DEMB Working Paper Series

N. 20

Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy

Graziella Bertocchi[1]
Alfonso Gambardella[2]
Tullio Jappelli[3]
Carmela A. Nappi[4]
Franco Peracchi[5]

October 2013

[1]University of Modena and Reggio Emilia and Department of Economics Marco Biagi
Viale Berengario 51, 41121 Modena, Italy. Phone: 39 059 2056856
e-mail: graziella.bertocchi@unimore.it
[2] Bocconi University
[3] University of Napoli Federico II, email: tullio.jappelli@unina.it
[4] ANVUR
[5] University of Rome Tor Vergata

ISSN: 2281-440X online

Bibliometric Evaluation vs. Informed Peer Review: Evidence from Italy

Graziella Bertocchi, Alfonso Gambardella, Tullio Jappelli, Carmela A. Nappi, Franco Peracchi

October 27, 2013

Abstract

A relevant question for the organization of large scale research assessments is whether bibliometric evaluation and informed peer review where reviewers know where the work was published, yield similar results. It would suggest, for instance, that less costly bibliometric evaluation might - at least partly - replace informed peer review, or that bibliometric evaluation could reliably monitor research in between assessment exercises. We draw on our experience of evaluating Italian research in Economics, Business and Statistics, where almost 12,000 publications dated 2004-2010 were assessed. A random sample from the available population of journal articles shows that informed peer review and bibliometric analysis produce similar evaluations of the same set of papers. Whether because of independent convergence in assessment, or the influence of bibliometric information on the community of reviewers, the implication for the organization of these exercises is that these two approaches are substitutes.

Keywords: Research Assessment, Peer Review, Bibliometric Evaluation, VQR

The authors have been, respectively, president of the panel evaluating Italian research in Economics and Statistics (Tullio Jappelli), coordinators of the sub-panels in 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 research assessment. We are also grateful to Dimitris Christelis for implementation of the multiple imputation model and comparison with the baseline model.

Authors
Graziella Bertocchi, University of Modena and Reggio Emilia; Alfonso Gambardella, Bocconi University; Tullio Jappelli, University of Napoli Federico II; Carmela A. Nappi, ANVUR; Franco Peracchi, University of Rome Tor Vergata

Corresponding author: Tullio Jappelli, University of Napoli Federico II, email: tullio.jappelli@unina.it

1. Introduction

Measuring research quality is a topic of growing interest to universities and research institutions. This has become central issue in relation to the efficient allocation of public resources which, in many countries and especially those 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 university research (Rebora and Turri, 2013). 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) include the British Research Excellent Framework (REF) and the Italian Evaluation of Research Quality.

The main criteria for evaluating research performance combine, in various ways, bibliometric indicators and peer review. Bibliometric indicators typically are based on the number of citations that a research paper receives, which may be considered 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.) poses 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 specific research paper.

Peer review, in principle, is a better way of evaluating the quality of a research 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. In fact, there is a conflict between the quality and quantity of peer review in the search for qualified peer reviewers and in the attention that each may devote to the evaluation of a research paper, in particular in the context of large-scale research assessments. In addition, peer review may be subject to conflicts of interest, and the assessments may not be uniform across research papers, disciplines and research topics. Moreover, the number and the nature of the criteria that reviewers are asked to take into account in their evaluation are issues of extensive discussion

(Rinia et al., 1998). Finally, peer review is much more costly and demanding of time than bibliometric evaluation.

Since no evaluation method appears to dominate, it is important to understand how and to what extent bibliometric indices and peer review can be efficiently combined to assess research quality. 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, involved all public Italian universities, all private universities awarding recognized academic degrees, and all public research institutions under the supervision of the Italian Ministry of Education, University and Research (MIUR).[1] Typically, Researchers affiliated to these institutions were asked to submit three papers (chosen among journal articles, books, book chapters, conference proceedings, etc.) published between 2004 and 2010.[2] The evaluation process was conducted by 14 Groups of Experts of Evaluation (GEV), one for each research area, coordinated by the National Agency for the Evaluation of University and Research (ANVUR).[3] Evaluation of the research papers was carried out using a combination of bibliometric analysis and peer review, in proportions that varied across research areas under the legal constraint that, overall, at least half of the papers were to be assigned to peer review.

