differ by at most one class. Extreme disagreement (difference of 3 classes) occurs in only 2% of the cases, and a milder disagreement (difference of 2 classes) in only 9% of the cases.

Table 12 cross-tabulates the evaluations of the two external referees. In 45% of the cases they agree on the same evaluation, and in 82% of the cases their evaluation differs by at most one class. Note that referees agree on an "A" evaluation in about half of the cases. It is interesting also to compare *F* and *P* evaluations by sub-area. Disagreement by more than one class occurs in 19% of the cases for

[20] The VQR defined a set of rules for allocating papers to panelists and referees in order to avoid conflicts of interest with authors and authors' affiliations. Referee independence was ensured by paying attention to research collaborations and, where possible, to nationality.

[21] Referees were selected according to standards of scientific quality, impact in the international scientific community, experience in evaluation, expertise in their respective areas of evaluation, and considering their best three publications in terms of *h*-index. Half of the referees were affiliated to non-Italian institutions.

[22] The evaluation form is available in Appendix B.

[23] Labeling the two referees as "P1" or "P2" is purely a convention that only reflects the order in which the referees accepted to review the paper.

[24] The conversion grid is as follows: 23–27: Excellent (A); 18–22: Good (B); 15–17: Acceptable (C): 3–14: Limited (D).

[25] One possible reason is that, according to the VQR rules, class A includes about 20% of the journals, which is likely to be greater than the fraction of journals a typical referee would consider as "top journals".

Table 13
Kappa statistic for the amount of agreement between *F* and *P* scores.

	Total sample	Economics	History	Management	Statistics
	(1)	(2)	(3)	(4)	(5)
F and *P*, linear weight kappa	0.54 (18.11)**	0.56 (11.94)**	0.32 (2.95)**	0.49 (8.91)**	0.55 (9.41)**
F and *P*, VQR weighted kappa	0.54 (17.29)**	0.56 (11.53)**	0.29 (2.56)**	0.50 (8.37)**	0.55 (9.18)**
P1 and P2, equal weights	0.40 (12.93)**	0.44 (9.06)**	0.18 (1.49)	0.33 (5.90)**	0.33 (5.47)**
P1 and P2, VQR weights	0.39 (12.06)**	0.42 (8.28)**	0.15 (1.29)	0.33 (5.55)**	0.32 (5.17)**

Note: The table reports the kappa statistic and the associated *z*-value in parenthesis for the total sample and by research sub-area.
* Indicates significance at the 5% level.
** Indicates significance at the 1% level.

History, but only in 10% of the cases for the other three sub-areas. The lower frequency of "A" and the higher frequency of "B" in the informed peer review, compared to the bibliometric analysis, occur for all sub-areas except History, where 10 papers are classified as "A" by the informed peer review and 9 by the bibliometric analysis. In this case, however, the sample is small (only 37 observations), so cell-by-cell comparison might not be reliable.[26]

6.2. *Comparison between F and P*

When comparing informed peer review and bibliometric analysis, two criteria may be considered. The first is the degree of agreement between *F* and *P*, that is, whether *F* and *P* tend to agree on the same score. The second is the presence of systematic difference between *F* and *P*, measured by the average score difference between *F* and *P*.

Of course, perfect agreement would imply no systematic difference, but the reverse is not true and, in general, these two criteria highlight somewhat different aspects. Consider for instance a distribution with a high level of disagreement between *F* and *P* (for many papers the *F* and *P* evaluations are different). It could still be that, on average, *F* and *P* provide a similar evaluation. This distribution has low agreement and low systematic differences. Adopting one of the two evaluations (for instance, *F*) would result in frequent misclassification of papers according to the other criterion (e.g., many papers with good *F* but poor *P* evaluations, and vice versa).

Alternatively, consider a case of close (but not perfect) agreement between *F* and *P*. It could still be that, for instance, *F* assigns a higher class more often than *P*. This distribution has high agreement but large systematic differences, as the average *F* score differs from the average *P* score in a systematic way. Adopting one of the two evaluations would result in over-evaluation (or under-evaluation) compared to the other criterion; that is, on average papers receive a higher (or a lower) score using the *F* or *P* evaluations.

From a statistical point of view, the level of agreement between *F* and *P* can be measured using Cohen's kappa[27], while systematic differences between sample means can be detected using a standard *t*-test for paired samples.[28]

6.3. *Degree of agreement*

Table 13 reports the kappa statistic for the entire sample and by sub-area. The kappa statistic is scaled to be zero when the level of agreement is what one would expect to observe by pure chance, and to be one when there is perfect agreement. The statistic is computed using standard linear weights (1, 0.67, 0.33, 0) to take into account that cases of mild disagreement (say, disagreement between "A" and "B") should receive less weight than cases of stronger disagreement (say, disagreement between "A" and "C", or between "A" and "D").

Overall, kappa is equal to 0.54 and statistically different from zero at the 1% level. For Economics, Management and Statistics, the value of kappa is close to the overall value for the sample, while History has a lower kappa value (0.32). For each sub-area, kappa is statistically different from zero at the 1% level.

