Characteristics of a Megajournal: A Bibliometric Case Study

C. Sean Burns *

College of Communication and Information
School of Information Science
University of Kentucky, USA
E-mail: sean.burns@uky.edu

ABSTRACT

The term *megajournal* is used to describe publication platforms, like *PLOS ONE*, that claim to incorporate peer review processes and web technologies that allow fast review and publishing. These platforms also publish without the constraints of periodic issues and instead publish daily. We conducted a yearlong bibliometric profile of a sample of articles published in the first several months after the launch of *PeerJ*, a peer reviewed, open access publishing platform in the medical and biological sciences. The profile included a study of author characteristics, peer review characteristics, usage and social metrics, and a citation analysis. We found that about 43% of the articles are collaborated on by authors from different nations. Publication delay averaged 68 days, based on the median. Almost 74% of the articles were coauthored by males and females, but less than a third were first authored by females. Usage and social metrics tended to be high after publication but declined sharply over the course of a year. Citations increased as social metrics declined. *Google Scholar* and *Scopus* citation counts were highly correlated after the first year of data collection (Spearman *rho* = 0.86). An analysis of reference lists indicated that articles tended to include unique journal titles. The purpose of the study is not to generalize to other journals but to chart the origin of *PeerJ* in order to compare to future analyses of other megajournals, which may play increasingly substantial roles in science communication.

Keywords: Megajournals, Bibliometrics, Gender Differences, Peer Review, Open Access, Case Study

Open Access

1. INTRODUCTION

"What is a journal?" (Garfield, 1977, p. 6)

The term *megajournal* (Björk & Solomon, 2013; MacCallum, 2011; Solomon, 2014) is used to describe publication platforms that post large numbers of author pay, open access articles, that do not publish by issue but rather continuously, that involve a claim to objective editorial criteria, that take advantage of web technologies to reduce publication delay and experiment with pre- and post-publication peer review, and that highlight newer post-publication gate-keeping metrics such as article level metrics and altmetrics. Peter Binmore (2013), founder of *PeerJ*, has articulated various terms in an attempt to capture the essence of megajournals such as *PeerJ, PLOS ONE, BMJ Open, Sage Open*, and others, and these terms include *non-selective, impact neutral*, or *rigorous but inclusive review*. In *PeerJ's* case, this means referees are encouraged to focus only on the "soundness" of research and not its presumed "importance" (PeerJ, 2015) when reviewing manuscripts.

With manuscript acceptance rates of around 70% for *PLOS ONE* and *PeerJ* (PLOS ONE, 2014; PeerJ, 2014), megajournals position themselves as inclusive channels of scientific communication. This positioning suggests an interaction between journal prestige and the commercial viability of author pay publishing models. In particular, Lipworth and Kerridge (2011) find that "editors of less prestigious journals […] need to work harder to improve manuscripts since they did not have the liberty of using rejection of manuscripts as a quality control mechanism, and need to be more careful not to disenfranchise potential future authors" (p. 105). For megajournals then, it seems causal that high acceptance rates ordain less prestige, if prestige is related to manuscript quality and measured by impact in the field. However, since megajournals are open access publications and publish large quantities of articles, they may benefit from an open access citation advantage (Eysenbach, 2006) that may counter this interaction, despite manuscript quality. That is, it seems possible to receive a citation advantage when the publication model is based on bulk and the published material is openly accessible.

Megajournals also accentuate theoretical tensions between the normative view of science (Merton, 1973a; Merton, 1973b; Merton, 1973c) and the social constructivism view of science, as outlined in Bornmann (2008) and Bornman & Marx (2012). Bornmann (2008) states that one assumption of the social constructivism view is that "scientific work is a social construction of the scientist under review and the reviewers" [emphasis original] (Bornmann, 2008, p. 31). Interestingly, authors and reviewers who participate in the *PeerJ* publication process have the option of making their peer review histories publicly available. Those reviews that are public are also citable and afforded so by being assigned digital object identifier (DOI) reference URLs to the published article's DOI. Therefore, not only do publicly available peer review histories warrant empirical studies in the quality of peer review, but research that examines the citation potential of specific reviews will raise theoretical questions about contribution, authorship, and scientific norms (Cronin & Franks, 2006).

The existence of megajournals tacitly suggests dramatic changes in scientific and scholarly communication. If this is so, then a number of issues that have been explored in traditional journals need to be explored in megajournals in order to understand how particular differences are expressed. These issues include open access impact or citation advantage (Björk & Solomon, 2012; Eysenbach, 2006), open access author publication fees (Solomon & Björk, 2012), submission and acceptance rates (Opthof, Coronel, & Janse, 2000), the underrepresentation of women in science (Ceci & Williams, 2011), alternative peer review models (Birukou et al., 2011), gender author disparity (West, Jacquet, King, Correll, & Bergstrom, 2013) and order (Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013), gender bias in peer review (Lloyd, 1990; Paludi & Bauer, 1983) and other forms of bias (Lee, Sugimoto, Zhang, & Cronin 2013), publication delays (Bornmann & Daniel, 2010; Björk & Solomon, 2013; Luwel & Moed, 1998; Pautasso & Schafer, 2010), journal internationality (Calver, Wardell-Johnson, Bradley, & Taplin, 2010), length or word count of review as a proxy for the quality of reviews and of journals (Bornmann, Wolf, & Daniel, 2012), inter-reviewer agreement, author and institutional prestige, and professional age (Petty, Fleming, & Fabrigar, 1999), manuscript corrections (Amat, 2008), and reviewer experience (Blackburn & Hakel, 2006).

Given that megajournals position themselves separately from traditional journals, studies of megajournals are needed to understand how their existence influences scholarly behavior and scholarly information use and to understand how they might function as disruptive forces (Cope & Kalantzis, 2009; Ewing, 2004), if they do. The increasing number of megajournals marks the present as a good time to document the beginnings of these platforms. In this spirit, the purpose of this study is (1) to examine the first articles published on *PeerJ*, which was launched in February 2013 as an open access and peer reviewed publishing platform in the medical and biological sciences, in order to document their characteristics (peer review, authorship, social metrics, and citations), and (2) to chart how the impact of the articles developed over the course of twelve months.

