Are peer-review activities related to reviewer bibliometric performance? A scientometric analysis of Publons

José Luis Ortega
Cybermetrics Lab, Madrid, Spain,
jortega@orgc.csic.es

Abstract

This study attempts to analyse the relationship between the peer-review activity of scholars registered in Publons and their research performance as reflected in Google Scholar. Using a scientometric approach, this work explores correlations between peer-review measures and bibliometric indicators. In addition, decision trees are used to explore which researchers (according to discipline, academic status and gender) make most of the reviews and which of them accept most of the papers, assuming that these are reasonable proxies for reviewing quality. Results show that there is a weak correlation between bibliometric indicators and peer-review activity. The decision tree analysis suggests that established male academics made the most reviews, while young female scholars are the most demanding reviewers. These results could help editors to select good reviewers as well as opening a new source of data for scientometrics analyses.

Keywords: Publons, Google Scholar Citations, peer-review, manuscript acceptance, scientometrics

Introduction

One of the foundations of the academic publishing system is the peer-review of manuscripts submitted to scientific journals. Based on the principles of falsifiability and replicability of science, the peer-review process is a mechanism to discuss and check the methods and analyses expressed in a paper, as a way to verify the reliability of the results. This procedure acts as a quality filter, selecting the most relevant and rigorous articles as well as of being a control instrument for reporting irreproducible and fraudulent articles. Even though the peer-review process is settled in the academic culture, including the evaluation of projects and the recruitment of scholars, this filtering mechanism was promoted by journal editors in light of the exponential growth of scientific literature during the 20th century (Price, 1963; Burnham, 1990; Kronick, 1990). In this way, peer-review may be viewed more as an editorial solution to select the most relevant papers than as a mechanism for the self-regulation of the scholarly community. This revision procedure introduced a competition mechanism among journals for publishing the best works, generating elevated rejection rates. Since then, these ratios have been used until now as indicators of editorial quality.

However, the time and effort that each researcher spends on this activity have not been recognised by the academic reward system. The number of reviewed papers and the importance of the journals that ask for these reviews have not been considered for promotions, positions and funding. This lack of acknowledgement, in many instances, could be due to the difficulty of confirming this activity and the quality of these tasks. But perhaps the most important impediment is that much of this information is not made public by the publishers because the current peer-review system implies the anonymity of the reviewers.

From a scientometric point of view, this information lack has caused that this activity has not been analysed in-depth. As a consequence, it has not been possible to quantify this activity in context with other metrics related to research activity, production and impact. However, Publons (http://home.publons.com) now offers a way to explore this academic activity. This online platform attempts to give credit to scholars for the reviewing activity they do for academic journals. This service checks and publishes the editorial activity of each reviewer, which is verified by the editors of the journals. Then, a profile is set up for quantifying the activity using several indicators and charts. In this way, the goal of this site is to visualize the contribution of these scientists to the academic publishing system and value the quality of their editorial tasks. Publons creates a new space that makes possible the assessment of scholars by journal editors when they come to recruit reviewers. This article attempts to explore the relationship between the peer-review activity gathered on Publons and the production and impact as reflected in Google Scholar, with the aim of observing the connection between both academic activities.

The peer-review practice is, however, not free from shortcomings. Kassirer and Campion (1994) enumerated the main limitations of this system. In many cases, there is a strong subjective element which produces particular biases and prejudices. Mahoney (1977) found that reviewers were biased against manuscripts which showed results contrary to their theoretical perspective. Peters and Ceci (1982) verified that there was a bias in favour of prestigious and highly productive departments. Travis and Collins (1991) identified a phenomenon of cognitive particularism in the peer-review of grant applications. In fact, many studies have shown strong disagreements between reviewer's judgements. Cole and Simon (1981) concluded that there was a high degree of disagreement among reviewers. Rothwell and Martyn (2000) also found that the agreement between reviewers in clinical neuroscience was little greater than would be expected by chance alone. Other studies have pointed out the difficulty of this process for detecting methodological flaws. Godlee et al. (1998) confirmed that the blinding of reviewers did not report differences in the peer-review process. Another problem is the slowness of the review process, which could delay the publishing of important results. This problem is very relevant in disciplines with high obsolescence rates. The peer-review system also has limitations to detect fraudulent misconducts. Lerner (2003) detailed several fraudulent cases undetected by reviewers. Haug (2015) warned about fabricated peer reviews. In spite of these problems, peer review provides important advantages, mainly filtering novel and valuable papers and improving the quality of original manuscripts (Purcell, 1998). Bornmann and Daniel (2008) observed that the papers accepted by a prestigious journal achieved almost the double the number of citations compared to rejected papers published in other journals. Pautasso and Schäfer (2009) pointed out that the rejection percentage was directly related to the impact factor of the journal, which was caused by the high submission rates of these journals.

However, bibliometric-based research evaluation has been proposed as an alternative method to the peer-review due to its economic and temporal advantages. Thomas and Watkins (1998) found high correlations between peer-review and citations-based rankings. Abramo and D'Angelo (2011) shows that for the natural and formal sciences, the bibliometric methodology is by far preferable to peer-review. Many studies have explored the relationship between bibliometric indicators and the results of the review process. Opthof et al. (2002) observed that reviewers' recommendations and editor's ratings were positively correlated with citations. Aksnes and Taxt (2004) investigated the relationship between bibliometric indicators and the outcomes of peer-reviews, finding positive but relatively weak correlations. Van Raan (2006), on the contrary, provided comparable results between the h-index and the scores given to several research groups by a revision panel. Patterson and Harris (2009) found a low but statistically significant correlation between citations and quality scores.

Most of the studies have focused on the results of the peer-review process and the importance of this task for the improvement of research articles. Nevertheless, few papers have studied the role of the reviewers in this process, exploring the characteristics of these referees and detecting which of them provide the best reviews (Weller, 2001). A pioneering study on the quality of reviewers' opinions and their academic positions found that high-status reviewers only agreed to review a small number of papers, while the low-status reviewers usually provided a positive report (Stossel, 1985). Similarly, Oxman et al. (1991) compared the review scores of three groups (research assistants, clinicians and experts) without finding any significant differences among them. Evans et al. (1993) analysed 200 researchers and their qualifications as reviewers. They described the good reviewer as young, from strong academic institutions and well-known by the editors. Black et al. (1998) surveyed more than 700 reviewers and they only detected that the reviewers' age was associated with high-quality reviews. A few years later, a similar result was found by Kliewer et al. (2005) exploring the quality score of 800 reviewers and their attributes (sex, age, position and speciality), showing that the age was the only variable significantly associated with high scores. According to importance of the incentives for peer-review, many authors have defended the implementation of this step (Kumar, 2014; Nguyen et al. 2015). However, Squazzoni et al. (2013) demonstrated that offering material rewards to reviewers tends to decrease the quality of the reviewing process.

