Journal of Informetrics 5 (2011) 275–291

Contents lists available at ScienceDirect

Journal of Informetrics

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

The first Italian research assessment exercise: A bibliometric perspective

Massimo Franceschet[a,*], Antonio Costantini[b]

[a] Department of Mathematics and Computer Science, University of Udine, Via delle Scienze 206, 33100 Udine, Italy
[b] Department of Agriculture and Environmental Sciences, University of Udine, Via delle Scienze 206, 33100 Udine, Italy

ARTICLE INFO

Article history:
Received 4 October 2010
Received in revised form 6 December 2010
Accepted 8 December 2010

Keywords:
Research assessment
Peer review
Bibliometrics

ABSTRACT

In December 2003, seventeen years after the first UK research assessment exercise, Italy started up its first-ever national research evaluation, with the aim to evaluate, using the peer review method, the excellence of the national research production. The evaluation involved 20 disciplinary areas, 102 research structures, 18,500 research products and 6661 peer reviewers (1465 from abroad); it had a direct cost of 3.55 millions Euros and a time length spanning over 18 months. The introduction of ratings based on ex post quality of output and not on ex ante respect for parameters and compliance is an important leap forward of the national research evaluation system toward meritocracy. From the bibliometric perspective, the national assessment offered the unprecedented opportunity to perform a large-scale comparison of peer review and bibliometric indicators for an important share of the Italian research production. The present investigation takes full advantage of this opportunity to test whether peer review judgements and (article and journal) bibliometric indicators are independent variables and, in the negative case, to measure the sign and strength of the association. Outcomes allow us to advocate the use of bibliometric evaluation, suitably integrated with expert review, for the forthcoming national assessment exercises, with the goal of shifting from the assessment of research excellence to the evaluation of average research performance without significant increase of expenses.

1. Introduction

In December 2003 Italy started up its first-ever research assessment exercise, called *Valutazione Triennale della Ricerca* (VTR), with the aim to evaluate the excellence of research activities performed by universities and other research institutions under Ministry of Education, University, and Research funding. VTR covered the research of 20 disciplinary areas during the three-year period from 2001 to 2003. It involved the evaluation of 102 research structures including 77 universities, 12 public research agencies, and 13 private research agencies, which submitted 18,500 research products for evaluation. Peer-reviewing the submitted products involved 6661 experts (1465 from abroad), with a direct cost of 3.55 millions Euros and a time length of 18 months.

Evaluation activities in Italian universities traditionally favored a bureaucratic approach based on an ex ante check of the respect for input, processes, or compliance with provisions of the law (Minelli, Rebora, & Turri, 2008). The introduction of ratings based on the ex post quality of output and not on ex ante respect for parameters and compliance is an important

* Corresponding author. Tel.: +39 0432558754; fax: +39 0432558499.
E-mail address: massimo.franceschet@uniud.it (M. Franceschet).
URL: http://users.dimi.uniud.it/~massimo.franceschet/ (M. Franceschet).

1751-1577/$ – see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2010.12.002

cultural leap forward (Bleiklie, 1998; Neave, 1998). Furthermore, the rankings comparing the peer review ratings obtained by the universities in the different disciplinary areas were posted on the Web.[1] This apparently plain decision is in fact unprecedented in the setting of Italian evaluation systems in the state sector, including universities, which is characterized by a general lack of courage and a production of rankings that are based on criteria giving loose indication of merit (Calzà & Garbisa, 1995; Minelli et al., 2008).

The VTR evaluation is fully based on peer review evaluation method: each submitted research product was assessed by a pool of experts who expressed a qualitative judgement that is then mapped to a quantitative categorial rating. Reale, Barbara, and Costantini (2007) show that the VTR exercise was carried out on the basis of assessment criteria proposed in the literature for peer-reviewing (rationality, reliability, impartiality, efficiency, and effectiveness), controlling the presence and the relevance of bias of the peer judgements as prestige of institutions and reputation of scientists (Chubin & Hackett, 1990; Martin & Irvine, 1983). Hence, for the purposes of this study, we assume that peer reviewers expressed a reliable judgement on the products submitted at VTR and that this rating reflects the intrinsic *quality* of the product.

Submitted products were autonomously selected by research institutions in the measure of at most one product every four researchers (universities) or every two researchers (research agencies) choosing among the entire production over a three-year period. In order to maximize peer rating, each structure selected the products deemed to be of highest quality. It turned out that, for areas in which journal publication is the routine, most of the submitted products are journal articles and most of these articles appear in journals indexed in databases of Thomson Reuters, formerly known as ISI. For each article covered by Thomson Reuters, we have at disposal an *article citation rating*, measuring the number of citations that the article received from other papers in the database, and a *journal citation rating*, evaluating the impact factor of the journal in which the article appears, which corresponds to the average number of recent citations received by papers published in the journal (Garfield & Sher, 1963). Furthermore, for relatively large publication sets, we may compute the popular Hirsch (h) index, which attempts to assess both production and impact in a single figure (Ball, 2007; Hirsch, 2005).

This opens the unprecedented opportunity to perform a large-scale multi-disciplinary comparison of peer review and bibliometric indicators for the Italian research system. This is the aim of the present contribution. More specifically, we pose the following *research questions*:

1. Are peer review judgements and (article and journal) bibliometric indicators independent variables?
2. If not, what is the strength of the association?
3. In particular, is the association between peer judgement and article citation rating significantly stronger than the association between peer judgement and journal citation rating?

Answering these questions is of crucial importance to evaluate the opportunity of using bibliometrics in the next research assessment exercises.

In Section 2 we concisely describe the VTR assessment exercise. In Section 3 we address the posed questions with a careful analysis comparing peer review and bibliometric indicators at both levels of research disciplines (Section 3.3) and research structures within disciplines (Section 3.2). Related work is amply surveyed in Section 4. Finally, in Section 5 we draw our conclusions.

2. An overview of VTR

VTR was managed by the Committee for the Evaluation of Research (CIVR) and was designed as an ex post assessment exercise based on *peer review*. Its plan can be summarized as follows. CIVR divided the national research system into 20 scientific-disciplinary areas, including 6 interdisciplinary sectors, and set up an evaluation panel responsible for the assessment of each area. Panels were composed by high level experts (panelists), which number fluctuated from 5 to 17 according to the area size and disciplinary complexity. The exercise was then articulated in three phases, that were in charge of research structures, panels and CIVR, respectively.

In the initial phase, research institutions submitted to panels a set of autonomously selected research products. Types of products admitted to submission are: journal articles, books, book chapters, proceedings of national and international conferences, patents, designs, performances, exhibitions, manufactures and art operas. The only mandatory principle of selection stated that products of research should not exceed 50% of the full-time-equivalent researchers in the institution.[2] The research structures submitted an overall sample of 18,500 products partitioned as follows: journal articles 72%, books 17%, book chapters 6%, patents 2% and the remaining typologies 3%. Evaluated products were more than 17,300 (there are products submitted by more than one institution). Research structures were also demanded to transmit to CIVR data and indicators about human resources, international mobility of researchers, funding for research projects, patents, spin-off and partnerships, allowing to reveal impact on employment.

[1] http://vtr2006.cineca.it.

[2] A full-time-equivalent researcher represents 0.5 researchers in universities, where researchers teach as well, while it corresponds to 1 researcher in research agencies. Hence, universities were allowed to submit a maximum number of products corresponding to 25% of the three-year average permanent academic staff.

In the second phase of the exercise, which was carried out with the aid of a web platform, panelists assigned research products to external referees. Each product was assessed by at least two referees who peer-reviewed it according to four aspects of merit: quality (the opinion of peer on the scientific excellence of the product compared to the international standard), importance, originality and internationalization. Referees also expressed a final score on the following four-point scale:

1. *excellent*: a product within the top 20% of the value in a scale shared by the international scientific community;
2. *good*: a product in the 60–80% segment;
3. *acceptable*: a product in the 40–60% segment;
4. *limited*: a product within the bottom 40%.

For every evaluated product panels drew up a consensus report where panelists re-examined the peer judgments and fixed the final score. Furthermore, CIVR weighted the peer review scores as follows: 1 (excellent), 0.8 (good), 0.6 (acceptable), and 0.2 (limited). The numeric formulation made it possible to sum product scores, in order to obtain a mean rating for single research structures providing a proxy for the value of the institution research performance and the possibility to compile corresponding rankings of structures. Rankings were compiled for each disciplinary area and within groups of structures of comparable sizes: mega structures (more than 74 products), large structures (25–74 products), medium structures (10–24 products), and small structures (less than 10 products). Panels provided a final report including ranking lists of the institutions in the surveyed area, highlighting strength and weakness points of the research area, and proposing possible actions of improvement.