2. METHODS

2.1. Data Collection

On July 20, 2013, we took a small, random sample of *PeerJ's* published papers in order to focus on a breadth of factors that include review time, peer review history, peer referee anonymity, author internationality, author gender, citations, social metrics, and cited references. We collected a small sample in order to collect repeated measurements of many of these data points over the course of a year. The random sample included 49 of the 108 then published articles on *PeerJ*. Thirty-three (67%) of these articles had publicly available peer review histories.

The peer review histories included a document trail that contained the original manuscript submissions, the responses by the referees, the responses by the editors, the authors' rebuttals, and final, revised submissions. Only peer review histories of accepted articles are public. Referees here are labeled first, second, and third by order listed on the public peer review history pages. Additional data included the number / count of reviewers, the count of referees who remained anonymous, the number of revisions, the standing after the first round of reviews (i.e., minor or major revisions needed), the standing after the second round of reviews, and the dates for manuscript submission, acceptance, and publication.

Author data points included author affiliations and gender. *PeerJ* lists the affiliation of each author of an article; we only noted whether the authors' institutions were from a single nation or from multiple nations. Gender data was collected by looking up author profiles on *PeerJ* and by conducting web searches for each author in the absence of author profiles. We only noted binary gender composition for each article (male / female, all male, all female) and the gender of the first author.

We collected usage data and referral data for each article in the sample. These data are displayed on each article's web page and constitute the cumulative counts of downloads, unique visitors, page views, social referrals, and top referrals and were collected throughout a one year period with a baseline count on August 20, 2013 and follow up counts every three months after: November 20, 2013, February 20, 2014, May 20, 2014, and August 20, 2014.

Reference lists for each of the sampled articles were harvested in order to analyze citing behavior. Each article's web page was scraped using the scrapeR (Acton, 2010) library for the R programming language (R Core Team, 2014). The scraping process involved retrieving the articles in the sample and saving the results in an R list object. The list object was then parsed using XPath queries for journal and book titles in the reference lists. To parse these source types, we analyzed the source code of the sampled article web pages and noted that *PeerJ* specifies journal titles using the HTML node and attribute *//span[@class='source']* and book titles using the node and attribute *//a[@class='source']*. To retrieve all the references for a single author, the XPath query would search for the node and attribute *//li[@class='ref']*.

To acquire a conventional view of the impact of these articles, citation counts were retrieved for each of the articles in *Google Scholar* and in *Scopus* on February 20, 2014, May 20, 2014, and August 20, 2014.

2.2. Statistical analysis

As an exploratory profiling, statistical analysis mainly involved descriptive statistics, Chi-square cross tabulation, and rank sum comparisons. The significance level for all tests is *alpha* = 0.05. The usage and referral data from *PeerJ* and the citation counts from *Scopus* and *Google Scholar* were collected at

different time periods, and percentage changes were used to compare growth over the annual period of investigation. Since the citation window for the study was small and the accumulation of citations low, rank sum repeated block tests were used to measure differences in citation counts among the citation data collection points. Word counts were computed on the GNU/Linux command line using the GNU wc core utility to count words. The *pdftotext* and the *docx2txt* GNU/Linux command line utilities were used to convert author rebuttals that were submitted in PDF and DOCX formats into text files.

The majority of the analysis was conducted using the R programming language (R Core Team, 2014). The R programming language (R Core Team, 2014) libraries that were used can be categorized into three parts: data gathering and preparation, data analysis, and data visualization. To gather and prepare the data, the following libraries were used: the *scrapeR* library (Acton, 2010) for scraping web pages, the *lubridate* library (Grolemund & Wickham, 2011) for handling dates, and the *reshape2* library (Wickham, 2007) and *dplyr* library (Wickham & Francois, 2014) for manipulating and preparing data frames. For data analysis, we used the *pastecs* library (Grosjean & Ibanez, 2014) for detailed descriptive statistics, the *psych* library (Revelle, 2014) for providing descriptive statistics by grouping variables, the *Hmisc* library (Harrell, 2014) for computing correlation matrices, the *gmodels* library (Warnes, 2013) for cross tabulation analysis, and the *pgirmess* library (Giraudoux, 2014) for providing *Friedman* rank sum test post hoc comparisons. The *ggplot2* (Wickham, 2009) library was used for data visualization.

3. RESULTS

3.1. Characteristics of Peer Review

The sample contained 33 articles with open peer review histories and 16 articles with closed peer review histories, and we can reject the null hypothesis that authors have no preference for the public status of their peer review histories ($\chi2 = 5.898$, *df* = 1, *p* = 0.015). However, fourteen out of the 16 articles with closed peer review histories were submitted to *PeerJ* before it launched on February 12, 2013, which may indicate authors' trepidations about opening the review history before the journal went live.

Among the publicly available peer review histories, referees varied in allowing their identities to be made public. Out of the 33 articles with publicly available review histories, nineteen referees were entirely anonymous for nine of the articles and 14 referees were entirely attributed for seven of the articles. For seventeen of the articles, the disclosure of 36 referee identities was mixed (both anonymous and public). However, we fail to reject the null hypothesis that no significant differences exist among reviewers preferences for attribution ($\chi2 = 5.091$, *df* = 2, *p* = 0.078).

Each article with an open peer review history was attended to by at least one referee. Thirty-two of the articles were attended to by at least two referees. Five articles received responses by three referees. Most of the articles with open peer review histories underwent at least one round of revision, but one article was accepted *as-is*. Nineteen articles were accepted after one round of revision and 13 articles were accepted after two rounds of revision. After removing the article that was accepted *as-is* as an outlier, the category of articles that were accepted after one round of revision and the category of articles that were accepted after two rounds of revision were not significantly different from their expected values ($\chi2 = 1.125$, *df* = 1, *p* = 0.289), indicating no observable pattern in whether an article required one or two rounds of revisions.

