Understanding the Peer Review Endeavor

Shenmeng Xu
University of North Carolina at Chapel Hill, United States
shenmeng@unc.edu

Guangyao Zhang
Dalian University of Technology, China
sdgyzhang@163.com

Yao Sun
Dalian University of Technology, China
sylviass11@163.com

Xianwen Wang
Dalian University of Technology, China
xianwenwang@dlut.edu.cn

ABSTRACT

Peer review plays an essential role in the scholarly publishing life cycle. Using the verified peer review records of researchers on Publons, this study investigated how this scientific endeavor was distributed throughout the world. In addition, we took an initial step toward understanding the length of reviews as a potential indicator of time and effort put into this “scientific gatekeeping” process. The proficiency of English, the scientific Lingua Franca, was found to have a significant effect on the length of reviews. In addition, there was a significant effect of a country’s economic development level on review length. Furthermore, we explored a subgroup of reviewers – reviewers in Singapore – in more depth, and found that reviewers in different genders, disciplines, economic and cultural backgrounds wrote reviews of significantly different lengths.

KEYWORDS

Peer Review; Publons; Scholarly Communication; Scholarly Publishing; Global Science

ASIS&T THESAURUS

Informetrics; Scientometrics; Scholarly Communication

INTRODUCTION

The peer review we focus on in this study is the mechanism for quality control before the acceptance and publication of scholarly articles (Borgman & Furner, 2002; Rowland, 2002). In this process, normally two or three reviewers (in some cases also called referees) provide the author(s) and editor(s) with written evaluation and an overall recommendation of the submission through one or more rounds of review. Since Le Journal des Sçavans (Journal of Scholars) was published in 1665, the practice of peer review has had a history of more than 300 years and has become the internationally accepted practice (Crawford & Stucki, 1990). However, the specific forms and processes of peer review differ depending on the publishing venues, publishers, norms in different disciplines and areas, and other factors. Given the private nature of the peer review practice, it has been a challenge to gain a large picture of the peer review landscape throughout the world.

The peer review process is highly regarded by scholars and considered to be essential to the communication of scholarly research (Mulligan, Hall, & Raphael, 2013). Acknowledged as an effective instrument for self-regulation in the evaluation and selection of scientific contributions (Abelson, 1980; Goodman, Berlin, Fletcher, & Fletcher, 1994; Shatz, 2004), peer review has been playing an important role in the evolution of scientific knowledge. At the same time, peer review is recognized as a flawed process (Eysenck & Eysenck, 1992). The unreliability of peer review has been demonstrated (Ceci & Peters, 1982; Ernst, Saradeth, & Resch, 1993). The validity of judgments in the peer review process is also often questioned (Armstrong, Idriss, Kimball, & Bernhard, 2008; Chubin & Hackett, 1990). Overall, researchers have shown mixed attitudes toward the peer review practice (Mulligan et al., 2013). Researchers from all over the world nowadays participate in the peer review process in different roles (authors, reviewers, and editors), marking it even more difficult to understand the many facets of peer review.

Publons was launched in 2012 as a website for academics to record and present their peer review and editorial activities for academic journals. Publons partners with academic journals so that the peer review activities that researchers choose to claim and showcase can be verified. By 2018, more than 500,000 reviewers have joined the site, adding more than one million reviews across 25,000 journals (Publons, 2018). In 2018, Publons released a Global State of Peer Review Report (Publons, 2018), addressing several issues including the “supply and demand chain” of peer review and academic publishing, the reviewer fatigue phenomenon, the efficiency of the peer review process, as well as the quality of peer review.

Aiming to provide complementary understandings to the Publons global report, driven by the following research questions, this study aims to draw back the curtain of the peer review endeavor in more depth:

1. How are peer reviewers distributed all over the world?
2. How are peer reviews distributed all over the world?
3. How long are reviews written by reviewers throughout the world?

82nd Annual Meeting of the Association for Information Science & Technology | Melbourne, Australia | 19–23 October, 2019

DOI: 10.1002/pra2.00026

4. Is there an effect of language (English) proficiency of reviewers on the length of reviews?
5. Based on a smaller sample of researchers affiliated with a Singapore organization, as a preliminary exploration: Do factors including gender, discipline, academic position, economic background, and cultural background of reviewers affect the length of reviews?

In this study, the length of reviews is studied as an initial step toward understanding it as a potential indicator of time and effort researchers devote to the peer review process.

METHODS AND DATA

Data Collection and Cleaning

On Publons.com, a researcher's peer review contributions for scholarly journals are verified and displayed on their Publons homepage. We used a Python program to collect the researchers' identity and affiliation information, as well as the counts of their peer review activities and the average length of their review texts. Our data were collected in April 2018. Starting from the aggregated page of all the countries (https://publons.com/country/), we included all reviewers that had at least one verified review in our dataset. In total, we collected 82,798 data points.

To investigate the relationship between review length and language proficiency, we collected the EF English Proficiency Index (EPI) for each country. The EF EPI indicators were calculated aiming to rank countries by the average level of English language skills among those adults who took the EF test. After excluding the countries with no EF score data available, we had 77,859 data points.

We then manually cleaned the data by removing duplicates and comparing the information on the Publons website with our data wherever erroneous data were spotted. After data cleaning, we obtained 76,370 data points. These data points are not unique Publons reviewers. Instead, each of them represents one reviewer with a unique affiliation. Considering that it is impossible to accurately identify the main affiliations of the reviewers on a large scale if they had multiple affiliations listed, those with two or more affiliations are counted multiple times in our dataset. In other words, if one reviewer has two affiliations, he or she accordingly appears as two data points. In total, there are 74,785 unique Publons reviewers (out of the 76,370 data points mentioned above) in this dataset. This dataset is our final dataset for the analysis of the distribution of reviewers and reviews.