In the final phase of the assessment exercise, CIVR produced a detailed analysis of requested data and indicators, integrating panel reports with collected data about human resources and project funding. The CIVR final report defines a first-ever comprehensive assessment of the national research system. In summer 2009, VTR outcomes have been used for the first time by Ministry of Education, University, and Research as one of the indicators to allocate a 7% share of the Ordinary Fund for Higher Education (FFO), together with other research and teaching quality related proxies.

3. A bibliometric analysis of VTR

Our analysis considers the following research areas:

1. mathematics and computer sciences (MCS);
2. physics (PHY);
3. chemistry (CHE);
4. earth sciences (EAS);
5. biology (BIO);
6. medical sciences (MED);
7. agricultural sciences and veterinary medicine (AVM);
8. civil engineering and architecture (CEA);
9. industrial and information engineering (IIE);
10. economics and statistics (ECS).

We excluded from our investigation the six interdisciplinary areas as well as the following four areas: philological-literary sciences, antiquities and arts; history, philosophy, psychology and pedagogy; law; political and social sciences. The number of submitted products in these areas that are covered by Thomson Reuters databases is too modest for a reliable application of bibliometrics.

In the following, we refer to a product contained in Thomson Reuters databases as a Thomson Reuters (TR) article. For each submitted product we have at disposal a peer review judgement. Moreover, for each TR article we computed the following bibliometric indicators:

1. *article citation rating*, counting the number of citations that the article received from other TR papers. We retrieved all citations recorded in Thomson Reuters Web of Science database received by more than 17,000 papers up to June 2006. Since papers refer to period 2001–2003, this means that we used a dynamic citation window ranging between 2.5 and 5.5 years.[3] These periods are generally sufficient for a paper to collect the peak of citations in each of the surveyed disciplines;
2. *journal citation rating*, evaluating the average number of recent citations received by papers published in the journal in which the article appears. We used the 2-year journal impact factor corresponding to the journal and the publication year of the paper.

[3] We also undertook the correlation analyses between peer ratings and bibliometric indicators separately for each publication year, hence using a static citation window. The results are not significantly different from the ones obtained using a dynamic citation window, and hence only the latter are reported.

Table 1
Analysis at the level of research discipline.

Area	Size	Cov	Auth	Own	Peer	Cites	IF	*h*
MCS	787	92%	2.26	69%	0.830 (0.831)	3.97 (3.54)	1.12	18
PHY	1767	89%	51.85	42%	0.879 (0.885)	24.66 (4.26)	5.79	87
CHE	1089	92%	5.10	68%	0.807 (0.813)	16.14 (3.14)	5.14	50
EAS	651	90%	4.17	64%	0.825 (0.836)	7.33 (2.44)	3.01	26
BIO	1575	96%	6.56	66%	0.826 (0.831)	24.58 (2.90)	8.48	83
MED	2639	96%	8.47	59%	0.776 (0.780)	26.65 (3.20)	8.34	106
AVM	750	89%	4.81	67%	0.712 (0.728)	8.20 (3.08)	2.66	27
CEA	758	45%	2.40	84%	0.750 (0.755)	3.58 (3.10)	1.16	14
IIE	1195	82%	3.48	77%	0.774 (0.779)	4.78 (2.98)	1.61	23
ECS	971	54%	1.86	76%	0.673 (0.799)	3.16 (3.63)	0.87	17

Furthermore, we computed the Hirsch (*h*) index over relatively large sets of papers. The *h* index for a publication set is the highest number *n* such that there are *n* papers in the set each of them received at least *n* citations (Hirsch, 2005). The *h* index immediately found interest in the public (Ball, 2007) and in the bibliometrics literature (see Bornmann & Daniel, 2007b for opportunities and limitations of the *h* index). In particular, it is currently computed by both Thomson Reuters Web of Science and Elsevier Scopus bibliometric data sources. The index is meant to capture both production and impact of a publication set in a single figure. It favors publication sets containing a continuous stream of influential works over those including many quickly forgotten ones or a few blockbusters. Moreover, the index is robust to self-citations: all self-citations to papers with less than *h* citations are irrelevant for the computation of the index, as are the self-citations to papers with many more than *h* citations. Hirsch originally defined the *h*-index for the assessment of individual careers, but the metric can be applied to any information production process consisting of a set of sources (e.g., authors and journals) that produce items (e.g., articles) (Alonso, Cabrerizo, Herrera-Viedma, & Herrera, 2009; Egghe, 2010; Egghe & Rousseau, 2006).

We aggregated peer review and bibliometric data at both levels of research disciplines (Section 3.1) and research structures within disciplines (Section 3.2).

3.1. Analysis at the level of research discipline

Table 1 contains, for each surveyed discipline, the following columns:

- *area*: the disciplinary area abbreviated as above;
- *size*: the number of submitted products.[4] This gives an indication of the size (number of researchers) of the field;
- *cov*: the fraction of submitted products that are covered in TR databases;
- *auth*: the mean number of authors per paper. We interpret this as a measure of discipline propensity of collaboration among scholars;
- *own*: degree of ownership. For a given paper submitted by a given structure, it is the number of paper authors that are affiliated to the structure that submitted the paper divided by the total number of paper authors. It demonstrates the discipline propensity of collaboration with scholars of different research structures (belonging to the same or different fields): the lower the degree of ownership, the higher the inter-structure collaboration propensity.
- *peer*: the average peer review rating. Within brackets we show the rating over TR articles only;
- *cites*: the average number of received citations. Within brackets we show the ratio between number of citations and impact factor;
- *IF*: the average impact factor of the journals publishing the papers;
- *h*: the *h* index for the set of submitted TR papers.

The largest area is MED, followed by PHY and BIO; small fields are EAS, AVM, CEA and MCS. All areas have a large TR coverage with two notable exceptions: CEA (45%) and ECS (54%); important sub-fields of these areas frequently publish on books, which are not covered by TR. More precisely, CEA groups civil engineering and architecture; the former mostly publish in journals and has a good TR coverage (75%). On the contrary, scholars in architecture frequently publish books and book chapters and hence the TR coverage is limited (3%). It follows that, for the purpose of this study, the output of area CEA is largely dominated by civil engineering products. As for ECS, it is mainly composed of economics, management, and mathematics. Scholars in mathematics and economics publish mostly in journals, but these are differently covered by TR (56% in economics versus 78% in mathematics). Scholars in management prefer books or book chapters, reducing the TR coverage to 22%. Computer scientists typically prefer conference proceedings to archival journals as a mean of publication but typically journals convey a higher impact (Franceschet, 2010a, 2010c). Although TR does not index conference proceedings (at least it did not at the time of the assessment exercise), TR coverage of MCS is reasonably high. This because computing

[4] Papers with authors affiliated to structures belonging to different areas are counted for each affiliation area.

structures submitted for evaluation mostly journal papers instead of the more frequent proceeding papers, probably because they perceived that these publications are of higher quality.

The mean number of authors varies across disciplines. PHY, MED, and BIO are the fields with the largest number of authors per paper, while ECS, MCS, and CEA are the areas with the lowest authorship propensity. Notice that PHY is a significant outlier: on average, papers in this discipline have more than 50 authors. A closer look to the authorship distribution reveals that it is highly skewed: there are many papers with few authors and few ones with a huge number of authors. The median number of authors is 5, meaning that at least 50% of the papers have at most 5 authors, a figure comparable with other disciplines. On the other hand, 13% of the papers have more than 100 authors, and there exists a hub paper with the impressive number of 1412 co-authors. This phenomenon, known as *hyperauthorship* and typical of certain areas of research including high energy physics, is investigated in Cronin (2001).

We observed a (not very surprising) negative correlation between authorship and ownership[5]: the larger the number of authors per paper, the lower the ownership degree of papers, indicating a stronger propensity to collaborate outside the home institution. For instance, more than half of the authors of papers in PHY belong to a different structure with respect to the submitting one. At the other extreme, authors in CEA and, to a less extent, those in IIE and ECS, prefer to work in small groups within their research structures.

Peer review judgements were, on average, quite high, reflecting the selection of the best papers only provided by each structure in each discipline. Moreover, the average judgement over all products corresponds to the mean judgement with respect to TR articles only, with the exception of area ECS: in this field TR articles have been evaluated significantly higher than non-TR products. The fields with the best peer ratings are PHY, MCS, BIO, and EAS in this order. The areas with the poorest peer judgements are ECS and AVM. In the case of economics and statistics, an explanation of the bad performance is the high frequency of non-TR products which received a low peer rating. Furthermore, Reale et al. (2007) claim that the lower levels of rating for this area are also associated with the higher disagreement of the panel consensus in this sector with respect to the others. As for AVM, the ratings of its sub-fields are: agronomy (0.678), entomology (0.681), veterinary science (0.684), food and nutrition (0.697), animal science (0.720), plant science (0.721), and agricultural chemistry (0.757). Based on available ratings, animal science, plant science and agricultural chemistry tend to be in line with situations of good scientific quality, but the other sub-fields rank below the national standard.