Again excluding as an outlier the single article that was accepted *as-is*, out of the 19 articles that were accepted after one round of revision, thirteen were accepted after referees suggested minor revisions, and six were accepted after the referees accepted major revisions. Out of the 13 articles that were accepted after two rounds of revision, six were accepted after the referees suggested minor revisions, and seven were accepted after the referees suggested major revisions. There was no statistically significant difference between the number of revisions and the magnitude (minor / major) of revisions in the sample ($\chi2 = 1.5867$, *df* = 1, *p* = 0.208), indicating that manuscripts requiring major revisions were just as likely to be accepted after one round of revision than two rounds and that manuscripts that required minor revisions were just as likely to undergo two rounds of revision than one round.

Most articles with publicly available reviews had

19

received comments from two referees, but five articles had received comments from three referees, and one had not received any comments although its peer review history was accessible. For round one of the revisions, the mean word count among all reviewers was 488.3 words per review ($n = 70$, $md = 326.0$, $sd = 417.99$). This included 33 comments from first reviewers, 32 comments from second reviewers, and five comments from third reviewers for all 33 articles (reviewer order based on *PeerJ* review history). For round two of the revisions, the mean word count among all reviewers was 558.38 words per review ($n = 13$, $md = 186.00$, $sd = 738.65$). This comprised nine comments from first reviewers and four comments from second reviewers for nine of the articles.

3.2. Speed of Review and Publication

To determine the publication delay, or the time between submission, acceptance, and publication, we used the time stamps from each article's web page to set the interval between the date of submission and the date of publication. We used the date of acceptance as a subinterval point. Furthermore, *PeerJ* published its first article on February 12, 2013, and the sampled articles included manuscripts that were submitted to *PeerJ* before and after the launch date. To control for the launch of the journal, we partitioned the data by this date. Thirty-six of the articles in the sample were submitted after *PeerJ* began accepting submissions, but before it launched, and 13 were submitted after the launch date.

Overall, the subinterval from date of submission to the date of acceptance comprised most of the post-submission time, indicating that the referee process took longer than the process to prepare the manuscript for publication. For all articles in the sample, the grand median time from submission to publication was 83 days ($n = 49$; $m = 88.27$; $sd = 33.22$). For those articles that were submitted to *PeerJ* before launch date, the median time from submission to publication was 89.5 days ($n = 36$; $m = 95.92$; $sd = 33.81$). For those articles that were submitted after launch date, the median time from submission to publication was 68 days ($n = 13$; $m = 67.08$; $sd = 20.49$).

There was no practical difference in the speed of review, or the time between submission to acceptance, between articles that were submitted before the launch date and articles that were submitted after launch date. For all articles in the sample, the grand median time from submission to acceptance was 47 days ($n = 49$; $m = 56.82$; $sd = 37.32$). For those articles that were submitted to *PeerJ* before launch date, the median time from submission to acceptance was also 47 days ($n = 36$; $m = 60.5$; $sd = 41.7$). For those articles that were submitted to *PeerJ* after launch date, the median time from submission to acceptance declined by one day to 46 days ($n = 13$; $m = 46.62$; $sd = 18.6$).

The buildup of submissions before the launch of *PeerJ* caused a delay in the publication of manuscripts after they were accepted. For all articles in the sample, the grand median time from acceptance to publication was 28 days ($n = 49$; $m = 31.45$; $sd = 17.16$). For those articles that were submitted to *PeerJ* before launch date, the median time from acceptance to publication was 32.5 days ($n = 36$; $m = 35.42$; $sd = 18.2$). For those articles that were submitted to *PeerJ* after launch date, the median time from acceptance to publication was shortened to 20 days ($n = 13$; $m = 20.46$; $sd = 5.8$).

3.3. Author Characteristics

One component of the international status of a journal is whether a journal appeals to international collaborators (Calver, Wardell-Johnson, Bradley, & Taplin, 2010). Here we look at coauthorship as a proxy for collaboration. There was a median of 4 authors per article ($n = 49$, $m = 4.9$, $min = 1$, $max = 12$, $sd = 2.8$). Based on each author's institutional affiliation, twenty-one (42.9%) of the articles had authors affiliated with at least two nations and 28 (57.1%) of the articles were written by authors affiliated with an institution or institutions in a single nation. There was no statistically significant difference between these two categories ($\chi2 = 1$, $df = 1$, $p = 0.3173$).

There was a statistically significant difference between articles that were authored only by men and articles that were authored by both men and women. Thirteen (26.5%) of the articles were authored only by men and 36 (73.5%) were coauthored by men and women ($\chi2 = 10.7959$, $df = 1$, $p = 0.001$). Women led authorship for thirteen of the articles, but no articles were written solely by women. Among the articles with mixed gender authorship ($n = 36$), there was no statistically significant difference between mixed authorship and the gender of first authorship ($\chi2 = 2.778$, $df = 1$, $p = 0.096$).

Overall, mixed gender authorship outnumbered male only authorship and mixed gender authorship was slightly more common when authorship was multinational. However, it was more common for females to be first author on articles affiliated to institutions in a single nation than on articles affiliated with institutions in multiple nations; but a chi-square test showed no statistically significant difference between single or multiple institutional affiliations and the gender of the first author ($\chi2 = 0.491$, $df = 1$, $p = 0.484$).

3.4. Usage and Social Referrals

PeerJ provides article level usage and referral data. These data were collected for each sampled article from *PeerJ*. Usage based statistics included the cumulative number of unique visitors and page views for each article, and as of the May 20, 2014 data collection date, *PeerJ* was providing the cumulative number of downloads for each article. Additionally, *PeerJ* provides the cumulative number of social and top referrals to each article. Social referrals were specifically referrals from *Twitter, Facebook, Google Plus, LinkedIn, Reddit*, and *Slashdot*. Top referrals were referrals from all URLs, including URLs from the above social referral sites and from emails. Only URLs that appear at least two times are listed in the top referral list.

The data for these metrics were collected five times. A baseline count was collected on August 20, 2013. To measure growth throughout the year, quarterly counts were collected on November 20, 2013, February 20, 2014, May 20, 2014, and August 20, 2014. The exception is the download statistics, which did not appear on the site for the first three data collection events.