Further Data Preparation

After we had obtained the full final dataset described above, we further processed the data to prepare for the analyses of review lengths and the factors that would potentially affect the length of reviews.

For the analyses focusing on review length, first, we removed all data points that had a review length of 50 words or less. This was because these data were considered incomplete or erroneous data. More explanations and discussions will be provided in the Discussion section of this paper. This step excluded 4,374 (5.72%) of the total data points. Secondly, we excluded data points that had a review count of one, because in the process of data cleaning described above, we found many erroneous data among these relatively inexperienced reviewers (or reviewers who have relatively incomplete and erroneous records on Publons). In this step, we deleted 24,437 data points. Our final dataset for the analyses relevant to review lengths contained 47,559 data points.

To further explore the potential factors that would affect the review lengths, we extracted all reviewers who were currently affiliated with an organization in Singapore for further analyses. Singapore has a diverse population with people from diverse cultural backgrounds and economies. We manually collected their gender, title, discipline, country background, and cultural background information. Table 1 presents the coding scheme used in this "Singapore dataset".

Gender is determined by the combination of name classification and the reviewers' online profile photos (on websites including Publons, Google Scholar, LinkedIn, and ResearchGate). Discipline was determined based on the Research Fields of the reviewers on Publons, including two large categories: the humanities and the social sciences were categorized as "Social Science"; the hard sciences were categorized as "Natural Science". Country background and cultural background were determined based on the country where the reviewers completed their undergraduate level education. The reason for this was because although these reviewers were currently affiliated with a research institution in Singapore, they came from different economic and cultural backgrounds. The country background variable was labeled according to the World Bank's classification mentioned in the following section; Cultural background was classified into three categories: Confucianism influenced culture, Christianism influenced culture, and others. The more detailed rationales and interpretations of this classification are provided in the Discussion section.

We obtained 199 reviewers affiliated with Singapore institutes who have sufficient reliable information online for us to label the complete variables listed in Table 1. This process was conducted by two authors. After checking for agreement and further examination, a complete agreement was reached for all reviewers coded in this dataset.

Data Analysis

In the Global State of Peer Review Report (Publons, 2018), countries were categorized into established countries and emerging countries "based on the typical classification of emerging economies" (p17). We adopted a similar approach. We applied the World Bank's latest classification of countries and regions based on income (retrieved from http://databank.worldbank.org/data/download/site-content/CLASS.xls) to label all the countries in our dataset.

Variable	Values
Gender	Male, Female
Discipline	Social Science, Natural Science
Title/Position	Assistant Professor, Associate Professor, Professor, Researchers, Senior Researcher
Country Background	Emerging, Established
Cultural Background	Confucianism, Christianism, Others
h-index	The value collected from the reviewers' Google Scholar homepages
Scopus	The value collected from Scopus by searching for the reviewers' publications

Table 1. Coding scheme of Singapore reviewers.

We performed statistical analyses using Python packages NumPy, SciPy, Pandas, and Regressor. We calculated the descriptive statistics to understand the distribution of reviewers, reviews, as well as the length of reviews. The geographical distribution of reviewers and reviews (sum and average) in different countries were visualized using the Python packages GeoPandas and Matplotlib. Linear Regression was run to explore the effect of English language proficiency on the length of reviews by reviewers from different countries. In our further investigation of the Singapore dataset, we used One-way ANOVA to test if gender, discipline, academic position, economic background, and cultural background are factors that would affect the length of reviews. We additionally performed Kruskal–Wallis tests to confirm the ANOVA results.

RESULTS

Distribution of Reviewers

In our dataset, there are 74,785 unique reviewers who have verified peer review records on Publons. These reviewers are from 70 countries over the world. The descriptive statistics of the number of reviewers by country is presented in Table 2 below, including the minimum, maximum, 1st, 2nd (median), and 3rd quantiles, as well as the standard deviation and skewness. These data are highly skewed, indicating a large number of reviewers are affiliated with research organizations in a small number of countries.

The countries and regions with the top 30 largest numbers of reviewers are shown in Table 3 below. Approximately half of all the reviewers are affiliated with the top five countries in this list.

Min	Q1	Median	Q3	Max	Mean	Sd.	Skewness
2	69.5	342.5	915	16716	1091	2316	4.85

Table 2. Descriptive statistics of the number of reviewers by country.

Rank	Country	No. of reviewers	% of reviewers
1	United States	16716	22.35%
2	United Kingdom	6684	8.94%
3	Australia	5567	7.44%
4	Italy	4318	5.77%
5	China	3857	5.16%
6	Spain	3660	4.89%
7	Canada	3046	4.07%
8	Japan	2699	3.61%
9	Germany	2439	3.26%
10	India	2272	3.04%
11	Brazil	2062	2.76%
12	Turkey	1636	2.19%
13	France	1566	2.09%
14	Iran	1454	1.94%
15	Sweden	1405	1.88%
16	Portugal	1207	1.61%
17	New Zealand	999	1.34%
18	Belgium	924	1.24%
19	Malaysia	888	1.19%
20	South Korea	825	1.10%
21	Switzerland	795	1.06%
22	South Africa	662	0.89%
23	Taiwan	645	0.86%
24	Poland	634	0.85%
25	Mexico	615	0.82%
26	Egypt	594	0.79%
27	Ireland	581	0.78%
28	Norway	562	0.75%
29	Greece	561	0.75%
30	Denmark	538	0.72%

Table 3. Top 30 countries with the largest number of reviewers.