As already mentioned, the computation of kappa in the first row of Table 13 uses linear weights. One may argue that, in the present context, the appropriate weights are the VQR weights based on the numerical scores associated with the qualitative evaluations (1 for A, 0.8 for B, 0.5 for C, and 0 for D). The second row in Table 13 reports the "VQR weighted" kappa. The resulting statistic is quite similar to the linearly weighted kappa, indicating fair to good agreement for the total sample (0.54) and for Economics, Management and Statistics, and poor agreement for History (0.29).[29]

The degree of agreement between bibliometric ranking (*F*) and informed peer review (*P*) is actually higher than between the two external referees (P1 and P2). This is shown in Table 13 which reports the kappa statistics for the degree of agreement between the two referees (P1 and P2) in the total sample and by sub-area. In the total sample, the linearly weighted kappa is equal to 0.40 (0.39 using VQR weights) and is lower than the corresponding kappa for the comparison of *F* and *P* (0.54 for both the linear and the VQR weights). For each sub-area, the pattern is similar to that observed when comparing *F* and *P*. For Economics, Management and Statistics there is more agreement between the referees than for History (for this sub-area, kappa is not statistically different from zero).

[26] Results are available from the authors upon request.

[27] Cohen's kappa is defined as $\kappa = [Pr(a) - Pr(e)]/[1 - Pr(e)]$, where $Pr(a)$ is the relative observed agreement among referees, and $Pr(e)$ is the probability of random agreement. If the referees are in complete agreement then $\kappa = 1$. If there is no agreement among the referees other than what would be expected by chance (as defined by $Pr(e)$), $\kappa = 0$. The test statistics $Z = \hat{\kappa}/\hat{s}$, where $\hat{s}$is the standard error of kappa, is assumed to be distributed $N(0,1)$.

[28] Even if the underlying variable is ordinal, both the sample mean and the difference between sample means are asymptotically normal. Given our sample size, this justifies using a *t*-test to perform group comparisons.

[29] Landis and Koch (1977) characterize the range of values 0–0.20 as "slight agreement", 0.21–0.40 as "fair agreement", 0.41–0.60 as "moderate agreement", 0.61–0.80 as "substantial agreement", and 0.81–1 as "almost perfect agreement". These guidelines are somewhat arbitrary and by no means universally accepted. Fleiss (1981) for instance characterizes kappas over 0.75 as "excellent", 0.40 to 0.75 as "fair to good", and below 0.40 as "poor". Kappa has also been shown to increase with the number of classes (only 4 in our case). Since the most common scales to subjectively assess the value of kappa mention "adequate" and "fair to good", these are the terms that we use in the paper to convey the meaning of the statistic when commenting the estimated kappas.

Table 14
Test for the difference between average *F* and *P* scores.

	Score P1	Score P2	Score *P*	Score *F*	Difference between *F* and *P*	Sample size	*t*-Test for difference between *F* and *P*	*p*-Value
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Economics	0.503	0.521	0.561	0.607	0.046	235	2.286	0.023
History	0.649	0.700	0.705	0.597	−0.108	37	−1.672	0.103
Management	0.335	0.421	0.386	0.441	0.054	175	1.999	0.047
Statistics	0.649	0.625	0.658	0.624	−0.034	143	−1.417	0.159
Total	0.498	0.528	0.542	0.561	0.019	590	1.417	0.157

Note: The table reports the average scores of the two referees (Score P1 and Score P2), the score resulting from the final evaluation by the consensus group (Score *P*) and the score of the bibliometric evaluation (Score *F*). The *F* and *P* scores are obtained by converting the four merit classes to numerical scores using the values established by the VQR rules: A = 1; B = 0.8; C = 0.5; D = 0. The *t*-test is computed for paired samples.

Furthermore, for each sub-area there is more agreement between *F* and *P* than between P1 and P2.

6.4. Systematic differences

Table 14 reports the average scores resulting from the *F* and *P* evaluations. Numerical scores are obtained converting the qualitative *F* and *P* evaluations using the weights assigned by the VQR to the four merit classes. Note again that, given the rules of the VQR, deviations between *F* and *P* do not carry the same weight: for instance, a difference between “D” and “C” has a weight of 0.5, while a difference between “A” and “B” has a weight of only 0.2.

Table 14 also reports the average numerical scores of the two referees (columns labeled “Score P1” and “Score P2”). Column (3) reports the average score of the informed peer review (“Score *P*”), which is equal to 0.542. The score is lower for Management (0.386) and higher for History (0.705) and Statistics (0.658). The difference across sub-areas in column (3) may be due to several reasons, including sampling variability, higher quality of the pool of papers in History and Statistics, or more generous referees compared to other sub-areas.

Column (4) shows the average score of the bibliometric evaluation, which is equal to 0.561. Similar to the *P* score, the *F* score tends to be lower for Management (0.441) and slightly higher for Statistics (0.624). Column (5) shows the difference between *F* and *P* scores, while column (7) shows the associated paired *t*-statistic. Overall, the difference is positive (0.019) and not statistically different from zero at conventional levels (the *p*-value is 0.157). However, there are differences across sub-areas. For Economics and Management, the difference is positive (0.046 and 0.054, respectively) and statistically different from zero at the 5% level (but not the 1% level). For Statistics and History, the difference is instead negative (−0.108 and −0.034, respectively) but not statistically different from zero.