3.5. Downloads, Unique Visitors, and Page Views

As of May 2014, the grand median number of cumulative downloads was 342 ($n = 49$, $m = 532.5$, $min = 135$, $max = 2,720$). By August 2014, the grand median number of cumulative downloads increased by 13.45% to 388 ($n = 49$, $m = 597.4$, $min = 170$, $max = 2,854$). Not surprisingly, articles ($n = 23$) that were published in the first publication month had accumulated more downloads than articles ($n = 26$) that were published more recently. However, as Table 1 shows, download rates decreased as articles aged. For those articles that were published in February 2013, the median percentage change in the download rate was 13.45% between data collected in May 2014 and in August 2014. For those articles published between March and July 2013, the median percentage change in download rate was higher at 16.30%.

Table 1. Download statistics for articles published *during* launch month (February 2013) and articles published *after* (March 2013 - July 2013). Levels indicate date of data collection. Percentage changes show the change between medians

Launch Month (February 2013)	*n*	*min*	*max*	*m*	*mdn*	Percentage change of *mdn*
During						
May-14	23	156	2720	615.4	342.0	
Aug-14	23	198	2854	680.9	388.0	13.45%
After						
May-14	26	135	1460	459.1	340.5	
Aug-14	26	170	1572	523.5	396.0	16.30%
All						
May-14	49	135	2720	532.5	342.0	
Aug-14	49	170	2854	597.4	388.0	13.45%

By August 20, 2014, the grand median number of unique visitors was 703 (n = 49, m = 1,294, min = 204, max = 14,272) per article, and the grand median number of page views was 1,042 (n = 49, m = 1,933, min = 295, max = 22,571) per article. This represents 1.48 page views per unique visitor, based on the grand medians. Also, the overall percentage increases for unique visitors and page views were comparable. The percentage increase between the grand median unique visitors in August 2013 and the grand median unique visitors in August 2014 was 167.3% (n = 49) and, for page views, the percentage increase was 161.81% for the same time period.

As suggested by the download statistics, interest appears to be strongest after articles are published and then declines sharply throughout the year. For the cumulative number of unique visitors, the percentage increase from median number of visitors in August 2013 to median number of visitors in November 2013 was 71.86%. This dropped to a percentage increase of 22.12% for the period spanning November 2013 to February 2014, to a median percentage increase of 18.12% for the period spanning February 2014 to May 2014, and to a median percentage increase of 7.82% for the period spanning May 2014 to August 2014.

Page view results were similar. For the cumulative number of page views, the percentage increase in median number of page views in August 2013 to the median number of page views in November 2013 was 68.84%. This fell to a percentage increase of 22.77% for the period spanning November 2013 to February 2014, to a median percentage increase of 15.88% for the period spanning February 2014 to May 2014, and

to a median percentage increase of 9.00% for the period spanning May 2014 to August 2014.

The Spearman rho correlations between downloads, unique visitors, and page views were highly correlated. The Spearman rho correlation between downloads and unique visitors as of August 2013 was 0.65 (p = 0). This grew to 0.83 (p = 0) by August 2014. The Spearman rho correlation between downloads and page views as of August 2013 was 0.67 (p = 0) and this grew to 0.81 (p = 0) by August 2014. For the relationship between unique visitors and page views, the Spearman rho correlation was 0.99 (p = 0) as of August 2013. This saw a very minor drop to 0.98 (p = 0) by August 2014.

3.6. Social and Top Referrals

Social referrals on *PeerJ* are referrals that come from *Twitter, Facebook, Google Plus, LinkedIn, Reddit*, and *Slashdot.* The articles consistently received a median of two site referrals from the set of six sites for all data collection dates, and these sites provided a median of 12.0 to 14.0 referrals for all data collection dates (e.g., an article that receives eight referrals from Facebook and four referrals from Twitter would result in two unique social referrals and 12 total social referrals). The sum cumulative number of unique social referrals (site referrals) increased from August 2013 (sum = 87) to May 2014 (sum = 103) but dropped by six in August 2014 (sum = 97).

Top referrals comprise article visits from any site, including social sites, blogs, web pages, and emails. In August 2013, a median of six sites contributed a median of 116.0 referrals. By May 2014, the median

Table 2. Median unique visitors and page views per day, controlling for days since publication

Data Collection	n	Median Unique Visitors / Day	Median Page Views / Day	Median Days since Publication
August 20, 2013	49	2.28	3.41	168
November 20, 2013	49	1.98	2.86	260
February 20, 2014	49	1.78	2.52	352
May 20, 2014	49	1.63	2.30	441
August 20, 2014	49	1.46	2.12	533

number of sites doubled to 12 and the median number of referrals tripled to 374. Three months later, as of August 2014, the median number of referral sites nearly tripled to 34 but the median number of referrals from these 34 sites only increased by a factor of 1.29, or from a median of 374 to a median of 483. This does not reveal a surge of new site referrals but only that top link referrals are not displayed on *PeerJ* article pages until a referral to an article appears at least two times. The Spearman *rho* correlation between total social referrals and total top referrals was moderate as of August 2013 (*rho* = 0.44, *p* = 0.002) but much higher by August 2014 (*rho* = 0.63, *p* = 0).

3.7. Bibliometric and Citation Analysis

3.7.1. Citing Sources

For all the sampled articles, published between February 2013 and July 2013, citation counts were collected from *Google Scholar* and *Scopus* on three collection dates: February 20, 2014, May 20, 2014, and August 20, 2014. Although the citation window was short and the median number of citations was low from both sources, differences were observable. Specifically, when comparing median citation count differences, data from *Google Scholar* show very slow growth relative to data collected from *Scopus*. However, when comparing the sum of the ranked citations, *Google Scholar* data show statistically significant differences between the three data collection dates whereas *Scopus* data only showed statistically significant differences between two of the data collection dates.

In February 2014, *Google Scholar* showed a median citation count of 1.00 (*n* = 49) and in May 2014, this doubled to a median citation count of 2.00 (*n* = 49). This remained stable for August 2014, when *Google Scholar* again showed a median citation count of 2.00 (*n* = 49). The citation count differences among the three data collection dates are more revealing when the sum of their ranks are considered. A nonparametric Friedman rank sum test (Field, Miles, & Field, 2012; Neuhäuser, 2012; Townend, 2002) for repeated measures showed that *Google Scholar* citation counts changed significantly over the six month time period ($\chi2$ = 55.12, *df* = 2, *p* < 0.000). Specifically, a Friedman *post hoc* test (Giraudoux, 2014) showed that citation counts changed significantly among all comparisons when considering the differences in the sum ranks.