Considering the highly skewed nature of the counts of reviewers in different countries, we calculated the log value of the counts to normalize the data visualized in Figure 1. In this map, the darker color indicates the larger number of reviewers in a country. Countries with higher number of reviewers are mainly concentrated in North America, Western Europe, Australia, and East Asia. The majority of missing data are countries in Africa.

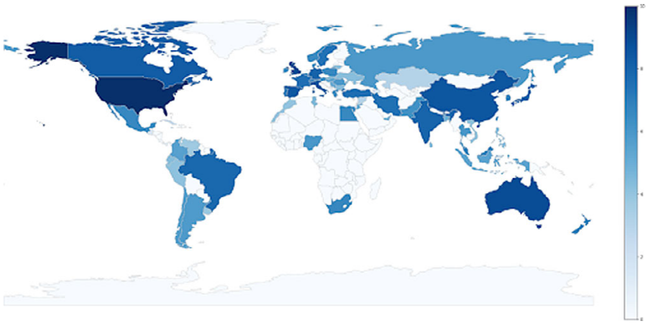

Figure 1. Choropleth map of reviewer distribution by country (log data).

Distribution of Reviews

In our dataset, 464,728 reviews have been verified on Publons. The descriptive statistics are shown in Table 4 below. The most productive reviewer in our dataset has reviewed 2,613 articles. On average, each reviewer has 6.09 verified reviews. The distribution is highly skewed. The median is one review with two verified review records.

Min	Q1	Median	Q3	Max	Mean	Sd.	Skewness
1	1	2	5	2613	6.09	18.79	47.51

Table 4. Descriptive statistics of the number of reviews by country.

After aggregating the counts of reviews by country, we ranked them according to the sum of reviews by reviewers in each country (3rd column). The average number of reviews in each country normalized by the number of reviewers are also included (4th column). The ranking of these values is different from the sum values. Out of the 30 countries in this list, Greece has the highest average number of reviews per reviewer, followed by Switzerland, Austria, and Portugal. However, Syria and Morocco have higher average numbers of reviews per reviewer, which is shown in Figure 3. but not in Table 5.

We calculated the log value of the total number of reviews by country and visualized them in Figure 2. Overall, this map shows a similar pattern to Figure 1. Countries with higher number of reviews are mostly located in North America, Western Europe, Australia, and East Asia.

Further, we mapped the average number of reviewers by country in Figure 3a. In Figure 3b, a different view is displayed. Syria and Morocco are two countries with significantly higher average number of reviews.

Rank	**Country**	**Sum of reviews**	**Avg.**	**Median**
1	United States	92303	5.52	2
2	United Kingdom	38263	5.72	2
3	Australia	33818	6.07	2
4	Italy	27056	6.27	2
5	Japan	21546	7.98	3
6	Spain	21321	5.83	2
7	Germany	19839	8.13	3
8	Canada	19562	6.42	2
9	China	16452	4.27	2
10	India	13052	5.74	2
11	Brazil	11594	5.62	2
12	France	11412	7.29	3
13	Portugal	10362	8.58	3
14	Sweden	9417	6.70	3
15	Turkey	8147	4.98	2
16	Iran	7702	5.30	2
17	Switzerland	7163	9.01	3
18	New Zealand	6934	6.94	3
19	Greece	6392	11.39	3
20	Belgium	5665	6.13	3
21	Egypt	4550	7.66	3
22	Malaysia	4524	5.09	2
23	Taiwan	4371	6.78	2
24	South Korea	4318	5.23	2
25	Austria	4276	9.00	3
26	Norway	4142	7.37	2
27	Denmark	4045	7.52	3
28	Poland	3688	5.82	2
29	South Africa	3450	5.21	2
30	Mexico	3394	5.51	3

Table 5. Top 30 countries with the largest number of reviews.

Length of Reviews

In this section, we mainly describe our exploration of the effect of language (i.e., English proficiency) on review length. The average length of reviews for each country was calculated. A linear regression was conducted using the EF EPI as the independent variable. A significant regression

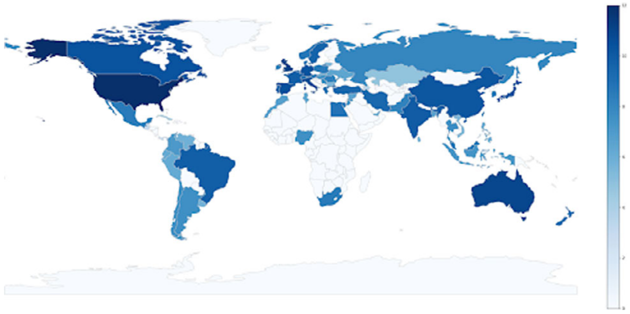

Figure 2. Choropleth map of review distribution (sum) by country (log data).

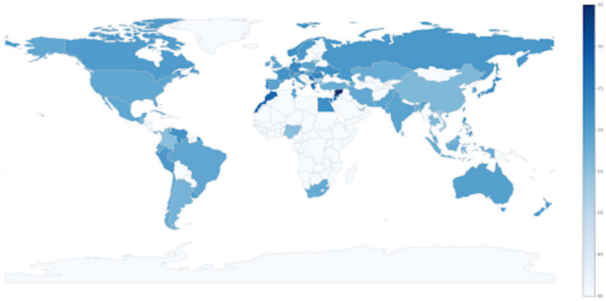

Figure 3a. Choropleth map of review distribution (average) by country (log data).

equation was found (p = .00), with an R^2 of 0.26 and an adjusted R^2 of 0.25. The predicted length of reviews is equal to 45.22 + 6.12 (EF EPI).