Other studies devote more attention to the reviewers' abilities and attitudes to perform high-quality reviews. Yankauer (1990) interviewed the *American Journal of Public Health*'s reviewers and he observed that the review time was longer and inversely related to the number of papers reviewed. However, Callaham and Tercier (2007) concluded that there are no easily identifiable types of formal training that predict reviewer performance. Other works addressed how subjective reasons could determine the reviewers' behaviour. Tite and Schroter (2007) observed that reviewers are more likely to accept to review a manuscript when it is relevant to their own research line; and Schriger et al. (2016) found that reviewers were more favourable to accepting a paper with citations to their own work. From a longitudinal view, Callaham and McCulloch (2011) detected that the quality of reviewers' reports fell over time. However, no studies until now have described the relationship between the scholarly performance of researchers and their reviewing activity from a bibliometric view.

Objectives

This study attempts to analyse the relationship between the peer-review activity of scholars registered in Publons and their research performance as reflected in Google Scholar. Using a scientometric approach, this work explores correlations between peer-review measures and bibliometric indicators. Additionally, this study tries to identify what types of researchers, according to discipline, position and gender, carry out most of the review reports and which of these researchers' groups are the toughest. Several research questions were formulated to meet these objectives:

- Are decision trees a feasible technique to characterize the most prolific reviewers according to their personal features (i.e. gender, age, academic position and research area)?
- Using the same technique, is it also possible to identify the reviewers that accept or reject the largest number of articles according to their gender, age, status and research interests?

And as secondary objectives:

- Is there any important and positive correlation between the bibliometric indicators (i.e. citations, h-index and number of publications) and the peer-review activity (i.e. number of reviews, acceptance ratios and the reviewed journals) of Publons users?
- According to retrieved data, could Publons be a suitable instrument for the quantitative study of the peer-review activity? And therefore, could it be used for research evaluation?

Methods

Data sources

Publons

Publons is a web platform created by Andrew Preston and Daniel Johnston in New Zealand in 2013. The service is addressed to the scholarly community and its purpose is to create an open site that may improve the peer-review system, making it faster, more efficient and effective. In this way, Publons offers, on the one hand, a service where journal editors can select appropriate reviewers and, on the other hand, a way for scholars to get credit for their review activity. Yet, this platform aspires to be a meeting point between scholars and research journals.

For this task, each user can build a personal profile where he/she may include the number of review reports, the name of journals involved and if the review finished with a rejection or an acceptance. This information has to be confirmed through email messages or by the journal’s own editors. Besides, members may include information on their institutional affiliations, their academic positions and research areas, which allows to group reviewers by region and discipline. From these data, Publons calculates some in-house indicators on the peer-review activity of their users:

- Number of Reviews: number of reviews carried out by the user and checked by the journal’s editor. This quantitative indicator expresses the commitment of the users in the peer-review process. However, the list of reviews could not be exhaustive because it is possible that some researchers remove or lose old reviews.
- Number of Reviews (last 12 months): this variation of Number of Reviews only computes the review reports from last year and it could be an indicator of the current peer-review activity.
- Acceptance Rate: percentage of review reports that have finished with acceptance. For Publons, a manuscript is accepted when it has been published online (a DOI has been assigned). This fact introduces a time delay bias because the publication can occur months after the review, and therefore, the acceptance rate is always lower than the real one. This rate informs us about the thoroughness of a reviewer because previous studies pointed out that the quality of reviewers is inversely correlated to their acceptance rates (Callaham et al., 1998; Kurihara and Colletti, 2013).
- Openness: percentage of reviews that are published in Publons.
- Merit: this indicator calculates the degree of participation of the members on the website. Publons assigns a base of three points per verified review, and reviews can get more merit if members of the community endorse them.

For the purposes of this analysis, Openness and Merit were not used as these indicators have only internal validity and they do not give important information. In contrast, other indicators were calculated from the extracted data:

- Number of journals: Number of different journals for which the user wrote a review. This metric allows to see the degree of collaboration that a scholar has in the review of manuscripts for different journals. This measurement could be understood as a quality index, as the researcher could be more appreciated by the journal editors if he is involved in a larger number of reviews for a diverse group of different journals.
- Average Impact of the journals: Average of the SJR (Scimago Journal Rank) of the journals for which the researcher wrote a report. This metric informs us on the quality of the reviewed journals and therefore it would be a quality indicator on the importance and thoroughness of the researchers’ reviews.

Google Scholar Citations

Google Scholar Citations (GSC) is a Google Scholar's service created in November 2011 that allows users to build a brief curriculum vitae where they list their publications indexed in Google Scholar. In addition, this website shows several bibliometric indicators that describe the scientific impact and production of each researcher. This bibliometric data source was used due to several advantages. The first one is the comprehensiveness of this search engine, which covers the largest amount of academic sources (Khabsa and Giles, 2014). The second one is that GSC profile is managed by the own author, who directly verifies the publication list. This fact favours the accuracy of the data. The third reason is that this information is public and easily retrievable using a web crawler.

Data extraction

In November 2015, a crawler was designed to extract the largest ever sample of Publons profiles[1]. This crawler harvested the discipline, organization and reviewed journals of each profiles as well as his/her peer-review indicators in Publons. A total of 266,391 (43%) profiles were extracted from an estimated population of 600,000 users. From this sample, only those users that had ten or more reviews were selected. This cutoff was set to get a consistent group of reviewers with a sufficient number of reviews. This criterion cuts down the sample to 1,968 (.7%) profiles, obtaining a much reduced set. This drastic reduction demonstrates that most of the registered users do not fill out their profiles, perhaps, because each review report has to be validated by the journal's editor and therefore many users avoid doing that.

