Do Open Access Articles Have Greater Citation Impact?

A critical review of the literature

Iain D. Craig,[1] Andrew M. Plume,[2] Marie E. McVeigh,[3] James Pringle[3] and Mayur Amin[2]

[1] Wiley-Blackwell, 9600 Garsington Road, Oxford, OX4 2DQ, UK

[2] Elsevier, The Boulevard, Langford Lane, Kidlington, Oxford, OX5 1GB, UK

[3] Thomson Scientific, 3501 Market Street, Philadelphia, PA 19104, USA

prc
publishing
research
consortium

Contents

Executive overview

1 The last few years have seen the emergence of several Open Access options in scholarly communication which can broadly be grouped into two areas referred to as 'Gold' and 'Green' Open Access (OA). In this article we review the literature examining the relationship between OA status and citation counts of scholarly articles, and take no position on the relative value or sustainability of these communication models.

2 Early studies showed a correlation between the free online availability or OA status of articles and higher citation counts.

3 The authors of many of these studies implied that this correlation was causal, without due consideration of potential confounding factors.

4 More recent investigations have applied sophisticated bibliometric methods to dissect the nature of the relationship between article OA status and citations.

5 Three non-exclusive postulates have been proposed to account for the observed citation differences between OA and non-OA articles: an Open Access postulate, a Selection Bias postulate, and an Early View postulate.

   a. The Open Access (OA) postulate suggests that authors are more likely to read, and thus cite, articles that are made available in an OA model.

   b. The Selection Bias (SB) postulate suggests that the most prominent (and thus most citable) authors are more likely to make their articles available in an OA model, and that they are more likely to do so with their most important (and thus most citable) articles.

   c. The Early View (EV) postulate relates only to articles posted before final journal publication, and suggests that the period between the early posting of an article (either pre-print or post-print) and the appearance of the cognate published journal article allows for earlier accrual of citations. Failing to account for this effect must necessarily give a biased result.

6 The most rigorous study to date, conducted in the field of condensed matter physics, showed that after controlling for a clearly demonstrated Early View postulate, the remaining difference in citation counts between OA and non-OA articles is explained by the Selection Bias postulate. No evidence was found to support the OA postulate *per se*; i.e. article OA status alone has little or no effect on citations.

7 As citation practices vary widely by discipline, further studies using a similarly rigorous approach are required to determine the generality of this finding in other fields of research. Such studies must account for the heterogeneous distribution of citations across any group of articles and establish the date of earliest availability of each article in the study, as citation accumulation is time sensitive.

Introduction

With the advent of the internet and electronic publishing, new models of scholarly communication have emerged that simultaneously complement and challenge established systems. Although the term 'Open Access' (OA) is taken broadly to mean that accessing, downloading, and reading material is free to the entire population of internet users, several options for the provision of that access have emerged. These can be grouped into two broad models: 'Gold' and 'Green'. The Gold model uses a traditional journal publication system, but shifts the economic/financial model. Instead of a subscriber paying to read the final version of a peer-reviewed article, an author or sponsor pays to publish the article, and reading the article is free to anyone wishing to do so. A journal may operate wholly under this model, or may use a hybrid model combining subscription and article sponsorship. The Green model of OA relies on posting the author's manuscript of an article into an institutional or subject-based electronic archive, either in the form of a pre-print (as submitted to a journal for peer review) or as a final copy of the peer-reviewed edited full text (a post-print).[1] A less rigorous but increasingly common form of archiving is the use of individual author webpages, outside of a structured archive. Articles that are posted as pre-prints may be subsequently accepted for publication in a journal and may then also be archived as post-prints, sometimes, but not universally, replacing the pre-print version. Economically, Green OA relies on the sustainability of the existing journal system as, unlike Gold OA, it does not provide any financial support for journals.