According to the classification of World Bank described in the Data Analysis section, 35 countries out of these 70 countries were classified as established countries; The other 35 countries were classified as emerging countries. In Figure 3b, the established countries are labeled as orange dots and the emerging countries are labeled as blue ones.

To test if there existed a statistical difference between these two types of economies, a one-way ANOVA test was conducted to compare the effect of economy types on review length between established countries and emerging countries. After normalizing the length data for reviews, they met the normal distribution assumption (established countries p-value = .51, emerging countries p-value = .24) and the homogeneity of variance assumption (Levene's Test p-value = .91).

There was a significant effect of economy types on review length at the $p < .05$ level (p = .00). Nonparametric comparison also indicated that the mean length for the established countries was significantly different than that in the emerging countries (Kruskal–Wallis test p-value = .00).

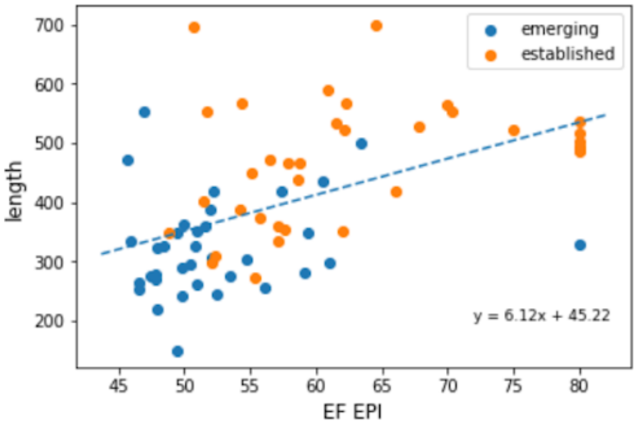

Figure 3b. Choropleth map of review distribution (average) by country (log data).

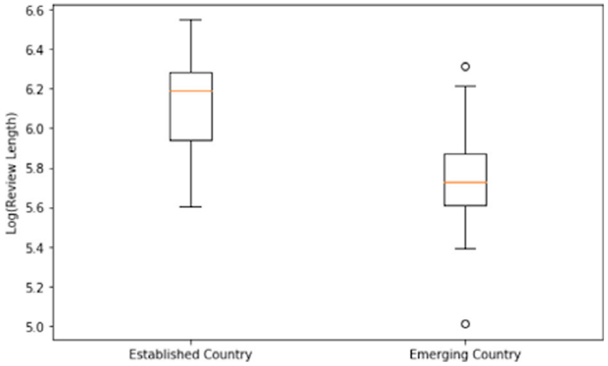

Figure 4. Boxplot of review length data in established and emerging countries.

Exploration of Factors in Review Length: Further Analyses Based on the Singapore Data

As described in the Data Analysis section, reviewers in Singapore were selected and further coded for our initial exploration of factors other than language that would potentially affect the length of reviews. This resulted in a complete dataset of 199 reviewers affiliated with Singapore organizations. The potential factors that we explored included gender, academic position, discipline, economic background, and cultural background. For each of the comparison, we visualized the different groups of data using boxplots (Figure 6, 8, 10, 12, and 14), and conducted one-way ANOVA tests and Kruskal–Wallis tests where appropriate.

Gender

Among these 199 reviewers affiliated with Singapore organizations, women are significantly underrepresented.

In order to test if there existed a statistical difference between male and female reviewers, a one-way ANOVA test was performed. After normalizing the length data for reviews by male and female reviewers, they met the normal distrbution

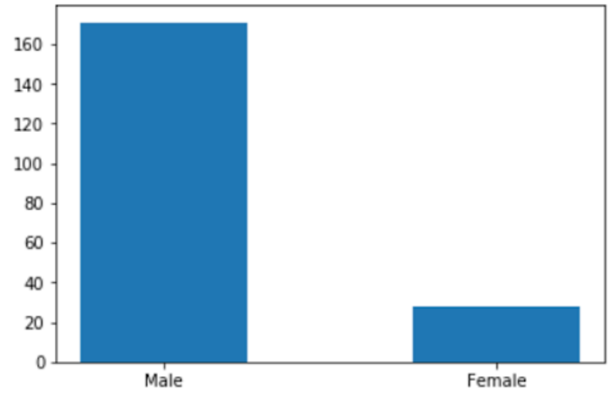

Figure 5. Distribution of gender.

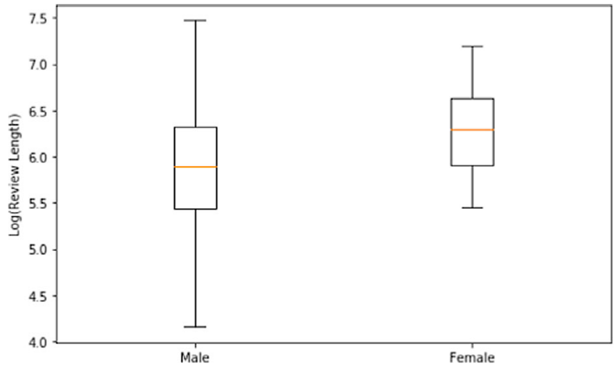

Figure 6. Boxplot of review length data by male and female reviewers in Singapore.

assumption (male *p*-value = .67, female *p*-value = .30) and the homogeneity of variance assumption (Levene's Test *p*-value = .07).