The statistically significant differences are seen in the baseline count to the second collection of citation counts (*observed difference* = 28.0, *critical difference* = 23.70, *p* < 0.05), from the baseline count to the third collection of citation counts (observed difference = 57.5, *critical difference* = 23.70, *p* < 0.05), and from the second collection of data to the third collection of citation counts (*observed difference* = 29.5, *critical difference* = 23.70, *p* < 0.05). Only ten articles had not received a citation by February 2014. This remained the same by May 2014. By August 2014, only six articles had not received a citation.

Unlike *Google Scholar, Scopus* showed median citation count differences for each data collection date but no statistically significant difference in all the *post hoc* tests. By February 2014, the median citation count was 0.00 (*n* = 49). This increased to a median of 1.00 (*n* = 49) by May 2014 and then doubled to a median of 2.00 (*n* = 49) by August 2014. A Friedman rank sum test for repeated measures showed that the rank accumulation of *Scopus* citation counts was statistically significant over the six month time period ($\chi2$ = 54.64, *df* = 2, *p* < 0.000). However, despite the median citation count increasing by one unit for each three data collection dates, the Friedman *post hoc* test only showed significant changes in citation counts between the baseline count and the third collection of citation counts (*observed difference* = 53.5, *critical difference* = 23.70, *p* < 0.05) and between the second collection of citation counts to the third collection of citation counts (*observed difference* = 33.5, *critical difference* = 23.70, *p* < 0.05). There was no statistically significant difference between the baseline count and the second collection of citation counts (*observed difference* = 20.0, *critical difference* = 23.70). This indicates the initial slower rate of citation accumulations in *Scopus* compared to *Google Scholar*, more detectable by the rank sum tests than by comparisons based on medians. As of February 2014, over half of the articles (*n* = 25) had not received a citation in *Scopus*. By May 2014, seventeen articles had not received a citation and this dropped to eleven by August 2014.

Despite *Scopus'* more conservative citation counting relative to *Google Scholar*, the relationship between *Google Scholar* citation counts and *Scopus* citation counts grew increasingly and positively linear as additional data was collected. As of February 2014, the

Spearman *rho* correlation coefficient was 0.55 ($p < 0.000$). The correlation increased to 0.68 ($p < 0.000$) by May 2014 and to 0.86 ($p < 0.000$) by August 2014.

3.7.2. Cited Sources

The *cited reference identity* or the *citation density* is based on "the mean number of references cited per article" (Garfield, 1999, p. 979; McVeigh & Mann, 2009). There were 2,253 total journal titles ($m = 45.98$) listed in the 49 reference lists. This included 973 unique journal titles for the entire sample ($m = 19.86$ / article) when calculating uniqueness relative to the whole sample. When allowing journal titles to repeat across *PeerJ* articles, the sample contained 1,395 unique journal titles ($m = 28.47$ / article).

We applied the *citation density* to examine the distribution of unique journal titles in the sample. This *cited title density* is the ratio of the number of times a journal title appears in the sample to the number of times the journal title appears in separate reference lists. The goal was to measure the scattering or concentration of cited journal titles among these lists in order to understand shared journal title commonality among *PeerJ* authors. Consequently and on a per title basis, the higher the cited title density the more the titles are distributed among a greater number of *PeerJ* articles; the lower the value the more the titles are concentrated among fewer *PeerJ* articles. For example, 24 *PeerJ* articles referenced *PLOS ONE* 54 times for a cited title density of 0.444 (24 / 54). Therefore, *PLOS ONE* is a highly distributed title. Two *PeerJ* articles referenced *Molecular and Cellular Biology* 18 times for a cited title density of 0.111 (2/18). Thus, *Molecular and Cellular Biology* is not a highly distributed title. Table 3 outlines the most cited journal titles for titles that appear ten or more times, with no journal title appearing exactly ten times.

Figure 1 illustrates the distribution of unique journal titles among the references. The LOESS regression line shows that as the rank of a journal title decreases (that is, the overall number of times the title appearing in the reference lists decreases), the more likely the unique title will appear in a greater number of different article reference lists. However, articles that appear only once (singletons) in the entire sample will have a cited title density of 1, which is also equal to articles that, for example, appear 25 times in 25 separate arti-

cles. Given the number of singletons ($n = 775$), Figure 2 only includes titles that appear more than once ($n = 198$). In order to see the distribution without the pull of journal titles with a cited title density of 1, Figure 3 plots all unique journal titles that appear more than once and that have a cited title density less than 1.

4. DISCUSSION

We examined a sample of articles published in the first months of *PeerJ*, a publishing platform that, like *PLOS ONE, Sage Open*, and others, is being called a megajournal (Björk & Solomon, 2013; MacCallum, 2011; Solomon, 2014). The purpose was to study the characteristics of these articles and to monitor their usage over the course of a year. The motivation comes from *prima facie* and stated differences between megajournals and traditional journals (e.g., see MacCallum, 2011).

In this study, we examined the characteristics of peer review, publication delay, and author characteristics. We found evidence that there is some preference for an open peer review history among authors publishing at *PeerJ*, and that this was especially true for those authors who submitted to *PeerJ* after its launch (and thus were able to see *PeerJ* in operation). Referee word counts were of moderate length. Bornmann, Wolf, and Daniel (2012) report comparable mean word counts from community comments on manuscripts submitted to the journal *Atmospheric Chemistry and Physics*. However, the word counts from designated reviewers in their study averaged several hundred words more than the referee word counts in the *PeerJ* sample. Whether this functions as an indicator of the quality of reviews requires further study, and one limitation here is that only word counts of accepted manuscripts are examined.