6.5. Robustness checks

As noted in Section 2, we compare *P* and *F* following the decisions made by the panel to use the maximum between AIS and IF5 to rank journals. Since the *h*-index is available for all journals, an alternative is to use it directly in the bibliometric evaluation. Thus, we may classify as “A” all journals falling in the top 20% of the *h*-index distribution, as “B” those falling in the next 20%, as “C” those falling in the next 10%, and as “D” those falling below the median value of the *h*-index. The advantage of this procedure is that no imputation is required. On the other hand, in Google Scholar measurement error is more pervasive.

Table 15 reports Cohen’s kappa (using linear weights and VQR-weights) between informed peer review *P* and a bibliometric evaluation based only on the *h*-index, *F*(*h*). Results for the total sample indicate the same agreement (0.54) between the *F* and *P* distributions and between the *F*(*h*) and *P* distributions using VQR weights, and agreement of 0.52 using linear weights. Results by sub-area are also quite similar.

Table 16 reports, for the total sample and for each sub-area, tests for the differences between *P* and *F*(*h*). The overall difference is 0.005, to be compared with the 0.019 difference between *F* and *P* in Table 14, which shows that the *F* evaluation is slightly more generous than the *F*(*h*) evaluation. By sub-areas, the pattern of the differences between *F*(*h*) and *P* is similar to that of the *F* and *P* differences in Table 14 (positive for Economics and Management, and negative for History and Statistics). In terms of significance, the differences between *F*(*h*) and *P* are statistically significant only for History.

6.6. Informed peer review

As already stressed, the VQR relies on informed peer review. This is because referees know not only the identity of the authors of an article, but also its final publication outlet, together with the bibliometric indicators associated with the journal.[30] This raises the question of whether informed peer review and bibliometric analysis are independent. That is, to what extent does the perceived quality of a journal, which may in turn be based on bibliometric indicators, affect a referee’s evaluation? In other words, is the opinion about the quality of a paper disconnected from the opinion about the quality of the journal that published the paper? If preliminary knowledge of bibliometric data matters a lot for a referee’s evaluation, then our comparative analysis would not be meaningful. However, the aim of our research is not to isolate the two components, but to discover whether the two approaches yield similar results regardless of whether the correlation stems from independent assessments or because the community of reviewers trusts bibliometric information. As noted in the Introduction, this is an important caveat. In particular, our analysis is not meant to produce an intrinsic comparison of peer review and bibliometrics, or to establish the intrinsic validity of the latter against the benchmark of peer review.

As a matter of fact, to check whether the perceived quality of a journal carries a disproportionate weight in the evaluations, we employ additional background information about the refereeing process. So far, we have analyzed the VQR process as it was actually carried out. At this stage, we introduce additional information extracted from the VQR database in order to shed light on this specific issue. The referee evaluation form used by the VQR includes three questions about the originality, relevance and internationalization of a paper. The form is available in Appendix B. While the

[30] The referees were provided with the panel journal classification list and the actual or imputed values of IF, IF5 and AIS. See Sgroi and Oswald (2013) for a discussion of the combined use of bibliometric indicators and peer review within the context of the UK Research Excellence Framework 2014.

Table 15
Kappa statistic for the amount of agreement between $F(h)$ and P scores.

	Total sample	Economics	History	Management	Statistics
	(1)	(2)	(3)	(4)	(5)
$F(h)$ and P, linear weight kappa	0.52 (17.10)**	0.54 (11.34)**	0.21 (1.94)*	0.53 (9.42)**	0.46 (7.89)**
$F(h)$ and P, VQR weighted kappa	0.54 (17.02)**	0.56 (11.12)**	0.20 (1.83)*	0.55 (9.22)**	0.50 (8.31)**

Note. The table reports the kappa statistic and the associated z-value in parenthesis for the total sample and by research sub-area.
* Indicates significance at the 5% level.
** Indicates significance at the 1% level.

Table 16
Test for the difference between average $F(h)$ and P scores.

	Score P	Score $F(h)$	Difference between $F(h)$ and P	Sample size	t-Test for difference between $F(h)$ and P	p-Value
Economics	0.561	0.597	0.035	235	1.624	0.106
History	0.705	0.511	−0.195	37	−2.692	0.011
Management	0.386	0.401	0.014	175	0.543	0.588
Statistics	0.658	0.652	−0.006	143	−0.217	0.828
Total	0.542	0.547	0.005	590	0.336	0.737

Note. The table reports the average scores resulting from the final evaluation by the Consensus Group (Score P) and the score of the bibliometric evaluation based on the h-index (Score $F(h)$). The P and $F(h)$ scores are obtained by converting the four merit classes to numerical scores using the values established by the VQR rules: A = 1; B = 0.8; C = 0.5; D = 0. The t-test is computed for paired samples.

Table 17
Pearson's correlation coefficients of reviewers' questions.

	Originality	Relevance	Internationalization	Overall score
Originality	1.00			
Relevance	0.87	1.00		
Internationalization	0.82	0.85	1.00	
Overall score	0.94	0.95	0.95	1.00

Note: The table reports the correlation matrix of the overall score assigned by the referees with the score assigned to each of the three questions (relevance, originality or innovation, and internationalization or international standing).

first two questions refer directly to the paper's quality, the third explicitly refers to its international reach and potential for future citations. Therefore, the responses to this third question are more likely to be influenced by the referee's assessment based on journal rankings. However, the correlation coefficients reported in Table 17 show that the three dimensions along which referees were asked to rank papers tend to be highly correlated.[31] This suggests that the reviewers were likely influenced by their knowledge of the publication outlet, and particularly by the bibliometric indices of the journals. However, their perceptions are also highly correlated with other indicators of the quality of the paper and are not the leading factor in the overall informed peer review assessment.

