

Public accessibility of biomedical articles from PubMed Central reduces journal readership—retrospective cohort analysis

Philip M. Davis¹

Phil Davis Consulting, Ithaca, New York, USA

ABSTRACT Does PubMed Central—a government-run digital archive of biomedical articles—compete with scientific society journals? A longitudinal, retrospective cohort analysis of 13,223 articles (5999 treatment, 7224 control) published in 14 society-run biomedical research journals in nutrition, experimental biology, physiology, and radiology between February 2008 and January 2011 reveals a 21.4% reduction in full-text hypertext markup language (HTML) article downloads and a 13.8% reduction in portable document format (PDF) article downloads from the journals' websites when U.S. National Institutes of Health-sponsored articles (treatment) become freely available from the PubMed Central repository. In addition, the effect of PubMed Central on reducing PDF article downloads is increasing over time, growing at a rate of 1.6% per year. There was no longitudinal effect for full-text HTML downloads. While PubMed Central may be providing complementary access to readers traditionally underserved by scientific journals, the loss of article readership from the journal website may weaken the ability of the journal to build communities of interest around research papers, impede the communication of news and events to scientific society members and journal readers, and reduce the perceived value of the journal to institutional subscribers.—Davis, P. M. Public accessibility of biomedical articles from PubMed Central reduces journal readership—retrospective cohort analysis. *FASEB J.* 27, 2536–2541 (2013). www.fasebj.org

Key Words: digital repositories • downloads • open access • scientific publishing

RESEARCHERS RECEIVING MONIES from the U.S. National Institutes of Health (NIH) are required to deposit copies of peer-reviewed journal manuscripts into PubMed Central (PMC), a digital repository of biomedical literature operated by the National Library of Medicine. The NIH policy requires that these papers must be accessible to the public no later than 12 mo from final publication (1). Many journal publishers deposit these manuscripts (or copies of the final published articles) on behalf of their authors.

Abbreviations: HTML, hypertext markup language; NIH, U.S. National Institutes of Health; PDF, portable document format; PMC, PubMed Central

Free access to the research literature has been shown to increase readership but has no effect on article citations (2, 3). An experiment in depositing peer-reviewed manuscripts into institutional repositories reported that public accessibility increases downloads from the publishers' websites in the short term (4). In a preliminary study of physiology articles by a single publisher, we reported that free access to articles from PMC was responsible for a 14% reduction in full-text downloads from the journals' websites (5).

The purpose of this study is to expand on those preliminary results by enlarging our study to include a broader array of biomedical journals covering diverse fields such as nutrition, experimental biology, physiology, and radiology over an expanded period of time, and by measuring the effect of free access from PMC on portable document format (PDF) article downloads, as well as full-text article downloads.

MATERIALS AND METHODS

The data set consists of 13,223 original articles and reviews published in 14 society-run biomedical research journals in nutrition, experimental biology, physiology, and radiology between February 2008 and January 2011 (Table 1). Other document types (news, editorials, book reviews, errata, perspectives, clinical cases, among others) were excluded from the analysis. The data set also excludes 188 articles that were made freely available on publication, for whatever reason, as their publication trajectory was dissimilar from other articles in the study.

All articles included in the study were available to subscribers for the first 12 mo from final publication, after which they became freely available to all readers from the participating journals' websites. For the 14 participating journals, this is the normal path of publication.

Forty-five percent (5999 of 13,223) of the study articles declared some form of NIH funding, were deposited by the

¹ Correspondence: Phil Davis Consulting, 432 Mitchell St., Ithaca, NY 14850, USA. E-mail: pmd8@cornell.edu

