Scientometrics (2013) 96:395–410
DOI 10.1007/s11192-013-0969-9

Correlations between bibliometrics and peer evaluation for all disciplines: the evaluation of Brazilian scientists

Jacques Wainer · Paula Vieira

Received: 15 September 2012 / Published online: 13 February 2013

Abstract This paper correlates the peer evaluations performed in late 2009 by the disciplinary committees of CNPq (a Brazilian funding agency) with some standard bibliometric measures for 55 scientific areas. We compared the decisions to increase, maintain or decrease a scientist's research scholarship funded by CNPq. We analyzed these decisions for 2,663 Brazilian scientists and computed their correlations (Spearman rho) with 21 different measures, among them: total production, production in the last 5 years, production indexed in Web of Science and Scopus, total citations received (according to WOS, Scopus, and Google Scholar), *h*-index and *m*-quotient (according to the three citation services). The highest correlations for each area range from 0.95 to 0.29, although there are areas with no significantly positive correlation with any of the metrics.

Keywords Scientist evaluation · Peer evaluation · Bibliometric measures · Brazil

Introduction

Science evaluation, and in particular scientists evaluation takes into consideration may aspects, including the production, the productivity, and the impact of the scientist's work. Many metrics have been proposed to measure these aspects. Standard metrics of production include: total number of articles written, total number of articles indexed in Web of Science (WOS), fractional count of the articles written, and so on. Productivity measures are the production measures calculated over a recent time interval, for example, total number of WOS indexed papers in the last 5 years. Impact measures usually refer to the citations received by the papers published by the scientist, such as, total number of citations received, average number of citations by year, and so on. Impact measures can also be evaluated over a recent time interval, such as number of citations received on papers published in the last 5 years. Finally, there is a plethora of hybrid measures that combine

J. Wainer (✉) · P. Vieira
Institute of Computing—UNICAMP, Av. Albert Einstein, 1251, Campinas, SP 13083-852, Brazil
e-mail: wainer@ic.unicamp.br

production (or more rarely productivity) with impact, such as the *h*-index, the *m*-quotient (Hirsch 2005) and variations (some of them reviewed by Alonso et al (2009)).

The gold standard for scientist evaluation is the peer evaluation. One would like to believe that peers can not only weight correctly the numerical, objective evidence regarding a scientist production, productivity, and impact, but also consider more intangible aspects, such as "potential" or "prestige".

This paper should be understood as a descriptive research: we will determine the correlation of many different *traditional* bibliometric measures to peer evaluations of scientists working in 55 different scientific fields. We do not aim at defining a new measure that better correlates with the peer evaluations, in one or many of the scientific fields. This research should be compared with others that measured the correlation of peer evaluation with bibliometric measures, such as Rinia et al (1998), van Raan (2006), Li et al (2010), and Franceschet and Costantini (2011). But these works focus on a single scientific field, with the exception of Franceschet and Costantini (2011) who deals with 10 different scientific fields. We compute the correlations for 55 different fields. Also, most of these works deal with the peer evaluations of research groups, while our work focus on the peer evaluation of the researcher himself.

Brazilian system of scientist evaluation

The Brazilian National Council for Scientific and Technological Development (CNPq) is one of the main research funding bodies in Brazil. One of CNPq programs is the *scholarship program*, which provide to each awarded scientist a small stipend. The scholarships are divided into levels named 2, 1D, 1C, 1B, and 1A. The levels 1D to 1A are collective known as level 1. Scientists on a level 2 scholarship receive a monthly, tax-free stipend of about US$500.00; level 1 scientists receive about double that value, with increasing values for higher level scholarships.

Each scholarship lasts for 3 years (for level 2), 4 years (level 1D to 1B), or 5 years (level 1A). Before one's scholarship ends, one must apply again for a "new" scholarship, in an yearly call for proposals (CFP). All scientists whose scholarships are ending and all scientists who do not hold a scholarship apply to that CFP. To apply one must write a standard research proposal but the most important component in evaluating the application is the scientist's curriculum vitae (CV), in a standardized, publicly available format called Lattes CV. The Lattes CV (discussed, for example, by Oliveira et al (2012) and Hicks (2011)) lists, among other things, all of a scientist's production, including books, books chapters, conference papers, journal papers, reports, patents, and so on. The scientist submits his or her proposal to one of 55 scientific areas, which are organized into 6 general domains (agricultural sciences, biological sciences, exact sciences, social sciences, health sciences, applied social sciences, engineering, and language studies). Each scientific area has a "disciplinary committee" (or CA) of scientists that work in the area, chosen from the research community. The CAs may ask other scientists to evaluate both the proposal and the candidate's CV, but ultimately it rests on the CA to make the decisions regarding assigning or not a scholarship to the candidate, and at which level.

The scholarship level is the only researcher evaluation system in Brazil[1] and the researcher level is used as the *de facto* quality evaluations of the researcher. For example, some research grants can only be applied for by scientist that have a level 1 scholarship; very large multi-institutional grants require the main investigator to have at least

[1] There is another evaluation system sponsored by CAPES for graduate programs.

a level 1B. Also, a faculty scholarship level is sometimes used by internal university committees as part of promotion evaluations.

But since the CNPq scholarship level is tied to receiving the scholarship grant, and because there are federal government constraints regarding the total scholarship grant budget, it cannot be seen as a "pure" evaluation system. For example, the budget for each CA is fixed, so to promote a scientists to a 1C level, the CA must demote one of the 1C scientist who applied for the CFP in that year. There are other considerations regarding the minimum numbers of years since receiving a doctoral degree to be eligible for the different level grants, and so on. Therefore, one cannot claim that the level a researcher holds is just the result of the researcher evaluation by her peers, but the result of peer evaluation, of history, of the evaluation of other scientists in the area in that same year, of budget constraints and so on. This is not unlike the 41st chair issue discussed by Merton (1968). For example, Oliveira et al (2012) showed that there are very few bibliometric measures that are significantly different for scientists with scholarships in Clinical Medicine. And in most cases, the significant differences are only present when comparing the bibliometric indicators of level 2 and level 1A scientists.