Further insights about the correlation between informed peer review and bibliometric evaluation can be gained comparing the correlation between bibliometric evaluation and informed peer review using the approach of Allen et al. (2009) and Eyre-Walker and Stoletzki (2013).[32] In our dataset the simple correlation coefficient between the scores of the two referees is 0.36. The correlation between the sum of the two scores (as an overall indicator of informed peer review not filtered by the consensus groups) and the bibliometric indicators is lower for IF5 and h-index (0.24 and 0.13, respectively), but higher for AIS (0.39). Thus, we cannot rule out that the evaluation of the two referees was affected by bibliometric indicators (in particular, by AIS).

Digging deeper in this issue, we compute the partial correlations between the scores of the two referees after removing the effect of all three bibliometric indicators (IF5, AIS, and h-index). Confirming our previous results, we find that the partial correlation between the scores of the two referees is 0.27 (down from the simple correlation of 0.36), indicating that the referees' evaluation is indeed likely to be affected by knowledge of the journal where the paper was published. These results support Eyre-Walker and Stoletzki (2013) findings in a rather different context, discipline and sample design, suggesting that post-publication research

[31] Since the referees originally provided a numerical score (from 1 to 9) for each of the three questions, Table 17 computes the correlation matrix of the overall score with the question-specific scores. The matrix shows that the correlations are quite high. In particular, the correlations between the overall and the question-specific scores are 94% for originality, 95% for relevance and 95% for internationalization. Further, the correlations between the three question-specific scores are also quite high (between 82% and 87%). Although we cannot replicate the overall score assigned by peer review (which includes the opinion of the two referees weighted by the Consensus Group) it is very likely that the outcome of the peer review process would have been quite similar had we excluded the third question from the referee's evaluation form.

[32] Allen et al. (2009) use data from biomedical papers to compare post-publication experts' assessment and their bibliometric measures (IF and citations) collected three years after the publication. They find a positive and significant correlation between experts' opinion and subsequent performance of the papers according to IF (0.625) and citations (0.445). Eyre-Walker and Stoletzki (2013) use the same dataset and investigate whether the correlation between the two assessors' score is influenced by the journal in which the paper was published. They also conclude that assessors' scores are influenced by the journal's impact factor.

evaluations based on bibliometric evaluations will tend to be correlated with informed peer review.

7. Conclusions

This article contributes to the debate on bibliometric and informed peer review evaluation in two ways. First, it proposes a method for using bibliometric analysis in an area characterized by partial coverage of bibliometric indicators for journals. Second, it compares the results of two different evaluation methods – bibliometric and informed peer review – using a random sample of journal articles assessed using both. This comparison represents an important contribution to the literature.

Our results reveal that, in the total sample, there is fair to good agreement between bibliometric and informed peer review. Furthermore, there is no evidence of systematic differences between the average scores provided by the two evaluations, although peer review assigns a lower number of papers to the top class relative to bibliometric analysis. However, most of the papers "downgraded" by the informed peer review are still assigned to the class immediately below the top, and deviations from the two upper classes do not carry a large weight in the VQR.

If we compare the sub-areas we analyzed (i.e., Economics, Management, Statistics, and History), the degree of agreement is lower for History. Systematic differences between the average scores for the four sub-areas are generally small and not always of the same sign: they are positive and statistically significant at the 5% level for Economics and Management, and negative but not statistically different from zero for Statistics and History.

Our results have important implications for the organization of large-scale research assessment exercises, like those that are becoming increasingly popular in many countries. First and foremost, they suggest that the agencies that run these evaluations could feel confident about using bibliometric evaluations and interpret the results as highly correlated with what they would obtain if they performed informed peer review, at least in the disciplines studied in this paper and for research output published in ranked journal articles. Of course, we are suggesting neither that bibliometric evaluations should replace peer review nor that national assessment exercises could now be performed through less costly bibliometric evaluation. However, our results entail that bibliometric evaluation could be used to monitor the research output of a nation or a community on a more frequent basis than national research assessments, which involve huge amounts of time and effort to organize and therefore take place only every few years. Bibliometric evaluation can be organized more flexibly and at less cost than large-scale informed peer review evaluation, so it could be employed between national evaluations to allow more frequent monitoring of the dynamics of research outcomes. It could also be used by individual departments or institutions as a managerial tool to monitor their outcomes knowing that bibliometric assessments can be good predictors of the performance of the department or institution in the future large-scale informed peer evaluations.

We also recommend that formal evaluation exercises still include a sizeable share of articles assessed by informed peer review. Apart from preserving the richness of both methods, the agencies could run experiments similar to ours by allocating some research papers to both informed peer review and bibliometric evaluation. This would further contribute to testing the similarities between them.