An increasing amount of research on the effects of OA models on scholarly communication has emerged in recent years, and it is clear that the methods for performing this research have been developing alongside the new models that they study. One of the foremost questions asked is: 'Do Open Access research articles have a greater citation impact?' Another way of asking this question at the most personal level for the authors of journal articles is: 'Will my research paper(s), and therefore will I, get a citation benefit from the Gold and Green Open Access models?' In this article we survey the original research literature on this topic to date, with a particular emphasis on methodological issues, and highlight areas in which further research is required. As this is an evolving area of research, the terminology has yet to become fixed; we follow the terminology used by the authors of each article under discussion.

[1] Although sometimes Green OA is used in a narrower context to refer to post-print archiving only, we use the broader definition here, inclusive of both pre-print and post-print archiving, because citation studies referenced here are generally broader in their definition of what is termed OA.

Methodological issues in citation analysis

A citation is defined as the listing of a previously published article in the reference section of a current work; this is usually taken to imply the relevance of the cited article to the current work. Information about articles and the citations between them are collected in databases known as citation indexes. The best-known example of a citation index is Thomson Scientific's Web of Science®, which now contains about 40 million bibliographic records and over 550 million citations from the past century. Other indexes include Scopus™, Google Scholar™, CiteSeer, and NASA's Astrophysics Data System. Citation analysis is a core tool in the research discipline known as bibliometrics, defined as the quantitative analysis of the units of scientific communication (e.g. articles, book chapters, etc.) and the citations that connect them. This field consists of a worldwide community of research bibliometricians, with their own learned societies, journals, and active online discussion lists.

Bibliometrics is a specialized and often complex field of study, and far transcends a simple counting of citations. Three measurement problems in bibliometrics are of particular relevance to the relationship between OA and citations. Firstly, it can be difficult to match the development of citations to publication dates or article posting dates. Citations accrue to articles over time, and thus older articles typically have greater citation counts than new articles. To overcome this age effect citations must be counted over a fixed period of time after publication or posting to allow useful comparisons between articles published at different times. Secondly, comparisons of the average properties of two sets of articles (grouped for example by journal, subject area, nationality, or OA status) must be interpreted very carefully, as such aggregates usually represent a heterogeneous population with a skewed distribution of citations. Finally, subject variations must be acknowledged, because dissemination of research findings via journal articles is not the primary channel of scholarly communication in all disciplines (see Table 7.3 in Moed, 2005)[2], and because citation practices themselves differ greatly between subject fields,[3] making it difficult to translate observations from one subject directly to the prediction of effects in another subject.

[2]Moed, H.F. (2005) Citation Analysis in Research Evaluation (Dordrecht: Springer).

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

Correlations of online availability and increased citations

The first research that showed a correlation between articles made available online and higher citations was carried out by Lawrence.[4,5] Lawrence based his study solely on conference proceedings articles in computer sciences and related disciplines that were listed in the Digital Bibliography & Library Project Computer Science Bibliography. He assessed the availability of a corresponding full-text article online and the number of citations (excluding author self-citations) received to date using the ResearchIndex (often referred to as CiteSeer) autonomous indexing database. Importantly, he made no distinction between the different methods of making the full-text articles available online. Online availability can occur in many ways, only some of which can be correctly termed 'Open Access'. There is also ambiguity about the definition of 'online' and 'offline' in this study – it is not clear whether the latter includes articles that are available online, but only via a subscription. A recent investigation comparing CiteSeer to Thomson Scientific's Web of Science and to GoogleScholar suggested that CiteSeer may not be ideal for citation studies[6]. However, it is not clear how this apparent over-reporting of citations might affect within-subject, and within-citation-index, comparisons.

Lawrence demonstrated a correlation between the likelihood of online availability of the full-text article and the total number of citations to date for articles published in non-overlapping consecutive pairs of years from 1989 to 1999. He further showed that the relative citation counts for articles available online are on average 336% higher than those for articles not found online but that were presented at the same conference (or 'publication venue').