Our study focuses on the evidence available for one of the 14 areas covered by the VQR, namely Area 13 (Economics and Statistics). This area is particularly interesting because, at least in Italy, it lies in between the "hard" sciences, where most research is covered by bibliometric databases, and the humanities and social sciences, where bibliometric databases are incomplete or almost entirely missing. While the subjects of Economics and Statistics do have bibliometric databases, they tend to be incomplete because many journals (published in Italy and elsewhere) are not indexed. Thus, prior to the bibliometric evaluation, we compiled a list of the academic journals in which Italian researchers published in 2004-2010.

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

[2] The VQR Call sets the number of research products to be submitted for the evaluation at 3 for university personnel, and 6 for research institution personnel. Reductions to these numbers are calculated according to year of recruitment and take account of periods of maternity and sickness leave.

[3] 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).

We describe the construction of this list and the statistical procedures used to impute missing bibliometric indicators in order to allow a uniform classification. We then compare the results of the two evaluation methods – bibliometric evaluation and 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.

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

It should be noted 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 journals peer review submitted papers). Since the referees know in which journal the paper has been published, they may also know a number of bibliometric indicators associated with that journal. These indicators are likely to shape their opinions about the quality of the journal. which, in turn, are likely to influence their evaluation. Thus, comparing informed peer review with bibliometric analysis raises the issue of whether the two evaluations are independent. To check whether the perceived quality of a journal carries a disproportionate weight in the evaluation process, we employ background information about the refereeing process. We find that even were reviewers likely to be influenced by the perceived quality of the publication outlet, their perceptions are highly correlated with other indicators of the quality of the paper, and are not the leading factor in the overall peer review assessment.

The remainder of the paper is organized as follows. Section 2 describes the construction of the journal list database, and presents descriptive statistics. Section 3 deals with the imputation of missing values in the bibliometric indicators, describing a simple two-step procedure and a more elaborate procedure based on multiple imputations. Section 4 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 5 describes the comparison of peer review and bibliometric evaluation for the random sample. Section 6 concludes. Two appendices provide information on the technical aspects of the multiple imputation procedure, and the referees' evaluation forms.

2. The journal list

The initial journal list was based on ISI-Thomson Reuters Web of Science (WoS) and included all journals in the ISI-JCR Social Science Edition[4] belonging to the subject categories relevant to Economics and Statistics, plus other journals in the ISI-JCR Science Edition.[5] This initial list was expanded using the U-GOV dataset, to include all journals in which Italian researchers in the area of Economics and Statistics published in 2004-2010.[6] Each journal was then assigned to one of five sub-areas: Business, management and finance (hereafter Management); Economics; Economic history and history of economic thought (hereafter History); Statistics and applied mathematics (hereafter Statistics); and an additional sub-area comprising three general-interest journals, namely *Science*, *Nature*, and *Proceedings of the National Academy of Sciences*. To avoid different rankings for the same journal across different sub-areas, each journal was uniquely assigned to a single sub-area.

Several studies within the social sciences[7] have concluded that there is a high degree of agreement between the bibliometric indicators from WoS and Google Scholar, and that the rankings of the journals for which both sets of indicators are available tend to be similar, especially if the objective is classification into broad categories – as in the VQR – rather than comparison across individual journals or articles.

---

[4] 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).

[5] 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).

[6] U-GOV provides a dataset which includes the products of all researchers employed by the Italian public university system. From the U-GOV dataset we excluded the following publication outlets: journals that clearly fall outside the area of Economics and Statistics; working paper series and collections/reports of Departments/Faculty/Research Institutions; journals for which Google Scholar's *h*-index is missing for the period 2004-2010 (or shorter periods for recent journals); journals for which the *h*-index was less than 3 in 2004-2010; and journals that are 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*.

[7] See, e.g. Mingers et al. (2002) 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).