This research will not use the scientist's scholarship *level* as the result of the peer evaluations; we will use *changes* in the scientist's level, at a particular year as the result of the peer evaluation. Let us consider the scientists of a particular area that hold a level 1C, and that applied to renew their scholarship at a particular year. Some of them had their scholarship renewed at the same level, some dropped one or more levels, and some were raised to the 1B or higher levels. Clearly those who were promoted one or more levels were considered "better" than those that remained at level 1C, which in turn were considered "better" than those that were demoted one or more levels. We will use this ordering as the ordering that reflects the peer evaluation. There will be such an ordering for all scholarship levels. For level 2, we will state that those who were promoted to 1D are "better" than those that kept the level 2, which are better than those who lost the scholarship. At level 1A, the only ordering is that those who kept their 1A scholarships will be considered "better" than those who lost one or more levels. In this research, we will calculate the correlation of these orderings with standard bibliometric measures.

Related research

There has been some research on the correlation of peer evaluation and bibliometric measures. The research by van Raan (2006) compared the *h*-index and their own measure, the crown indicator, with the peer evaluation of 147 Dutch research groups in chemistry. van Raan (2006) did not provide the correlations, but a followup of that paper Waltman et al (2011) reported that the Spearman correlation for the crown indicator was 0.45.

Aksnes and Taxt (2004) studied 34 research groups in mathematics and natural sciences from the University of Bergen, Norway. They used Pearson correlation, and found, among others a correlation of 0.34 for the fractional number of papers per person, and a correlation of 0.31 for the number of citations per paper. The highest correlation was 0.57, for the impact-productivity factor (relative subfield citedness times the fractional number of papers per person). Aksnes and Taxt (2004) considered these correlations as "weak" but they believed that the limitations of the peer evaluation itself was a important reason for the low correlations.

Rinia et al (1998) compared some bibliometric measures and peer evaluation for 56 research groups in condensed matter physics in the Netherlands, divided into two categories: application-oriented and basic research groups. They correlated the peer evaluation of these

groups with 11 different bibliometric indicators, using Spearman rho. They found that for the application-oriented category, 7 correlations were significant (with 99 % confidence). The highest correlation in this category was 0.57. For the basic research category, there were 6 positive and one negative significant correlations; the highest correlation was 0.68.

Franceschet and Costantini (2011) studied the result of the Italian first national research evaluation, and correlated the peer evaluation of "research structures" (universities, groups) in 10 disciplinary areas with two bibliometric measures: publication citations rating and journal citation ratings. The significant correlations (95 % confidence) range from 0.42 to 0.81 (Spearman rho) for the article citation ratings, and from 0.38 to 0.85 for the journal citation ratings.

Li et al (2010) correlated *h*-index and other indicators and peer evaluation of 101 LIS researchers using data from WOS, Scopus, and Google Scholar. The Spearman rho for the *h*-index was 0.50, 0.51, and 0.46 for data from Google Scholar, Scopus, and WOS respectively.

Bornmann and Daniel (2005) found that the *h*-index for successful applicants for post-doctoral research fellowships (in the area of biomedicine) was consistently higher than for non-successful applicants. Bornmann et al (2008) proposed a new index (the *m*-index) and showed that it had a higher correlation to the post-doctoral fellowship decision than the *h*-index. The correlation was measured by Cramer's V, ranged from 0.83 to 0.97 (across different years) for the *m*-index, and from 0.32 to 0.61, for the *h*-index.

Oliveira et al (2012) described some bibliometric measures for all (411) Brazilian researchers with CNPq scholarship in medicine (in the period 2006–2008). Among the metrics collected were total number of publications, publications in the last 5 years, total number of citations received, citations per year of the scientific career, *h*-index and m quotient. Oliveira et al (2012) did not compute correlations between these indicators and the scholarship levels; instead, they analyzed the results in terms of significant differences: are the measures for one group significantly different that those of the other groups? In general, they found significant differences only between researchers in the extremes of the scale, that is, among the researchers with level 2 and those with level 1A.

There are a number of research on *publication* peer evaluation and its correlation to citation metrics (Patterson and Harris 2009; Reale et al 2007; Franceschet and Costantini 2011) or *journal* peer evaluation (Korevaar 1996; Franceschet and Costantini 2011), which will not be further discussed here.

Data and methods

We received from CNPq the list of all researchers that received a scholarship in 2010, with information whether the scholarship was renewed, canceled, or if it changed level in beginning of 2010, reflecting the evaluations that took place by the end of 2009. We collected the Lattes CV for all the scientists that were evaluated in the end of 2009. In particular, we collected the title, date, and publication name of all the scientist's journal and conference papers, books and book chapters. We call this the researcher's total production. We also collected the year the scientist received his or her doctorate degree, and consider it the start of the researcher's career.

From this set we removed all scientists that had a level 1 but lost their scholarships in 2010. There are two reasons for a scientists to lose the scholarship: either because of the negative evaluation of the corresponding CA, or because the researcher is no longer eligible to receive the grant (the scientist retired, is in a sabbatical leave abroad, or did not apply for the scholarship for some other reason). In fact, from the list of scientists that lost

their level 1 scholarship in 2010, the authors knew that two were due to the peer evaluation, but two others were due to these external reasons. Thus, to be safe, we removed these scientists from the calculations. But we included all level 2 scientists that lost their scholarship: we believe that since level 2 are more likely attributed to junior scientists, it was unlikely that such exogenous reasons applied to them.