In his analysis Lawrence assumed that 'articles published in the same venue are likely to be of similar quality', and that, by inference, articles of similar quality

[4]Lawrence, S. (2001a) Online or invisible? Available at http://citeseer.ist.psu.edu/online-nature01/ (link verified 23rd April 2007)

[5]Lawrence, S. (2001b) Free online availability substantially increases a paper's impact. *Nature* 411: 521.

[6]Bar-Ilan, J. (2006) An ego-centric citation analysis of the works of Michael O. Rabin based on multiple citation indexes, *Information Processing and Management* 42: 1553-1566.

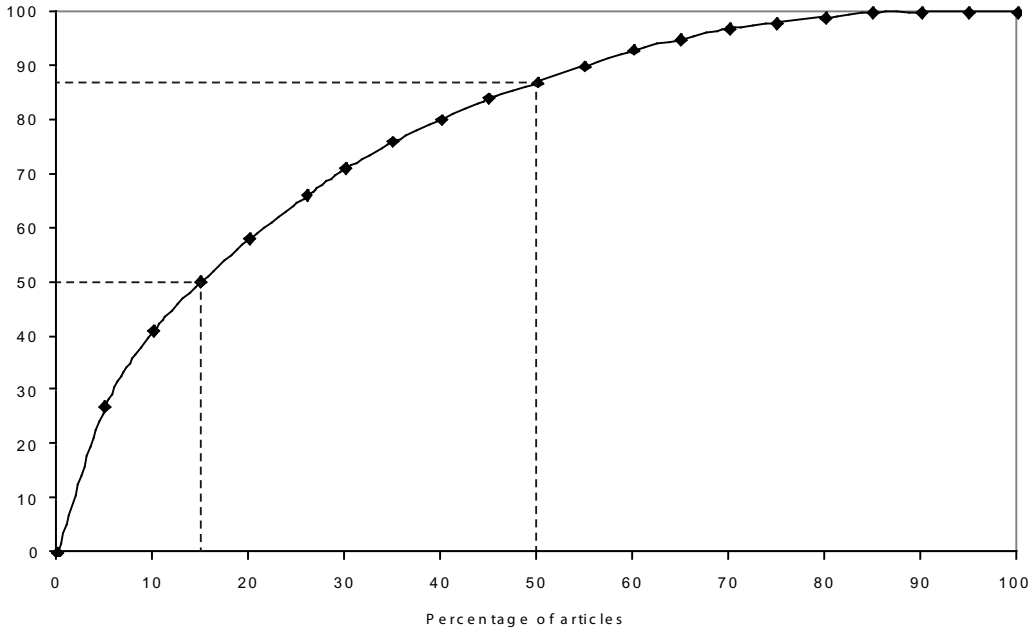


Figure 1

Typical skewed distribution of citations across a grouping of articles. Just 15% of the articles attract 50% of citations and almost 90% of citations were to 50% of the articles. The selection bias hypothesis suggests that Open Access articles are represented disproportionately in the former sub-group. Adapted from Seglen[6].

would receive similar numbers of citations. Neither of these key assumptions was supported by previous quantitative analyses. Seglen[7] reported that within the 'publication venue' of a single journal, 15% of articles accounted for 50% of citations, and almost 90% of citations were to just 50% of the articles (Figure 1). Lawrence did acknowledge that there may be an element of what would later be called Selection Bias at work; that is, that the 'higher quality articles are more likely to be made available online'. When Lawrence attempted to account for this bias, by limiting the analysis by publication venue to 20 conferences with stringent selection criteria, the relative citation counts of online articles reduced to 286%, while still not ensuring a population of uniformly citable articles. In other words, some articles will naturally be more cited than others, irrespective of any other attribute or condition of that article. In his conclusions, Lawrence (quite rightly) did not claim that this observed correlation was proof of causality, as correlation alone cannot prove a causal link.

[7]Seglen, P.O. (1992) The skewness of science. *Journal of the American Society for Information Science* 43: 628-638.