Bibliometric indicator scores wildly vary across fields. This field effect is a well known phenomenon in bibliometrics (see, e.g., Althouse, West, Bergstrom, & Bergstrom, 2008). This is mainly due to the different field publication coverages of the underlying bibliographic databases and to the different field citation habits, including number of references per paper and citation speed. The ratio between article citation and impact factor scores is supposed to mitigate the field effect. It tells us something about the ability of the institutions from different fields to select the papers with the highest potential impact. In this respect, PHY was the best area and EAS was the worst.

Table 3 in Appendix A analyze the variables size (number of papers), average number of citations per paper, average journal impact, and *h* index across sets of papers characterized by different levels of (peer review) quality. The size factor gives more insight into the area overall peer judgement (column peer in Table 1). For instance, papers in PHY received the highest peer review judgements (0.879 on average). Indeed, more than half (52%) of them have been judged excellent, while only 1% of them have been considered limited products. On the other hand, peer reviewers were very critical with respect to products in ECS (the average rating is 0.673): only 17% of the products in this area are excellent works, and a higher share, 18%, are considered limited contributions. Notice that, for all areas but PHY, the most popular referee opinion is good.

Article citations are positively correlated with the categorial peer review judgement: generally, the average number of citations per paper decreases as the peer rating declines. Excellent papers always receive the highest average number of citations, well above the discipline mean, while acceptable and limited contributions received an average citation impact lower than the discipline mean. Nevertheless, some exceptions to positive correlation exist, namely the impact of limited products in EAS (+2 positions in the categorial ranking), MED (+1), CEA (+1), and IIE (+2).

Journal impact factors are also positively correlated with peer assessment: on average, the impact factor of publishing journals drops as the peer evaluation decreases. The association is, however, not as strong as the one noticed for article citations. Indeed, there are more exceptions to positive association, namely the acceptable papers in MCS (+1 positions in the categorial ranking) and EAS (+1), and the limited products in PHY (+2), MED (+2), CEA (+1), and IIE (+2).

The *h* index discriminates very well between different peer review ratings with only two exceptions: excellent and good papers in fields AVM and CEA. In particular, the *h* index neatly separates the lower judgements acceptable and limited, on which the discrimination power of both article and journal citation measures is weaker. Take, for example, the sets of acceptable and limited papers in MED. Both the average number of paper citations and the average journal impact factor for limited articles are above the same measures for acceptable papers. On the other hand, the *h* index of acceptable papers (36) largely dominates that of limited articles (21). Indeed, the sorted citation sequence for acceptable publications features a longer stream of influential papers while that for limited papers is headed by two blockbusters, which are responsible for the relatively high mean citation values, but then it quickly decreases.

[5] Spearman coefficient −0.82, *p*-value 0.007.

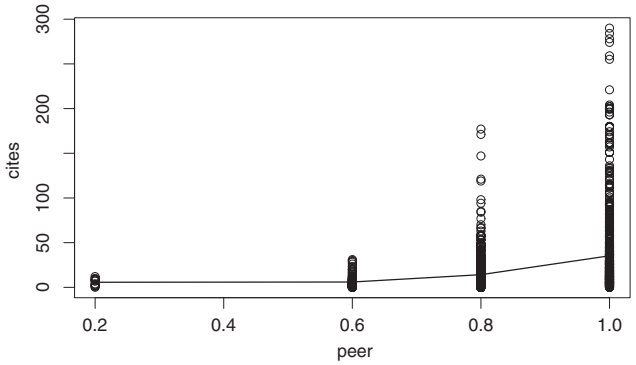

Fig. 1. Categorial scatter plot showing citations received by papers of different peer-assigned qualities for research area BIO. The solid line connects the mean number of citations for each group. Papers of higher quality generally receive more citations.

The relationship between peer judgements and bibliometric indicators, in particular article and journal citation indices, has been further investigated. Within each discipline, we expressed the discrete variable article citation as a categorial variable by splitting the distribution into quartiles to obtain a four-point scale for the variable. We did the same for journal impact factor. Then, we prepared, for each discipline, a contingency table displaying the categorial variables peer judgement and article citation (Table 4 in Appendix A) and a similar table for peer judgement and journal impact factor (Table 5 in Appendix A). Each table cell contains the joint relative frequency for the conditional distribution of the bibliometric variable (either article citation or journal impact factor) given the peer judgement variable. For example, Table 5, discipline BIO, shows that excellent papers in the discipline are split into citation quartiles as follows: 11.3% in the 1st quartile, 18.4% in the 2nd quartile, 25.3% in the 3rd quartile, and 45.0% in the 4th quartile.

It turns out that, with very few exceptions, the majority of excellent papers are associated with the highest bibliometric quartile (the 4th one), while the majority of limited products belong to the lowest bibliometric quartile (the 1st one). Good and acceptable papers distribute over the four quartiles, with a preference for the lower segments, in particular for acceptable products. If bibliometric and peer assessments were independent variables, we would expect that the relative frequency of each cell would be the product of its marginals (the row and column relative frequencies). Hence, we can test the independence of the bibliometric and peer review variables by comparing the observed frequencies with the expected ones in case of independent variables (this is the well-known Pearson χ^2 test for independence). The output of the test is that, for all disciplines, peer judgement and bibliometric indicators are *not* independent variables (with a significance level less than 0.001) with the unique exception of journal impact factor for area MCS. The strength of the association between peer opinion and article citation variables, measured with Spearman's rank-order coefficient, ranges from 0.187 for IIE to 0.403 for PHY. All values are significantly different from 0 (p-value < 0.001). The association between peer judgement and journal impact factor ranges from 0.197 for IIE to 0.529 for AVM. All values except that for MCS are significantly different from 0 (p-value < 0.001).

To conclude the investigation of association between peer review and bibliometrics, we performed a probabilistic analysis (Table 6 in Appendix A). Namely, for each pair of adjacent peer judgments X and Y, we computed the probability $P(c(X) > c(Y))$ (respectively, $P(c(X) = c(Y))$) that for two randomly drawn papers P and Q rated X and Y, respectively, the number of citations of P is greater than (respectively, equal to) the number of citations of Q. This probability is also related to the area under curve (AUC) statistic, commonly used in medical decision making as well as in machine learning and data mining (Fawcett, 2006; Kraemer et al., 2003; Stringer, Sales-Pardo, & Amaral, 2008). If peer judgments are positively correlated with article citations, an educated guess would be that, if rating X is above Y, then $P(c(X) > c(Y))$ is larger then $P(c(Y) > c(X))$. It holds that $P(c(X) > c(Y))$ can be expressed as the following ratio:

$$P(c(X) > c(Y)) = \frac{|\{(P, Q).r(P) = X \text{ and } r(Q) = Y \text{ and } c(P) > c(Q)\}|}{|\{(P, Q).\ r(P) = X \text{ and } r(Q) = Y\}|}$$

where $r(P)$ is the rating of P, $c(P)$ is the number of citations received by P, and $|\cdot|$ is the cardinality of a set. Clearly, we have that

$$P(c(X) > c(Y)) + P(c(Y) > c(X)) + P(c(X) = c(Y)) = 1$$

Similarly we computed the probabilities $P(IF(X) > IF(Y))$ and $P(IF(X) = IF(Y))$ for the journal impact factor.

We observe that for pairs of judgements (E,G) and (G,A), the number of pairs of articles whose citations are concordant with the judgements is always greater than the number of discordant pairs of papers: the higher the peer rating, the higher the probability of finding highly cited papers as well as that of finding papers published in journals of high impact. For the rating pair (A,L) the situation is more controversial: in four cases over ten, the exploited bibliometric indicators are less accurate at distinguishing acceptable papers from limited ones. By way of example, Fig. 1 illustrates the found association between citations and peer assessment for research area BIO. The probability that an excellent paper receives more citations

Table 2
Rank-order correlation between structure rating variables: peer review rating of TR articles (peer) is compared to article citation rating (cites) and to journal citation rating (IF). We show the Spearman rank-order correlation coefficient (σ) and the significance of the test (p-value).

Area	Peer vs. cites		Peer vs. IF	
	σ	p-Value	σ	p-Value
MCS	0.46	0.015	0.52	0.005
PHY	0.81	<0.001	0.29	0.088
CHE	0.60	<0.001	0.85	<0.001
EAS	0.79	<0.001	0.34	0.140
BIO	0.69	<0.001	0.74	<0.001
MED	0.56	<0.001	0.60	<0.001
AVM	0.52	0.015	0.52	0.015
CEA	0.32	0.124	0.41	0.043
IIE	0.58	<0.001	0.38	0.036
ECS	0.42	0.006	0.45	0.003

than a good one is 0.68 (as opposed to 0.30 for the probability of the opposite event), the probability that a good paper collects more citations than an acceptable one is 0.64 (as opposed to 0.33), and the probability that an acceptable paper harvests more citations than a limited one is 0.59 (as opposed to 0.35). Furthermore, in 88% of the cases a paper rated excellent receives more citations than a paper judged limited, while in only 10% of the cases the opposite happens (2% of the times the two papers receive the same number of citations).