We also collected the citations received by each of the papers in the researcher's CV using a set of programs that queried WOS, Scopus, and Google Scholar. The program collected all the pages returned by using the researcher's name as the query, searched the returned pages for the publication titles (as they appeared in the Lattes CV), and collected the corresponding citation counts. If none of the researcher publication was found in the returned pages, we considered the data missing, and did not include the researcher in the calculations. From the citation counts, and the papers listed in the Lattes CV we computed the following metrics for each researcher:

prodtot	total production, that is the number of journal and conference papers, books, and book chapters as indicated in the researcher's Lattes CV
prodwos	researcher's number of entries in WOS
prodscp	researcher's number of entries in Scopus
prod5	total production in the last 5 years (from Lattes)
prod5wos	researcher's production in the last 5 years according to WOS
prod5scp	researcher's production in the last 5 years (Scopus)
citwos	total number of citations to the researcher's publications according to WOS
citscp	total number of citations to the researcher's publications according to Scopus
citsch	total number of citations to the researcher's publication according to Google Scholar
citwosyr	average number of citations receiver per year according to WOS
citscpyr	average number of citations receiver per year according to Scopus
citschyr	average number of citations received per year according to Google Scholar
hwos	*h*-index according to WOS
hscp	*h*-index according to Scopus
hsch	*h*-index according to Google Scholar
mwos	*m*-quotient (Hirsch 2005) according to the WOS. The *m*-quotient is the *h*-index divided by the number of years of scientific activity.
mscp	*m*-quotient according to the Scopus
msch	*m*-quotient according to Google Scholar

We then grouped the researchers by area and by the original scholarship level they had at the end of 2009. We eliminated the groups with less than 8 researchers or where less than 4 researchers changed level. We computed the correlation of the peer evaluation and the bibliometric measures for each group. We then combined the correlation measures for all levels within the same area (see Sect. Statistical analysis) Notice that we did not compare a researcher from one area with researchers from different areas, nor we compared researchers that were originally at different levels.

For researchers that appear to have no citations (in WOS or Scopus), we distinguish two cases. We say that the researcher indeed has a total of 0 citations if all papers listed in the CV that were in the return result of the query had each 0 citations, or when the answer to our query to the different services returned a page stating that the researcher was not found. But if the service returned several pages and we did not find any of the researcher's papers in them, we considered the data missing and did not use that data in the computations. The reason is that for researchers with "common" last names, we could not be sure that the researcher's

data was among the pages returned, given the limits each service places on the answers to a query.

Statistical analysis

We used Spearman's rho to measure the correlation between the CA peer evaluations and the different metrics for each group (researchers in the same area and the same original level). To combine across the different levels of the same area the different correlations for a particular metric we used a technique developed in meta-analysis, a set of statistical methods to combine the results of many experiments (usually in medicine) into a single result. In this case, each correlation measure is the result of the experiment, which need to be combined. We used the fixed effect model (Hedges and Vevea 1998; Field 2001) for combining correlation measures. The fixed effect model method has the following steps:

– convert the correlation figures to a normalized measure (Fischer r-to-z conversion):

$$z_i = \frac{1}{2}\log_e \frac{1 - r_i}{1 + r_1}$$

– compute a weighted average of the z_i

$$z = \frac{\sum w_i z_i}{\sum w_i}$$
$$w_i = n_i - 3$$

where n_i is the number of researchers in group i

– the significance of the z is computed using the standard error as:

$$SE(z) = \frac{1}{\sqrt{\sum_i (n_i - 3)}}$$

– finally convert the z measure back to a correlation

$$r = \frac{e^{2z} - 1}{e^{2z} + 1}$$

Statistical significance is evaluated at 95 % confidence.

Results

The data received from CNPQ for all 2009 scholarship decision listed 4,172 scientists. As discussed, we removed all scientists that had a level 1 and lost the scholarship, a total of 237, resulting in 3,935 scientists. We grouped the scientists by area and original level, and removed those groups with less than 8 scientists or where less than 4 scientists changed their level. That resulted in 96 groups, listed in Table 2 of the Appendix, for a total of 2,663 scientists. The data in the appendix lists the areas, the original level of the scientists, how many were demoted, how many kept their level, how many were promoted, the total number of scientists in that level, and how many of them had no citation data in WOS, Scopus, and Scholar.

Table 1 lists the main results of this research. The table lists all areas, grouped by general area (according to the CNPq definition), and the Spearman rho for all metrics. If the correlation is not statistically significant with 95 % confidence it is not listed. In *bold*, the highest correlation for each area, but notice that although all correlations shown are statistically different than 0 (with 95 % confidence), no test was performed to verify if the highest correlation was significantly different that some of the other correlations. prodtot is the total production; prod5 the total production for the last 5 years, prodx is the highest correlation (for each area) among prodwos and prodscp; prod5x is the highest correlation among prod5wos and prod5scp; cit is the highest correlation among citwos, citscp, and citsch; h is the highest among the three *h*-index; m the highest among the three *m*-quotients; and cityear the highest among the three citations per year. Table 1 summarizes the correlation results for each area. The full correlation data is displayed in Tables 3 and 4 in the Appendix.

Discussion

There are some general conclusion one derives from the results in Table 1. The first one is that some of the correlations are extremely high in comparison with results published in the literature (Sect. Related research). If we use Cohen (1988) standard mapping of correlation values to linguistic terms, two of the correlation are "very strong" (correlations ranging from 0.9 to 1.0 are considered very strong), two are "high" (from 0.7 to 0.9), and 18 are "moderate" (from 0.4 to 0.7). All the previous published correlation values of peer evaluation and quantitative metrics are in the range of "moderate".

It is not unreasonable that some of the correlations should be high, maybe higher than other published results. Notice that the CNPq evaluation is ongoing and the members of the CA have to repeat the evaluations (of different scientists) every year. Thus, it is possible that they would rely on bibliometric measures more heavily than if the evaluation was a one-time event.

Another important conclusion is that there are large differences in the highest correlation among different scientific disciplines. One could point out, for example, that morphology peer evaluation has a high correlation with the *h*-index, much higher than the reported correlation for LIS researchers (Li et al 2010). But that high correlation cannot be transferred to other disciplines, not even within the biological sciences. Zoology, not only has no significant correlation for the *h*-index, but has a similar valued correlation to the *m*-quotient. This strongly suggest that generalizing from previous research in peer evaluation correlations for one discipline to other, even similar, disciplines is very risky.