At around the same time, Anderson *et al.*[8] reported on the consequences of making selected articles, accepted for publication in *Pediatrics* between 1997 and 1999, freely available online-only at the journal's website. This is article-level Gold OA, but without the requirement for sponsorship of the costs of publication by the author or another party. The analysis was complicated by the fact that up to the half-way point in this period (July 1998), articles not made 'Open Access' in this way were published solely in the print journal, with no online availability. After the mid-point of the study, both OA and non-OA articles were published online, although their access model differed. Thus, accepted papers made OA in the first half of this period were posted online immediately, while those destined for print only were subject to additional printing and distribution delays. Anderson and colleagues investigated citation counts for each article (cumulative to the end of 2000), using an unspecified version of Thomson Scientific's Science Citation Index®. The authors noted a citation *dis*advantage for the OA articles of more than twofold. Articles published online-only were selected not by the submitting authors themselves, but rather by the *Pediatrics* editor. These online-only articles were selected to give 'preference to articles of broader international interest', but not necessarily their relative quality or scientific importance, compared with those accepted for print publication. The removal of the possible influence of an author Selection Bias and its replacement with an editorial Selection Bias may have been sufficient to account for the observed differences in mean citation counts.

[8]Anderson, K., Sack, J., Krauss, L., and O'Keefe, L. (2001) Publishing online-only peer-reviewed biomedical literature: three years of citation, author perception, and usage experience. *Journal of Electronic Publishing* 6.

The outcomes of the Lawrence and Anderson *et al.* studies were exactly opposite to each other, but together illustrate some of the methodological problems in determining whether OA affects citations. The former showed a citation advantage, the latter a disadvantage. Both suffered from lack of clarity about the precise method of citation counting, and both suffered from Selection Bias in some form; in Lawrence's case an author Selection Bias, and in Anderson *et al.*'s case an editorial Selection Bias. Lawrence accounted for an important potential confounder in removing subsequent self-citations by one or more of the same authors in subsequent articles, whereas Anderson *et al.* did not. Authors' self-citations have been acknowledged as a source of distortion in the analysis of scientific communication and are more likely to be made to papers with multiple authors.[9] It is reasonable to expect that articles with numerous authors would also have a greater likelihood of being self-archived by one or more of the authors, or to be otherwise made available online.

[9]Glänzel, W. and Thijs, B. (2004) Does co-authorship inflate the share of self-citations? *Scientometrics* 61: 395-404.

Correlations of Open Access and increased citations

The first study to assess the effect of Green OA relating to published journal articles (and not simply to conference proceedings and 'online availability') was published by Harnad and Brody,[10] and elements of these data were included in later papers.[11,12] Over 95,000 pre-print manuscripts in physics and mathematics deposited in the subject-based repository arXiv (http://www.arXiv.org) were matched with the final published journal article indexed in Thomson Scientific's Web of Science and were termed 'Open Access'. Citation counts to these articles were then compared with those for all other articles (termed 'Non-Open Access') published in the same journal and the same year (between 1992 and 2003) and the ratio of the two values derived. Articles with a corresponding pre-print version deposited in arXiv had higher citation counts than those that did not. The derived Open Access/Non-Open Access ratio varies by subject field, year of publication, and whether certain factors have been controlled or not (such as removal of author self-citations or comparison with articles published in the same journal).

An important methodological issue in this study is that it ignored the potential skewness of the distribution of citations within each group of articles. Coupled with the fact that a small proportion of articles had a corresponding pre-print version in arXiv, this means that distortions due to non-uniform sampling were even more likely than if a more representative sample were available. Moreover, expression of the citation count differences as a ratio obscures the actual magnitude of the effect: in the example of very small sample sizes of Open Access and Non-Open Access articles, large ratios may be due to a small number of additional citations for the Open Access articles, an effect that is readily attributable to many other factors and indiscernible from common citation 'background noise', an effect that is readily attributable to many other factors and indiscernible from common citation 'background noise'. Furthermore, these articles may have been published at any time during a 12-month period, so that they were available for citation for very different durations, however, this effect diminishes with larger counting intervals. In this study, the authors also drew a direct causal link between Open Access and citations, based on an observed correlation between them, but without substantiating this claim.