3.2. Analysis at the level of research structures

In this section we investigate the structure rankings within each discipline compiled with respect to peer review judgements and bibliometric indicators. For the sake of statistical significance, for each discipline, we included in this analysis only research entities that submitted at least 10 products belonging to the discipline. For each structure in each discipline we compute the following ratings:

- *peer review rating*: this is the average peer review judgement of the products submitted by the structure; we also consider the peer review judgement restricted to TR articles;
- *article citation rating*: this is the average number of citations received by TR articles submitted by the structure;
- *journal citation rating*: this is the average impact factor of journals that published the TR articles submitted by the structure.

Universities were allowed to submit a maximum number of products corresponding to (only) 25% of the three-year average permanent academic staff. Research institutions were partitioned according to the number of submitted products in mega structures (over 74 submitted products), large structures (from 25 to 74 products), medium structures (from 10 to 24 products), and small structures (less than 10 products). Except for mega structures, the numbers of submitted products are, in general, not sufficient for a reliable computation, at the structure level, of the h index, whose score, by definition, is bounded by the number of papers in the evaluation set. For this reason, we do not consider the h index in the present analysis at the level of research structures.

We performed a rank-order correlation analysis to compare the structure compilations according to peer review and bibliometric ratings. We tested the hypothesis that the Spearman correlation coefficient is different from null and, when it holds, we investigated the strength of the correlation. Table 2 gives the main outcomes for the analysis. The used peer rating refers to TR articles only. The outcomes are interpreted as follows using guidelines suggested in Kraemer et al. (2003):

- There is an overall positive correlation between peer rating and article citation rating at the structure level. Areas PHY and EAS show a correlation coefficient larger than or equal to 0.70 (much larger than typical), while areas CHE, BIO, MED, AVM, and IIE have a correlation strength between 0.50 and 0.70 (larger than typical). Finally, for disciplines MCS, CEA, and ECS the correlation is between 0.30 and 0.50 (typical).
- The correlation between peer rating and journal citation rating is weaker: CHE and BIO have a correlation coefficient larger than or equal to 0.70 (much larger than typical), while MCS, MED, and AVM have a correlation strength between 0.50 and 0.70 (larger than typical). Areas EAS, CEA, IIE, and ECS show a typical correlation coefficient (between 0.30 and 0.50), while PHY has a smaller than typical correlation strength (below 0.30). Interestingly, PHY is the discipline with the largest correlation between peer judgements and article citations as well as the area with the smallest correlation between peer assessments and journal impact factors.

By way of example, Fig. 2 contains a rank plot comparing peer and article citation ratings for structures in research area PHY (35 structures that submitted at least 10 products). In general, the structure rank in the citation compilation increases as the structure rank in the peer compilation rises. The median change of rank is 4 (11% of the compilation length). Peer review, compared to citation rating, mostly favors structures Milano (+15 positions with respect to the citation compilation), Trento

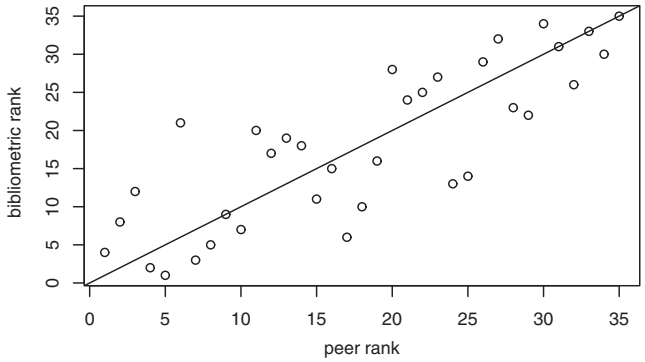

Fig. 2. Rank plot comparing peer and article citation ratings for structures in research area PHY. For each structure, the rank of the structure according to peer rating is plotted against the structure rank according to article citation rating. Peer rating favors structures above the solid bisector line and hampers those below, while those on the line do not change their ranks in the two compilations.

(+9), L'Aquila (+9), and Roma Tre (+8). On the other hand, structures that are most advantaged by using the bibliometric ranking are Genova (+11 positions with respect to the peer compilation), Pavia (+11), and Trieste (+11). Institutions Roma La Sapienza, Palermo, ENEA, and Parma do not change their positions in the two listings.

3.3. Discussion

Bibliometric assessment of research performance using citation-centric indicators is based on the central assumption that citations represent the acknowledge of intellectual debt and the witness of the use of ideas. Assuming this hypothesis, citations can be used as an indicator of the impact of academic publications. Nevertheless, citations might reflect different factors, some of them are related to the subjective needs and idiosyncrasies of the citer (MacRoberts & MacRoberts, 1989). For instance, we are aware that multi-authored publications tend to attack more citations, in particular if the authors belong to different institutions and countries (Franceschet & Costantini, 2010). In our analysis, we did not control for such factors that might potentially distort the use of citations as a pure indicator of impact.

Furthermore, citation rates vary significantly from field to field, since different disciplines generally have different publication and citation practices, including the average number of authors per paper and the average number of references per paper. Our association analyses are performed within disciplines, and not across them, and hence the outcomes are not biased by the field effect in bibliometrics. Citation rates, however, might moderately vary also across different sub-fields of the same discipline (Bornmann & Daniel, 2009; Radicchi, Fortunato, & Castellano, 2008). Hence, the calculated correlations between peer ratings and bibliometric indicators could be biased by the different citation potentials within different sub-fields. To control this possible bias, we performed the following additional experiment. We normalized citations with respect to the main sub-fields in three large disciplines, namely chemistry, biology, and economics and statistics. We used two recently proposed normalization methods: the relative citation indicators $cf = c/\mu$ (Radicchi et al., 2008) and $z\text{-}score = (c - \mu)/\sigma$ (Bornmann & Daniel, 2009), where c is the number of article citations, μ is the average number of citations for all papers in the sub-field of the article, and σ is the sub-field standard deviation. Then, we performed the association analyses between peer judgements and article citation ratings on the normalized citations scores. The results are summarized in Tables 7–9 in Appendix A. We noticed no significant discrepancies between the computed correlations in the normalized and unnormalized cases, as well as between the two normalization methods. We conclude that, within our dataset, the sub-field effect in bibliometrics does not bias the correlation of citation indicators with peer review judgements.

4. Related work

The most famous and discussed European national research evaluation is the research assessment exercise (RAE) in Great Britain, which is a peer review evaluation that started in 1986. For RAE 2008, research structures were invited to submit four research products for each full-time researcher (as opposed to one product every four researchers in the Italian VTR). RAE will be replaced with the Research Excellence Framework (REF), which is the new exercise for assessing research in UK higher education institution. The REF will be a process of expert review, informed by citation-based indicators where appropriate. The quality of research outputs will be assessed by panels of experts considering the international standards of excellence. Each sub-panel will decide whether to use citation information to inform its review of outputs, eventually explaining clearly how it will use bibliometrics in advance of submissions. A pilot exercise was conducted in 2009 by the Higher Education Funding Council for England (HEFCE) to test the potential use of citation analysis in the REF, concluding that bibliometrics are not sufficiently mature to be used formulaically or to replace expert review, but it can be used to inform and supplement the review in some science-based disciplines. There is general consensus that the use of citation

analysis is not appropriate in the arts, humanities and most of the social sciences, and some suggested that their possible use should not even be considered by panels in these fields. REF is due to be completed in 2014.

In the US, evidence suggests that publication and citation metrics are more readily accepted and more liberally applied. Cronin (1996) cites the following example to illustrate the greater tolerance of evaluative bibliometrics in North America:

> In a recent legal action initiated by a female assistant professor of biology, who had been denied tenure at Vassar College, the plaintiffs lawyer brought forward as evidence of discrimination the fact that her untenured client had a higher citation count than some tenured male staff in the same department. Although the female candidate's case was overturned subsequently on appeal (in part, and ironically, as a result of errors in the citation data submitted as evidence), the legal admissability and potential courtroom impact of citations are worthy of note.

In Australia research output will be assessed conducting an innovative exercise, the Excellence in Research for Australia (ERA), which has been recently launched, in which citation analysis will play a significant role. In ERA 2010 evaluation is carried out by the Research Evaluation Committees in eight discipline clusters. Unit of evaluation is the field of research by institution, and fields of research have been defined by the Australian and New Zealand Standard Research Classification. Evaluation will be informed by a dashboard of indicators, including lists of journals ranked in four quality tiers, citation analysis, peer review, and peer-reviewed Australian and international research income. In particular, citation analysis will be used for science-based fields of research (six discipline clusters) with the exception of subsectors as "Pure mathematics" and "Built environment & design", while peer review will be used for humanities and creative arts and for social sciences (two clusters) with the exception of psychology domains. Citation analysis of papers will consist of relative citation impact against world and Australian benchmarks and percentile analysis.