Then the names of these 1,968 users were searched for matching profiles in Google Scholar Citations (GSC). But it was found that 3,005 profiles shared full names with the reviewers, so a manual checking was performed for identifying the correct profiles using the organization name, disciplines and, as the first criterion, the profile photo. In case of doubt, the author was removed from the study. At the end, for the 1,968 Publons' users[2] only 571 (29%) could be identified in GSC database. In this way, the Publons' metrics could be matched with the bibliometric data of GSC. Citations, h-index and publications were harvested from GSC's profiles. Besides, a list of journals with their corresponding SJR indicator was downloaded from SCImago Journal & Country Rank to calculate the Average Impact of the journals. This measure was selected because it is calculated for a larger number of journals than the Journal Impact Factor (JIF) of Web of Science. Another reason is that it is discipline independent, which means that this indicator is suitable for analysing journals from different

[1] However, this practice could affect the efficiency of the service and the owners ask to be contacted for future studies or use the publicly available API (https://publons.com/api/).

[2] Data on this study are publicly available in http://hdl.handle.net/10760/29799

disciplines. Finally, the last reason is that it is publicly accessible (www.scimagojr.com/journalrank.php).

Decision trees

This statistical technique, widely used in data mining and decision sciences, groups objects characterized by a variable (dependent) according to the values of other independent variables (predictors). Its aim is to trace significant variations in the distribution of the dependent variable with regard to the other independent variables, characterizing which factors have more influence on the detection of homogeneous groups. This process is developed through a reasoning process implemented in different algorithms (CHAID, CRT, QUEST, etc.).

The Exhaustive CHAID (Chi-square automatic interaction detector) algorithm was used because it is the most generalized and restrictive in its results. This algorithm uses the chi-square test to generate new nodes, detecting significant differences in the distribution of variables (McCarty and Hastak, 2007). The CHAID model can be viewed as an inverted tree, which is split into different branches and sub-branches. The model starts from an initial trunk with all the sample's elements. Then, Chi-square test is done and the p-values are calculated. If the p-values are statistically significant, then the algorithm splits the respective predictor categories, creating the first branching of the tree. Next, this procedure continues until the groups get the highest purity, this is, each group contains only the highest proportion of a unique value of the target variable (Ritschard, 2014). This technique is suited for nominal or ordinal variables because it is easier to observe how the presence or absence of certain variable values may affect the distribution of the sample.

This technique is used to characterize the most active reviewers and those that accept more manuscripts for publication. Decision tree is a proper method to describe which personal features (sex, position and discipline) are the most associated with a high performing group. The advantages of this technique are its promptness to detect the most significant features and to distinguish the lowest and highest performers. The visual representation helps to take a clear picture of which qualities are the most associated with good reviewers. From a statistical point of view, this method is not sensitive to non-linear relationships and it does not need the normalization of variables. This makes easy the interpretation of the results and its use with different distributions and data types. In this sense, decision tree is suitable for this study because many of the variables are at different scales (i.e. average SJR, acceptance, etc.) or their distributions follow different trends (i.e. citations, h-index, etc.). However, the major weakness of this method is its lack of statistical representativeness in very large trees because the model could fraction the sample in very small groups, introducing more randomness during the group creation (Lantz, 2015). To avoid this problem, tree's branches were pruned to no more than three levels.

Results

Correlations

A correlation matrix was built to explore the relationship between bibliometric and peer-review indicators. All the variables were transformed into a logarithmic scale and then Pearson's correlation coefficient was used.

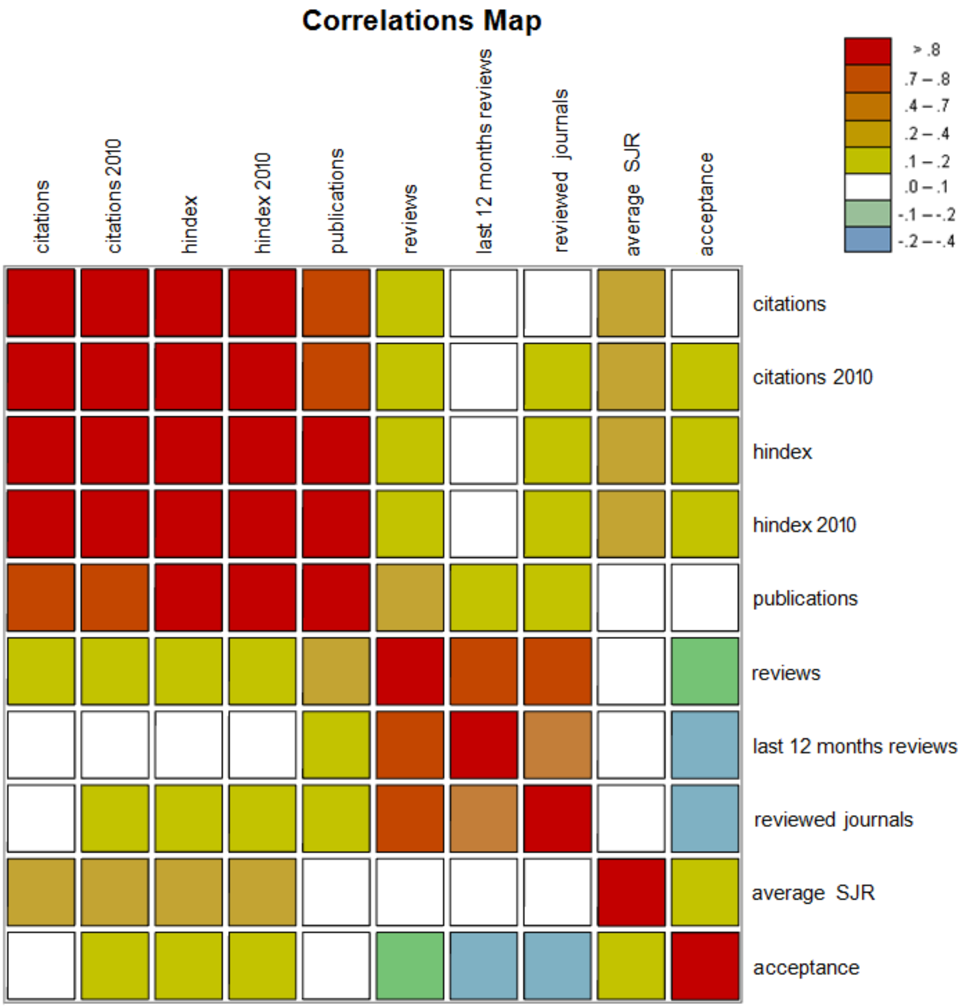

Figure 1. Correlation heat map between bibliometric and peer-review metrics (Pearson's correlation coefficient)