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/3.0/us/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

doi: 10.1096/fj.13-229922

TABLE 1. Description of the data set

Journal	Articles (n)	Deposited in PMC		Date range (mo/yr)
		n	%	
<i>The American Journal of Clinical Nutrition</i>	1,210	387	32	04/08–01/11
<i>American Journal of Physiology—Cell Physiology</i>	773	427	55	07/08–01/11
<i>American Journal of Physiology—Endocrinology and Metabolism</i>	750	352	47	07/08–01/11
<i>American Journal of Physiology—Renal Physiology</i>	854	489	57	07/08–01/11
<i>American Journal of Physiology—Gastrointestinal and Liver Physiology</i>	691	399	58	07/08–01/11
<i>American Journal of Physiology—Heart and Circulatory Physiology</i>	1,251	762	61	07/08–01/11
<i>American Journal of Physiology—Lung Cellular and Molecular Physiology</i>	499	344	69	07/08–01/11
<i>American Journal of Physiology—Regulatory, Integrative and Comparative Physiology</i>	1,034	490	47	07/08–01/11
<i>The FASEB Journal</i>	1,128	431	38	06/08–01/11
<i>Journal of Applied Physiology</i>	1,135	449	40	07/08–01/11
<i>Journal of Neurophysiology</i>	1,521	857	56	07/08–01/11
<i>Journal of Nutrition</i>	959	278	29	04/08–01/11
<i>Physiological Genomics</i>	283	157	55	07/08–11/10
<i>Radiology</i>	1,135	177	16	02/08–01/11
Total	13,223	5999	45	02/08–01/11

publisher into PMC, and were made freely available from PMC 12 mo after final publication. These articles formed the treatment group. The remaining 7224 articles formed the control group.

Readership was measured for each article by the number of full-text hypertext markup language (HTML) and PDF downloads from the journals' websites, aggregated by month and extending 24 mo from final publication. Abstract views were omitted as a measure of readership since they are provided free to the reader on publication from the journal websites and from the PubMed index. Full-text HTML downloads from PMC were gathered by month for each journal from the PMC publisher reporting system.

Final publication is defined as the date when a journal issue is released to the public. Because most study articles were published online in PDF format several weeks before final issue publication, we excluded these initial readership counts from the data set since the 12-mo embargo in PMC is calculated from the date of final publication. All articles in the data set have aged ≥ 24 mo from the date of final publication. To ensure consistency in readership statistics across journals (6), all participating journals were hosted on the same online publishing platform (HighWire Press; <http://highwire.stanford.edu>).

STATISTICAL METHODS

We were chiefly interested in the relative performance of treatment articles to control articles in their second year of publication—the period during which treatment articles became freely available from the PMC archive.

A linear regression model was constructed on a pretest-posttest design (7), where pretest was defined as the total number of downloads for each article during its first year of publication, and posttest was defined as the number of downloads received for each article during its second year of publication. The posttest scores of each article (full-text HTML and PDF) served as our dependent variables, while pretest scores served as baseline performance covariates for each article.

Other covariates in the statistical model were PMC (an indicator variable identifying treatment articles; PMC=1, control=0) and article age (yr). Interactions of the PMC variable with article age and baseline performance variables were also included in the model to test whether the PMC effect was constant across the data set. Variation in article downloads across journals was controlled by including the journal as a random variable in the model. Similarly, variation among article types was controlled by including the journal section as a random variable, which was nested within each journal (8). In both cases, we were not interested in estimating the effect of each section within each journal on article downloads, specifically, but to control for their effects in the model, generally.