Another conclusion is that some areas, in particular the general areas of language and social sciences (with the exception of sociology) seem follow evaluation practices that are qualitative and not based on quantitative measures. The problem with these general areas goes beyond the known issue that their publications are not well indexed in most bibliometric services. Scientists in these areas have been arguing that most of their publications are books (and not journal papers). Furthermore these books are written in Portuguese, about specific aspects of the Brazilian society or culture that are not necessarily of interest in other countries. Even Google Scholar, which indexes books (as opposed to WOS and Scopus), has a problem of counting the citations to books because, as far as we could determine, it does not count citations made by *other books*. If most of the production is published as books, most of the citations are done by books, which are not counted. Thus, it is reasonable that peers evaluating other scientists in these areas would not rely on citation based metrics. But our set of bibliometric measures included many non-citation based

Table 1 The correlations between peer evaluation and the bibliometric measures

General area	Area	Prodtot	Prod5	Prodx	Prod5x	Cit	*h*	*m*	Cityear
Agricultural	Agricultural eng.	**0.66**	0.51	0.63	0.65	0.48	0.63	**0.66**	0.55
	Agronomy	0.12	0.25	0.27	0.30	0.26	0.28	**0.31**	0.26
	Animal sci.			**0.29**		0.25	0.23	0.26	**0.29**
	Fisheries sci.			**0.69**	0.64	0.51	0.60	0.52	0.50
	Food sci.			0.51	**0.57**	0.30	0.30	0.46	0.35
	Veterinary	0.22	0.35	0.39	**0.46**	0.40	0.33	0.29	0.37
Biological	Biochemistry		0.27		**0.38**		0.28		
	Biotechnology								
	Botanics		0.30	0.60	**0.62**	0.50	0.48	0.51	0.53
	Ecology		0.46	**0.54**	0.53	0.43	0.46	0.49	
	Genetics		0.21	0.30	**0.45**	0.30	0.35		0.37
	Microbiology		0.35			**0.49**	0.40		
	Morphology			0.61	0.61	0.65	**0.71**	0.57	0.63
	Parasitology								
	Pharmacology		**0.30**						
	Physiology		**0.33**			0.31			
	Zoology					**0.67**			
Exact	Astronomy	0.60	0.60			**−0.90**			
	Chemistry		0.27	0.44	0.45	0.52	0.52	0.46	**0.58**
	Computer sci.	0.27	0.39	0.36	**0.42**	0.23	0.29		
	Geosciences		0.30	0.24	**0.42**	0.31	0.27	0.37	0.40
	Mathematics		0.29	0.38	**0.60**	0.42	0.40	0.50	0.45
	Oceanography			0.42			**0.45**		
	Physics	0.23	0.26	0.35	0.36	0.39	0.40	0.32	**0.42**
Social	Anthropology								
	Education			**−0.35**					
	History					**−0.27**			
	Philosophy								
	Psychology			**0.27**					
	Sociology			0.37		0.36	**0.55**		
Health	Medicine	0.32	0.29	0.38	**0.45**	0.40	0.40	0.33	0.41
	Nursing								
	Nutrition					0.50	**0.67**		
	Odontology	0.41	0.59	0.55	0.57	0.49	**0.63**	0.58	0.61
	Pharmacy			0.38	**0.45**	0.31	0.36		
	Physical education								
	Physical therapy		**−0.56**						
App soc sci	Architecture	0.54	0.43			**0.95**			
	Communications								
	Economy		**0.47**	0.41	0.46	0.33			
	Law								
	Management								
	Political sci.								
	Social service								
	Urban planning			**0.95**	0.87	0.62	0.88	0.92	0.78

Table 1 continued

General area	Area	Prodtot	Prod5	Prodx	Prod5x	Cit	*h*	*m*	Cityear
Engineering	Biomedical eng.								
	Chemical eng.	0.36	0.47	0.52	**0.71**	0.38			
	Civil eng.			0.28	**0.34**				
	Electric eng.	0.22	0.20	0.42	**0.51**	0.35	0.34	0.30	0.37
	Material sci.	0.35	0.35	0.35	0.36	0.37	0.33	0.37	**0.43**
	Mechanical eng.		0.37	0.63	**0.68**	0.48	0.58	0.61	0.59
	Production eng.			**0.51**			0.42		
	Sanitation eng.	0.37	**0.56**						
Language	linguistics								
	Literature								

Areas are grouped by great area. Correlations (Spearman rho) that are not statistically significant with 95 % confidence are not shown. The number in bold indicates the highest correlation for the line. See text for the explanation on the meaning of the columns

metrics, and we expected that some of them would be correlated to the peer decisions, in particular prodtot and prod5. Notice that both prodtot and prod5 include books and book chapters, and it would be reasonable that the peer evaluation would be correlated on the total production or the productivity of the researcher.

Beyond the lack of significant positive correlation for language and social sciences, one has to explain the significant negative correlations. Furthermore, one should also explain the significant negative correlations for astronomy, physical therapy, biomedical engineering, and production engineering. We believe that some of these correlations are "accidents." Of course, one would expect that the statistic significance test would remove all such possible "accidents", but at 95 % confidence one would expect an error in every 20 decisions, and there are many such decisions involved in constructing Table 1. For example, the data in Table 2 shows that for astronomy, the calculations included only one group (at level 2) and that group had only 9 researchers, 3 of them demoted and one promoted. The group satisfied the inclusion criteria (at least 8 researchers and at least 4 changed level) but it is possible that the high (and significant) negative correlations for cit and cityear were due to luck. The same could be true for biomedical eng., physical therapy, and philosophy, each with one group with only 12 researchers in them. We must also point out the high number of researchers for which we had no citation data for the philosophy group, thus the correlations for the impact metrics were calculated using only few researchers. This explanation is less convincing for the other areas: education, history, linguistics, and production eng., each had group sizes of more than 21 researchers. Education had two groups (level 2 and 1D), but there was also a large number of researchers with missing citation data for that area. Thus, for these areas there may be a real negative correlation of the decisions and some of the metrics.