The literature offers more than a few contributions dedicated to the comparison of peer review and bibliometric evaluation methodologies. The following is a (necessarily incomplete) selection. See Bornmann (2011) for a recent overview on studies dealing with the relationship between peer ratings and bibliometric indicators.

The use of citation metrics in place of, or as a supplement to, the UK RAE has been considered extensively. For instance, Oppenheim and Norris (2003) observe a statistically significant correlation between the 2001 RAE result and citation counts for archeology, and contains references to other studies that have found positive associations for other fields and exercises. Butler and McAllister (2009), using data from the 4400 submissions to the RAE 2001 political science panel, obtained OLS regression estimates suggesting that citations are the most important component in predicting the outcome of the RAE. Their discussion focused on the cost saving derivable from a metric-based model, adopting objective and transparent indicators.

van Raan (2006) investigates the statistical correlation between different bibliometric indicators, including the *h* index and the 'crown indicator' (a citation average normalized to world average, a measure developed and implemented by the author's group at Leiden) with peer review judgement for university chemistry research groups in the Netherlands. Results show that the *h* index and the crown indicator both relate in a quite comparable way with peer judgements. In particular, both indicators discriminate very well between highly rated groups and poorly rated ones, but less well between good and excellent judgements.

Bornmann and Daniel (2007a) investigate the convergent validity of decisions for awarding long-term fellowships to post-doctoral researchers as practiced by the Boehringer Ingelheim Fonds – an international foundation for the promotion of basic research in biomedicine – by using the *h* index. Grant and fellowship peer review is principally an evaluation of the potential of the proposed research. The *h* indices of approved applicants are on average consistently higher than those of rejected applicants. Nevertheless, the distributions of the *h* indices partly overlap: some rejected applicants have a *h* index that is substantially higher than that of approved applicants, and some approved applicants have a *h* index that is substantially lower than that of rejected applicants.

Rinia, van Leuween, van Vuren, and van Raan (1998) study the correlation between bibliometric indicators and the outcomes of peer judgements of research programmes made by expert committees of condensed matter physics in the Netherlands. In particular, a breakdown of correlations to the level of different peer review criteria has been made. The authors draw a number of interesting conclusions. Positive and significant but no perfect correlations are found between a number of bibliometric indicators (in particular average number of citations per publication and the above mentioned crown indicator) and peer judgements of research programmes. The impact of publication journals, as reflected by the mean journal citation rates, does not correlate well with the quality of these programmes as perceived by peers. A negative correlation is found between the percentage of self-citations and jury ratings. Correlations between bibliometric indicators and expert judgements are higher in the case of 'curiosity driven' basic research than in the case of 'application driven' research. Finally, at the level of specific criteria used by juries, the highest correlation is found between ratings for bibliometric indicators and the criterion 'team' – the assessment of the competency of researchers and of the research team.

Asknes and Taxt (2004) investigate the relationship between bibliometric indicators and the outcomes of peer reviews based on a case study of research groups within the natural sciences at the University of Bergen, Norway. The analysis shows positive but relatively weak correlations. Groups obtaining the highest citedness indices were all rated as very good or excellent. On the other hand, groups cited below the world average obtained rather heterogeneous ratings. The authors conclude that peer review and bibliometric methods should be used in combination. In particular, in cases where there is a significant deviation between the two evaluation outcomes, the panel should investigate the reasons for these discrepancies.

The preceding comparisons are limited to only a few disciplinary sectors or to just one sector, or even to a single institution. By contrast, our investigation spans over 10 disciplinary areas in the sciences and social sciences and involves the output

of more than a hundred public and private research structures. Two contributions mostly relate to ours. Reale et al. (2007) analyze the output of Italian VTR for four areas: chemistry, biology, economics and humanities. The authors find a general consensus between expert advice (but weaker in economics) and show that peer review was not biased toward prestige of institutions or reputation of scientists. On the other hand, they notice a bias linked to the interdisciplinary (non-conventional) research. Furthermore, the authors perform a Spearman correlation analysis as well as an ordinal regression one to compare peer judgements of papers with the impact factor of journals publishing the papers for chemistry, biology, and economics areas. They find a statistically significant association, although not strong, and conclude that "*this reinforces the idea that impact factor is a good predictor of the quality of journals – not for the quality of articles published in a particular journal*". Finally, they suggest that "*further developments of VTR should go toward a larger use of the bibliometric indicators, in conjunction with peer review*".

Abramo, D'Angelo, and Caprasecca (2009) provide a broader investigation on Italian VTR outcomes for eight disciplines, the ten disciplines we have used in our study with the exclusion of civil engineering and architecture (CEA) and economics and statistics (ECS), for which the database coverage is less important. The authors correlate, at the research structure level, peer quality opinions on papers with metrics based on the impact factor of the journals publishing the papers, normalized across scientific disciplinary sectors within disciplinary areas. They conclude that the two evaluation methods (peer review and bibliometrics) significantly overlap for the surveyed fields, and that "*bibliometrics currently offer levels of potential and methodological maturity that should induce a reconsideration and revision of their role*." Furthermore, the study shows that, with the benefit of hindsight, Italian universities, in the main, did not identify and submit for evaluation their best publications in terms of citational impact. Finally, the authors give evidence that research structures indicated as being of top quality by VTR are not necessarily also the most productive ones.

The main difference between the two mentioned previous studies and ours is the set of bibliometric indicators we have contrasted to peer judgements. Besides the journal impact factor, we used the number of citations collected by individual papers, which directly relates to the potential impact of papers, and not to that of publishing sources, as well as the *h* index for relatively large publication sets. It is worth remembering that the journal impact factor was conceived as a measure of journal status, and not of impact of single papers published within it (see Garfield, 2006; Pendlebury, 2009 for recent additions to this incessant debate). In particular, citation distributions considered in the computation of journal impact factors are always severely skewed, meaning that the majority of the papers in the journal are cited much less than the mean represented by the impact factor (Campbell, 2008; Seglen, 1992). Furthermore, we provided investigation both at the level of research disciplines (the ratings of papers) and at the level of research structures (the ratings of institutions submitting the papers). We included in the analysis also civil engineering as well as economics and statistics, for the not irrelevant fractions of submitted products that are covered by Thomson Reuters data sources. Finally, we performed different types of correlation analysis, including an intuitive probabilistic investigation.

Finally, in Franceschet and Costantini (2010) we study, using the same dataset of this paper, how scholar collaboration varies across disciplines in science, social science, arts and humanities as well as the effects of author collaboration on impact and quality of co-authored papers. We observe that collaboration intensity neatly varies across disciplines and we measure a general positive association between cardinality of the author set of a paper and citation count as well as peer quality of the contribution. There exist, however, notable and interesting counter-examples.

5. Conclusion

We recall the research questions posed in Section 1 and we propose answers based on the current investigation of the Italian research system:

1. *Are peer review judgements and (article and journal) bibliometric indicators independent variables?*
   Both article citation and journal impact are *not* independent of peer review assessment, but the correlation is positive in both cases: the higher the peer review opinion on a paper, the higher the number of citations that the paper and the publishing journal receive. Furthermore, the recently proposed *h* index appears to discriminate very well between sets of papers assessed with different peer judgements. It might be a viable indicator of the impact of research structures in the next editions of the evaluation exercise as soon as the average number of submitted products per structure significantly increases.
2. *What is the strength of the association?*
   The correlation strength between peer assessment and bibliometric indicators is statistically significant, although not perfect. Moreover, the strength of the association varies across disciplines, and it also depends on the discipline internal coverage of the used bibliometric database (the higher the discipline coverage, the higher the reliability of citation measures). Notwithstanding, the skeptical has at disposal a few examples of papers that receive a positive peer judgement but do not collect a significant number of citations or that even sleep uncited (van Raan, 2004). Furthermore, there are papers that obtain a poor judgement from peers but that rally when citations are taken into account. Even more exceptions are available when comparing peer conclusions and impact factors of journals. Nevertheless, using words of Moed (2005), a methodology, even if provides invalid outcomes in individual cases, may be beneficial to the scholarly system as a whole.
3. *Is the association between peer judgement and article citation rating significantly stronger than the association between peer judgement and journal citation rating?*

A somewhat surprising finding of the present investigation is that the difference between the correlation strengths of article citation and journal impact factor with respect to peer assessment, although perceivable, is not as strong as one might expect.[6] It is worth noticing that, during the evaluation process, peer reviewers had access to the impact factors of journals that published the assessed papers, but they did not have enough information about the number of citations collected by the evaluated papers, since most of these citations were not yet mature at the time of reviewing. Therefore, peer quality opinions cannot be biased toward highly cited papers and the association between peer review and article citation is authentic.