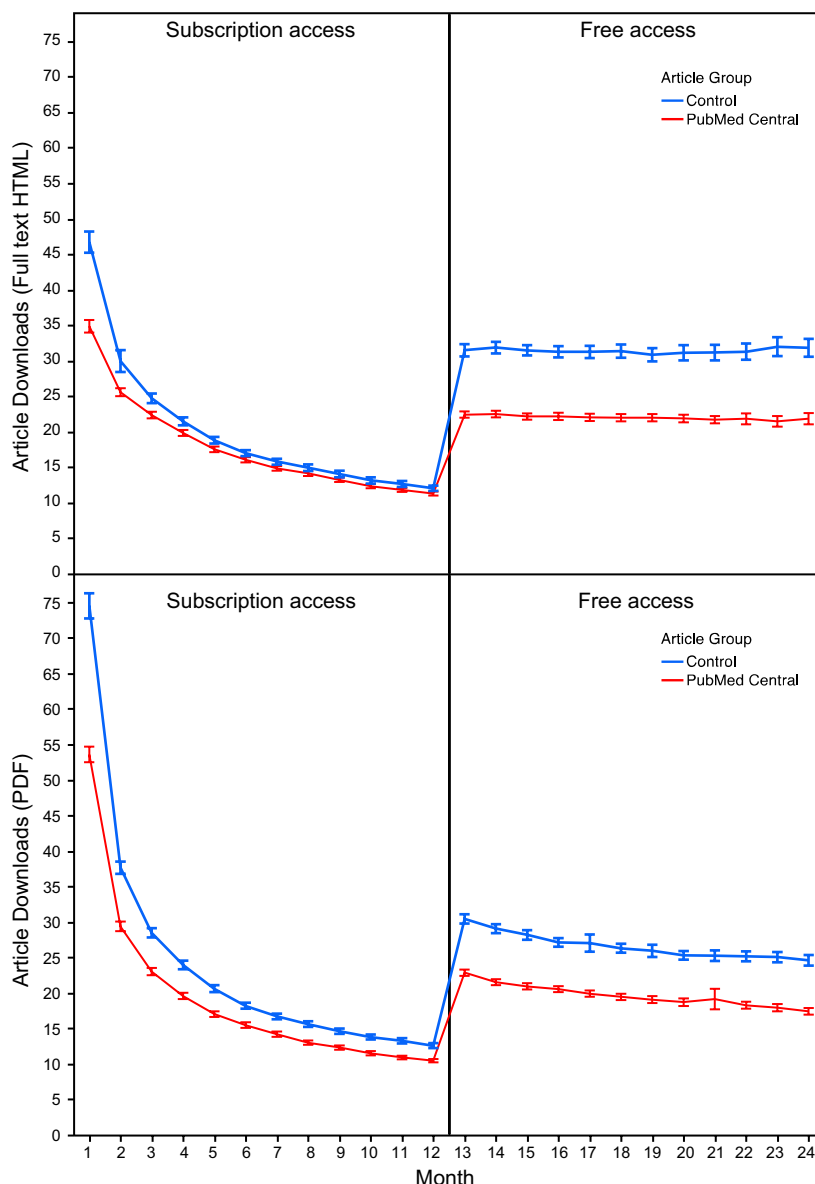
Finally, since the distribution of article downloads was highly skewed, it was necessary to normalize article download counts by taking the natural log of the pretest and posttest observations. To interpret the results of the regression equation, the estimates of independent variables were exponentiated in order to arrive at their multiplicative effect. For example, if the estimate for PMC were -0.25 , the multiplicative effect of this estimate would be $e^{-0.25}$ or 0.779, representing a 22.1% reduction in article downloads. For simplicity in interpretation, the percentage effect is reported in the results table. All analyses were performed using JMP 10.0 (SAS Institute, Cary, NC, USA).

RESULTS

Figure 1 plots the longitudinal performance of articles by cohort for the first 24 mo after final publication. While article download numbers vary greatly across journals and article types, control articles generally outperformed treatment articles for the first 12 mo of publication.

Beginning with month 13, as all articles became freely available from the journal websites, both treat-

Figure 1. Mean full-text and PDF article downloads ($\pm 95\%$ CI) from the journals' websites for the first 24 mo of publication. All articles were available to subscribers for the first 12 mo, after which they became freely available to all readers from the journals' websites. Articles declaring NIH funding were deposited into PMC and became freely available to readers 12 mo after final publication.



ment and control articles received a large boost in monthly article downloads. However, treatment articles—those articles made freely available from PMC—received a much smaller increase in downloads.

Controlling for their baseline performance in year 1, article age, journal, and journal section, treatment articles received 21.4% fewer full-text HTML downloads and 13.8% fewer PDF downloads in year 2 compared to control articles (Tables 2 and 3). Higher-performing articles, as measured by their baseline performance in year 1, were more strongly affected by being available from PMC. Lastly, there is evidence that PMC is exerting a greater effect in reducing PDF article downloads over time, growing at a rate of 1.6%/yr. There was no discernable longitudinal effect for full-text HTML downloads.