A significant difference was found in the lengths by male and female reviewers (*p* = .00). In addition, nonparametric comparison also indicated a significant differece (Kruskal–Wallis test *p*-value = .00).

Academic Position

Among these subgroups of reviewers, Assistant Professors and Associate Professors account for nearly 60% of all the reviewers.

A one-way ANOVA test was conducted to compare the effect of different academic positions on review length. After normalizing the length data for reviews by reviewers in a diverse range of academic positionss, they met the normal distrbution assumption (associate professor *p*-value = .53, assistant professor *p*-value = .29, professor *p*-value = .90, researcher *p*-value = .77, senior research *p*-value = .40). The homogeneity of variance assumption was also met (Levene's Test *p*-value = .34).

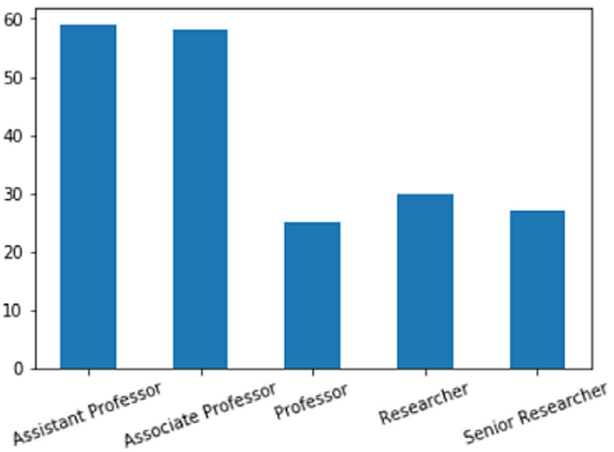

Figure 7. Distribution of academic positions.

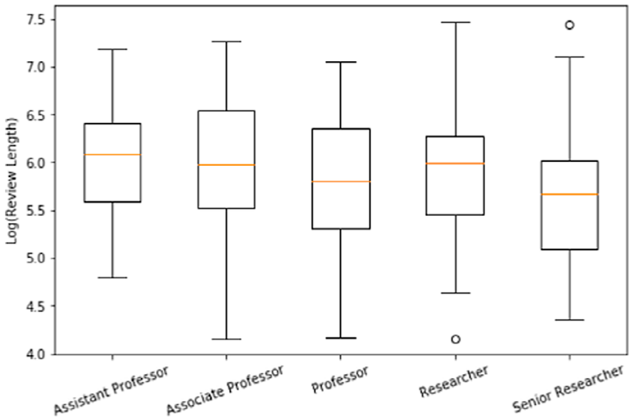

Figure 8. Boxplot of review length data by reviewers in different academic positions in Singapore.

There was no significant difference in the length of reviews by reviewers in different academic positions (*p* = .31). To confirm this result, we conducted a Kruskal-Wallis test, which indicated the same result (*p*-value = .19).

Discipline

Compared to reviewers in the natural sciences, i.e., Science, Technology, Engineering and Medical (STEM) disciplines, reviewers in the social sciences and humanities are underrepresented in this subgroup of reviewers.

Similarly, a one-way ANOVA test was conducted to investigate whether there was a difference between the length of reviews written by reviewers in natural science disciplines and the social sciences and humanities. After the length data was normalized, the length of reviews by social sciences and humanities reviewers was not in a normal distribution (*p*-value = .03), while review length data by natural science reviewers were normally distributed (*p*-value = .62). Because the normal distribution assumption was violated, we did not

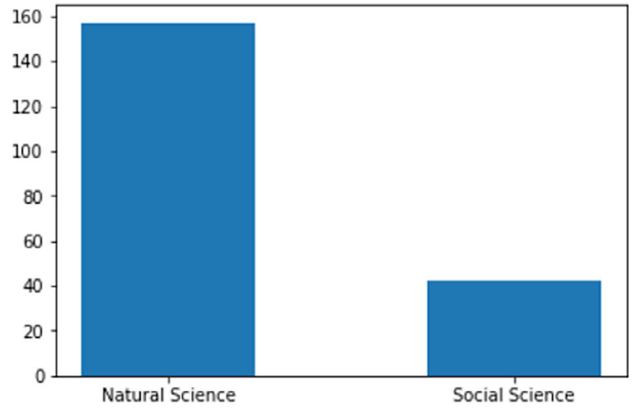

Figure 9. Distribution of disciplines.

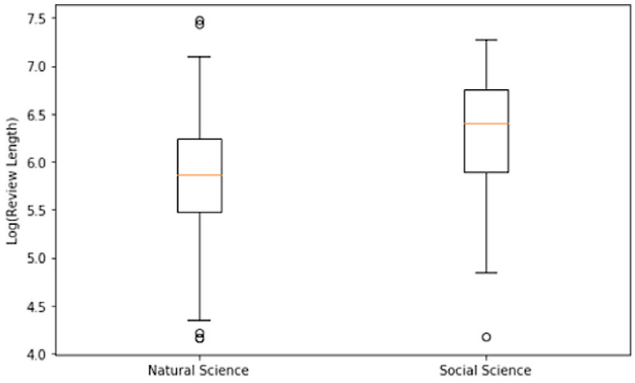

Figure 10. Boxplot of review length data by reviewers in different disciplines in Singapore.

conduct the one-way ANOVA test. Instead, the nonparametric Kruskal–Wallis test *p*-value was .00, indicating a significant difference between the natural science disciplines and the social sciences and humanities, even if the social sciences and humanities length data were still skewed after being log-transformed. In fact, from the boxplot (Figure 7), it can be seen that a large number of reviews in social sciences and humanities tend to give relatively longer reviews, leading to the skewness.