It is worth observing that, as already pointed out by Asknes and Taxt (2004), peer judgements and bibliometric performance measures can be expected to be positively correlated only if the aspects assessed by the peers correspond to those reflected through bibliometric indicators. The notion of *quality* assessed during peer review is perceived as a broad concept with different aspects; some of these aspects, but not necessarily all, are captured by bibliometrics. Moreover, different bibliometric measures reflect different aspects of quality, for instance, productivity, popularity, and prestige (Franceschet, 2010b).

In summary, we found a compelling body of evidence that judgements given by domain experts and bibliometric indicators are significantly positively correlated. Therefore, bibliometric indicators may be considered as *approximation measures* of the inherent quality of papers, which, however, remains fully assessable only with aid of human unbiased judgement, meditation, and elaboration. We advocate the integration of peer review with bibliometric indicators, in particular those directly related to the impact of individual articles, during the next national assessment exercises. The cost effectiveness of bibliometric evaluation compared to that of peer review would allow the evaluation of a larger sample of the universe under investigation without significant increase of costs, which is a major requirement due to the chronic national deficit and the pressing necessity of controlling public expenses in Italy.[7] This would allow a shift from the assessment of research excellence to a more balanced evaluation of average research performance. Larger samples would, in turn, enhance the reliability of bibliometric indicators.

Acknowledgements

The authors would like to thank CIVR and its President, Prof. Franco Cuccurullo, for making available data used in this paper, through the agreement protocol between CIVR and PhD course in “Strumenti e metodi per la valutazione della Ricerca” of the University of Chieti-Pescara. The first author is partially supported by PRIN 2008 project “Innovative and multi-disciplinary approaches for constraint and preference reasoning” (20089M932N).

Appendix A.

See Tables 3–9 .

Table 3
Peer judgement and bibliometric indicators. Rating: peer review rating (E = Excellent, G = Good, A = Acceptable, L = Limited), size: number of products with the given peer rating (with percentage with respect to all products), cites: average number of citations of articles with the given peer rating (with ratio with respect to the average over all articles), IF: average impact factor of journals of articles with the given peer rating (with ratio with respect to the average over all articles), *h*: *h* index of articles with the given peer rating (with ratio with respect to the index over all articles).

Rating	Size	Cites	IF	*h*
		Part I		
MCS				
E	284 (36%)	5.52 (1.39)	1.15 (1.02)	16 (0.89)
G	381 (48%)	3.31 (0.83)	1.10 (0.98)	13 (0.72)
A	101 (13%)	2.61 (0.66)	1.18 (1.05)	7 (0.39)
L	21 (3%)	2.18 (0.55)	0.91 (0.81)	3 (0.17)
PHY				
E	914 (52%)	35.30 (1.43)	6.97 (1.20)	85 (0.98)
G	676 (38%)	14.19 (0.58)	4.71 (0.81)	40 (0.46)
A	158 (9%)	5.98 (0.24)	3.10 (0.54)	14 (0.16)
L	19 (1%)	5.69 (0.23)	5.59 (0.97)	7 (0.08)
CHE				
E	342 (32%)	24.72 (1.53)	6.84 (1.32)	42 (0.84)
G	513 (47%)	13.54 (0.84)	4.67 (0.91)	34 (0.68)
A	200 (18%)	8.84 (0.55)	3.57 (0.69)	18 (0.36)
L	34 (3%)	7.57 (0.47)	3.47 (0.67)	9 (0.18)

[6] As noticed above, the journal impact factor is a measure of journal status and not of the impact of individual papers published in the journal.
[7] To be sure, the cost of bibliometric evaluation is lower than that of peer review; nonetheless, every experienced bibliometrician knows that the cost to produce a reliable large-scale bibliometric assessment is far from null.

Table 3 (*Continued*)

Rating	Size	Cites	IF	*h*
EAS				
E	220 (34%)	10.37 (1.42)	4.13 (1.37)	22 (0.85)
G	324 (50%)	6.10 (0.83)	2.39 (0.79)	18 (0.69)
A	91 (14%)	4.12 (0.56)	2.60 (0.87)	9 (0.35)
L	16 (2%)	6.37 (0.87)	2.16 (0.72)	4 (0.15)
BIO				
E	519 (33%)	40.86 (1.66)	12.01 (1.42)	75 (0.90)
G	802 (51%)	17.97 (0.73)	7.09 (0.84)	49 (0.59)
A	222 (14%)	11.59 (0.47)	5.47 (0.64)	21 (0.25)
L	32 (2%)	5.65 (0.23)	5.02 (0.59)	6 (0.07)
		Part II		
MED				
E	667 (25%)	47.72 (1.79)	11.73 (1.41)	89 (0.84)
G	1314 (50%)	21.98 (0.82)	7.49 (0.90)	65 (0.61)
A	492 (19%)	13.83 (0.52)	6.17 (0.74)	36 (0.34)
L	166 (6%)	14.72 (0.55)	7.67 (0.92)	21 (0.20)
AVM				
E	76 (10%)	16.54 (2.02)	6.41 (2.41)	18 (0.67)
G	393 (52%)	8.67 (1.06)	2.54 (0.96)	24 (0.89)
A	218 (29%)	5.15 (0.62)	1.77 (0.66)	15 (0.56)
L	63 (9%)	3.21 (0.39)	1.28 (0.48)	6 (0.22)
CEA				
E	166 (22%)	5.43 (1.52)	1.88 (1.62)	11 (0.79)
G	329 (43%)	3.58 (1.00)	1.04 (0.90)	10 (0.71)
A	217 (29%)	2.29 (0.64)	0.80 (0.70)	7 (0.50)
L	46 (6%)	2.50 (0.70)	0.81 (0.70)	4 (0.29)
IIE				
E	248 (21%)	7.16 (1.50)	2.03 (1.27)	19 (0.83)
G	612 (51%)	4.57 (0.96)	1.56 (0.97)	18 (0.78)
A	300 (25%)	3.18 (0.67)	1.33 (0.83)	11 (0.49)
L	35 (3%)	4.74 (0.99)	1.65 (1.03)	5 (0.22)
ECS				
E	168 (17%)	5.55 (1.76)	1.31 (1.51)	14 (0.82)
G	365 (38%)	2.77 (0.88)	0.75 (0.87)	11 (0.65)
A	265 (27%)	1.06 (0.33)	0.58 (0.67)	4 (0.24)
L	173 (18%)	0.67 (0.21)	0.48 (0.56)	2 (0.12)

Table 4
Contingency table displaying the conditional distribution of article citation given peer rating. Peer judgments are abbreviated as follows: E (Excellent), G (Good), A (Acceptable), L (Limited). The Pearson χ^2 test statistic and p-value are also shown.

Rating	1st quartile	2nd quartile	3rd quartile	4th quartile	Total
			Part I		
MCS					
E	27.3%	13.3%	27.3%	32.1%	100%
G	40.0%	15.7%	24.9%	19.4%	100%
A	48.9%	18.1%	21.3%	11.7%	100%
L	47.0%	35.3%	5.9%	11.8%	100%
χ^2 (9, N = 717) = 37.85, p-value = 0.000					
PHY					
E	16.1%	20.6%	25.7%	37.6%	100%
G	36.7%	29.2%	22.2%	11.9%	100%
A	63.4%	25.2%	9.1%	2.3%	100%
L	46.2%	53.8%	0.0%	0.0%	100%
χ^2 (9, N = 1566) = 274.83, p-value = 0.000					

Table 4 (*Continued*)