Point estimates for each journal varied somewhat across the data set (Fig. 2). PMC reduced full-text HTML article downloads by as much as 26.3% for the *American Journal of Physiology—Heart and Circulatory Physiology* (95% CI -28.8% to -23.7%) to as little as 11.2%

for the *American Journal of Physiology—Regulatory, Integrative and Comparative Physiology* (95% CI -15.0% to -7.2%). As reported, the reduction in PDF downloads was somewhat lower. PMC reduced PDF article downloads by as much as 18.4% for the *Journal of Nutrition* (95% CI -21.3% to -15.3%) to as little as 2.6% for the *American Journal of Physiology—Regulatory, Integrative and Comparative Physiology* (95% CI -5.7% to 0.5% , not significant).

DISCUSSION

There is strong evidence to suggest that mandated deposit of published articles in a public repository is responsible for drawing significant numbers of readers away from journal websites when those articles become freely available after a 12-mo embargo period. Furthermore, there is evidence to suggest that the effect of PMC is growing over time.

TABLE 2. *Linear mixed model estimating the effect of free access from PMC on full-text HTML article downloads [response log (HTML downloads)] from the journals' websites*

Fixed effect	Estimate	Lower 95%	Upper 95%	<i>t</i> ratio	<i>P</i> > <i>t</i>
PMC	−21.4%	−22.4%	−20.4%	−37.05	<.0001
Age (yr)	−8.6%	−9.6%	−7.6%	−16.43	<.0001
Baseline HTML	102.4%	99.7%	105.1%	103.94	<.0001
PMC × article age	−0.7%	−2.3%	0.9%	−0.91	0.3624
PMC × baseline HTML	−1.9%	−3.7%	−0.1%	−2.09	0.0364

Random effect	Variance ratio	Variance component	95% lower	95% upper	Percentage of total
Journal	0.217	0.027	0.013	0.086	15.3
Section [journal]	0.198	0.025	0.018	0.034	14.0
Residual		0.124	0.121	0.127	70.7
Total		0.176	0.153	0.204	100

$N = 13223$, $R^2 = 0.707$, mean response = 5.553. PMC is the difference in the number of downloads (expressed as a percentage) between the articles deposited in PMC and control articles. Article age, baseline HTML, and baseline PDF are centered in the interaction effects. Variance within section is nested within each journal.

While the participating journals only deposit full-text copies of their articles into PMC, PMC does provide readers with a printer-friendly PDF version of the full-text article, structured in the semblance of an article's final printed copy. Provision of the printer-friendly version to readers may explain the decline in publisher PDF views. Giving preferential visibility to the PMC copy over a link to the journal website from PubMed search results may also partially explain the decline.

The persistent reductions in article downloads from journal websites also suggests that PMC is, in part, competing directly with the journal for readers of the biomedical literature. As PMC-deposited articles become freely available simultaneously from the publisher's website, the reduction in reader traffic cannot be explained by differential access barriers. Secondly, it suggests that the printer-friendly PDF rendering of the full-text article may provide a viable substitute to the publisher's PDF for many readers of the scientific literature.

The relationship between free access and subscription cancellation behavior is not well understood. As

librarians base cancellation decisions, in large part, on publisher-provided article download data (9–13), a significant reduction in usage as a result of reader access from PMC is likely to have unintended consequences for those publishers who participate in the direct article deposit program. Usage traffic for articles on PMC has been increasing steadily since the participating journals began depositing article content (Fig. 3).