This paper inevitably has some limitations. First and foremost, as noted repeatedly in this paper, the influence exerted on the reviewers by the information on the publication outlet implies that, in our study, assessment by bibliometric analysis and peer review are not independent. As a result, we can say nothing about the correlation between the primary factors that link them. The goal of the VQR exercise is to evaluate published work of Italian academics between 2004 and 2010, and therefore this limitation is imposed on us by the very structure and goals of the VQR exercise. Future VQR waves could compare bibliometric analysis and peer review using anonymized published material so that neither the publication outlet nor the name of the authors are revealed to the reviewers. The researchers could then identify the correlation produced by the two independent evaluations, after removing the impact of the information about the publication source. To be sure, even if properly anonymized, this exercise will not be straightforward, as reviewers can figure out the identity of the published paper and authors. Moreover, anonymizing papers requires pre-publication texts that may only be provided by the authors or the journals.

Second, it is difficult to generalize our results to other disciplines. Even within our sub-areas, we found important differences between the two approaches to research evaluation. These differences could arise for a number of reasons. First, there may be differences in refereeing style across subject areas: for example, referees may be less generous in some areas than in others. Second, the reliability of journal ranking may differ across areas: for example, the ranking of journals may be more generous (e.g., placing more journals in the top class) in Economics and Management relative to other sub-areas. Finally, the available sample size may limit the power of statistical tests, as in the case of History. Future research could focus on the analysis of these differences and possibly control better for heterogeneity using larger sample size.

Despite these caveats, we believe that the Italian research assessment exercise offers the unusual opportunity of employing a very rich set of data to evaluate the relationships between bibliometric analysis and informed peer review. As national or large-scale research assessments gain momentum, a better understanding of the relationships between the two approaches should help provide more efficient evaluations. We hope that future work will uncover other aspects of these processes and address some of the limitations in the present study.

Appendix A. Detailed description of the MIM

The imputation methodology that we use is the fully conditional specification method (FCS) of van Buuren et al. (2006, henceforth BBGR), and the exposition from this point on follows closely theirs.[33]

Let $Y = (Y_1, Y_2, \ldots Y_k)$ be an $n \times K$ matrix of K variables (all potentially containing missing values) for a sample of size n. In our case $K = 3$, as we are imputing the logarithms of IF, IF5 and AIS. Y has a multivariate distribution characterized by a parameter vector θ, denoted by $P(Y; \theta)$. The objective of the imputation procedure is to generate imputed values for the missing part of Y (denoted by Y_{mis}) that, combined with the non-missing part Y_{obs}, will reconstitute as closely as possible the joint distribution $P(Y; \theta)$.

One way to proceed would be to assume a fully parametric multivariate density for Y, and starting with some priors about θ to generate imputations of Y_{mis} conditional on Y_{obs} (and on any other vector of variables X that are never missing, like the h-index in our case).

An alternative to specifying a joint multivariate density is to predict any given variable in Y, say Y_k, conditional on all remaining variables in the system (denoted by Y_{-k}) and a parameter vector θ_k. We apply this procedure to all K variables in Y in a sequential manner, and after the last variable in the sequence has been imputed then a single iteration of this process is considered to be

[33] The exposition in this Appendix is based on Christelis (2011).

completed. In this way, the K-dimensional problem of restoring the joint density of Y is broken into K one-dimensional problems of conditional prediction. This has two principal advantages over the joint approach. First, it can readily accommodate many different kinds of variables in Y (e.g., binary, categorical, and continuous). This heterogeneity would be very difficult to model with theoretical coherence using a joint distribution of Y. Second, it easily allows the imposition of various constraints on each variable (e.g., censoring), as well as constraints across variables.

The principal drawback of this method is that there is no guarantee that the K one-dimensional prediction problems lead to convergence to the joint density of Y. Because of this potential problem, BBGR ran a number of simulation tests, often complicated by conditions that made imputation difficult, and found that the FCS method performed very well. Importantly, it generated estimates that were generally unbiased, and also good coverage of the nominal confidence intervals.

As the parameter vector θ of the joint distribution of Y is replaced by the K different parameter vectors θ_k of the K conditional specifications, BBGR propose to generate the posterior distribution of θ by using a Gibbs sampler with data augmentation.

Let us suppose that our imputation process has reached iteration t, and that we want to impute variable Y_k. We first estimate a statistical model[34] with Y_k as the dependent variable (using only its observed values) and the variables in Y_{-k} as predictors. For every element of Y_{-k} that precedes Y_k in the sequence of variables, its values from iteration t are used (i.e., including the imputed ones). On the other hand, for every element of Y_{-k} that follows Y_k in the sequence, its values from iteration t-1 are used. After obtaining the parameter vector θ_k from our estimation, we make a draw $\theta_k{}^*$ from its posterior distribution[35], i.e., we have

$$\theta_k^{*(t)} \sim P\left(\theta_k | Y_1^{(t)}, \quad \ldots, Y_{k-1}^{(t)}, \quad Y_{k,obs}, \quad Y_{k+1}^{(t-1)}, \quad Y_k^{(t-1)}\right) \tag{A.1}$$

The fact that only the observed values of Y_k are used in the estimation constitutes, as BBGR point out, a deviation from most Markov Chain Monte Carlo implementations, and it implies that the estimation sample used for the imputation of any given variable will include only the observations with non-missing values for that variable.