Although there was a greater tendency for reviews to be public, we did not find that referees possessed a strong preference for attribution, and they often stayed anonymous. There was no statistical association between the number and the magnitude (minor/major) of revisions articles required. Publication delay was affected by the launch of the journal, and despite its overall speed, total delay remained a function of the peer review process (Amat, 2008). The reduced pub-

Table 3. Cited title density for the most cited journal titles in *PeerJ* articles

Journal Title	Count of Instances	No. of *PeerJ* Articles Citing	Cited Title Density
PLOS ONE	54	24	0.444
Proceedings of the National Academy of Sciences	48	24	0.500
Nature	47	21	0.447
Science	36	20	0.556
Journal of Biological Chemistry	33	12	0.364
Nucleic Acids Research	32	10	0.313
Bioinformatics	25	10	0.400
Ecology Letters	21	6	0.286
Molecular and Cellular Biology	18	2	0.111
Journal of Experimental Biology	18	7	0.389
Vision Research	17	3	0.176
RNA	17	2	0.118
Journal of Molecular Biology	17	5	0.294
Oecologia	16	3	0.188
Ecology	16	4	0.250
Journal of Neuroscience	15	7	0.467
Biochemical and Biophysical Research Communications	14	7	0.500
Embo Journal	13	4	0.308
Proteins	13	4	0.308
Proceedings of the Royal Society B: Biological Sciences	13	4	0.308
Cell	13	6	0.462
Neuron	12	7	0.583
Molecular Cell	12	6	0.500
Molecular Biology and Evolution	12	2	0.167
Journal of Vertebrate Paleontology	12	3	0.250
Journal of the Acoustical Society of America	11	1	0.091

lication delay for the articles that were submitted post launch date highlighted a buildup of submissions before the launch date and the shortening of the overall post-submission process once *PeerJ* launched. However, whether considering the overall interval or the subintervals, the time from submission to acceptance ranged from slightly lower in some studies (Bornmann & Daniel, 2010) to substantially lower in others (Björk
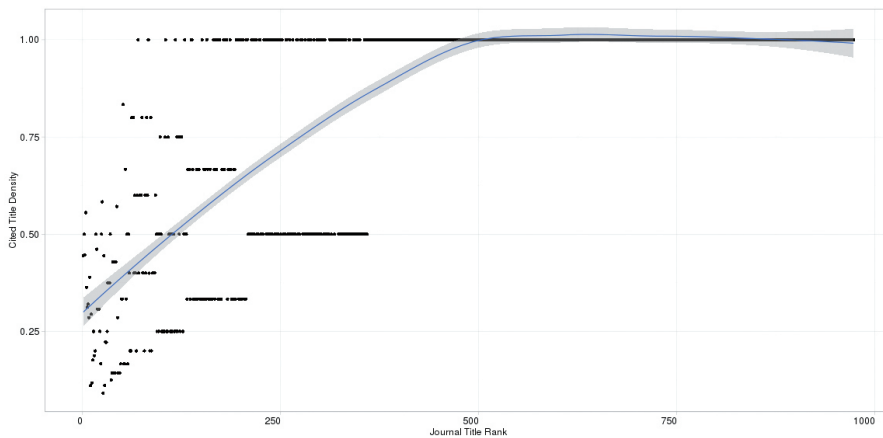
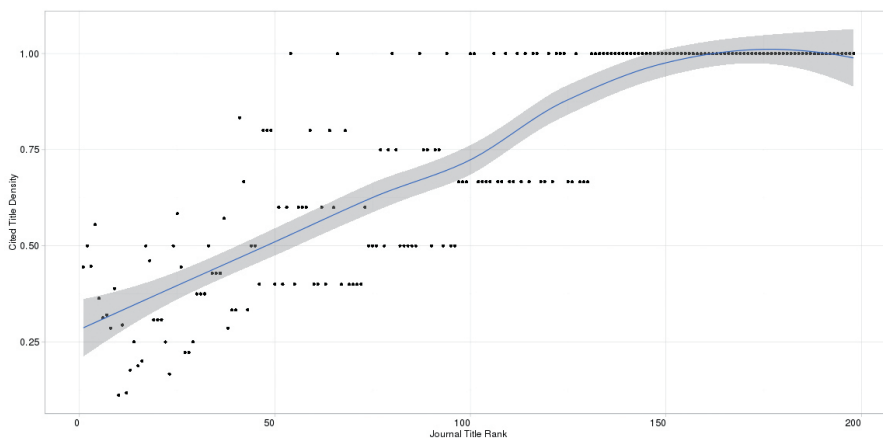**Fig. 1** Distribution of cited title density by rank of journals cited. Includes entire sample

Fig. 2 Distribution of cited title density by rank of journals cited. Only includes journal titles that appear more than once
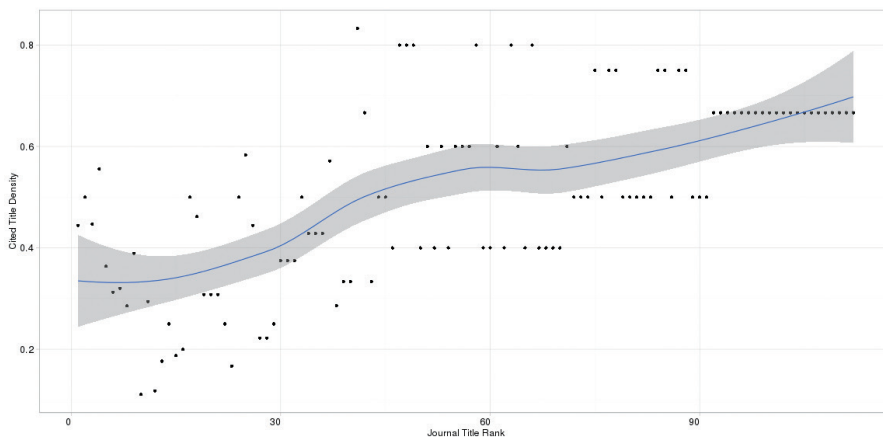

Fig. 3 Distribution of cited title density by rank of journals cited. Only includes journal titles that appear more than once and that have a cited title density less than 1

& Solomon, 2013; Pautasso & Schäfer, 2010).