Figure 1 shows a colour map of the correlations between the bibliometric indicators from GSC and the peer-review metrics from Publons. The first impression is that there are not significant correlations between these two groups of metrics. The highest correlations do not exceed the 30% (ρ<.3), which evidences that the peer-review activity is independent of the publishing performance. Even so, *average SJR* is the variable that better correlates with *citations 2010* (ρ=.277) and *citations* (ρ=.268). This fact indicates a moderate connection between the research impact of reviewers and the bibliometric quality of journals that they review. In addition, the production of a researcher is positively associated with the number of reviews (ρ=.204), suggesting that authors with a higher number of publications might be more interested in reviewing manuscripts.

Other possible explanation could be that journal editors would tend to select as reviewers those authors with an important production.

Analysing the relationship among peer-review indicators, the correlation matrix shows a relevant association between *reviews* and *reviewed journals* (ρ=.771), confirming that the more the peer-review activity increases; the more diverse is the number of reviewed journals. To a lesser extent, it is interesting to notice the inverse relationship between *acceptance* and *reviews* (ρ=-.134), *last 12 month reviews* (ρ=-.375) and *reviewed journals* (ρ=-.284). These results suggest that the more active is the peer-review activity of an author, the less is the rate of manuscript acceptance, showing that the demand for review of manuscripts might be linked to certain training degree. The strong correlation of *acceptance* with *last 12 month reviews* could be motivated by the aforementioned bias in measuring the acceptance rate. Manuscripts that were reviewed recently are less likely to have completed the publication process, which it would have an impact in the correlation.

Decision Trees

Another relevant aspect of this study is the detection of what kind of researchers, according to their disciplines, positions and gender, reviews more papers and which of them are the most demanding when they review those articles. Publons uses the same academic categories as Elsevier Group to classify their users. These categories can be grouped in four large research areas: Health Sciences, Physical Sciences and Engineering, Life Sciences and Social Sciences and Humanities. Gender was defined from the first name and the picture of each profile in Publons and Google Scholar. Positions were obtained from each personal home page and classified in six academic statuses: PhD Student, Research Fellow, Assistant Professor, Associate Professor, Lecturer and Professor. To better use and interpret this tool, the number of reviews and the acceptance rate were transformed to categorical variables, grouping the values by quartile. In this way, 1st quartile corresponds to the authors' group that have more reviews or acceptance rate higher than 75% of the sample, while the 4th quartile shows the authors set which values are below 25% of the sample. Table 1 shows the mean and range of each quartile for reviews and acceptance.

	Reviews		**Acceptance**	
Quartile	Mean	Range	Mean	Range
Q1	11.81	10-14	6.6%	0%-14.9%
Q2	19.35	15-24	21%	15%-27%
Q3	31.06	25-41	32.6%	27.3%-39.7%
Q4	115.34	42-2,709	62.5%	40%-100%

Table 1. Distribution of reviews and acceptance rates by quartiles

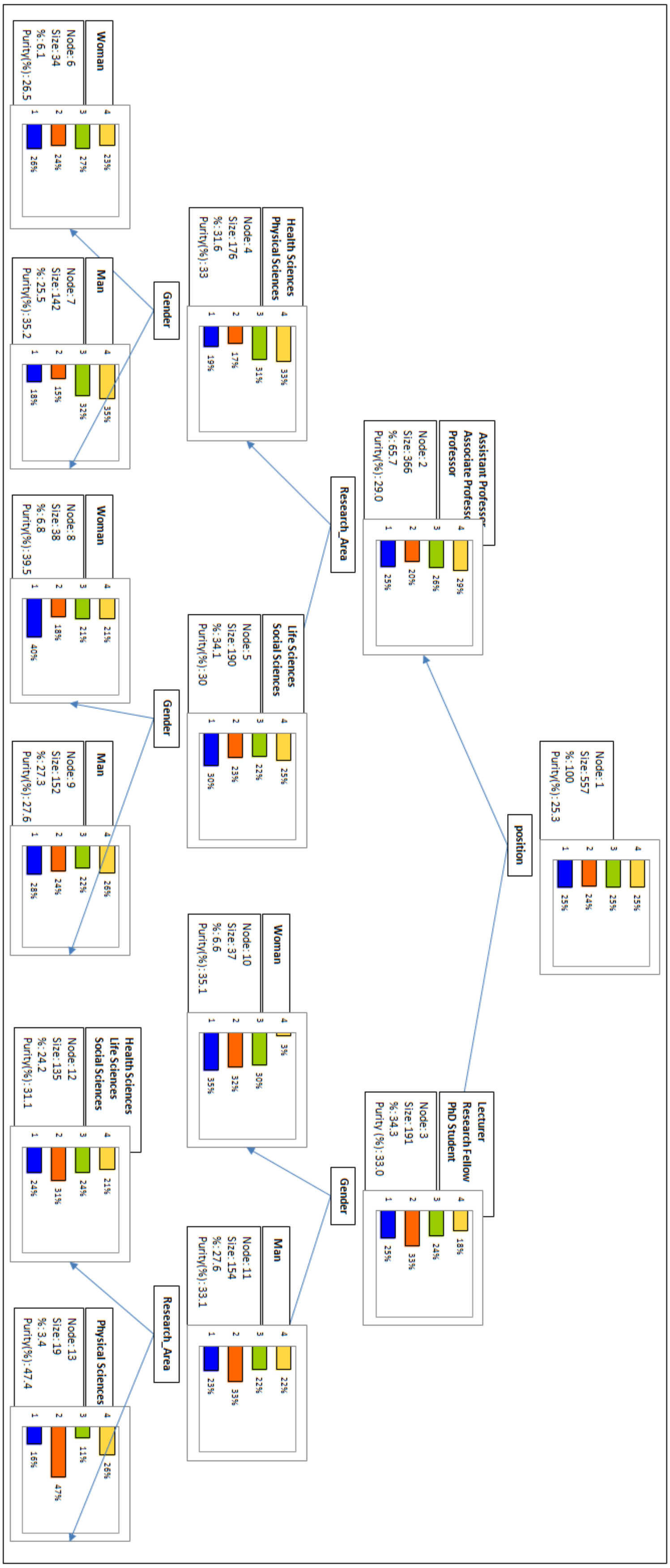
position
Node: 1
Size: 557
%: 100
Purity(%): 25.3
25%
24%
25%
25%
Assistant Professor
Associate Professor
Professor
Node: 2
Size: 366
%: 65.7
Purity(%): 29.0
25%
20%
26%
29%
Lecturer
Research Fellow
PhD Student
Node: 3
Size: 191
%: 34.3
Purity (%): 33.0
25%
33%
24%
18%
Research_Area
Health Sciences
Physical Sciences
Node: 4
Size: 176
%: 31.6
Purity(%): 33
19%
17%
31%
33%
Life Sciences
Social Sciences
Node: 5
Size: 190
%: 34.1
Purity(%): 30
30%
23%
22%
25%
Gender
Woman
Node: 10
Size: 37
%: 6.6
Purity(%): 35.1
35%
32%
30%
3%
Man
Node: 11
Size: 154
%: 27.6
Purity(%): 33.1
23%
33%
22%
22%
Gender
Woman
Node: 6
Size: 34
%: 6.1
Purity(%): 26.5
26%
24%
27%
23%
Man
Node: 7
Size: 142
%: 25.5
Purity(%): 35.2
18%
15%
32%
35%
Gender
Woman
Node: 8
Size: 38
%: 6.8
Purity(%): 39.5
40%
18%
21%
21%
Man
Node: 9
Size: 152
%: 27.3
Purity(%): 27.6
28%
24%
22%
26%
Research_Area
Health Sciences
Life Sciences
Social Sciences
Node: 12
Size: 135
%: 24.2
Purity(%): 31.1
24%
31%
24%
21%
Physical Sciences
Node: 13
Size: 19
%: 3.4
Purity(%): 47.4
16%
47%
11%
26%
1
2
3
4