Having obtained the parameter draw $\theta_k^{*(t)}$ at iteration t we can use it, together with $Y_{-k}^{(t)}$ and the observed values of Y_k, to make a draw from the conditional distribution of the missing values of Y_k. That is, we have

$$Y_k^{*(t)} \sim P\left(Y_{k,\text{miss}} | Y_1^{(t)}, \quad \ldots, \quad Y_{k-1}^{(t)}, \quad Y_{k,obs}, \quad Y_{k+1}^{(t-1)}, \quad Y_k^{(t-1)}; \quad \theta_k^{*(t)}\right) \tag{A.2}$$

As an example, let us assume that Y_k represents the logarithm of the value of a particular bibliometric indicator, and that we want to impute its missing values at iteration t via ordinary least squares, using the variables in $Y_{-k}^{(t)}$ as predictors. We perform the initial estimation, and obtain the parameter vector $\theta_k^{(t)} = \left(\beta_k^{(t)}, \sigma_k^{(t)}\right)$, with $\beta_k^{(t)}$ denoting the regression coefficients of $Y_{-k}^{(t)}$, and $\sigma_k^{(t)}$ the standard deviation of the error term. After redrawing the parameter vector $\theta_k^{*(t)}$ using (1), we first form a new prediction that is equal to $Y_{-k}^{(t)}\beta_k^{*(t)}$. Then, the imputed value $Y_{k,i}^{*(t)}$ for a particular observation i will be equal to $Y_{-k,i}^{(t)}\beta_k^{*(t)}$ plus a draw of the error term (assumed to be normally distributed with a standard deviation equal to $\sigma_k^{*(t)}$).[36] The error draw for each observation with a missing value for Y_k is made in such a way as to observe any bounds that have been already placed on the admissible values of Y_k for that particular observation. These bounds can have many sources, e.g., overall minima or maxima imposed for the particular variable.

The process described in (A.1) and (A.2) is applied sequentially to all K variables in Y, and after the imputation of the last variable in the sequence (i.e., Y_k) iteration t is considered complete. We thus end up with an example of a Gibbs sampler with data augmentation (Tanner and Wong, 1987) that produces the sequence $\left\{\left(\theta_1^{(t)}, \ldots, \theta_k^{(t)}, Y_{\text{mis}}^{(t)}\right) : \quad t = 1, 2, \ldots\right\}$. The stationary distribution of this sequence is $P\left(Y_{\text{mis}}, Y_{\text{obs}}; \quad \theta\right)$, provided that convergence of the imputation process is achieved.

As pointed out by Schafer (1997), a sufficient condition for the convergence to the stationary distribution is the convergence of the sequence $\left\{\theta_1^{(t)}, \ldots, \theta_k^{(t)}\right\}$ to the conditional distribution of the parameter vector $P\left(\theta | Y_{\text{obs}}\right)$ or, equivalently, the convergence of the sequence $\left\{Y_{\text{miss}}^{(t)}\right\}$ to the conditional distribution of the missing values $P(Y_{\text{mis}}|Y_{\text{obs}})$. Hence, in order to achieve convergence to the stationary distribution of Y, we iterate the Gibbs sampler till we have a number of iterations indicating convergence of the distributions of the missing values of all the variables in our system.

One important feature of the FCS method (shared with several other similar approaches found in the imputation literature)[37] is that it operates under the assumption that the missingness of each variable in Y depends only on other variables in the system and not on the values of the variable itself. This assumption, commonly known as the missing at random (MAR) assumption, is made in the vast majority of imputation procedures applied to micro datasets. It could be argued, however, that it is unlikely to hold for all variables: for example, missingness in AIS could depend on whether the journal might have a high or low citation count and thus high or low potential AIS. This would be a case of data missing not at random (MNAR) and, if true, would present major challenges for the construction of the imputation model.

Some evidence on the consequences of the violation of the MAR assumption comes from the results of one of the simulations run by BBGR, which exhibits a NMAR pattern. In addition, BBGR use in this simulation conditional models that are not compatible with a single joint distribution. Even in this rather pathological case, however, the FCS method performs reasonably well, and leads to less biased estimates than an analysis that uses only observations without any missing data. As a result, BBGR conclude that the FCS method (combined with multiple imputation) is a reasonably robust procedure, and that the worry about the incompatibility of the conditional specifications with a joint distribution might be overstated.

One further issue to be addressed is how to start the iteration process given that, as described above, in any given iteration one needs to use imputed values from the previous iteration. In other words, one needs to generate an initial iteration, which will constitute an initial condition that will provide the lagged imputed values to the first iteration. This initial iteration is generated by imputing the first variable in the system based only on variables that

[34] In our case, the statistical model is always linear, but in other cases nonlinear models can be used (e.g., probit, multinomial logit) depending on the nature of Y_k.

[35] The formulas used for redrawing the parameter vector can be found in Appendix A of BBGR.

[36] As already discussed in the text, the estimation of all models of amounts is done in logarithms in order to make our conditional specifications more compatible with the maintained assumption of normality.

[37] A similar imputation procedure is proposed by Lepkowski et al. (2001). See also BBGR for references to a number of other approaches that have significant similarities to theirs.

are never missing (namely the logarithm of the *h*-index and the English language indicator), then the second variable based on the first variable (including its imputed values) and the non-missing variables, and so on, till we have a complete set of values for this initial condition. Having obtained this initial set of fully imputed values, we can then start the imputation process using the already described procedures, as denoted in Eqs. (A.1) and (A.2).