Most of the articles were coauthored by both men and women, but only a fraction were first authored by women, supporting the fractionalized authorship findings in the much larger study conducted by Larivière, Ni, Gingras, Cronin, and Sugimoto (2013). *PeerJ* showed indication of being an international publication, in terms of co-authorship, with about 43% of the articles authored by collaborators from at least two nations. However, this result only considers whether articles are jointly written by authors from different nations and does not consider the count of different national collaborations. Therefore, the international status of *PeerJ* could be much higher when taking this into account.

As usage metrics declined, more conventional metrics increased. Downloads, unique visitors, and page views tended to be very high after an article's publication but rates of increase dropped sharply over the course of the year. However, as the rate of usage declined, the articles began receiving citations in both *Google Scholar* and *Scopus*, and both citation counts were highly correlated after one year. *Post hoc* nonparametric rank sum tests were better able to discern measurement differences between *Google Scholar* and *Scopus* than were comparisons based on medians. Measurements based on medians indicated faster citation accumulations in *Scopus*, but rank sum *post hoc* tests were able to detect more subtle differences, and showed that citations in *Google Scholar* accumulated faster. In either database, very few articles were not cited by the end of the study.

Out of the 973 unique journal titles, we found that as unique journal titles were cited less, their cited title density increased. Inversely, unique journal titles that were cited multiple times were often cited multiple times by fewer *PeerJ* articles. Thus, while authors frequently draw from common titles, such as *PLOS ONE, Nature, Proceedings of the National Academy of Sciences*, and similar high impact journals, each *PeerJ* article is likely to cite a number of unique journal titles. One interpretation is that *PeerJ* articles are highly varied in topics covered. If so, this would mean that *PeerJ* might appeal to a broad readership within the medical and biological sciences. However, this would also mean that discovering relevant *PeerJ* articles will be more dependent on search engines and bibliographical da-

tabases if users do not associate *PeerJ* with any specific sub-disciplines.

5. CONCLUSION

This purpose of this study was to conduct an exploratory case study of one megajournal, and therefore it does not lend itself to a generalization of all journals or of journals limited to specific fields or topics. Future research should conduct cross comparisons among megajournals as well as traditional journals in order to tease out differences in article characteristics, and such studies should entail increased sample sizes and a limited number of variables. However, this study should be useful to those projects since it highlights some of the unique parameters of the megajournal.

Megajournals as born digital scholarly publication platforms provide an opportunity to understand more about scientific norms and values. That is, scientometricians can use megajournals as devices that can help reveal whether issues in scholarly communication are a function of, for example, web-based publication technologies or of scientific norms or constructs. In particular, as a device, megajournals may help information scientists discover whether the technological advantages that megajournals have, by re-imagining scholarly publishing given present day technological affordances, over traditional journals, which are often based on a print paradigm, result in different patterns of authorship, peer review, and other characteristics.

ACKNOWLEDGEMENTS

REFERENCES

Acton, R. M. (2010). scrapeR: Tools for scraping data from HTML and XML documents. R package version 0.1.6. Retrieved from http://CRAN.R-project.org/package=scrapeR

Amat, C. B. (2008). Editorial and publication delay of

papers submitted to 14 selected food research journals. Influence of online posting. *Scientometrics, 74*(3), 379-389. doi:10.1007/s11192-007-1823-8

Binmore, P. (2013). Open access megajournals: Have they changed everything? Keynote presentation presented at University of British Columbia UBC Open. Retrieved from http://creativecommons.org.nz/2013/10/open-access-megajournals-have-they-changed-everything

Birukou, A., Wakeling, J. R., Bartolini, C., Casati, F., Marchese, M., Mirylenka, K., Osman, N., Ragone, A., Sierra, C., & Wasser, A. (2011, December). Alternatives to peer review: Novel approaches for research evaluation. *Frontiers in Computational Neuroscience.* doi:10.3389/fncom.2011.00056

Björk, B., & Solomon, D. (2012). Open access versus subscription journals: A comparison of scientific impact. *BMC Medicine, 10*(73). doi:10.1186/1741-7015-10-73

Björk, B., & Solomon, D. (2013). The publishing delay in scholarly peer-reviewed journals. *Journal of Informetrics, 7*(4), 914-923. doi:10.1016/j.joi.2013.09.001

Blackburn, J. L., & Hakel, M. D. (2006). An examination of sources of peer-review bias. *Psychological Science, 17*(5), 378-382. doi:10.1111/j.1467-9280.2006.01715.x

Bornmann, L. (2008). Scientific peer review: An analysis of the peer review process from the perspective of sociology of science theories. *Human Architecture: Journal of the Sociology of Self-Knowledge, 6*(2), 23-38.

Bornmann, L., & Daniel, H.-D. (2010). How long is the peer review process for journal manuscripts? A case study on *Angewandte Chemie International Edition. Chimia, 64*, 72-77. doi:10.2533/chimia.2010.72

Bornmann, L., & Marx, W. (2012). The Anna Karenina principle: A way of thinking about success in science. Journal of the *American Society for Information Science and Technology, 63*(10), 2037-2051. doi:10.1002/asi.22661

Bornmann, L., Wolf, M., & Daniel, H.-D. (2012). Closed versus open reviewing of journal manuscripts: How far do comments differ in language use? *Scientometrics, 91*(3), 843-856. 10.1007/s11192-011-0569-5

Calver, M., Wardell-Johnson, G., Bradley, S., & Taplin, R. (2010). What makes a journal international? A case study using conservation biology journals. *Scientometrics, 85*, 387-400. doi:10.1007/s11192-010-0273-x

Ceci, S. J., & Williams, W. M. (2011). Understanding current causes of women's underrepresentation in science. *Proceedings of the National Academy of Sciences, 108*(8), 3157-3162. doi:10.1073/pnas.1014871108

Cope, W. W., & Kalantzis, M. (2009). Signs of epistemic disruption: Transformations in the knowledge system of the academic journal. *First Monday, 14*(4). doi:10.5210/fm.v14i4.2309

Cronin, B., & Franks, S. (2006). Trading cultures: Resource mobilization and service rendering in the life sciences as revealed in the journal article's paratext. *Journal of the American Society for Information Science and Technology, 57*(14), 1909-1918. doi:10.1002/asi.20407

Ewing, J. (2004). The orthodoxy of open access. *Nature Focus.* Retrieved from http://www.nature.com/nature/focus/accessdebate/32.html

Eysenbach, G. (2006). Citation advantage of open access articles. *PLOS Biology, 4*(5), e157. doi:10.1371/journal.pbio.0040157

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R.* Los Angeles: Sage.