Figure 2. Decision tree according to the number of reviews[3]

Figure 2 shows the decision tree according to the number of reviews. The most distinguishing variable is academic position. Thus, *Assistant Professors/Associate Professors/Professors* are the researchers that review the largest number of manuscripts because 29% of them are in the 4th quartile and 26% in the 3rd quartile, while most of the *Lecturer/Research Fellow/PhD Student* researchers are in 1st (25%) and 2nd (33%) quartiles, demonstrating that young scholars could tend to review fewer papers. In fact, *Professors* carry out in average the double of reviews (mean=65.4) than *PhD Students* (mean=30.5) or *Lecturers* (mean=30.4).

In the case of Node 2, *Health Sciences/Physical Sciences* group the authors with most of the reviews (Q4=33%; Q3=31%), and *Life Sciences/Social Sciences*, gather the less active reviewers (Q2=23%; Q1=30%). This is confirmed by the Kruskal-Wallis test (p-value=.003), where the average of reviews in *Health Sciences* (mean=54) and *Physical Sciences* (mean=68.8) are much larger than in *Life Sciences* (mean=34.1) and *Social Sciences* (mean=33).

Finally, at the third level, the model detects significant gender differences. Thus, in all the branches, *Man* category has more users in 4th and 3rd quartiles than *Woman*, which it suggests that men might review more manuscripts than women. This is confirmed by the Mann-Whitney test (p-value=.002). The Node 3, set up by the youngest researchers, is split by gender. The new nodes (10, 11) showing again that *Man* cluster (Q1=23%; Q2=33%) has lower percentages than *Woman* (Q1=35%; Q2=32%). Finally, from Node 11 (*Man*), the disciplinary group of *Physical Sciences* (13) shows the highest percentages of reviewers with low-performance (Q1=16%; Q2=47%), while *Health Sciences/Life Sciences/Social Sciences*' group (12) has the worst percentages of researchers with few reviews (Q1=24%; Q2=31%). Therefore, according to these results, established (*Assistant Professors/Associate Professors/Professors)* men researchers from Health Sciences and Physical Sciences set up the group (Node 7) with the largest percentage of highly productive reviewers. On the contrary, young men scholars from Physical Sciences (Node 13) are those that produce fewer reviews.

[3] A full-size version of the figures 2 and 3 can be downloaded from here http://hdl.handle.net/10760/30071