In April 2012, *h*-index data from Google Scholar were collected for all journals in the list.[8] At the end of April 2013, a preliminary version of the list was published for comments and suggestions from the scientific community. The final list, published at the end of July 2012, reflects several changes based on these comments.[9] It includes a total of 1,906 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 list, but the fraction of ISI journals varies by sub-area, ranging from 40% for History to 42% for Management, 52% for Economics, and 56% for Statistics (Table 1).

Table 2 presents the basic statistics for our four bibliometric indicators: Impact Factor (IF); 5-year Impact Factor (IF5); Article Influence Score (AIS); and the *h*-index from Google Scholar. The IF is computed by ISI using the same methodology as for IF5, but over a two-year period; AIS excludes journal self-citations and gives more weight to citations received from higher ranked journals.

IFs are available for all 912 ISI journals, with a mean of 1.19 and a standard deviation of 0.97. The average IF varies by sub-area, and is highest for Management (1.47) and lowest for History (0.49). IF5 and AIS are available only for a subset of 648 ISI journals. Averages and percentile data for these two indices show important differences in citation patterns among sub-areas, with the lowest values for History journals. Apart from History, the distribution of AIS appears to be more similar across sub-areas than the distribution of IF5. Correlation coefficients are reported after converting the indicators to a logarithmic scale; in Section 3, we use logarithms to reduce heteroskedasticity and make the distribution of the indicators closer to a Gaussian distribution. The correlation between the three bibliometric indicators available in WoS is very high: for instance, the correlation between log(IF) and log(IF5) is above 0.9 for all sub-areas, and the correlation between log(IF5) and log(AIS) is higher than 0.8 for all sub-areas.

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, while the highest is for Management. The *h*-index is strongly and positively

[8] 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 2004-2010. Data were collected in April 2012 and checked throughout May 2012.

[9] Comments concerned: misclassification of journals across the four sub-areas; misreported presence of journals in WoS; misreported values of the *h*-index; inclusion of journals that meet the GEV classification requirements; exclusion of journals published after 2008 or pertaining to other disciplines; errors in the name or ISSN of journals.

correlated with the other three bibliometric 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 IF) to 0.72 (for IF5), while for Statistics it ranges from 0.65 (for AIS) to 0.73 (for IF5) (Table 3). These values make us confident that the *h*-index represents a strong predictor to use for imputing missing values of IF, IF5 and AIS.

3. Imputation of bibliometric indicators

We now describe the procedure applied to impute missing values of the three ISI bibliometric indicators, namely IF, IF5 and AIS.

The fraction of missing values for all three indicators is shown in Table 4, 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 IF, and columns (4) and (5) give the same information for IF5 and AIS (these two indicators have identical patterns of missingness - the AIS can be defined only if IF5 is also defined). The fraction of missing values is notable for all three indicators, but especially for IF5/AIS. In relation to sub-areas, journals in History and Management are the most affected by missingness, while journals in Statistics are the least affected.

It is useful to inspect the distribution of non-missing values of the bibliometric indicators, because it is relevant for the choice of imputation model. The kernel density estimates for IF5 and AIS (not reported for brevity) reveal strong asymmetry. In particular, the distribution of both indicators is skewed to the right with a substantial right tail, and this is true for all four sub-areas. 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). Asymmetry is confirmed by the indices of skewness and kurtosis, displayed in columns (1)-(4) of Table 5, which can be compared with those of a Gaussian distribution, equal to zero and 3 respectively. The worst affected sub-areas are Economics and Management, while the least affected is History (possibly because of the small sample size).

Such large skewness and kurtosis in the distribution of the bibliometric indicators makes estimation of regression models in levels problematic since the resulting estimates are likely

to be unduly influenced by the outliers in the long right tail of the distribution. Therefore, we chose to estimate our models in logarithms rather than levels. The logarithmic transformation (which is strictly increasing and thus preserves rankings) makes the distribution of non-missing values much more symmetric and closer to Gaussian, as can be seen from the values of the indices of skewness and kurtosis in columns (5)-(8) of Table 5.