Economic Background

The difference of review length between different economic backgrounds was then examined. The log data of the review length were normally distributed (established countries *p*-value = .07, emerging countries *p*-value = .18). The assumption of homogeneity of variance was met (Levene's Test *p*-value = .34). The result of the one-way ANOVA test indicated that there was a significant difference in the length of reviews written by reviewers from different economic backgrounds (*p*-value = .00). Similarly, the Kruskal-Wallis test also indicated a significant difference between these two groups of reviewers (*p*-value = .00).

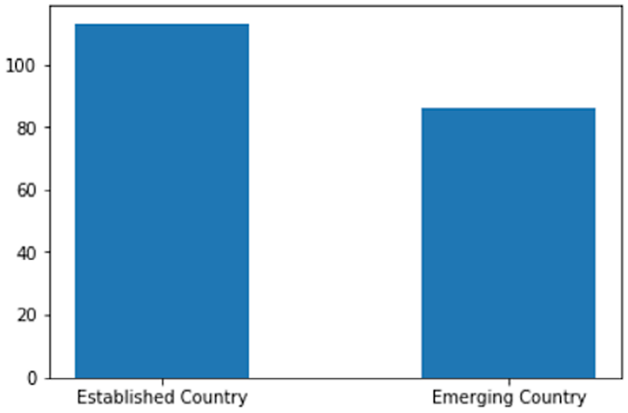

Figure 11. Distribution of economic backgrounds.

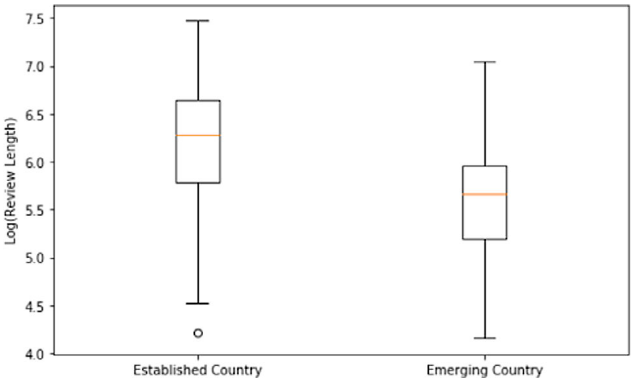

Figure 12. Boxplot of review length data by reviewers with different economic backgrounds in Singapore.

Cultural Background

Reviewers from a Confucianism influenced cultural background take up the largest proportion of reviewers in Singapore.

Last, the difference of review lengths provided by reviewers in different cultural backgrounds were investigated. After normalizing the length data, the length of reviews written by reviewers from a Confucianism influenced cultural background and other background met the normal distribution assumption (Confucianism *p*-value = .44, Others *p*-value = .08), but the length data of reviewers from a Christianism influenced cultural background (*p*-value = .04) was not normally distributed. As a result, the nonparametric Kruskal–Wallis test was conducted. The result suggested that reviewers from a Christianism influenced cultural background wrote significantly longer reviews than Confucianism influenced and other cultural backgrounds.

DISCUSSIONS

Interpreting the Factors

In our exploration, we found several potential factors that might affect the length of reviews, including gender,

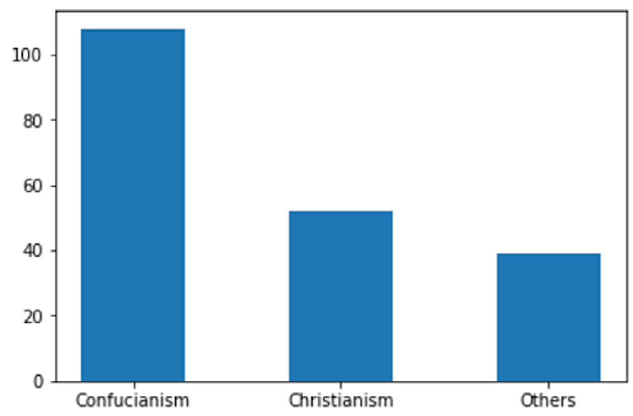

Figure 13. Distribution of cultural backgrounds.

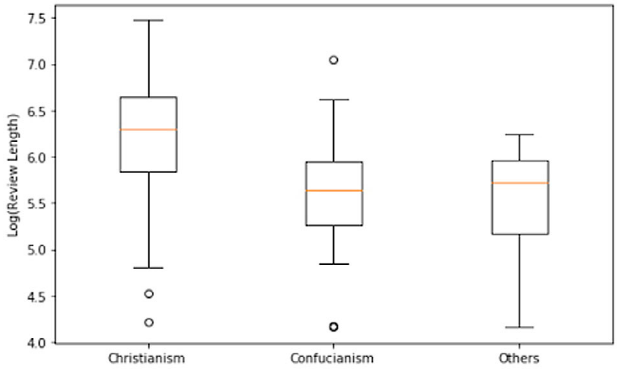

Figure 14. Boxplot of review length data by reviewers with different cultural backgrounds in Singapore.

discipline, English language proficiency, economic development level, and cultural background of a country. Among these factors, the latter three factors are relevant at the country level. They overlap with each other and might have compounding effects on each other. In terms of their effects on the length of reviews, each of these factors may play a role to a different extent that can hardly be measured.