Rating	1st quartile	2nd quartile	3rd quartile	4th quartile	Total
CHE					
E	14.3%	16.2%	25.9%	43.6%	100%
G	35.2%	22.2%	22.8%	19.8%	100%
A	42.4%	27.3%	23.8%	6.5%	100%
L	57.1%	17.9%	21.4%	3.6%	100%
χ^2 (9, N = 1007) = 133.74, p-value = 0.000					
EAS					
E	17.0%	24.0%	21.5%	37.5%	100%
G	33.5%	27.0%	20.4%	19.1%	100%
A	45.1%	29.6%	12.7%	12.6%	100%
L	37.5%	12.5%	25.0%	25.0%	100%
χ^2 (9, N = 583) = 43.29, p-value = 0.000					
BIO					
E	11.3%	18.4%	25.3%	45.0%	100%
G	27.7%	28.7%	27.3%	16.3%	100%
A	50.2%	30.3%	11.0%	8.5%	100%
L	74.0%	13.0%	13.0%	0.0%	100%
χ^2 (9, N = 1513) = 278.46, p-value = 0.000					
			Part II		
MED					
E	12.5%	17.5%	25.4%	44.6%	100%
G	25.4%	25.2%	28.3%	21.1%	100%
A	43.4%	25.4%	19.4%	11.8%	100%
L	60.5%	17.7%	14.3%	7.5%	100%
χ^2 (9, N = 2541) = 349.79, p-value = 0.000					
AVM					
E	8.3%	15.3%	26.4%	50.0%	100%
G	24.7%	25.2%	23.8%	26.3%	100%
A	38.4%	27.0%	22.2%	12.4%	100%
L	52.3%	26.2%	16.7%	4.8%	100%
χ^2 (9, N = 664) = 69.65, p-value = 0.000					
CEA					
E	18.5%	13.6%	30.9%	37.0%	100%
G	39.9%	14.0%	23.1%	23.0%	100%
A	54.1%	14.3%	21.4%	10.2%	100%
L	55.0%	15.0%	10.0%	20.0%	100%
χ^2 (9, N = 342) = 33.28, p-value = 0.000					
IIE					
E	25.5%	16.8%	25.0%	32.7%	100%
G	31.9%	24.1%	21.0%	23.0%	100%
A	44.6%	23.8%	17.1%	14.5%	100%
L	52.2%	17.4%	8.7%	21.7%	100%
χ^2 (9, N = 985) = 40.17, p-value = 0.000					
ECS					
E	14.7%	16.7%	27.3%	41.3%	100%
G	33.7%	21.5%	23.6%	21.2%	100%
A	54.7%	15.7%	21.3%	8.3%	100%
L	73.3%	6.7%	13.3%	6.7%	100%
χ^2 (9, N = 519) = 74.66, p-value = 0.000					

Table 5
Contingency table displaying the conditional distribution of journal impact factor given peer rating. Peer judgments are abbreviated as follows: E (Excellent), G (Good), A (Acceptable), L (Limited). The Pearson χ^2 test statistic and p-value are also shown.

Rating	1st quartile	2nd quartile	3rd quartile	4th quartile	Total
			Part I		
MCS					
E	20.6%	26.2%	27.0%	26.2%	100%
G	25.5%	26.6%	24.4%	23.5%	100%
A	34.7%	18.9%	22.2%	24.2%	100%
L	29.4%	41.2%	11.8%	17.6%	100%
χ^2 (9, N = 717) = 12.51, p-value = 0.186					
PHY					
E	15.9%	21.3%	28.4%	34.4%	100%
G	31.7%	31.2%	25.6%	11.5%	100%
A	54.2%	31.3%	9.2%	5.3%	100%
L	38.5%	28.5%	7.7%	15.3%	100%
χ^2 (9, N = 1566) = 214.49, p-value = 0.000					
CHE					
E	8.4%	16.1%	34.6%	40.9%	100%
G	26.5%	32.9%	24.1%	16.5%	100%
A	48.3%	31.4%	18.6%	1.7%	100%
L	53.6%	25.0%	17.9%	3.5%	100%
χ^2 (9, N = 1007) = 212.87, p-value = 0.000					
EAS					
E	16.5%	25.5%	22.5%	35.5%	100%
G	28.6%	21.7%	29.6%	20.1%	100%
A	45.2%	24.7%	17.8%	12.3%	100%
L	25.0%	37.5%	25.0%	12.5%	100%
χ^2 (9, N = 583) = 41.05, p-value = 0.000					
BIO					
E	7.5%	16.8%	26.2%	49.5%	100%
G	26.4%	31.0%	27.5%	14.1%	100%
A	56.2%	25.9%	11.9%	6.0%	100%
L	60.9%	21.7%	8.7%	8.7%	100%
χ^2 (9, N = 1513) = 387.54, p-value = 0.000					
			Part II		
MED					
E	6.8%	17.3%	26.7%	49.2%	100%
G	24.0%	29.0%	29.0%	18.0%	100%
A	45.5%	29.4%	14.9%	10.2%	100%
L	51.0%	12.9%	12.3%	23.8%	100%
χ^2 (9, N = 2536) = 499.31, p-value = 0.000					
AVM					
E	1.4%	2.8%	19.4%	76.4%	100%
G	15.8%	25.9%	31.6%	26.7%	100%
A	42.0%	33.6%	17.0%	7.4%	100%
L	70.5%	20.4%	9.1%	0.0%	100%
χ^2 (9, N = 664) = 225.98, p-value = 0.000					
CEA					
E	11.1%	7.4%	27.2%	54.3%	100%
G	25.0%	24.3%	32.6%	18.1%	100%
A	34.7%	38.7%	18.4%	8.2%	100%
L	38.1%	28.6%	23.8%	9.5%	100%
χ^2 (9, N = 342) = 79.00, p-value = 0.000					
IIE					
E	13.5%	24.0%	25.5%	37.0%	100%
G	25.5%	23.6%	26.3%	24.6%	100%
A	33.6%	29.5%	21.5%	15.4%	100%
L	39.1%	8.7%	26.1%	26.1%	100%
χ^2 (9, N = 985) = 46.55, p-value = 0.000					
ECS					
E	5.3%	14.0%	34.7%	46.0%	100%
G	27.2%	27.2%	27.2%	18.4%	100%
A	25.5%	32.7%	8.2%	13.6%	100%
L	53.3%	33.3%	6.7%	6.7%	100%
χ^2 (9, N = 519) = 115.03, p-value = 0.000					

Table 6
Probability analysis of peer judgement and bibliometric indicators. For each pair of adjacent peer ratings, we compute probabilities $P(c(X) > c(Y))$, $P(c(X) < c(Y))$, $P(c(X) = c(Y))$ and $P(IF(X) > IF(Y))$, $P(IF(X) < IF(Y))$, $P(IF(X) = IF(Y))$. Peer judgments are abbreviated as follows: E (Excellent), G (Good), A (Acceptable), L (Limited).

Ratings	Cites >	Cites <	Cites =	IF >	IF <	IF =
			Part I			
MCS						
E–G	0.54	0.35	0.11	0.55	0.45	0.00
G–A	0.49	0.36	0.15	0.53	0.47	0.00
A–L	0.40	0.41	0.19	0.57	0.43	0.00
PHY						
E–G	0.69	0.29	0.02	0.66	0.32	0.02
G–A	0.66	0.29	0.05	0.66	0.33	0.01
A–L	0.40	0.53	0.07	0.36	0.64	0.00
CHE						
E–G	0.67	0.31	0.02	0.71	0.27	0.02
G–A	0.57	0.39	0.04	0.68	0.31	0.01
A–L	0.53	0.41	0.06	0.56	0.43	0.01
EAS						
E–G	0.61	0.33	0.06	0.60	0.38	0.02
G–A	0.56	0.35	0.09	0.61	0.38	0.01
A–L	0.34	0.57	0.09	0.41	0.57	0.02
BIO						
E–G	0.68	0.30	0.02	0.74	0.25	0.01
G–A	0.64	0.33	0.03	0.68	0.32	0.00
A–L	0.59	0.35	0.06	0.57	0.43	0.00
			Part II			
MED						
E–G	0.65	0.33	0.02	0.72	0.28	0.00
G–A	0.61	0.36	0.03	0.65	0.34	0.01
A–L	0.60	0.35	0.05	0.51	0.49	0.00
AVM						
E–G	0.67	0.29	0.04	0.84	0.16	0.00
G–A	0.58	0.35	0.07	0.72	0.28	0.00
A–L	0.55	0.35	0.10	0.69	0.31	0.00
CEA						
E–G	0.60	0.31	0.09	0.72	0.27	0.01
G–A	0.54	0.31	0.15	0.62	0.37	0.01
A–L	0.39	0.41	0.20	0.48	0.52	0.00
IIE						
E–G	0.54	0.38	0.08	0.59	0.41	0.00
G–A	0.53	0.36	0.11	0.69	0.41	0.00
A–L	0.45	0.40	0.15	0.44	0.56	0.00
ECS						
E–G	0.59	0.28	0.13	0.75	0.25	0.00
G–A	0.50	0.25	0.25	0.63	0.37	0.00
A–L	0.37	0.20	0.43	0.55	0.44	0.01

Table 7
Peer judgement and sub-field normalized citations. Rating: peer review rating (E = Excellent, G = Good, A = Acceptable, L = Limited), size: number of products in the main sub-fields of the discipline that received the given peer rating (with percentage with respect to all products), cf: relative citation indicator as in Radicchi et al. (2008), *z*-score: relative citation indicator as in Bornmann and Daniel (2009).