Whereas PMC may be providing access to readers traditionally underserved by scientific journals, the loss of readership from journal websites may have negative consequences for readers. As PMC draws readership away from the journal, the journal editor loses the opportunity to direct readers to related articles, editorials, letters and commentary surrounding the article of interest. The journal also loses the opportunity to deliver news, educational material, advertisements (job announcements, grant and travel opportunities, products and services) and society events (conferences and workshops) to the reader. In summary, the reader loses access to, and ability to participate in, the context,

TABLE 3. *Linear mixed model estimating the fixed effect of free access from PMC on PDF article downloads [response log (PDF downloads)] from the journals' websites*

Fixed effect	Estimate	Lower 95%	Upper 95%	<i>t</i> ratio	<i>P</i> > <i>t</i>
PMC	−13.8%	−14.6%	−12.9%	−29.54	<.0001
Age (yr)	−13.0%	−13.7%	−12.3%	−33.19	<.0001
Baseline PDF	124.7%	122.2%	127.2%	142.59	<.0001
PMC × article age	1.6%	0.3%	2.8%	2.5	0.0126
PMC × baseline PDF	−2.5%	−3.9%	−1.1%	−3.44	0.0006

Random effect	Variance ratio	Variance component	95% lower	95% upper	Percentage of total
Journal	0.340	0.025	0.013	0.070	23.8
Section [journal]	0.093	0.007	0.005	0.010	6.5
Residual		0.074	0.072	0.076	69.8
Total		0.106	0.088	0.130	100

$N = 13223$, $R^2 = 0.804$, mean response = 5.428. PMC is the difference in the number of downloads (expressed as a percentage) between the articles deposited in PMC and control articles. Article age, baseline HTML, and baseline PDF are centered in the interaction effects. Variance within section is nested within each journal.