After applying the logarithmic transformation to all bibliometric indicators, we consider two different imputation methods:

i) a baseline imputation method (BIM), in which the logarithm of each of the three bibliometric indicators (IF, IF5, AIS) is regressed on a constant and the logarithm of the *h*-index.[10] 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), instead of producing a single imputation, creates multiple imputed values for each missing observation. The principle of multiple imputation, first introduced by Rubin (see e.g. Rubin, 1987), is currently 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 better 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 other indicators. The estimation sample again consists of the observations with non-missing values for the indicator of interest, but now the predictors can have imputed values. For example, to impute IF we use as predictors IF5 and AIS, which can have imputed values in the sample of non-missing observations for IF. Given the high correlation of IF with IF5 and AIS, including these two indicators should increase the predictive power of the regression model.[11] In addition to the level of the

[10] Standard errors are computed using the "robust" option in Stata.
[11] The particular implementation of MIM that we used is from van Buuren et al. (2006). Details can be found in Appendix 1.

observed or imputed bibliometric indicators, we include their squares to allow for possible nonlinearities. We also include an indicator for whether a journal is published in English because this affects the probability that the journal is included in WoS. To reduce the influence of outliers, in the MIM estimation sample, we only retain observations with values of the dependent variable above the 1st percentile and below the 99th percentile. As a result, estimation samples for MIM are slightly smaller than for BIM.

MIM is run 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. 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 estimation sample for the sub-area H is very small, we did not use the MIM method in that case.

The estimation results from both BIM and MIM show that for both AIS and 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).

4. 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 IF5, AIS and *h*-index for each sub-area, using both the non-imputed and 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 that there is substantial agreement between BIM and MIM; it also shows the differences in journal ranking between the two imputation methods. For example, we note that for AIS in sub-area E there are 40 journals with a difference in ranking equal to minus 1, i.e., they are ranked one level lower by BIM compared to MIM. In Statistics, there are 28 journals with a difference in ranking by IF5 equal to plus 1, i.e., they are ranked one level higher by BIM compared to MIM.

To better judge the difference in rankings between the two methods, for each sub-area/indicator combination we compute the percentage of journals for which the difference in ranking is between minus 1 and plus 1. These are journals for which the two imputation methods produce rankings that are "not too dissimilar". It turns out that the vast majority of journals (on average 95%, the lowest percentage being 92% for AIS in Management) have a ranking difference of at most one level in absolute value. In effect, most journals are ranked exactly the same.

We conclude from these results that, 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, we then looked at the differences in journal rankings between pairs of indicators. Most journals again are ranked the same no matter which indicator is used. This emerges clearly in Table 7, which shows the distribution of the differences in ranking between pairs of indicators. Most journals again are ranked very similarly by all three indicators. The differences are largest for AIS and the *h*-index for the sub-area S. However, even in this case, the percentage of journals with a ranking difference of at most 1 in absolute value is 94.8%, while the percentage of journals ranked the same is 74.6%. Again, this is not surprising since all indicators are strongly positively correlated and the *h*-index is a crucial predictor when imputing IF, IF5 and AIS.

The strong correlation between the various indicators implies that, in principle, any of them could be used for the classification. Given these considerations, we decided to base the final classification of journals on the maximum between their AIS and IF5 rank. We also decided to make the final classification of each journal article dependent on the individual citations it received in WoS. Specifically, a journal article that received at least five citations

per year in 2004-2010 was upgraded one level, with no correction for articles not in WoS journals because of lack of reliable citation data.[12]

Table 8 shows the final journal classification by sub-area. Overall, 48.7% of the journals are in class D, 9.4% in class C, 18.5% in class B, and 23.4% in class A. The slightly different proportions compared to the VQR guidelines (50/10/20/20) reflect the rule of the maximum between AIS and IF5 ranks, the presence of ties in the imputed values of AIS and IF5, and the decision to upgrade some Italian journals to class C.[13] In class D, the fraction of papers is similar for all sub-areas. In class A, the fraction is slightly higher than 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 ranked in class A (172), followed by Economics (152), Statistics (112), and History (10).