Antelman[13] took a novel approach to examining the relationship between online availability of full-text articles (not strictly Green Open Access) and citation counts. Using a method that mimics the information-seeking behaviour of researchers, she manually searched online for randomly selected articles published in leading journals across four distinct disciplines (mathematics, electrical and electronic engineering, political science, and philosophy). The articles considered were published in 2001 and 2002 (1999 and 2000 for philosophy), and citation counts until 2003 (excluding author self-citations and citations from within the same journal issue) were collected from Thomson Scientific's Web of Science. Full-text articles freely available online (at a location other than the publisher's website) that had the same title as the selected articles were deemed 'Open'; the remainder were termed 'Not Open'. Only in mathematics was a substantial number of 'Open' articles made available through a subject-based electronic archive (which is a major aspect of the Green mode of Open Access). In the other disciplines, posting on the author's own website was the primary means of online availability. Antelman calculated mean citation counts of Open and Not Open articles and showed that the percentage difference between the means of the two cohorts varied by discipline, from 45% in philosophy up to 91% in mathematics. It is difficult to generalize about these results, owing to societal differences in the citation behaviours of

[10] Harnad, S. and Brody, T. (2004) Comparing the impact of Open Access (OA) vs. non-OA articles in the same journals. *D-Lib Magazine* 10.

[11] Brody, T. (2004) Citation analysis in the Open Access world. Available at http://eprints.ecs.soton.ac.uk/10000 (link verified 23rd April 2007)

[12] Harnad, S., Brody, T., Vallières, F., Carr, L., Hitchcock, S., Gingras, Y., Oppenheim, C., Stamerjohanns, H., and Hilf, E.R. (2004) The access/impact problem and the green and gold roads to Open Access. *Serials Review* 30: 310-314.

[13] Antelman, K. (2004) Do Open-Access articles have a greater citation impact? *College & Research Libraries* 65: 372-382.

authors in these disparate academic communities.

Antelman recognized small sample sizes and the skewness of citation distributions amongst articles appearing in the same journal as confounding factors, and made an attempt to account for this in the statistical analysis of the data. Like Harnad *et al*.[12], she did not suggest or explore any potential reasons for the observed correlation, beyond the assumption that online availability leads to increased citation counts. In a subsequent response[14] to a letter to the editor about this paper,[15] Antelman was careful to state that her data did not support any notion of causality. Instead she noted that additional unpublished research indicates a significant Selection Bias effect (at least in the social sciences).

Following a similar approach, Hajjem *et al*.[16] used a robot to search online for more than 1.3 million articles published in Thomson Scientific-indexed journals across ten distinct disciplines (biology, psychology, sociology, health, political science, economics, education, law, business, and management). The articles were published between 1992 and 2003, and citation counts were cumulative to the end of this period, including author self-citations. Full-text articles available online that had the same title and first author name as the selected articles were deemed 'Open'; the remainder were termed 'Not Open'. The magnitude of the derived Open Access/Not Open Access ratio varied by subject field and year of publication between 25% and 250%. Again, citation practices in these areas are so divergent as to confound generalization across subject areas.