Once we have obtained the imputed values from the last iteration, we end up with five hundred imputed values for each missing one, i.e., with five hundred different complete datasets that differ from one another only with respect to the imputed values. We then need to consider how to use the five hundred implicate datasets in order to obtain estimates for any magnitude of interest (e.g., descriptive statistics or coefficients of a statistical model).

Let $m = 1, \ldots, M$ index the implicate datasets (with M in our case equal to 500) and let $\hat{\beta}_m$ be our estimate of the magnitude of interest from the mth implicate dataset. Then the overall estimate derived using all M implicate datasets is just the average of the M separate estimates, i.e.:

$$\bar{\hat{\beta}} = \frac{1}{M}\sum_{m=1}^{M}\hat{\beta}_m \tag{A.3}$$

The variance of this estimate consists of two parts. Let V_m be the variance of $\hat{\beta}_m$ estimated from the mth implicate dataset. Then the within-imputation variance WV is equal to the average of the M variances, i.e.:

$$WV = \frac{1}{M}\sum_{m=1}^{M}V_m \tag{A.4}$$

One would like each implicate run to explore as much as possible the domain of the joint distribution of the variables in your system; indeed, the possibility of the Markov Chain Monte Carlo process defined in (A.1) and (A.2) to jump to any part of this domain is one of the preconditions for its convergence to a joint distribution. This would imply an increased within variance, other things being equal.

The second magnitude one needs to compute is the between-imputation variance BV, which is given by

$$BV = \frac{1}{M-1}\sum_{m=1}^{M}\left(\hat{\beta}_m - \bar{\hat{\beta}}\right)^2 \tag{A.5}$$

The between variance is an indicator of the extent to which the different implicate datasets occupy different parts of the domain of the joint distribution of the variables in our system. One would like the implicate runs to not stay far apart but rather mix with one another, thus indicating convergence to the same joint distribution. Therefore, one would like the between variance to be as small as possible relative to the within one.

The total variance TV of our estimate $\bar{\hat{\beta}}_m$ is equal to:

$$TV = WV + \frac{M+1}{M}BV \tag{A.6}$$

As pointed out by Little and Rubin (2002), the second term in (A.6) indicates the share of the total variance due to missing values. Having computed the total variance, one can perform a *t*-test of significance using the following formula to compute the degrees of freedom:

$$df = (M-1)\left(1 + \frac{1}{M+1}\frac{WV}{BV}\right)^2 \tag{A.7}$$

Appendix B. The referee evaluation form

ANVUR—Assessment of the research quality 2004–2010: Assessment form (one form to be filled for each research product): In the following research output or work means: journal article, book chapter, monograph, conference proceeding. For each of the 3 criteria (relevance, originality/innovativeness, international reach/impact) a non exhaustive list of questions is provided to clarify its meaning.

Q1. *Relevance.* Are the research questions addressed by the work of general, narrow or limited interest? Are they likely to spur additional work? Are the methods, the data or the results likely to be used by other researchers?

Please grade the research output in terms of its relevance, expressing a score between 1 and 9, with **1 and 9 indicating minimal and maximal relevance**, respectively.

1	2	3	4	5	6	7	8	9

Q2. *Originality/innovativeness.* Does the work advance knowledge in some dimension? Does it pose new questions, provide new answers, use new data or methods?

Please grade the research output in terms of its originality, expressing a score between 1 and 9, with **1 and 9 indicating minimal and maximal originality/innovativeness**, respectively.

1	2	3	4	5	6	7	8	9

Q3. *International reach/impact:* Was the work able to reach an international audience, or does it have the potential to do so? Was it cited, quoted or reviewed by other researchers, or do you expect it will be in the future? Is it likely to leave a mark in the international scientific community? Did the work consider the relevant international contributions on the same or related issues?

Please grade the research output in terms of its international reach and impact, expressing a score between **1 and 9, with 1 and 9 indicating minimal and maximal international reach/impact**, respectively.

1	2	3	4	5	6	7	8	9

Q4. Optional (max. 1000 char.) Free format explanations of the grades:

Relevance:

Originality/Innovativeness:

International reach/Impact:

References

Allen, L., Jones, C., Dolby, K., Lynn, D., Walport, M., 2009. Looking for landmarks: the role of expert review and bibliometric analysis in evaluating scientific publication outputs. PLoS One 4, e5910.

Bartolucci, F., Dardanoni, V., Peracchi, F., 2013. Ranking scientific journals via latent class models for polytomous item response data. In: EIEF Working Paper No. 13/13, Available at ⟨http://www.eief.it/faculty-visitors/faculty-a-z/franco-peracchi/⟩.

Bornmann, L., 2011. Scientific peer review. Annual Review of Information Science and Technology 45, 199–245.

Bornmann, L., Daniel, H.-D., 2008. What do citation counts measure? A review of studies on citing behavior. Journal of Documentation 64, 45–80.

Burger, M., Frankfort, J.G., van Raan, A.F.J., 1985. The use of bibliometric data for the measurement of university research performance. Research Policy 14, 131–149.