5. Comparison between peer review and bibliometric evaluation

The set of articles published in one of the journals considered by the VQR for Area 13 includes 5,681 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.[14]

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 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 peer review process for our sample of articles was managed as for a scientific journal with two independent editors. First, each article was assigned to two GEV members with expertise in the article's specific field of research. Each assigned the article to an

[12] The practical effect of the upgrading was negligible, since (except for 6 papers) the few articles that received a sufficient number of citations already appeared A-class journals.

[13] The GEV 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.

[14] The sample was drawn before starting the peer review process using a random number generator. Over-sampling was applied for History because of the small size of its population.

independently chosen peer reviewer.[15] Overall 610 referees were selected on the basis of their academic curricula and research interests.[16] Each peer reviewer was asked to evaluate the article according to three criteria: relevance, originality or innovation, and internationalization or international standing. To express their evaluation, referees were provided with a form prepared by the GEV containing three broad questions each referring to different dimensions of the quality of the papers, and an open field.[17] Based on the peer reviews, the GEV then produced a final evaluation through a Consensus Group consisting of the two GEV members in charge of the article, plus a third when needed.

For each article included in our sample, the following variables are available: the bibliometric indicator (F) based on the number of citations to the article 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);[18] and the final evaluation of the Consensus Group (P). Each of these variable is mapped into one of the 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;[19] the other two (F and P) are directly expressed in the four-class format. When necessary, the four merit classes are converted to numerical scores following 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 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.

[15] The allocation of papers to panel members and referees avoided any conflicts of interest with authors and authors' affiliations according to the rules established by the VQR. Referee independence was ensured by paying attention to research collaborations and, where possible, to nationality.

[16] 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 and h-index. Half of the referees are affiliated to non-Italian institutions, and a third are from Anglo-Saxon countries.

[17] The evaluation form is available in Appendix 2.

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

[19] The conversion grid involves the following correspondence: 23-27: Excellent (A); 18-22: Good (B); 15-17: Acceptable (C): 3-14: Limited (D).

5.1. The F and P distribution

Table 11 presents the distribution of the F and P indicators; Table 12 presents the distribution of P1 and P2. The elements on the main diagonal in Table 11 correspond to cases where peer review and bibliometric evaluations coincide. The off-diagonal elements correspond to cases of disagreement between P and F, 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 shows that the main source of disagreement between F and P is that peer review classifies fewer papers (only 116) as "A", than bibliometric analysis (198 papers). Peer review classifies as "A" only 49% of the 198 papers classified as "A" by bibliometric analysis.[20] Table 11 shows also that 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 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 opinions of the two external referees. In 45% of cases they agree on the same evaluation, and in 82% of 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 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 peer review compared to the bibliometric analysis, occurs for all sub-areas except History, where 10 papers are classified as "A" by the peer review and 9 by the bibliometric analysis. In this case, however, the sample is relatively small (37 observations), and cell-by-cell comparison might not be reliable.

[20] One possible reason is that, by construction, our journal classification places in the top class A, a larger number of journals than many reviewers consider to be among the population of "top" journals.

5.2. Comparison between F and P

When comparing peer review and bibliometric analysis, we can consider two criteria: first, the degree of agreement between F and P, that is, if F and P tend to agree on the same score; second, 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 (many papers receive different evaluations according to the F and P criteria). It could still be that, on average, F and P provide a similar evaluation. The distribution is characterized by low agreement and low systematic differences. Adopting one of the two evaluations (for instance, the F evaluation) 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 is characterized by high agreement, but large systematic differences since 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) if measured with reference 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, while systematic differences between sample means can be detected using a standard *t*-test for paired samples.

5.3. 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 1 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").

In the total sample, kappa is equal to 0.54 and statistically different from zero at the 1% level. For Economics, Management and Statistics the agreement is close to the value for the sample as a whole, 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. It can be argued that, in the present context, the appropriate weights are the VQR weights. These compute the distance between the evaluations using the numerical scores (1, 0.8, 0.5, 0) associated with the qualitative evaluations (A, B, C, D). The second row in Table 6 reports the "VQR weighted" kappa. The resulting statistic is quite similar to the linearly weighted kappa, indicating good agreement for the total sample (0.54) and for Economics, Management and Statistics, and a lower value for History (0.29).