The use of English has always been a complex issue over the world and in history. In Australia, New Zealand, United Kingdom, and United States, English is used as a de facto official language; In some other countries, English is used as a de jure official language. In these countries, English is in some cases the primary language, in some cases only used in official and educational uses, and in some cases used as Lingua Franca. In this Publons dataset, the peer reviews are typically performed for international journals, the majority of which are published in English. As a result, the so-called linguistic imperialism (Phillipson, 1992) phenomenon exists in this context. Considering some countries do not use English as their official languages or commonly used languages, it is likely that reviewers in those countries tend to write less in volume in a language that they are not proficiently familiar with (Hinkel, 2011). Empirical studies have also provided evidence that critical thinking is more difficult in a second language (e.g., Floyd, 2011).

Institutional factors relevant to the national scientific systems and norms should also be taken into consideration in the interpretation of our findings. The form of peer review is a relatively newer practice to researchers in the economically and scientifically emerging countries. For instance, even one decade ago, the main refereeing scheme for many Chinese academic journals was a system called “three-level review,” which was carried out mainly by journal editors (Fang, Xu, & Lian, 2008). The three-level review is based on the hierarchy of Chinese publishing houses: the first review is done by the editor, the second review by the head of the editorial department, and the final review by the editor-in-chief of the publishing house (Fang et al., 2008). As a result, researchers who do not serve as editors might be less familiar with peer review guidelines and have less exposure to discussions and trainings about the scientific norms and best practices in the peer review process.

Publons’ approach turning peer review into a visible and potentially measurable research output is a relatively new and not well received practice, probably particularly in the scientifically emerging countries where the focus of researcher evaluation is on the output and impact of scientific publications (King, 1987; King, 2004). In contrast, in western countries, it is more widely accepted that academics can use their review and editorial records as evidence of their standing and influence in their field (Bornmann, 2011). For instance, conducting peer review is not counted as research outputs or research impact in the current research evaluation criteria in the majority of Chinese research universities and institutions. This would result in researchers paying less time and effort in reviewing others’ works as their time and effort are not recognized and rewarded.

In our hypotheses, culture also plays a role in peer review. For instance, east Asian countries’ high-context culture prefers indirect communication over direct communication (Hall, 1976; Kim, Pan, & Park, 1998), which is believed to be contributed by the impact of Confucianism (Yum, 1988). Particularly, in the peer review process, providing criticism for authors involves issues relevant to modesty, politeness, and conflict management. The confrontation-avoiding disposition (Kim et al., 1998) in eastern cultures might lead to a shorter length of criticism.

In this section of discussion, we do not intend to provide speculations about the reasons behind the different lengths of review. Instead, we would like to bring more relevant factors into our horizon so that they will not be ignored in the interpretation of our findings or in future studies on peer review. It is also important to note that in this paper we do not intend to associate the length of reviews with the quality of reviews.

Limitations

There are several limitations in this study. First, the data in this study were collected in April of 2018. Given the increasing number of users and their records, these data need to be updated, potentially with more granularity. Ideally, we also hope to seek a way to confirm the data with Publons in the future. Currently, it is hard to compare our results to the Publons' Peer Review Report (Publons, 2018) due to different time frames.

The second limitation is in the length data and is rooted in the fact that the length data on Publons were calculated based on the review text submitted by reviewers themselves. This submission process was not mandatory. As a result, some reviewers do not submit review text onto Publons; Some reviewers submit the review text for some reviews but not others. In addition, the lack of a clear description about what should be submitted (i.e., the complete review, the first round of review, the last round of review, or the review decision sentence, etc.). This might lead to reviewers submitting different content to Publons. An additional concern about the length data is the ease of gaming. Unlike the review records, which are verified by Publons or the publishers, the actual review text does not need to go through any verification. This might provide an ease of gaming for researchers who would like to boost their review contribution by length data. Currently, the allegedly "average length" of reviews by one reviewer displayed on the Publons website is very likely to be calculated based on the total number of words divided by the total number of reviews that is not null in the "review text" field (instead of the total number of words in the reviews divided by the total number of reviews). We agree that this is the most optimal strategy based on the current constraint, but would like to point out that caution needs to be used when interpreting statistics based on these statistics.

Thirdly, as mentioned in the Methods and Data section, one reviewer might be affiliated with more than one organization. This is due to the difficulty in determining the main affiliation of reviewers. The strategy we adopted to handle this difficulty was that we treated reviewers with multiple affiliations as multiple reviewers. In total, 1,585 out of 76,370 datapoints in our dataset (2.08%) were duplicated reviewers due to multiple affiliations.

Fourthly, the classification of cultures can be improved and supported by more literature investigation. The reason we chose to classify the cultures into the current three categories was because we found the eastern/western classification of countries dichotomous. It is also not enough to simply classify them into for example, cultures that are directly or indirectly critical, or cultures with high or low context in communication (Hall, 1976; Kim et al., 1998). Since Confucianism is more commonly regarded as a social and ethical philosophy rather than a religion, we could not claim that this classification is based on religion either.

Fifthly, the gender information in our data is not self-reported and can be biased. Also, the binary classification is not inclusive enough to represent all genders.

Finally, this dataset is based on Publons data. It is possible that the self-reported review records are incomplete because of inactiveness of users. In addition, it is not clear to what extent can the verification process eliminate gaming. This is an important point when interpreting our data (for instance, 32% of the reviewers had reviewed only one article, and that the most productive reviewer had reviewed 2,613 articles.)

Future Works

In this study, reviewers in Singapore were studied as an initial step. In the future works, we plan to expand this in-depth analysis of the potential factors that would have effects on review length to a larger sample of reviewers from diverse backgrounds. This at the same time will help us better understand the meaning and indication of the length of reviews.