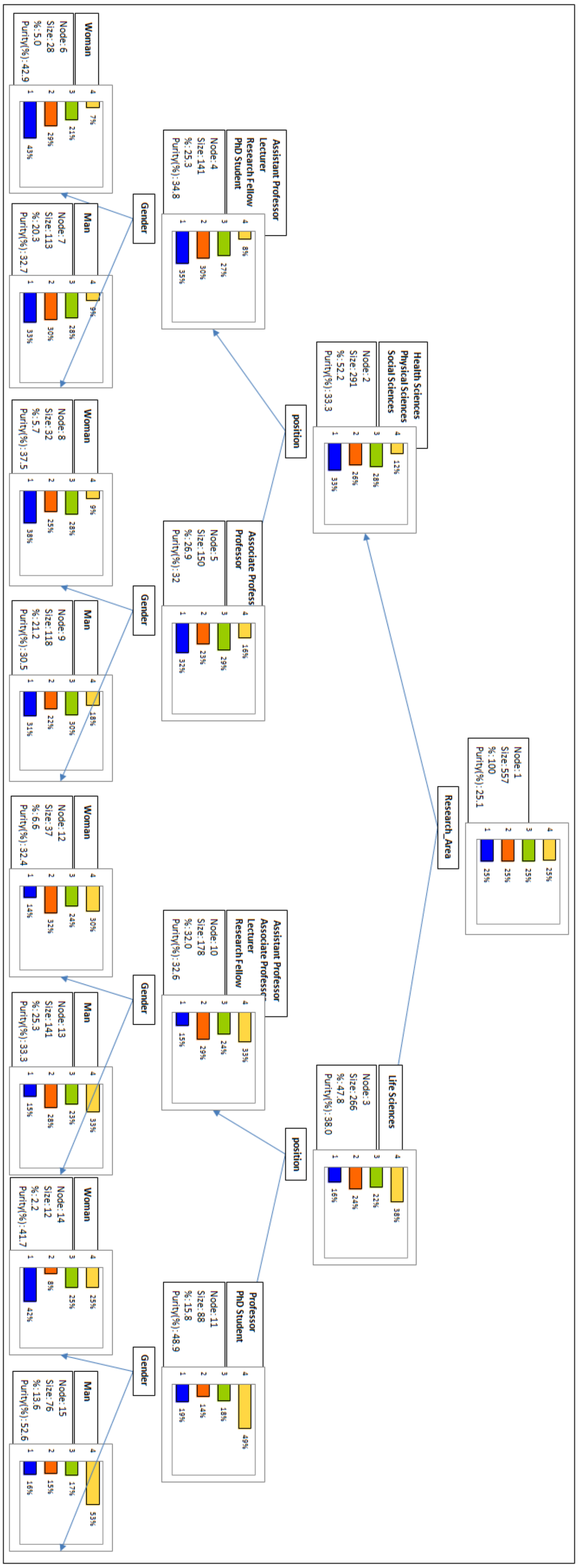
Node: 1
Size: 557
%: 100
Purity(%): 25.1
1 2 3 4
25% 25% 25% 25%
Research_Area
Health Sciences
Physical Sciences
Social Sciences
Node: 2
Size: 291
%: 52.2
Purity(%): 33.3
33% 26% 28% 12%
position
Life Sciences
Node: 3
Size: 266
%: 47.8
Purity(%): 38.0
16% 24% 22% 38%
position
Assistant Professor
Lecturer
Research Fellow
PhD Student
Node: 4
Size: 141
%: 25.3
Purity(%): 34.8
35% 30% 27% 8%
Gender
Associate Profess
Professor
Node: 5
Size: 150
%: 26.9
Purity(%): 32
32% 23% 29% 16%
Gender
Assistant Professor
Associate Professo
Lecturer
Research Fellow
Node: 10
Size: 178
%: 32.0
Purity(%): 32.6
15% 29% 24% 33%
Gender
Professor
PhD Student
Node: 11
Size: 88
%: 15.8
Purity(%): 48.9
19% 14% 18% 49%
Gender
Woman
Node: 6
Size: 28
%: 5.0
Purity(%): 42.9
43% 29% 21% 7%
Man
Node: 7
Size: 113
%: 20.3
Purity(%): 32.7
33% 30% 28% 9%
Woman
Node: 8
Size: 32
%: 5.7
Purity(%): 37.5
38% 25% 28% 9%
Man
Node: 9
Size: 118
%: 21.2
Purity(%): 30.5
31% 22% 30% 18%
Woman
Node: 12
Size: 37
%: 6.6
Purity(%): 32.4
14% 32% 24% 30%
Man
Node: 13
Size: 141
%: 25.3
Purity(%): 33.3
15% 28% 23% 33%
Woman
Node: 14
Size: 12
%: 2.2
Purity(%): 41.7
42% 8% 25% 25%
Man
Node: 15
Size: 76
%: 13.6
Purity(%): 52.6
16% 15% 17% 53%

Figure 3. Decision tree according to the acceptance rates

Figure 3 shows the classification of researchers according to their acceptance rates. Thus, reviewers included in the 4th and 3rd quartiles are those with the highest acceptance percentages and therefore with a less rigorous attitude toward the review of papers. On the contrary, researchers in 1st and 2nd quartiles are the most demanding reviewers with the lowest acceptance rates. The variable with the highest discrimination power is the research discipline. In this case, decision tree shows that *Life Sciences* researchers are more prone to accept papers (Q3=22%; Q4=38%) than scholars from the remaining disciplines (Q3=28%; Q4=12%). Kruskal-Wallis test confirms this observation (p-value<.0001), and the Dunn's post-test also points that *Life Sciences* researchers (mean=37.9%) accept more papers than the researchers from the other disciplines (mean=22%-32.9%).

The second level corresponds to the academic positions. Then, from the previous *Life Science*'s group (Node 3), *Professor/PhD Student* (11) is the cluster that accepts higher number of manuscripts (Q3=18%; Q4=49%), while the rest of the academic statuses (10) shows a more balanced ratio of acceptance (Q3=24%; Q4=35%). From the other branch hanging from Node 2, *Assistant Professor/Lecturer/Research Fellow/PhD Student* (4) is the group that rejects more papers (Q1=30%; Q2=30%), in contrast with *Associate Professor/Professor* (5) (Q1=32%; Q2=23%), which has a larger proportion of acceptance rates. However, Kruskal-Wallis test did not find significant differences (p-value=.794) among academic positions, although *PhD Students* have in average the lowest acceptance rate (mean=28%) and *Professor* the highest one (mean=33.9%).

Finally, the last discriminant variable is gender and, in all the cases, *Woman* shows lower acceptance rates than *Man*, which it is confirmed by the Mann-Whitney test (p-value=.029). In summary, the reviewers that accept a higher number of papers are male professors and PhD students from Life Sciences (Node 15, Purity=52.6%), while the most demanding reviewers are female with starting academic positions from Health, Physical and Social Sciences (Node 6, Purity=42.9%).

Discussion

The analysis of the correlations between peer-review and bibliometric indicators has shown that the connection between these different academic activities is weak. But even, these slight relationships inform about some connections between the peer-review activity and the bibliometric outputs. Hence, for example, the research production is slightly associated with the number of reviews (ρ=.204) and the academic impact of reviewers is linked to the quality of the reviewed journals (ρ=.277). However, these low correlations demonstrate that the peer-review actions are a different and independent activity from the publishing performance of researchers. Thus, the ability and commitment of scholars to review manuscripts is not related to the scientific performance and impact of their own outputs. From an editorial point of view, these results show that the best reviewers are not necessarily those who have the best bibliometric scores, but rather that the quality of reviewers would reside in other attributes (Snell and Spencer, 2005). On the other hand, correlations among peer-review indicators have brought other interesting results on this task. Thus, for example, the acceptance rate is inversely related to the number of reviewed manuscripts. In this way, the more papers are revised, the more demanding become the researchers, which it suggests that the quality of reviewers might be based on their training and experience as well.