Rating	Size	cf	*z*-Score
BIO			
E	437 (36%)	1.54	0.41
G	640 (52%)	0.74	−0.19
A	141 (11%)	0.56	−0.33
L	13 (1%)	0.22	−0.56
CHE			
E	254 (34%)	1.51	0.42
G	342 (46%)	0.81	−0.15
A	124 (17%)	0.58	−0.35
L	21 (3%)	0.43	−0.53
ECS			
E	142 (30%)	1.72	0.46
G	217 (46%)	0.88	−0.08
A	100 (21%)	0.34	−0.42
L	14 (3%)	0.29	−0.46

Table 8
Contingency table displaying the conditional distribution of the normalized article citation according to the relative citation indicator of Radicchi et al. (2008). Peer judgments are abbreviated as follows: E (Excellent), G (Good), A (Acceptable), L (Limited).

Rating	1st quartile	2nd quartile	3rd quartile	4th quartile	Total
BIO					
E	11.0%	17.8%	27.2%	44.0%	100%
G	28.3%	29.2%	26.3%	16.2%	100%
A	51.8%	28.4%	11.3%	8.5%	100%
L	76.9%	15.4%	7.7%	0.0%	100%
χ^2 (9, N = 1231) = 219.13, p-value = 0.000					
CHE					
E	11.0%	18.9%	25.6%	44.5%	100%
G	31.0%	27.8%	23.1%	18.1%	100%
A	37.9%	31.5%	22.5%	8.1%	100%
L	57.2%	19.0%	23.8%	0.0%	100%
χ^2 (9, N = 741) = 111.98, p-value = 0.000					
ECS					
E	15.5%	16.9%	23.9%	43.7%	100%
G	34.1%	19.4%	23.0%	23.5%	100%
A	55.0%	13.0%	29.0%	3.0%	100%
L	71.4%	7.1%	14.4%	7.1%	100%
χ^2 (9, N = 473) = 78.65, p-value = 0.000					

Table 9
Contingency table displaying the conditional distribution of the normalized article citation according to the relative citation indicator of Bornmann and Daniel (2009). Peer judgments are abbreviated as follows: E (Excellent), G (Good), A (Acceptable), L (Limited).

Rating	1st quartile	2nd quartile	3rd quartile	4th quartile	Total
BIO					
E	11.7%	18.5%	25.9%	43.9%	100%
G	27.7%	28.0%	28.0%	16.3%	100%
A	50.4%	31.9%	9.2%	8.5%	100%
L	69.2%	23.1%	7.7%	0.0%	100%
χ^2 (9, N = 1231) = 210.31, p-value = 0.000					
CHE					
E	15.3%	13.8%	27.2%	43.7%	100%
G	30.1%	28.4%	22.8%	18.7%	100%
A	31.5%	39.4%	21.0%	8.1%	100%
L	52.4%	19.0%	28.6%	0.0%	100%
χ^2 (9, N = 741) = 107.92, p-value = 0.000					
ECS					
E	15.5%	16.9%	24.6%	43.0%	100%
G	31.8%	25.3%	19.8%	23.0%	100%
A	52.0%	19.0%	26.0%	3.0%	100%
L	64.3%	14.3%	14.3%	7.1%	100%
χ^2 (9, N = 473) = 75.36, p-value = 0.000					

References

Abramo, G., D'Angelo, C. A., & Caprasecca, A. (2009). Allocative efficiency in public research funding: Can bibliometrics help? *Research Policy*, *38*(1), 206–215.

Alonso, S., Cabrerizo, F., Herrera-Viedma, E., & Herrera, F. (2009). *h*-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, *3*(4), 273–289.

Althouse, B. M., West, J. D., Bergstrom, C. T., & Bergstrom, T. (2008). Differences in impact factor across fields and over time. *Journal of the American Society for Information Science and Technology*, *60*(1), 27–34.

Asknes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: A comparative study at a Norwegian university. *Research Evaluation*, *13*(1), 33–41.

Ball, P. (2007). Achievement index climbs the ranks. *Nature*, *448*, 737.

Bleiklie, I. (1998). Justifying the evaluative state: New public management ideals in higher education. *European Journal of Education*, *33*(3), 299–318.

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

Bornmann, L., & Daniel, H.-D. (2007a). Convergent validation of peer review decisions using the *h* index: Extent of and reasons for type I and type II errors. *Journal of Informetrics*, *1*(3), 204–213.

Bornmann, L., & Daniel, H.-D. (2007b). What do we know about the *h* index? *Journal of the American Society for Information Science and Technology*, *58*(9), 1381–1385.

Bornmann, L., & Daniel, H.-D. (2009). Universality of citation distributions. A validation of Radicchi et al.'s relative indicator $cf = c/c_0$ at the micro level using data from chemistry. *Journal of the American Society for Information Science and Technology*, *60*(8), 1664–1670.

Butler, L., & McAllister, I. (2009). Metrics or peer review? Evaluating the 2001 UK research assessment exercise in political science. *Political Studies Review*, *7*(1), 3–17.

Calzà, L., & Garbisa, S. (1995). Italian professorships. *Nature*, *374*, 492.

Campbell, P. (2008). Escape from the impact factor. *Ethics in Science and Environmental Politics*, *8*, 5–7.

Chubin, D. E., & Hackett, E. J. (1990). Peer review in theory and practice. In *Peerless science: Peer review and U. S. science policy*. Albany: State University of New York Press., pp. 17–48

Cronin, B. (1996). Rates of return to citation. *Journal of Documentation*, *52*(2), 188–197.
Cronin, B. (2001). Hyperauthorship: A postmodern perversion or evidence of a structural shift in scholarly communication practices. *Journal of the American Society for Information Science and Technology*, *52*(7), 558–569.
Egghe, L. (2010). The Hirsch-index and related impact measures. *Annual Review of Information Science and Technology*, *44*, 65–114.
Egghe, L., & Rousseau, R. (2006). An informetric model for the Hirsch-index. *Scientometrics*, *69*(1), 121–129.
Fawcett, T. (2006). An introduction to roc analysis. *Pattern Recognition Letters*, *27*(8), 861–874.
Franceschet, M. (2010a). A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar. *Scientometrics*, *83*(1), 243–258.
Franceschet, M. (2010b). The difference between popularity and prestige in the sciences and in the social sciences: A bibliometric analysis. *Journal of Informetrics*, *4*(1), 55–63.
Franceschet, M. (2010c). The role of conference publications in computer science: A bibliometric view. *Communications of the ACM*, *53*(2), 129–132.
Franceschet, M., & Costantini, A. (2010). The effect of scholar collaboration on impact and quality of academic papers. *Journal of Informetrics*, *4*(4), 540–553.
Garfield, E. (2006). The history and meaning of the journal impact factor. *Journal of the American Medical Association*, *295*(1), 90–93.
Garfield, E., & Sher, H. (1963). New factors in the evaluation of scientific literature through citation indexing. *American Documentation*, *14*, 195–201.
Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of United States of America*, *102*(46), 16569–16572.
Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry*, *42*(12), 1524–1529.
MacRoberts, M. H., & MacRoberts, B. R. (1989). Problems of citation analysis: A critical review. *Journal of the American Society for Information Science*, *40*(5), 342–349.
Martin, B. R., & Irvine, J. (1983). Assessing basic research. Some partial indicators of scientific progress in radio astronomy. *Research Policy*, *12*(2), 61–90.
Minelli, E., Rebora, G., & Turri, M. (2008). The structure and significance of the Italian research assessment exercise (VTR). In C. Mazza, P. Quattrone, & A. Riccaboni (Eds.), *European universities in transition: Issues, models and cases* (pp. 221–236). Edward Elgar Publishing.
Moed, H. F. (2005). *Citation analysis in research evaluation*. Springer.
Neave, G. (1998). The evaluative state reconsidered. *European Journal of Education*, *33*(3), 265–285.
Oppenheim, C., & Norris, M. (2003). Citation counts and the research assessment exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, *56*(6), 709–730.
Pendlebury, D. A. (2009). The use and misuse of journal metrics and other citation indicators. *Archivum Immunologiae et Therapiae Experimentalis*, *57*(1), 1–11.
Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences of the United States of America*, *105*(45), 17268–17272.
Reale, E., Barbara, A., & Costantini, A. (2007). Peer review for the evaluation of academic research: Lessons from the Italian experience. *Research Evaluation*, *16*(3), 216–228.
Rinia, E. J., van Leuween, T. N., van Vuren, H. G., & van Raan, A. F. J. (1998). Comparative analysis of a set of bibliometric indicators and central peer review criteria. Evaluation of condensed matter physics in the Netherlands. *Research Policy*, *27*(1), 95–107.
Seglen, P. O. (1992). The skewness of science. *Journal of the American Society for Information Science*, *43*(9), 628–638.
Stringer, M. J., Sales-Pardo, M., & Amaral, L. A. N. (2008). Effectiveness of journal ranking schemes as a tool for locating information. *PLoS ONE*, *3*(2), e1683.
van Raan, A. F. J. (2004). Sleeping beauties in science. *Scientometrics*, *59*(3), 467–472.
van Raan, A. F. J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics*, *67*(3), 491–502.