Butler, L., 2007. Assessing university research: a plea for a balanced approach. Science and Public Policy 34, 565–574.

Christelis, D., 2011. Imputation of missing data in Waves 1 and 2 of SHARE. In: CSEF Working Paper No. 278, Available at ⟨http://www.csef.it/WP/wp278.pdf⟩.

Eyre-Walker, A., Stoletzki, N., 2013. The assessment of science: the relative merits of post-publication review, the Impact Factor, and the number of citations. PLoS Biology 11, e1001675.

Fagerberg, J., Landström, H., Martin, B.R., 2012. Exploring the emerging knowledge base of "the knowledge society". Research Policy 41, 1121–1131.

Fleiss, J.L., 1981. Statistical Methods for Rates and Proportions, second ed. John Wiley, New York, NY.

Geuna, A., Martin, B.E., 2003. University research evaluation and funding: an international comparison. Minerva 41, 277–304.

Harzing, A.-W., van der Wal, R., 2008. Comparing the Google Scholar *h*-index with the ISI Journal Impact Factor, Available at ⟨http://www.harzing.com/h_indexjournals.htm⟩.

Hicks, D., 2012. Performance-based university research funding systems. Research Policy 41, 251–261.

Hicks, D., 1999. The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics 44, 193–215.

Hicks, D., Wang, J., 2011. Coverage and overlap of the new social sciences and humanities journal lists. Journal of the American Society for Information Science and Technology 62, 284–294.

Jacobs, J.A., 2011. Journal rankings in sociology: using the *H* index with Google Scholar. In: PSC Working Paper No. 11-05, Available at ⟨http://repository.upenn.edu/psc_working_papers/29⟩.

Landis, J.R., Koch, G.G., 1977. The measurement of observer agreement for categorical data. Biometrics 33, 159–174.

Lepkowski, J.M., Raghunathan, T.E., Van Hoewyk, J., Solenberger, P., 2001. A multivariate technique for multiply imputing missing values using a sequence of regression models. Survey Methodology 27, 85–95.

Linnemer, L., Combes, P., 2010. Inferring missing citations: a quantitative multi-criteria ranking of all journals in economics. In: GREQAM Discussion Paper No. 2010-25, Available at ⟨http://www.vcharite.univ-mrs.fr/pp/combes/⟩.

Little, R.E., Rubin, D.B., 2002. Statistical Analysis of Missing Data, second ed. John Wiley & Sons, New York, NY.

Martin, B.R., 2011. The Research Excellence Framework and the "impact agenda": are we creating a Frankenstein monster? Research Evaluation 20, 247–254.

Martin, B.R., Whitley, R., 2010. The UK Research Assessment Exercise: a case of regulatory capture? In: Whitley, R., Gläser, J., Engwall, L. (Eds.), Reconfiguring Knowledge Production: Changing Authority Relationships in the Sciences and their Consequences for Intellectual Innovation. Oxford University Press, Oxford, pp. 51–80.

Mingers, J., Macri, F., Petrovici, D., 2012. Using the *h*-index to measure the quality of journals in the field of business and management. Information Processing and Management 48, 234–241.

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, 575–583.

Moed, H.F., 2005. Citation Analysis in Research Evaluation. Springer, Dordrecht.

Nicolaisen, J., 2007. Citation analysis. Annual Review of Information Science and Technology 41, 609–641.

OECD,2010. Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings. OECD Publishing, Paris.

Oswald, A.J., 2007. An examination of the reliability of prestigious scholarly journals: evidence and implications for decision-makers. Economica 74, 21–31.

Rafols, I., Leydesdorff, L., O'Hare, A., Nightingale, P., Stirling, A., 2012. How journal rankings can suppress interdisciplinary research: a comparison between Innovation Studies and Business & Management. Research Policy 41, 1262–1282.

Rebora, G., Turri, M., 2013. The UK and Italian research assessment exercises face to face. Research Policy 42, 1657–1666.

Rinia, E.J., van Leeuwen, Th.N., van Vuren, H.G., van Raan, A.F.J., 1998. Comparative analysis of a set of bibliometric indicators and central peer review criteria. Research Policy 27, 95–107.

Rubin, D.B., 1987. Multiple Imputation for Nonresponse in Surveys. John Wiley & Sons, New York, NY.

Schafer, J.L., 1997. Analysis of Incomplete Multivariate Data. Chapman and Hall, Boca Raton, FL.

Seglen, P.O., 1997. Why the impact factor of journals should not be used for evaluating research. BMJ 314, 498–502.

Seglen, P.O., 1992. The skewness of science. Journal of the American Society for Information Science 43, 628–638.

Sgroi, D., Oswald, A.J., 2013. How should peer-review panels behave? Economic Journal 123, F255–F278.

Stern, D.I., 2013. Uncertainty measures for economics journal impact factors. Journal of Economic Literature 51, 173–189.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 82, 528–550.

van Buuren, S., Brand, J.P.L., Groothuis-Oudshoorn, C.G.M., Rubin, D.B., 2006. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation 76, 1049–1064.

Waltman, L., Costas, R., 2013. F1000 Recommendations as a New Data Source for Research Evaluation: A Comparison with Citations (a*rXi*v:1303.3875). Available at ⟨http://arxiv.org/pdf/1303.3875v1⟩.