The metrics evaluated in this research could be considered "simple" or even naive. These metrics do not follow the more recent research lines of normalizing citations by the average of the scientific area, such as the research in Zitt et al (2005), Radicchi et al (2008), Iglesias and Pecharroman (2007), Alonso et al (2009). We believe that normalization is less relevant and much more difficult in this case. Normalization is less relevant because we do not compare researchers working in one area with researchers working in another. We only compare the researchers within each scientific field and thus, there is no need to normalize across each different area. But there may be differences of citation and production within a particular area. In some way, Rinia et al (1998) distinction of applied vs basic condensed matter research

when computing the correlations is an acknowledgment of this phenomena. The only other research we are aware of regarding subarea differences is Wainer et al (2012), which discuss differences in productivity and citations rates for different subareas of computer science. Thus, some normalization for different subareas within a particular scientific area may be called for, but there are two difficult problems that must be addressed: how one defines the subareas of each of the scientific areas, and how to assign a researcher to one or more of these subareas. We do not know yet how to address any of these problems.

Finally, there is the issue of whether the correlations presented in this research are specific to Brazil, and thus this research reveals particularities of the Brazilian evaluation system, or if this research reveals phenomena that is general. It is true that the Brazilian evaluation system has peculiarities (as discussed above), but we believe that by using the changes in the scholarship level, and not the levels themselves, we factored out most of these peculiarities. We believe that the correlations measured in this research should be taken as a reflection of the current state of scientists' evaluation, in the same spirit that, for example Aksnes and Taxt (2004) is not understood as describing particularities of the Norwegian science evaluation, or Franceschet and Costantini (2011), as describing particularities of the Italian system. In this case, the fact that some of the correlations measured herein are higher than the previous published research may indicate that globally, scientist evaluation is becoming more strongly correlated to some quantitative bibliometric measures. It would be interesting to test this hypothesis on counties for which there has been some previously measured correlations.

Appendix

Data on the groups

See Appendix Table 2.

Table 2 Data on the groups

Area	Original level	Demoted	Maintained	Promoted	Total	Missing WOS	Missing scopus	Missing scholar
Agricultural eng.	2	1	10	8	19	1	4	1
Agronomy	2	25	94	27	146	9	19	8
Agronomy	1D	1	36	7	44	3	6	3
Agronomy	1C	5	17	6	28	1	4	1
Agronomy	1B	4	17	4	25	0	2	0
Animal sci.	2	8	24	5	37	1	7	0
Animal sci.	1D	2	5	4	11	1	2	1
Animal sci.	1C	3	12	1	16	3	4	1
Animal sci.	1B	2	5	3	10	0	0	0
Anthropology	2	2	13	8	23	12	7	1
Architecture	2	6	10	2	18	13	16	4
Astronomy	2	3	5	1	9	1	2	1
Biochemistry	2	6	37	1	44	3	5	2
Biomedical eng.	2	2	5	3	10	0	1	0
Biotechnology	2	3	6	2	11	0	0	0

Table 2 continued

Area	Original level	Demoted	Maintained	Promoted	Total	Missing WOS	Missing scopus	Missing scholar
Botanics	2	6	30	9	45	1	4	2
Chemical eng.	2	4	18	6	28	2	4	3
Chemistry	2	15	50	28	93	5	10	7
Chemistry	1D	1	20	14	35	1	4	3
Civil eng.	2	5	24	4	33	2	5	1
Civil eng.	1D	0	13	4	17	1	4	1
Communications	2	4	17	0	21	16	17	3
Computer sci.	2	10	39	6	55	7	13	9
Computer sci.	1D	1	17	5	23	3	9	3
Ecology	2	3	11	5	19	0	0	0
Ecology	1C	6	2	5	13	0	1	1
Economy	2	8	27	0	35	13	6	5
Education	2	7	40	12	59	46	36	4
Education	1D	0	5	6	11	5	5	0
Electric eng.	2	11	35	12	58	3	14	4
Electric eng.	1D	0	9	6	15	0	2	1
Electric eng.	1C	1	6	6	13	3	3	1
Fisheries sci.	2	13	8	0	21	1	2	1
Food sci.	2	3	24	9	36	0	5	0
Genetics	2	7	30	7	44	1	5	4
Genetics	1D	4	11	10	25	4	4	0
Genetics	1C	4	3	4	11	1	0	0
Geosciences	2	9	33	17	59	7	12	7
Geosciences	1D	2	7	9	18	3	3	0
Geosciences	1C	3	4	4	11	0	0	1
History	2	5	40	2	47	29	32	6
Law	2	4	8	1	13	12	13	4
Linguistics	2	6	18	2	26	18	19	3
Literature	2	7	26	5	38	31	32	10
Management	2	6	13	3	22	15	13	1
Management	1D	2	2	6	10	3	3	0
Material sci.	2	7	28	1	36	2	6	2
Material sci.	1D	1	4	6	11	0	0	0
Material sci.	1C	2	3	3	8	0	1	0
Mathematics	2	13	23	7	43	5	11	11
Mathematics	1D	1	10	4	15	1	4	0
Mechanical eng.	2	13	18	11	42	2	4	2
Mechanical eng.	1D	3	5	2	10	3	4	1
Mechanical eng.	1C	3	5	1	9	0	0	0
Mechanical eng.	1B	2	6	4	12	3	3	3
Medicine	2	34	52	18	104	2	7	1
Medicine	1D	6	3	4	13	0	1	0
Medicine	1C	9	5	5	19	2	2	1