Nevertheless, the most useful results have been obtained from the decision trees. They have helped to define which types of reviewers tend to participate more in the revision of papers and to detect the most demanding types of researchers. Thus, for example, the cluster of senior male scholars from Health and Physical Sciences includes the most active reviewers, opposite to the set of young men scientists from Physical Sciences, which contains the group of less prolific reviewers. This result shows that the career duration could be an important factor to detect the most productive reviewers and that this activity might be the result of a cumulative training process. The fact that young scholars might be less known by journal editors, could be another reason for these differences.

According to the acceptance rate of manuscripts, decision trees show that young scholars tend to reject more papers than senior ones. This result could be biased by the already commented delay in the counting of accepted papers because some reviews from young scholars could not be accounted yet. In spite of this shortcoming, this result is in accordance with previous studies (Black et al., 1998; Callaham and Tercier, 2007; Kliewer et al., 2005), suggesting that bias is not so relevant. The fact that young scholars are the most demanding reviewers could be due to they are absorbed in a more competitive environment (i.e. funding, positions) as well as they are in touch with the most current research lines and methodologies (Donaldson et al., 2010). On the other hand, results on acceptance of manuscripts have also shown gender differences, pointing out that women might be stricter in their reviews than men (Gilbert et al., 1994; Wing et al., 2010). Taking into account these results, this study encourages journal editors the recruitment of young and women scholars because these researchers are more committed to the peer-review process, writing stricter reviews. Nevertheless, these results could be contradictory with the observed ones in correlations because they suggest that researchers with a smaller number of reviews, that is, the younger ones, should be less strict than veteran scholars, who have more experience as reviewers. A possible explanation might be that the number of reviews may be not evenly distributed over time, but rather most of the reviews could be concentrated in the starting academic years. Regardless, more studies on the distribution of reviews along the time would shed light on this problem.

However, these results have to be considered with caution because it is possible that the data on the peer-review activity of Publons' members could be incomplete and therefore their metrics can be biased. Although there is not any evidence of intentional manipulation, it is possible that some users only include information on their most recent revisions or reviews written for prestigious journals. This could affect the statistical analysis and would distort the results of correlations and decision trees. Other limitation is that this is a relatively young platform and the number of active users is very low in proportion to the total population. This fact makes the sample more sensitive to possible biases and might affect the statistical significance of the smallest branches of the Decision trees. Another important problem is the delay in the counting of reviews that could influence the variable *reviews*, and especially *last 12 month reviews* and *acceptance*. Although this problem affects every reviewer, it could be more important to young researchers. However, the obtained results on the acceptance are in line with previous findings, so it is possible that this bias has little effect. Correlations have to be applied with caution because these relationships do not imply causation, but a mutual interaction between both phenomena or, in many instances, the influence of a third factor (spurious correlations). In our case, correlations are just used to explore connections between both academic activities. Due to all these limitations, results are only circumscribed to the Publons' users and it is hard for now to generalize these results to the whole peer-review system. This work just aims to be a starting point for the quantitative study of the peer-review and new analyses that confirm these results will be welcomed. In addition, the exploration of other similar platforms such as F1000, PubPeer, Scholastica, etc., are badly needed for shedding light on this academic process and the new challenges that it faces today.

Additional comments about the several new indicators that have been proposed for this study are needed. The number of different reviewed journals (*reviewed journals*) is proposed as quality indicator because it is assumed that as more reviews are done for different journals, more experience a reviewer gains and could be more appreciated for different editors. In fact, the inverse correlation between *reviewed journals* and *acceptance* (ρ=-.133) allows us to think that reviewing for different journals could improve the strictness of a reviewer. The average impact of the journals is just a proxy to observe the importance of the reviewed journals. It is true that this impact highly varies between disciplines and that this variation affects in the same way to journals and authors. The positive correlation between this indicator and the bibliometric ones would explain this fact and it suggests that reviewers revise for those same journals in which they also publish, connecting the impact of the journal with the author's performance.

Conclusions

Several conclusions can be drawn from the results about the relationship between the bibliometric performance of the Publons' members and their peer-review activity. Correlations have shown that the relationship between both academic activities is weak and they suggest that the peer-review activity is an independent facet of the scholarly activity. However, correlations among peer-review metrics show that the ratio of manuscript acceptance is inverse to the number of reviews.

The use of decision trees has made possible the identification of which type of researcher reviews more articles. Senior male scholars are the members that produce most of the reviews because they have accumulated more reviews throughout their academic career path. In contrast, young female scholars have the strictest acceptance criteria.

Finally, a third conclusion is that Publons offers interesting information and metrics on the peer-review activity of their members and, just for this reason, this social platform may be considered a promising tool for the exploration of peer-review activities.

References

Abramo, G., & D'Angelo, C. A. (2011). Evaluating research: From informed peer review to bibliometrics. Scientometrics, 87(3): 499-514.

Aksnes, D. W., & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. Research evaluation, 13(1): 33-41.

Black, N., Van Rooyen, S., Godlee, F., Smith, R., & Evans, S. (1998). What makes a good reviewer and a good review for a general medical journal? Journal of the American Medical Association, 280(3): 231-233.

Bornmann, L., & Daniel, H. D. (2008). Selecting manuscripts for a high-impact journal through peer review: A citation analysis of communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. Journal of the American Society for Information Science and Technology, 59(11): 1841-1852.

Burnham, J. C. (1990). The evolution of editorial peer review. Journal of the American Medical Association, 263(10):1323–1329

Callaham, M. L., Baxt, W. G., Waeckerle, J. F., & Wears, R. L. (1998). Reliability of editors' subjective quality ratings of peer reviews of manuscripts. Journal of the American Medical Association, 280(3): 229-231.

Callaham, M., & McCulloch, C. (2011). Longitudinal trends in the performance of scientific peer reviewers. Annals of emergency medicine, 57(2): 141-148.

Callaham, M. L., & Tercier, J. (2007). The relationship of previous training and experience of journal peer reviewers to subsequent review quality. PLoS medicine, 4(1): 32.

Cole, S., & Simon, G. A. (1981). Chance and consensus in peer review. Science, 214(4523): 881-886.