The degree of agreement between the bibliometric ranking (F) and peer review (P) is actually higher than that 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). Furthermore, for each sub-area there is more agreement between F and P than between P1 and P2.

5.4. 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 (A, B, C or D) using the weights assigned by the VQR to the four merit classes (1, 0.8, 0.5, 0). 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 labelled "Score P1" and "Score P2"). Column (3) reports the average score of the 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) could 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), labelled "F", shows the average score of the bibliometric evaluation (0.561). Similar to the P score, the F score tends to be lower for Management (0.444) 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. In the overall sample 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 negative (-0.108 and -0.034, respectively) but not statistically different from zero.

5.5. Informed peer review

As previously stressed, the VQR relies on informed peer review. In other words, the referees know not only the identity of the authors of the published piece, but also the final publication outlet, together with the bibliometric indicators associated with the journal.[21] Thus, comparing informed peer review with bibliometric analysis raises the question of whether the two evaluations are independent. That is, to what extent is the referee's evaluation affected by the perceived quality of the journal (which may in turn be based on bibliometric indicators)? In other words, is opinion about the quality of the paper disconnected from opinion about the quality of the journal in which the paper is published? 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.

To check whether the perceived quality of a journal carries a disproportionate weight in the evaluation process, we employ additional background information about the refereeing process. The referee evaluation form used by VQR includes three questions about the originality, relevance and internationalization of a paper. The form is available in Appendix 2.

[21] The referees were provided with the GEV journal classification list including both the ISI and the 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.

While the first two questions refer directly to the quality of the paper, the third explicitly refers to its international reach and potential for future citations. The responses to this third question are more likely to be influenced by the referee's assessment based on journal rankings. The correlation coefficients reported in Table 15 show that the three dimensions along which referees are asked to rank papers tend to be highly correlated.[22] 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 peer review assessment.

6. Conclusions

This article contributes to the debate on bibliometric and 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 remarkable agreement between bibliometric and peer review evaluation. Furthermore, there is no evidence of systematic differences between the average scores provided by the two rankings. Although in aggregate there are no systematic differences between bibliometric and peer review evaluation, there is a lower number of papers assigned by referees to the top class relative to the bibliometric analysis. However, most of the papers “downgraded” by the 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.

[22] Since the original scores of the referees were provided in grades (from 1 to 9) for each of the three questions, in Table 15 we compute the correlation matrix of the overall score assigned by each referee with the score assigned to each of the three questions. The matrix shows that the correlations are quite high: in particular, the correlations between the score and the questions are 94% for originality, 95% for relevance and 95% for internationalization. Furthermore, the correlations between the three scores themselves 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 close had we excluded the third question from the referee's evaluation form.

Across sub-areas, the degree of agreement is somewhat 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 replacing bibliometric evaluations with peer review, at least in the disciplines studied in this paper and for research output published in ranked journal articles. Since bibliometric evaluation is less costly, this could ease the research assessment process and its cost. Nevertheless, we recommend that formal evaluation exercises should still include a sizeable share of articles assessed by 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 peer review and bibliometric evaluation This would further contribute to testing the similarities between them.

Our results also suggest that bibliometric evaluation would be reliable to monitor the research outputs of a nation or a community on a more frequent basis. National research assessments involve huge amounts of time and effort to organize and, therefore, take place only every few years. This paper suggests that bibliometric evaluations, which could be organized more flexibly and at less cost than large-scale peer review evaluation, could be employed between national evaluations to allow more frequent monitoring of the dynamics of research outcomes.

As argued in the Introduction, our results do not necessarily imply genuine convergence of evaluation between peer review and bibliometric analysis. We have shown that reviewers were perhaps influenced by their information on publication outlet and the bibliometric ranking of the articles. However, and given the correlation between this dimension of the peer review evaluation form and more genuine assessment of the quality of the papers by the referees, this suggests only that the community of reviewers broadly trusts bibliometric indicators, otherwise their assessments would have differed. This finding is interesting in itself. It implies that, particularly for Economics, Management and Statistics, there is substantial alignment between the value judgment of the academic community and the

indicators produced by bibliometric indicators. Again, the implication for the organization of research assessments is that for large numbers of research papers, bibliometric indicators could to some extent substitute for peer review evaluation.