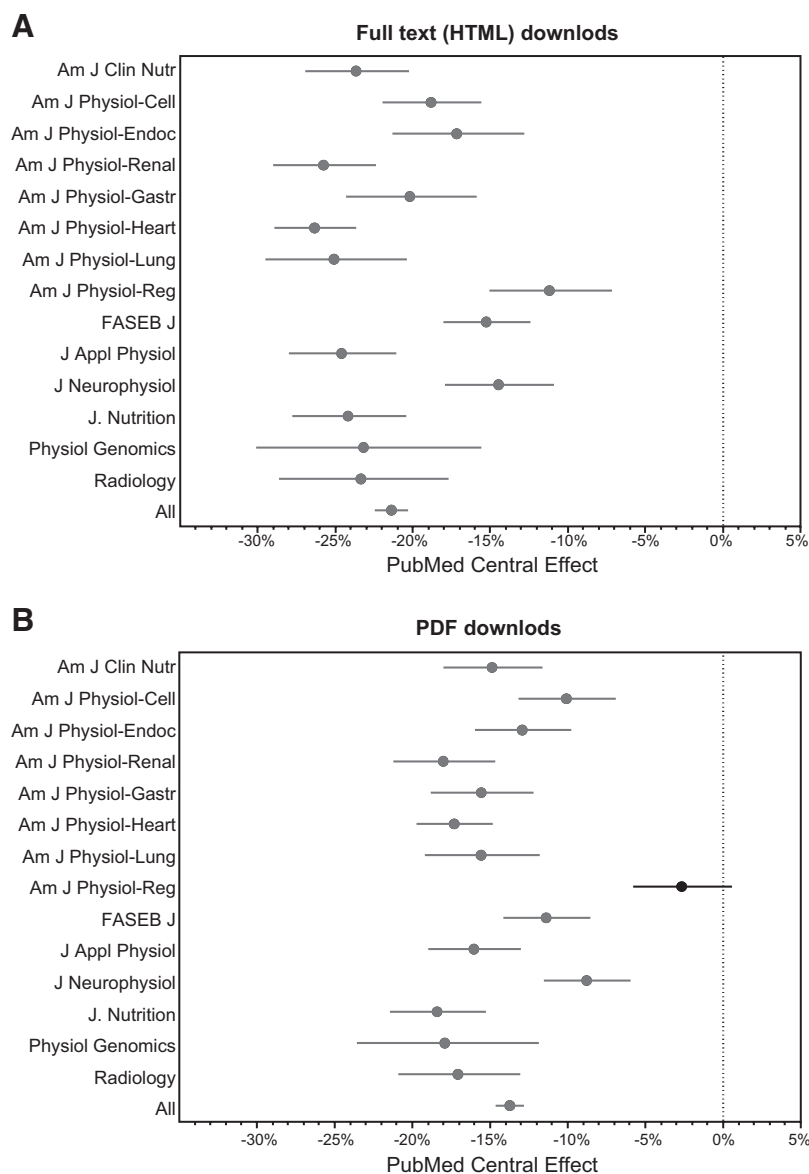


Figure 2. Journal estimates ($\pm 95\%$ CI) measuring the effect of article availability from PMC on full-text (HTML) downloads (A) and PDF downloads (B) from the journal website.

discourse, and community surrounding each article. This loss of context is not limited to subscription-access journals, but can be extended to all journals (including full open-access journals) publishing redundantly with PMC.

Limitations

The journals in this study employ a continuous publication model, whereby papers are published online, in PDF format, weeks in advance of final issue publication. As a result, the actual age of articles in each issue varies. The design of this study required us to ignore readership before final publication. However, since our study is comparative in nature, we were primarily interested in measuring the relative performance of articles in each cohort in our study rather than their absolute performance. As we assume that there is no temporal bias in publication, whereby treatment articles are published earlier (or later) than control articles, we believe that our methods form an unbiased estimate of treatment effect.

The usage profile of NIH-sponsored (treatment) articles is similar with, but not identical to, control articles (Fig. 1). These differences may be explained by the heterogeneity of biomedical journals included in the data set and is addressed by controlling for both article effects and journal effects in the statistical model. Homogeneity between treatment and control group characteristics may be achieved in a randomized controlled trial, although the NIH public mandate precludes such an experimental design from being conducted. In addition, the study does not include a comparison cohort of articles published before the study journals began depositing articles into PMC. It is possible that NIH-sponsored articles have a radically different usage profile, which manifests itself precisely and predictably after 12 mo, although this is highly unlikely.

Since we have no access to individual reader identifiers (such as IP addresses), we were unable to estimate the additional readership that is taking place on PMC. We are also unable to estimate the amount of reader traffic that is directed back to journal websites as a result of linking from the PMC record.

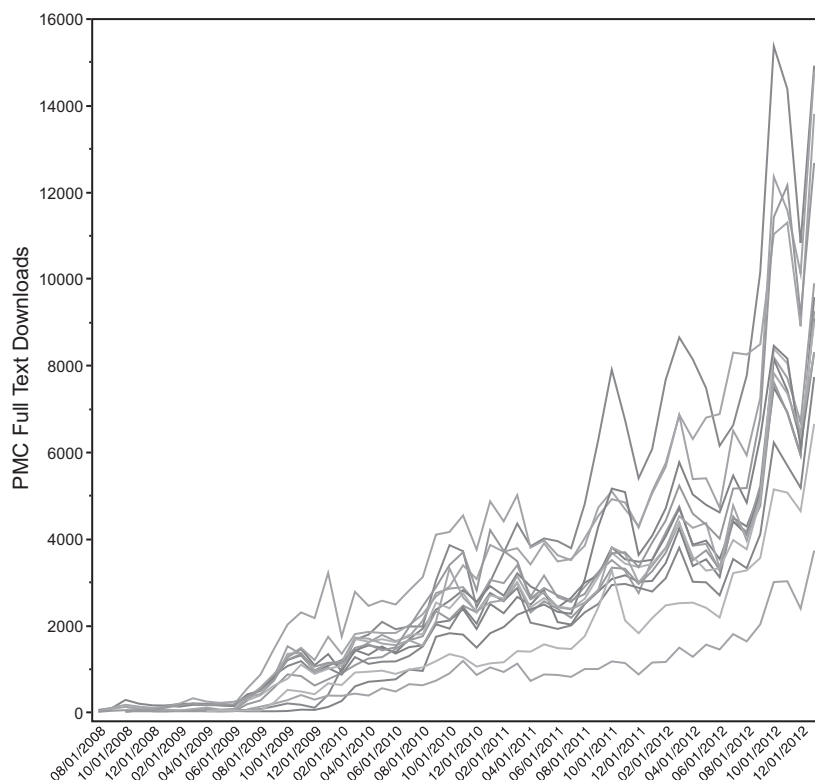


Figure 3. Total full-text downloads for NIH-sponsored articles deposited into PMC for 14 participating biomedical journals. Data from PMC.