The robot's ability to identify Open and Not Open articles correctly was subsequently re-analysed in a technical paper[17] and found to significantly overestimate Open articles. The data were re-analysed by the two authors who had contributed to both papers, and were published in another technical article that reversed this finding to agree again with the original article.[18] There are additional methodological issues concerning the apparent averaging of averages: Open and Not Open citation counts were generated by averaging for each group the citations per issue, then issues per journal, then journals per discipline. These data could be very susceptible to outlier values in a potentially skewed citation universe. Sample sizes across subject disciplines are also heterogeneous: the reported 'Open Access Advantage' (or OAA) in Biology, which comprised 49% of the articles in the study, was by far the lowest OAA reported at 36%. Conversely the highest value of OAA, 172%, occurred in Sociology, a discipline that accounted for about 8% of the articles studied. The calculated mean and median values of the OAA across all ten disciplines (83% and 77%, respectively) also took no account of the relative sizes of the disciplines involved, and so also represented averages of averages in a potentially skewed distribution. While the authors noted that the observed correlations did not demonstrate causality, they dismissed the possible influence of a Selection Bias without further analysis.

[14]Antelman, K. (2006) Letter to the Editor: Response to Philip Davis. *College & Research Libraries* 67: 105.

[15]Davis, P.M. (2006) Letter to the Editor: Do Open-Access articles have a greater citation impact? *College & Research Libraries* 67: 103-104.

[16]Hajjem, C., Harnad, S., and Gingras, Y. (2005) Ten-year cross-disciplinary comparison of the growth of Open Access and how it increases research citation impact. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering* 28: 39-47.

[17]Antelman, K., Bakkalbasi, N., Goodman, D., Hajjem, C., and Harnad, S. (2005) Evaluation of algorithm performance on identifying OA. Available at http://eprints.ecs.soton.ac.uk/11689 (link verified 23rd April 2007)

[18]Hajjem, C. and Harnad, S. (2006) Manual evaluation of robot performance in identifying Open Access articles. Available at http://eprints.ecs.soton.ac.uk/12220 (link verified 23rd April 2007)

When should citation counting begin?

None of the studies discussed so far have taken account of the critical dimension of temporal progression: that is, the time difference between when an article is made available online or deposited in an electronic archive and when it is published. The sole consideration is whether or not the article was freely available at the time of the study. Furthermore, the relative timing of publication and the counting of references to the article must be precisely defined (ideally imposing a fixed window of citation counting) to allow true measurement of citation effects (Figure 2). Fixed citation windows are a standard method in bibliometric analysis, in order to give equal time-spans for citation to articles published in different years, or at different times in the same year. In order to argue that online availability bears a causal relationship with subsequent citations, the duration of this online availability must be established and the time-course of citation accrual to Open and Not Open articles must be examined relative to their earliest availability in either form.