Donaldson, M. R., Hanson, K. C., Hasler, C. T., Clark, T. D., Hinch, S. G., & Cooke, S. J. (2010). Injecting youth into peer-review to increase its sustainability: a case study of ecology journals. Ideas in Ecology and Evolution, 3:1-7.

Evans, A. T., McNutt, R. A., Fletcher, S. W., & Fletcher, R. H. (1993). The characteristics of peer reviewers who produce good-quality reviews. Journal of general internal medicine, 8(8): 422-428.

Godlee, F., Gale, C. R., & Martyn, C. N. (1998). Effect on the quality of peer review of blinding reviewers and asking them to sign their reports: a randomized controlled trial. Journal of the American Medical Association. 280(3):237-240.

Haug, C. J. (2015). Peer-review fraud—hacking the scientific publication process. New England Journal of Medicine, 2015(373): 2393-2395.

Khabsa, M., & Giles, C. L. (2014). The number of scholarly documents on the public web. PloS one, 9(5): e93949.

Kassirer, J. P., & Campion, E. W. (1994). Peer review: crude and understudied, but indispensable. Journal of the American Medical Association, 272(2): 96-97.

Kliewer, M. A., Freed, K. S., DeLong, D. M., Pickhardt, P. J., & Provenzale, J. M. (2005). Reviewing the reviewers: comparison of review quality and reviewer characteristics at the American Journal of Roentgenology. American Journal of Roentgenology, 184(6): 1731-1735.

Kronick, D. A. (1990). Peer review in 18th-century scientific journalism. Journal of the American Medical Association, 263(10): 1321-1322.

Kumar, M. N. (2014). Review of the ethics and etiquettes of time management of manuscript peer review. *Journal of Academic Ethics*, *12*(4), 333-346.

Kurihara, Y., & Colletti, P. M. (2013). How do reviewers affect the final outcome? Comparison of the quality of peer review and relative acceptance rates of submitted manuscripts. American Journal of Roentgenology, 201(3): 468-470.

Lantz, B. (2015). Machine Learning with R. Birmingham: Packt Publishing

Lerner, E. J. (2003). Fraud shows peer review flaws. Industrial Physicist, 8(6): 12–17.

McCarty, J. A., & Hastak, M. (2007). Segmentation approaches in data-mining: A comparison of RFM, CHAID, and logistic regression. Journal of business research, 60(6): 656-662.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. Cognitive therapy and research, 1(2): 161-175.

Nguyen, V. M., Haddaway, N. R., Gutowsky, L. F., Wilson, A. D., Gallagher, A. J., Donaldson, M. R., Hammerschlag, N., & Cooke, S. J. (2015). How long is too long in contemporary peer review? Perspectives from authors publishing in conservation biology journals. PloS one, 10(8): e0132557.

Nicholas, D., Watkinson, A., Jamali, H. R., Herman, E., Tenopir, C., Volentine, R., Allard, S., & Levine, K. (2015). Peer review: still king in the digital age. Learned Publishing, 28(1): 15-21.

Opthof, T., Coronel, R., & Janse, M. J. (2002). The significance of the peer review process against the background of bias: priority ratings of reviewers and editors and the prediction of citation, the role of geographical bias. Cardiovascular research, 56(3): 339-346.

Oxman, A. D., Guyatt, G. H., Singer, J., Goldsmith, C. H., Hutchison, B. G., Milner, R. A., & Streiner, D. L. (1991). Agreement among reviewers of review articles. Journal of clinical epidemiology, 44(1): 91-98.

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.

Pautasso, M., & Schäfer, H. (2009). Peer review delay and selectivity in ecology journals. Scientometrics, 84(2): 307-315.

Peters, D. P., & Ceci, S. J. (1982). Peer-review practices of psychological journals: The fate of published articles, submitted again. Behavioral and Brain Sciences, 5(02): 187-195.

Price, D. (1963). Little science, big science. New York: Columbia University Press. 119 p.

Purcell, G. P., Donovan, S. L., & Davidoff, F. (1998). Changes to manuscripts during the editorial process: characterizing the evolution of a clinical paper. Journal of the American Medical Association, 280(3): 227–228.

Ritschard, G. (2014). CHAID and Earlier Supervised Tree Methods. In: McArdle, J. J. & Ritschard, G. Contemporary issues in exploratory data mining in the behavioral sciences. New York: Routledge

Rothwell, P.M., & Martyn, C.N. (2000). Reproducibility of peer review in clinical neuroscience: is agreement between reviewers any greater than would be expected by chance alone? Brain, 123(9): 1964-1969.

Schriger, D.L., Kadera, S.P., & Von Elm, E. (2016). Are Reviewers' Scores Influenced by Citations to Their Own Work? An Analysis of Submitted Manuscripts and Peer Reviewer Reports Presented as a poster at the Seventh International Congress on Peer Review and Biomedical Publication, September 2013, Chicago, IL. Annals of Emergency Medicine, 67 (3): 401-406.

Snell, L., & Spencer, J. (2005). Reviewers' perceptions of the peer review process for a medical education journal. Medical Education, 39(1): 90–97.

Squazzoni, F., Bravo, G., & Takács, K. (2013). Does incentive provision increase the quality of peer review? An experimental study. Research Policy, 42(1): 287-294.

Stossel, T. P. (1985). Reviewer status and review quality. New England Journal of Medicine, 312(10): 658-659.

Thomas, P. R., & Watkins, D. S. (1998). Institutional research rankings via bibliometric analysis and direct peer review: A comparative case study with policy implications. Scientometrics, 41(3): 335-355.

Tite, L., & Schroter, S. (2007). Why do peer reviewers decline to review? A survey. Journal of epidemiology and community health, 61(1): 9-12.

Travis, G. D. L., & Collins, H. M. (1991). New light on old boys: cognitive and institutional particularism in the peer review system. Science, Technology & Human Values, 16(3): 322-341.

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

Weller, A. C. (2001). Editorial Peer Review: Its Strengths and Weaknesses**.** Medford, NJ: Information Today, (ASIS&T Monograph Series). 342 p.

Yankauer, A. (1990). Who are the peer reviewers and how much do they review? Journal of the American Medical Association, 263(10): 1338-1340.