Last, the data set is based on articles that have been published for ≥ 24 mo. As the effect of free article availability from PMC appears to be growing over time, it is likely that articles deposited today would show a greater reduction of article downloads than articles deposited in 2008 when the direct deposit program began.

Further research

This study focused on how public access to final published versions of articles from PMC competes with free access to articles from the journal website. It is not known whether public access to peer-reviewed manuscripts has similar effects on journal readership. It is also not understood whether different publishing models (such open access publishing) are affected differentially.

Fj

The author was solely responsible for the design and conduct of the study, collection and management of the data, data analysis and interpretation, preparation and submission of the manuscript, and can take responsibility for the integrity of the data and accuracy of the analysis. The author thanks the American Physiological Society (APS), the American Society for Nutrition (ASN), The Federation of American Societies for Experimental Biology (FASEB), and the Radiological Society of North America (RSNA) for their participation in the study. The study was supported by the APS.

REFERENCES

1. U.S. National Institutes of Health (2009) National Institutes of Health Public Access. Retrieved from <http://publicaccess.nih.gov/>
2. Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., and Connolly, M. J. L. (2008) Open access publishing, article downloads and citations: randomised trial. *BMJ* **337**, a568
3. Davis, P. M. (2011) Open access, readership, citations: a randomized controlled trial of scientific journal publishing. *FASEB J.* **25**, 2129–2134
4. CIBER Research Ltd. (2012) PEER Usage Study: randomised controlled trial results. CIBER Research Ltd., Newbury, UK. Retrieved from http://www.peerproject.eu/fileadmin/media/reports/20120618_D5_3_PEER_Usage_Study_RCT.pdf
5. Davis, P. M. (2012) The effect of public deposit of scientific articles on readership. *Physiologist* **55**, 161–165
6. Davis, P. M., and Price, J. S. (2006) Ejournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER. *J. Am. Soc. Inf. Sci. Technol.* **57**, 1243–1248
7. Bonate, P. L. (2000) *Analysis of Pretest-Posttest Designs*, Chapman & Hall, New York
8. Singer, J. D., and Willett, J. B. (2003) *Applied Longitudinal Data Analysis*, Oxford University Pr, New York
9. Bennett, L. (2012) The potential effect of making journals free after a six month embargo, Association of Learned, Professional, and Society Publishers/Publishers Association, Bristol/London, UK. Retrieved from <http://www.publishingresearch.net/documents/ALPSPApotentialresultsofsixmonthembargo.pdf>
10. Cox, L. (2011) Librarians' use of usage statistics for journals and e-books. *Learn. Publ.* **24**, 115–117
11. Publishers Communication Group (2011) *Journal Subscriptions Renewal Trends 2006–2011*, p. 3, Publishers Communication Group, Cambridge, MA, USA. Retrieved from [http://www.pcgplus.com/pdfs/Journal Subscription Renewal Trends.pdf](http://www.pcgplus.com/pdfs/Journal%20Subscription%20Renewal%20Trends.pdf)
12. Ware, M. (2006) ALPSP survey of librarians on factors in journal cancellation. Association of Learned, Professional, and Society Publishers, Bristol, UK. Retrieved from http://www.alpsp.org/Ebusiness/Libraries/Publication_Downloads/libraryreport-summary.sflb.ashx?download=true
13. Boukacem-Zeghmouri, C., and Schöpfel, J. (2012) Statistics usage by French academic libraries: a survey. *Learn. Publ.* **25**, 271–278

Received for publication February 18, 2013.

Accepted for publication March 18, 2013.