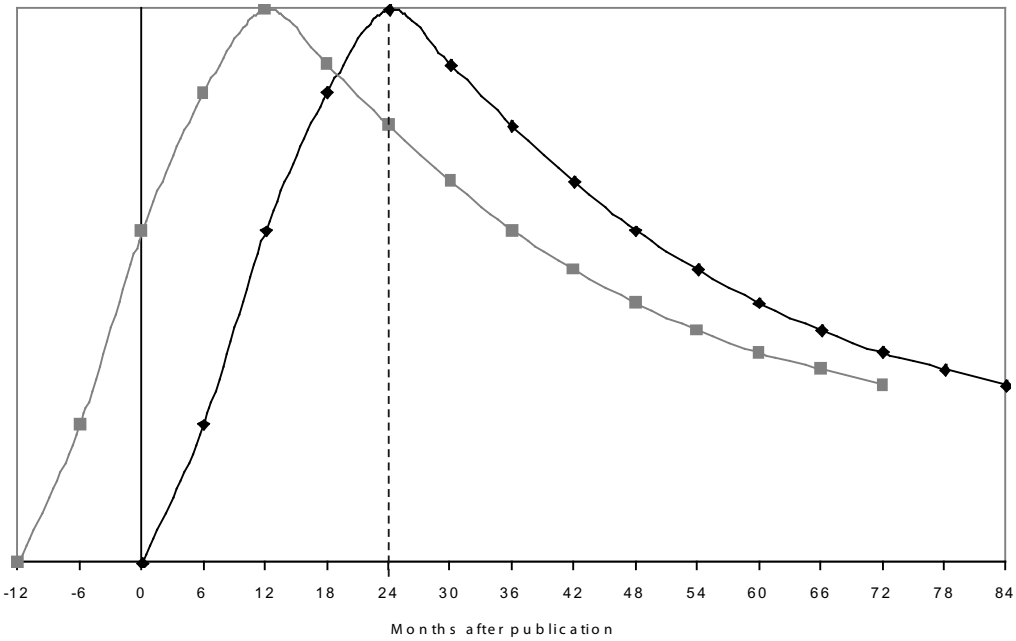


Figure 2

Typical citation time-course across a grouping of articles. The black line represents articles published in a journal at month 0, while the grey line represents articles deposited as a pre-print 12 months prior to publication in a journal at month 0. Citation counts at 24 months after journal publication therefore allows an additional 12 months of citation counting for the pre-print group – this is the basis of the Early View hypothesis. Adapted from Henneken[19] and Moed[20].

A breakthrough in the analysis of the relationship between citations and Open Access status came with the study performed by Schwarz and Kennicutt[21] on articles published in *Astrophysical Journal* (*ApJ*) in 1999 or 2002 with a matching pre-print version deposited in the astrophysics section of arXiv. These authors were among the first to recognize that citation counting begins earlier for articles deposited in the arXiv repository as pre-prints (i.e. before publication in a peer-reviewed journal) than articles without pre-prints deposited in arXiv. Schwarz and Kennicutt were certainly the first to attempt to account for this in the analysis of their findings. Comparing citations to these 'posted' articles within the limited citation data source of NASA's

[19] Henneken, E.A., Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., Murray, S.S. (2006) Effect of e-printing on citation rates in astronomy and physics. *Journal of Electronic Publishing* 9.

[20] Moed, H.F. (2007) The effect of "Open Access" upon citation impact: An analysis of ArXiv's Condensed Matter Section. *Journal of the American Society for Information Science and Technology* (in press). Available at http://arxiv.org/abs/cs.DL/0611060 (link verified 23rd April 2007).

[21] Schwarz, G.J. and Kennicutt, R.C. (2004) Demographic and citation trends in astrophysical journal papers and preprints. *Bulletin of the American Astronomical Society* 36: 1654-1663.

Astrophysics Data System (ADS) versus those of non-posted *ApJ* articles showed that the former had citation counts on average twice those of the latter (including self-citations). An (approximately) fixed citation window was imposed by counting cites to articles published in the second half of 1999 until a fixed point during 2003. Follow-up work by Metcalfe[22] expanded this analysis to a total of 13 journals (including the generalist titles *Nature* and *Science*) and broadly confirmed these findings for articles published in 2002.

Schwarz and Kennicutt produced citation histograms showing the accrual of citations to articles with a pre-print version and those without, and found a marked difference in the profiles. Pre-printed articles were available to be cited by an average of 12 months before the final *ApJ* published article, effectively starting the citation counting process earlier. The data suggested that this earlier citation counting does not affect the final magnitude of citations accrued to a journal article. At least part of the remaining difference may be attributable to the Selection Bias in posted articles that the authors described in their discussion. In related work, Metcalfe[23] analysed citation counts (again from ADS) for articles published in *Solar Physics* in 2003 with a matching pre-print version deposited in the astrophysics section of arXiv or in an independent solar physics archive at Montana State University (MSU). Despite very small sample sizes (170 articles, 13 of which were 'posted'), Metcalfe showed a significant difference in citation counts between 'unposted' articles, those 'posted' in arXiv, and those 'posted' in MSU. Like Schwarz and Kennicutt, Metcalfe also compared citation counts for 'posted' and 'unposted' conference proceedings articles. Whereas Schwarz and Kennicutt did so to determine the relationship between publication venue and citations, Metcalfe attempted to interpret these data as evidence for a lack of a Selection Bias effect. However, the fact that 'posted' proceedings papers follow the same higher citation patterns as 'posted' journal articles does not support this view.