Garfield, E. (1977). What is a journal? *Essays of an Information Scientist: Vol. 1. 1962-1973* (pp. 6-7). Philadelphia: ISI Press. Original work published 1964.

Garfield, E. (1999). Journal impact factor: A brief review. *CMAJ, 61*(8), 979-980. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1230709.

Giraudoux, P. (2014). pgirmess: Data analysis in ecology. R package version 1.5.9. Retrieved from http://CRAN.R-project.org/packages=pgirmess

Grolemund, G., & Wickham, H. (2011). Dates and times made easy with lubridate. *Journal of Statistical Software, 40*(3), 1-25. Retrieved from http://www.jstatsoft.org/v40/i03

Grosjean, P., & Ibanez, F. (2014). pastecs: Package for analysis of space-time ecological series. R package version 1.3-18. Retrieved from http://CRAN.R-project.org/package=pastecs

Harrell, Jr., F. E. (2014). Hmisc: Harrell miscellaneous. R package version 3.14-5. Retrieved from http://CRAN.R-project.org/package=Hmisc

Larivière, V., Ni, C., Gingras, Y, Cronin, B., & Sugimoto, C. (2013). Bibliometrics: Global gender disparities in science. *Nature, 504*(7479), 211-213. doi:10.1038/504211a

Lee, C. J., Sugimoto, C. R., Zhang, G., & Cronin, B. (2013). Bias in peer review. *Journal of the American Society for Information Science and Technology, 64*(1), 2-17. doi:10.1002/asi.22784

Lipworth, W., & Kerridge, I. (2011). Shifting power relations and the ethics of journal peer Review. *Social Epistemology, 25*(1), 97-121. doi:10.1080/02691728.2010.534567

Lloyd, M. E. (1990). Gender factors in reviewer recommendations for manuscript publication. *Journal of Applied Behavior Analysis, 23*(4), 539-543. doi:10.1901/jaba.1990.23-539

Luwel, M., & Moed, H. F. (1998). Publication delays in the science field and their relationship to the ageing of scientific literature. *Scientometrics, 41*(1-2), 29-40. doi:10.1007/bf02457964

McVeigh, M. E., & Mann, S. J. (2009). The Journal Impact Factor denominator: Defining citable (counted) items. *Journal of the American Medical Association, 302*(10), 1107-1109. doi:10.1001/jama.2009.1301

MacCallum, C. J. (2011). Why ONE is more than 5. *PLoS Biol, 9*(12):e1001235+. doi:10.1371/journal.pbio.1001235

Merton, R. K. (1973a). The normative structure of science. In N. W. Storer (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 267-278). Chicago: University of Chicago Press.

Merton, R. K. (1973b). The Matthew Effect in science. In N. W. Storer (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 439-459). Chicago: University of Chicago Press.

Merton, R. K. (1973c). Institutionalized patterns of evaluation in science. In N. W. Storer (Ed.), *The Sociology of Science: Theoretical and Empirical Investigations* (pp. 460-496). Chicago: University of Chicago Press.

Neuhäuser, M. (2012). Nonparametric statistical tests: A computational approach. Boca Raton, Florida: CRC Press.

Opthof, T., Coronel, R., & Janse, M. J. (2000). Submissions, editorial process and impact factor 1992-2000. *Cardiovascular Research, 47*(2), 203-206. doi:10.1016/s0008-6363(00)00132-2

Paludi, M. A., & Bauer, W. D. (1983). Goldberg revisited: What's in an author's name. *Sex Roles, 9*(3), 387-390. doi:10.1007/bf00289673

Pautasso, M., & Schafer, H. (2010). Peer review delay and selectivity in ecology journals. *Scientometrics, 84*(2), 307-315. doi:10.1007/s11192-009-0105-z

PeerJ. (2015). Editorial criteria. Retrieved from https://peerj.com/about/editorial-criteria

PeerJ. (2014). FAQ. Retrieved from https://peerj.com/about/FAQ

Petty, R. E., Fleming, M. A., & Fabrigar, L. R. (1999). The review process at PSPB: Correlates of interviewer agreement and manuscript acceptance. *Personality and Social Psychology Bulletin, 25*(2), 188-203. doi:10.1177/0146167299025002005

PLOS ONE (2014). PLOS ONE Journal Information. Retrieved from http://www.plosone.org/static/information.action

R Core Team. (2014). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*. Vienna, Austria. Retrieved from http://www.R-project.org

Revelle, W. (2014). psych: Procedures for personality and psychological research. R package version 1.4.8. Retrieved from http://CRAN.R-project.org/package=psych

Solomon, D. J. (2014). A survey of authors publishing in four megajournals. *PeerJ, 2*, e365. doi:10.7717/peerj.365

Solomon, D. J., & Björk, B. (2012). Publication fees in open access publishing: Sources of funding and factors influencing choice of journal. *Journal of the American Society of Information Science & Technology, 63*(1), 97-107. doi:10.1002/asi.21660

Townend, J. (2002). Practical statistics for environmental and biological scientists. West Sussex, England: John Wiley & Sons, LTD.

Warnes, G. R. (2013). gmodels: Various R programming tools for model fitting. R package version 2.15.4.1. Retrieved from http://CRAN.R-project.org/package=gmodels

West, J. D., Jacquet, J., King, M. M., Correll, S. J., & Bergstrom, C. T. (2013). The role of gender in

scholarly authorship. *PLOS ONE, 8*(7): e66212. doi:10.1371/journal.pone.0066212

Wickham, H. (2007). Reshaping data with the {reshape} package. *Journal of Statistical Software, 21*(12), 1-20. Retrieved from http://www.jstatsoft.org/v21/i12

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. New York: Springer.

Wickham, H., & Francois, R. (2014). *dpylr: A grammar of data manipulation*. R package version 0.3.0.3. Retrieved from http://CRAN-R-project.org/package=dpylr