This paper inevitably has some limitations. The most important is the difficulty of generalizing our results to other disciplines. Even within our sub-areas, we found important differences between the two approaches. In our case, these differences could be for a number of reasons. First, there might be differences among referees in different subject areas, exemplified by the possibility that referees might be less generous in some areas than in others. Second, journal ranking reliability might differ across areas; for instance, the ranking of journals might be more generous (e.g., placing larger number of journals in the top class) in Economics and Management relative to other sub-areas. Finally, the power of the statistical test might be limited if the sample size is not large, as in the case of some research areas, so that confidence intervals tend to be relatively large. Future research could improve on the analysis of these dimensions, and possibly control better for some of this heterogeneity using larger sample size.

Despite these caveats, we believe that the Italian research assessment exercise offers an unusual opportunity to employ 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 these relationships should help to 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.

References

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

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

Harzing, A.-W. and R. van der Wal (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.

Jacobs, J. A. (2011), "Journal rankings in sociology: Using the H Index with Google Scholar," PSC Working Paper No. 11-05. Available at http://repository.upenn.edu/psc_working_papers/29.

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

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

Little, R. E., and D. B. Rubin (2002), *Statistical Analysis of Missing Data*, 2nd Edition. New York, NY: John Wiley & Sons.

Mingers, J., F. Macri, and D. Petrovici (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.

M. Burger, J. G. Frankfort, and A. F. J. van Raan (1985), "The use of bibliometric data for the measurement of university research performance," *Research Policy*, 14:131–149.? Not in text

OECD (2010), *Performance-based Funding for Public Research in Tertiary Education Institutions: Workshop Proceedings*. Paris: OECD Publishing. Available at http://dx.doi.org/10.1787/9789264094611-en.

Rebora, G., and M. Turri (2013), "The UK and Italian research assessment exercises face to face," *Research Policy*, forthcoming. Available at http://dx.doi.org/10.1016/j.respol.2013.06.009.

Rinia, E. J., Th. N. van Leeuwen, H. G. van Vuren, and A. F. J. van Raan (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*. New York, NY: John Wiley & Sons.

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

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

Sgroi, D. and A. J. Oswald (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., and W. H. Wong (1987), "The calculation of posterior distributions by data augmentation (with discussion)," *Journal of the American Statistical Association*, 82: 528–550.

van Buuren, S., J. P. L. Brand, C. G. M. Groothuis-Oudshoorn, and D. B. Rubin (2006), "Fully conditional specification in multivariate imputation," *Journal of Statistical Computation and Simulation*, 76: 1049–64.

Appendix 1. Detailed description of the MIM

The imputation methodology that we use is the fully conditional specification method (FCS) of van Buuren, Brand, Groothuis-Oudshoorn and Rubin (2006, henceforth BBGR), and the exposition from this point on follows closely theirs.[23]

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 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[24] 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 θ_k^* from its posterior distribution[25], i.e., we have

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

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

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

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

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

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

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 $\sigma_k^{*(t)}$).[26] 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 (1) and (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 $\left\{\left(\theta_1^{(t)},...,\theta_k^{(t)},Y_{mis}^{(t)}\right):\ t=1,2,...\right\}$. The stationary distribution of this sequence is $P\left(Y_{mis},Y_{obs};\theta\right)$, 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 $\left\{\theta_1^{(t)},...,\theta_k^{(t)}\right\}$ to the conditional distribution of the parameter vector $P\left(\theta \mid Y_{obs}\right)$ or, equivalently, the convergence of the sequence $\left\{Y_{miss}^{(t)}\right\}$ to the conditional distribution of the missing values $P\left(Y_{mis} \mid Y_{obs}\right)$. 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.

[26] 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.

One important feature of the FCS method (shared with several other similar approaches found in the imputation literature)[27] 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 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 equations (1) and (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 m^{th} implicate dataset. Then the overall estimate derived using all M implicate datasets is just the average of the M separate estimates, i.e.,

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

[27] A similar imputation procedure is proposed by Lepkowski, Raghunathan, Van Hoewyk and Solenberger (2001). See also BBGR for references to a number of other approaches that have significant similarities to theirs.