Table 2 continued

Area	Original level	Demoted	Maintained	Promoted	Total	Missing WOS	Missing scopus	Missing scholar
Medicine	1B	5	5	3	13	1	1	0
Medicine	1A	4	10	0	14	0	1	0
Microbiology	2	7	20	0	27	2	6	3
Morphology	2	4	10	3	17	3	5	2
Nursing	2	3	22	4	29	4	2	0
Nutrition	2	2	10	4	16	2	3	1
Oceanography	2	6	14	0	20	0	2	0
Odontology	2	8	14	8	30	2	2	0
Odontology	1D	6	2	4	12	0	1	0
Odontology	1C	5	4	3	12	2	0	0
Parasitology	2	4	16	4	24	1	3	1
Pharmacology	2	10	22	3	35	0	1	0
Pharmacology	1D	0	5	4	9	1	2	0
Pharmacy	2	5	26	7	38	2	3	1
Philosophy	2	4	7	1	12	8	5	2
Physical education	2	6	4	4	14	2	0	0
Physical therapy	2	3	6	3	12	2	3	0
Physics	2	34	84	26	144	11	29	19
Physics	1D	4	39	7	50	5	7	8
Physics	1C	2	33	10	45	2	9	6
Physics	1B	0	17	4	21	0	3	1
Physiology	2	6	11	8	25	1	2	0
Physiology	1D	2	3	4	9	0	1	1
Political sci.	2	7	11	0	18	10	6	2
Production eng.	2	3	8	10	21	1	4	1
Psychology	2	11	24	6	41	11	10	5
Psychology	1D	1	12	5	18	7	6	3
Psychology	1C	1	8	4	13	1	3	0
Sanitation eng.	2	14	18	3	35	3	5	2
Social service	2	4	6	3	13	10	10	1
Sociology	2	5	18	3	26	13	8	1
Sociology	1B	1	6	4	11	6	2	1
Urban planning	2	5	7	1	13	5	8	2
Veterinary	2	15	22	8	45	1	4	5
Veterinary	1D	2	0	6	8	0	3	0
Veterinary	1C	3	16	5	24	0	0	0
Veterinary	1B	4	5	2	11	0	0	0
Zoology	1C	1	6	4	11	1	2	1
Total		531	1,599	533	2,663	421	564	198

Full correlation data

See Appendix Tables 3 and 4.

Table 3 Full correlation data—part 1

Area	Prodtot	Prod5	Prodwos	Prod5wos	Prodscp	Prod5scp	Citwos	Citscp	Citsch
Agricultural eng.	0.66	0.51	0.07*	0.17*	0.63	0.65	−0.09*	0.44	0.48
Agronomy	0.12	0.25	0.21	0.21	0.27	0.30	0.26	0.22	0.15
Animal sci.	0.00*	0.09*	0.05*	0.02*	0.29	0.05*	0.05*	0.21*	0.25
Fisheries sci.	0.10*	0.26*	0.45	0.35*	0.69	0.64	0.51	0.39	−0.02*
Food sci.	−0.06*	0.22*	0.32	0.45	0.51	0.57	0.30	0.29*	0.01*
Veterinary	0.22	0.35	0.39	0.46	0.26	0.31	0.38	0.27	0.40
Biochemistry	0.15*	0.27	0.13*	0.33	0.15*	0.38	0.21*	0.12*	0.06*
Biotechnology	−0.25*	−0.22*	−0.16*	−0.36*	0.21*	0.24*	0.33*	0.07*	−0.14*
Botanics	0.09*	0.30	0.52	0.52	0.60	0.62	0.50	0.31	0.31
Ecology	0.07*	0.46	0.54	0.53	0.39	0.47	0.43	0.23*	0.02*
Genetics	0.16*	0.21	0.30	0.45	0.19*	0.27	0.30	0.11*	0.22
Microbiology	0.24*	0.35	0.20*	0.30*	0.08*	0.22*	0.30*	0.23*	0.49
Morphology	0.33*	0.41*	0.35*	0.34*	0.61	0.61	0.48	0.52	0.65
Parasitology	−0.19*	−0.04*	−0.13*	0.08*	−0.27*	−0.12*	−0.11*	−0.15*	−0.24*
Pharmacology	0.23*	0.30	0.07*	0.21*	−0.05*	0.06*	0.23*	−0.05*	0.16*
Physiology	0.04*	0.33	0.21*	0.25*	0.09*	0.33*	0.07*	−0.11*	0.31*
Zoology	0.03*	0.19*	0.25*	0.39*	−0.19*	0.11*	0.67	0.21*	0.12*
Astronomy	0.60	0.60	−0.24*	−0.39*	−0.42*	0.27*	0.35*	−0.90	−0.87
Chemistry	0.12*	0.27	0.44	0.45	0.37	0.40	0.52	0.43	0.45
Computer sci.	0.27	0.39	0.30	0.42	0.36	0.41	0.18*	0.21*	0.23
Geosciences	0.12*	0.30	0.24	0.40	0.17*	0.42	0.20	0.16*	0.31
Mathematics	0.19*	0.29	0.38	0.60	0.34	0.51	0.42	0.37	0.20*
Oceanography	−0.25*	−0.09*	0.30*	0.04*	0.42	0.34*	0.35*	0.30*	0.09*
Physics	0.23	0.26	0.35	0.36	0.28	0.25	0.39	0.23	0.21
Anthropology	0.14*	0.21*	−0.31*	0.00*	0.18*	0.10*	0.19*	−0.44*	−0.08*
Education	0.15*	0.17*	0.43*	0.54*	−0.35	−0.18*	−0.02*	−0.13*	0.07*
History	−0.10*	0.00*	−0.14*	0.07*	0.39*	0.35*	−0.27	−0.16*	0.03*
Philosophy	0.19*	0.30*	0.50*		0.64*	0.58*	−0.06*		0.31*
Psychology	0.11*	0.05*	0.27	0.08*	0.09*	0.10*	0.01*	0.09*	−0.07*
Sociology	−0.13*	0.06*	0.02*	−0.12*	0.37	0.27*	0.25*	0.36	0.07*
Medicine	0.32	0.29	0.38	0.45	0.28	0.36	0.40	0.16	0.30
Nursing	0.05*	0.28*	0.08*	0.01*	0.04*	0.24*	0.11*	−0.02*	0.01*
Nutrition	0.06*	0.20*	0.44*	0.21*	−0.02*	0.18*	0.50	−0.14*	0.03*
Odontology	0.41	0.59	0.55	0.55	0.45	0.57	0.49	0.29	0.43
Pharmacy	0.20*	0.22*	0.26*	0.23*	0.38	0.45	0.21*	0.31	0.08*
Physical ed.	0.03*	0.33*	0.20*	0.22*	0.11*	0.00*	0.11*	−0.05*	−0.25*
Physical therapy	−0.31*	−0.56	0.22*	0.21*	0.16*	0.12*	−0.10*	0.48*	−0.22*
Architecture	0.54	0.43	0.73*	0.73*			0.41	0.95	0.17*
Communications	−0.22*	0.31*	0.30*	0.54*			−0.24*		−0.09*
Economy	0.23*	0.47	−0.07*	0.24*	0.41	0.46	0.33	−0.23*	−0.05*
Law	−0.22*	0.00*							0.09*