Another future direction of ours is to gain a more in-depth understanding of how Publons is used by scholars and researchers. This work would lay the groundwork for future research studies using Publons data because it is important to better understand if the data collected from Publons are representative of all peer reviewers.

CONCLUSION

Using the verified peer review records of researchers on Publons, this study investigated how this scientific endeavor was divided and distributed over the world. In terms of the absolute number of reviewers and the number of reviews performed, countries in North America, Western Europe, Australia, and East Asia played a primary role in this process. The data of average number of reviews performed by each reviewer in different countries displayed slightly different patterns.

On the global scale, the proficiency of English, the scientific Lingua Franca, was found to have a significant effect on the length of reviews. In addition, there was a significant effect of a country's economic development level on review length.

In our more in-depth explorations of reviewers in Singapore, we found that reviewers in different genders, disciplines, economic and cultural backgrounds wrote reviews of different lengths. Specifically, women, although underrepresented, write longer reviews. Secondly, although underrepresented, reviewers in social sciences and humanities write longer reviews than reviewers in the natural science disciplines. Thirdly, reviewers from economically established countries write longer reviews compared to those from economically emerging countries. Finally, peer reviewers in Western countries, which are countries with Christianism influenced cultures, write longer reviews than countries with Confucianism influenced cultures and other cultures.

To sum up, based on one of the largest sets of data about peer review so far, this study provided important complementary understandings to the Publons' Global State of Peer Review report (Publons, 2018). In addition to the detailed analysis of the global distributions of reviewers and reviews, we also took an initial step toward understanding the length of reviews as a potential indicator of time and effort researchers devote to the process of "scientific gatekeeping". Delving into the exploration of potential factors that would affect the length of peer review provided additional insight into how the endeavor of peer review is being accepted and performed.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Bradley M. Hemminger for his feedback on improving this research. We thank Dr. Yuan Chong for his support in the data analysis process. We gratefully acknowledge the grant from the National Natural Science Foundation of China (71673038) and the Project of Fundamental Research Funds for the Central Universities (DUT18JC32).

REFERENCES

Abelson, P. H. (1980). Scientific communication. *Science*, *209*(4452), 60–62.

Armstrong, A. W., Idriss, S. Z., Kimball, A. B., & Bernhard, J. D. (2008). Fate of manuscripts declined by the Journal of the American Academy of Dermatology. *Journal of the American Academy of Dermatology*, *58*(4), 632–635.

Borgman, C. L., & Furner, J. (2002). Scholarly communication and bibliometrics. *Annual Review of Information Science and Technology*, *36*(1), 2–72.

Bornmann, L. (2011). Scientific peer review. *Annual Review of Information Science and Technology*, *45*(1), 197–245.

Ceci, S. J., & Peters, D. P. (1982). Peer review: A study of reliability. *Change: The Magazine of Higher Learning*, *14*(6), 44–48.

Chubin, D. E., & Hackett, E. J. (1990). *Peerless science: Peer review and US science policy*. Suny Press.

Crawford, S., & Stucki, L. (1990). Peer review and the changing research record. *Journal of the American Society for Information Science*, *41*(3), 223–228.

Ernst, E., Saradeth, T., & Resch, K. L. (1993). Drawbacks of peer review. *Nature*, *363*(6427), 296.

Eysenck, H. J., & Eysenck, S. B. (1992). Peer review: Advice to referees and contributors. *Personality and Individual Differences*.

Fang, Q., Xu, L., & Lian, X. (2008). Peer-review practice and research for academic journals in China. *Journal of Scholarly Publishing*, *39*(4), 417–427.

Floyd, C. B. (2011). Critical thinking in a second language. *Higher Education Research & Development*, *30*(3), 289–302.

Goodman, S. N., Berlin, J., Fletcher, S. W., & Fletcher, R. H. (1994). Manuscript quality before and after peer review and editing at Annals of Internal Medicine. *Annals of Internal Medicine*, *121*(1), 11–21.

Hall, E. T. (1976). *Beyond culture*. NY: *Dobleday & Company*.

Hinkel, E. (2011). What research on second language writing tells us and what it doesn't. In *Handbook of research in second language teaching and learning* (Vol. *2*, pp. 523–538).

Kim, D., Pan, Y., & Park, H. S. (1998). High-versus low-context culture: A comparison of Chinese, Korean, and American cultures. *Psychology & Marketing*, *15*(6), 507–521.

King, D. A. (2004). The scientific impact of nations. *Nature*, *430*(6997), 311.

King, J. (1987). A review of bibliometric and other science indicators and their role in research evaluation. *Journal of information science*, *13*(5), 261–276.

Mulligan, A., Hall, L., & Raphael, E. (2013). Peer review in a changing world: An international study measuring the attitudes of researchers. *Journal of the American Society for Information Science and Technology*, *64*(1), 132–161.

Phillipson, R. (1992). Linguistic imperialism. *The Encyclopedia of Applied Linguistics*, 1–7.

Publons. (2018). *2018 Global state of peer review by Publons*. Retrieved from https://publons.com/static/Publons-Global-State-Of-Peer-Review-2018.pdf

Rowland, F. (2002). The peer-review process. *Learned publishing*, *15*(4), 247–258.

Shatz, D. (2004). *Peer review: A critical inquiry*. Rowman & Littlefield.

Yum, J. O. (1988). The impact of Confucianism on interpersonal relationships and communication patterns in East Asia. *Communications Monographs*, *55*(4), 374–388.