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

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

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 (1) and (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:

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

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:

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

As pointed out by Little and Rubin (2002), the second term in (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:

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

Appendix 2. The referee evaluation form

ANVUR – ASSESSMENT OF THE RESEARCH QUALITY 2004-2010

Assessment Form (one form to be filled for each research product)

Groups of Experts for Economics and Statistics - GEV 13.

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:

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

	Research sub-area				
	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	1,903
%	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. 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); 10^{th} , 25^{th}, 50^{th}, 75^{th} and 90^{th} percentiles (respectively p10, p25, p50, p75, p90); inter-quantile range (iqr).

Table 3. 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 logarithm 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 4. 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 5. 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)	
	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	Kurtosis	Skewness	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)).

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

Difference in Ranking 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		
Ranking difference = - 3	7	1.09%	7	1.09%
Ranking difference = - 2	18	2.80%	20	3.11%
Ranking difference = - 1	52	8.09%	40	6.22%
Ranking difference = 0	485	75.43%	494	76.83%
Ranking difference = + 1	66	10.26%	71	11.04%
Ranking difference = + 2	15	2.33%	10	1.56%
Ranking difference = + 3	0	0.00%	1	0.16%
Percentage of journals for which the difference in rankings is between - 1 and + 1	93.78%		94.09%	
		Management		
Ranking difference = - 3	5	0.65%	10	1.30%
Ranking difference = - 2	10	1.30%	25	3.26%
Ranking difference = - 1	66	8.61%	74	9.65%
Ranking difference = 0	607	79.14%	543	70.80%
Ranking difference = + 1	63	8.21%	86	11.21%
Ranking difference = + 2	16	2.09%	28	3.65%
Ranking difference = + 3	0	0.00%	1	0.13%
Percentage of journals for which the difference in rankings is between - 1 and + 1	95.96%		91.66%	
		Statistics		
Ranking difference = - 3	3	0.67%	6	1.35%
Ranking difference = - 2	8	1.80%	15	3.37%
Ranking difference = - 1	23	5.17%	28	6.29%
Ranking difference = 0	380	85.39%	338	75.96%
Ranking difference = + 1	28	6.29%	49	11.01%
Ranking difference = + 2	3	0.67%	9	2.02%
Ranking difference = + 3	0	0.00%	0	0.00%
Percentage of journals for which the difference in rankings 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.

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

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

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	1,903
	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	236	10
History	147	37	25
Management	1750	175	10
Statistics	1423	142	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	40.25
B	337	14.27	30	12.71
C	434	18.38	49	20.76
D	667	28.25	62	26.27
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.

Table 11. Comparison between F and P

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

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

Table 12. Comparison between P1 and P2

Peer #1	Peer #2				
	A	**B**	**C**	**D**	**Total**
A	53	43	7	11	114
	46.49	37.72	6.14	9.65	100.00
B	36	73	29	29	167
	21.56	43.71	17.37	17.37	100.00
C	8	34	21	29	92
	8.70	36.96	22.83	31.52	100.00
D	4	46	50	117	217
	1.84	21.20	23.04	53.92	100.00
Total	101	196	107	186	590
	17.12	33.22	18.14	31.53	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 peer reviewers agree on the evaluation. The off-diagonal elements correspond to cases of disagreement between the two peer reviewers. Note that labelling the two evaluations by the two peer reviewers as Peer#1 and Peer#2 is purely a convention, reflecting only the order in which the referees accepted to review the paper.

Table 13. Kappa statistic for the amount of agreement

	Total sample (1)	Economics (2)	History (3)	Management (4)	Statistics (5)
F and P, linear weighted 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. One star indicates significance at the 5% level; two stars indicate significance at the 1% level.

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

	Score P1 (1)	Score P2 (2)	Score P (3)	Score F (4)	Difference between F and P (5)	Sample size (6)	t-test for difference between F and P (7)	p-value (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.

Table 15. Correlation matrix of reviewers' questions

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

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