Table 3 continued

Area	Prodtot	Prod5	Prodwos	Prod5wos	Prodscp	Prod5scp	Citwos	Citscp	Citsch
Management	−0.28*	0.13*	−0.14*	−0.16*	0.43*	0.47*	0.04*	0.09*	−0.01*
Political science	0.19*	0.26*	0.52*	0.09*	0.16*	0.05*	−0.06*	−0.58*	0.04*
Social service	0.05*	−0.09*					0.31*		0.06*
Urban planning	−0.28*	−0.28*	0.49*	0.65*	0.95	0.87	0.54		0.62
Biomedical eng.	0.16*	0.10*	0.21*	0.41*	−0.39*	−0.08*	0.25*	−0.51*	−0.29*
Chemical eng.	0.36	0.47	0.52	0.71	0.23*	0.46	0.36	0.38	0.24*
Civil eng.	0.10*	0.09*	0.21*	0.34	0.28	0.21*	0.14*	0.15*	0.22*
Electric eng.	0.22	0.20	0.42	0.51	0.26	0.23	0.35	0.27	0.30
Material sci.	0.35	0.35	0.35	0.36	0.34	0.36	0.37	0.35	0.26
Mechanical eng.	0.19*	0.37	0.63	0.68	0.52	0.51	0.43	0.48	0.47
Production eng.	0.20*	0.32*	0.40	0.12*	0.51	0.11*	0.33*	−0.20*	0.29*
Sanitation eng.	0.37	0.56	−0.02*	0.04*	0.00*	0.11*	−0.15*	−0.04*	0.01*
Linguistics	0.03*	0.13*					0.07*	−0.44*	−0.20*
Literature	−0.07*	0.04*	−0.34*	−0.32*	−0.32*				0.20*

A "*" indicates that the correlation is not statistically significant. An empty entry indicates that there was not enough data to compute the correlation

Table 4 Full correlation data—part 2

Area	Hwos	Hscp	Hsch	Citwosyr	Citscpyr	Citschyr	Mwos	Mscp	Msch
Agricultural eng.	0.11*	0.60	0.63	−0.02*	0.41*	0.55	0.25*	0.66	0.53
Agronomy	0.28	0.18	0.15	0.26	0.22	0.21	0.31	0.20	0.22
Animal sci.	0.23	0.11*	0.05*	0.11*	0.18*	0.29	0.26	0.09*	0.13*
Fisheries sci.	0.55	0.60	0.12*	0.50	0.47	0.02*	0.51	0.52	0.15*
Food sci.	0.30	0.28*	−0.11*	0.35	0.34	0.11*	0.36	0.46	0.08*
Veterinary	0.33	0.32	0.25	0.34	0.24	0.37	0.24	0.23	0.29
Biochemistry	0.28	0.08*	0.09*	0.30	0.16*	0.10*	0.33	0.16*	0.21*
Biotechnology	0.14*	0.22*	−0.07*	0.50*	0.11*	0.24*	0.60	0.42*	0.42*
Botanics	0.44	0.48	0.32	0.53	0.29	0.37	0.47	0.51	0.34
Ecology	0.46	0.32	−0.13*	0.49	0.27*	0.11*	0.49	0.42	0.03*
Genetics	0.35	0.10*	0.17*	0.37	0.11*	0.20	0.26	0.06*	0.10*
Microbiology	0.35	0.19*	0.40	0.29*	0.33	0.46	0.30*	0.18*	0.32*
Morphology	0.48	0.71	0.55	0.51	0.35*	0.63	0.46	0.48*	0.57
Parasitology	−0.18*	−0.21*	−0.12*	−0.12*	−0.22*	−0.20*	−0.14*	−0.27*	−0.07*
Pharmacology	0.21*	0.00*	0.17*	0.21*	−0.01*	0.21*	0.26*	0.01*	0.31
Physiology	0.11*	−0.07*	0.23*	0.19*	−0.03*	0.34	0.23*	0.11*	0.32
Zoology	0.34*	0.10*	0.09*	0.78	0.41*	0.17*	0.70	0.62*	0.48*
Astronomy	0.00*	−0.67*	−0.36*	0.10*	−0.66*	−0.77	−0.15*	−0.54*	−0.21*
Chemistry	0.52	0.44	0.42	0.58	0.43	0.47	0.46	0.36	0.39
Computer sci.	0.21	0.29	0.26	0.20*	0.27	0.22	0.17*	0.25	0.16*
Geosciences	0.26	0.08*	0.27	0.26	0.17*	0.40	0.32	0.17*	0.37
Mathematics	0.36	0.40	0.37	0.45	0.41	0.25	0.41	0.50	0.42
Oceanography	0.26*	0.45	0.11*	0.38*	0.37*	0.19*	0.33*	0.57	0.19*
Physics	0.40	0.29	0.20	0.42	0.19	0.17	0.32	0.20	0.12
Anthropology	0.07*	−0.06*	0.14*	0.24*	−0.48*	−0.03*	0.07*	−0.02*	0.18*
Education	−0.06*	−0.20*	0.03*	0.23*	−0.24*	0.09*	−0.04*	−0.19*	0.07*

Table 4 continued

Area	Hwos	Hscp	Hsch	Citwosyr	Citscpyr	Citschyr	Mwos	Mscp	Msch
History	0.24*	−0.11*	0.05*	−0.12*		0.00*	0.24*	−0.08*	0.07*
Philosophy			0.25*	−0.63*		0.23*			0.26*
Psychology	0.18*	0.20*	−0.02*	0.20*	0.02*	−0.04*	0.17*	0.15*	−0.03*
Sociology	0.09*	0.55	−0.06*	0.12*	0.32*	0.09*	0.01*	0.53	−0.06*
Medicine	0.40	0.17	0.25	0.41	0.13*	0.28	0.33	0.12*	0.23
Nursing	0.14*	−0.02*	0.04*	0.22*	0.06*	0.22*	0.28*	0.09*	0.42
Nutrition	0.67	−0.15*	0.04*	0.63	−0.03*	0.05*	0.71	−0.01*	0.16*
Odontology	0.63	0.37	0.39	0.61	0.36	0.46	0.58	0.39	0.41
Pharmacy	0.14*	0.36	0.04*	0.21*	0.30	0.08*	0.16*	0.34	0.08*
Physical education	−0.22*	0.03*	−0.19*	0.22*	0.01*	−0.15*	−0.06*	0.16*	−0.02*
Physical therapy	0.18*	0.23*	−0.14*	0.23*	0.53*	−0.17*	0.20*	0.19*	−0.07*
Architecture	0.61*		0.10*	0.36*		0.13*	0.40*		−0.08*
Communications			−0.30*	0.00*		−0.10*			−0.12*
Economy	−0.18*	0.13*	0.24*	0.09*	−0.17*	0.07*	0.02*	0.30*	0.32
Law			0.00*			−0.09*			−0.01*
Management	0.16*	−0.14*	−0.03*	0.04*		0.11*	0.19*	0.18*	0.11*
Political sci.	0.37*	−0.21*	0.01*	0.45*	−0.52*	0.06*	0.15*	−0.09*	0.05*
Social service			−0.05*			0.04*			−0.02*
Urban planning	0.69	0.88	0.22*	0.78		0.67	0.72	0.92	0.22*
Biomedical eng.	0.31*	−0.54*	−0.53*	0.05*	−0.51*	−0.28*	0.07*	−0.67	−0.43*
Chemical eng.	0.14*	−0.06*	0.19*	0.13*	0.33*	0.02*	−0.03*	−0.01*	−0.06*
Civil eng.	0.06*	0.22*	0.16*	0.14*	0.13*	0.21*	0.10*	0.24*	0.14*
Electric eng.	0.34	0.32	0.26	0.37	0.32	0.26	0.30	0.30	0.23
Material sci.	0.31	0.27	0.33	0.38	0.43	0.27	0.35	0.37	0.37
Mechanical eng.	0.58	0.58	0.41	0.59	0.56	0.47	0.58	0.61	0.42
Production eng.	0.42	0.40*	0.22*	0.23*	−0.39*	0.21*	0.25*	0.24*	0.15*
Sanitation eng.	0.01*	−0.13*	0.05*	−0.17*	−0.09*	0.03*	−0.02*	−0.11*	0.09*
Linguistics			0.05*			−0.38			−0.10*
Literature			−0.03*			0.15*			−0.09*

A "*" indicates that the correlation is not statistically significant. An empty entry indicates that there was not enough data to compute the correlation

References

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

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.

Bornmann, L., & Daniel, H. (2005). Does the h-index for ranking of scientists really work?. *Scientometrics, 65*(3), 391–392.

Bornmann, L., Mutz, R., & Daniel, H. (2008). Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology, 59*(5), 830–837.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*, 2nd edn. London: Routledge Academic.

Field, A. P. (2001). Meta-analysis of correlation coefficients: A Monte Carlo comparison of fixed- and random-effects methods. *Psychological Methods, 6*(2), 161–180.

Franceschet M., & Costantini, A. (2011). The first Italian research assessment exercise: A bibliometric perspective. *Journal of Informetrics, 5*(2), 275–291.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta- analysis. *Psychological Methods, 3*(4), 486–504.

Hicks, D. (2011). Systemic data infrastructure for innovation policy. In: *Science and innovation policy, 2011 Atlanta Conference on, IEEE*, (pp 1–8).

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America, 102*(16), 569–16,572.

Iglesias, J., & Pecharroman, C. (2007). Scaling the h-index for different scientific ISI fields. *Scientometrics, 73*(3), 303–320.

Korevaar, J. (1996). Validation of bibliometric indicators in the field of mathematics. *Scientometrics, 37*(1), 117–130.

Li, J., Sanderson, M., Willett, P., Norris, M., & Oppenheim, C. (2010). Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. *Journal of Informetrics, 4*(4), 554–563.

Merton, R. (1968). The Matthew effect in science. *Science, 159*(3810), 56–63.

Oliveira, E., Colosimo, E., Martelli, D., Quirino, I., Oliveira, M., Lima, L., Simoes e Silva, A., & Martelli-Junior, H. (2012). Comparison of Brazilian researchers in clinical medicine: Are criteria for ranking well-adjusted?. *Scientometrics, 90*(2), 429–443.

Patterson, M., & Harris, S. (2009). The relationship between reviewers' quality-scores and number of citations for papers published in the journal physics in medicine and biology from 2003–2005. *Scientometrics, 80*(2), 343–349.

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, 105*(45), 17,268.

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., van Leeuwen, T., Van Vuren, H., & van Raan, A. (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.

van Raan, A. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. *Scientometrics, 67*(3), 491–502.

Wainer, J., Eckmann, M., Goldenstein, S., & Rocha, A. (2012). Differences in productivity and impact across the different computer science subareas. Technical Report IC-12-08, Institute of Computing, University of Campinas, http://www.ic.unicamp.br/∼reltech/2012/12-08.pdf Accessed December 2012.

Waltman, L., van Eck, N., van Leeuwen, T., Visser, M., & van Raan, A. (2011). On the correlation between bibliometric indicators and peer review: Reply to Opthof and Leydesdorff. *Scientometrics, 88*(3), 1017–1022.

Zitt, M., Ramanana-Rahary, S., & Bassecoulard, E. (2005). Relativity of citation performance and excellence measures: From cross-field to cross-scale effects of field-normalisation. *Scientometrics, 63*(2), 373–401.