[22] Metcalfe, T.S. (2005) The rise and citation impact of astro-ph in major journals. *Bulletin of the American Astronomical Society* 37: 555-557.

[23] Metcalfe, T.S. (2006) The citation impact of digital preprint archives for solar physics papers. *Solar Physics* 239: 549-553.

Deconstructing the Open Access citation effect

All of the studies discussed above were concerned with demonstrating a difference between average citation counts to articles that were made available online and those that were not. While some implied a causal relationship, most acknowledged Selection Bias as a possible explanation for the observed citation patterns, and some also noted differences in the effective citation life-times of the two groups. The articles discussed in this section represent a new phase in the development of the literature on this topic. They are concerned with systematically deconstructing the elements of the Open Access citation effect, which they recognize as being a complex, multi-dimensional phenomenon.

Kurtz *et al.*[24] were the first to formalize these possible explanations for the observed differences in citation patterns and to examine systematically their effects by controlling for each, one at a time: a general Open Access effect due to unrestricted ability to read and cite articles (the OA postulate); the Early View postulate (which they term the 'Early Access effect'), due to articles appearing sooner; and a Selection Bias due to more prominent authors posting their articles, and/or authors preferentially posting their better works (the Selection Bias postulate). To investigate the OA and Early View postulates the authors calculated the probability that an article will cite another article previously published in a defined window of time, using citations only within and between a set of seven core astrophysics journals in the ADS database. The results clearly showed that there is no *general* OA effect: the massive increase in online availability of articles in both arXiv and ADS in the 1990s was not correlated with any subsequent increase in citations to these articles. Evidence suggesting a strong Early View effect was suggested by an increase from the late 1990s in the probability of subsequent citation of an article within the first six months after publication (whether it was also deposited in arXiv or not). While this period correlates with the rise in popularity of arXiv, this correlation cannot be taken as proof of cause: other factors affecting the citation behaviour of authors in this field could be influencing the age of their cited references.

To investigate the Selection Bias postulate, Kurtz *et al.* (2005) took an approach similar to that of Schwarz and Kennicutt (2004), looking at articles published in *ApJ* in 2003 with a matching pre-print version deposited in the astrophysics section of arXiv. Comparing citations (again taken from ADS) of these 'posted' articles with those of non-posted *ApJ* articles showed that the former had citation counts on average twice those of the latter (including self-citations). A strong Selection Bias effect was demonstrated by the observation that articles with a version posted in arXiv had a greater probability of occurring in the top 200 most cited articles published in *ApJ* in 2003. A follow-up study by most of the same authors[19] used similar methods to confirm the Selection Bias effect in *ApJ* and four other astrophysics journals: articles with a version posted in arXiv were over-represented in the top 100 most cited articles published since the late 1990s. To re-examine the Early View postulate the authors examined articles published in *ApJ* between 1997 and 1999 and counted citations accrued each month after publication for 5 years. The resultant citation curves for articles with or without a version posted in arXiv were very similar in shape. Like Schwarz and Kennicutt (2004), Henneken and colleagues failed to account for the clear Early View effect of about 12 months of earlier citation counting (the average length of early visibility for *ApJ* pre-prints found by Schwarz and Kennicutt). Were they to do so, the remaining differences would become more easily visible and open for explanation in terms of the Selection Bias postulate.

Using these same three postulates (but with Selection Bias re-termed 'Quality

[24] Kurtz, M.J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E, and Murray, S.S. (2005) The effect of use and access on citations. *Information Processing and Management* 41: 1395-1402.

Differential'), Davis and Fromerth[25] analysed the articles published in four mathematics journals between 1997 and 2005, either with or without matching versions deposited in the mathematics section of arXiv. Using citation counts from the limited MathSciNet database (which includes self-citations and, uniquely, citations of the pre-print versions), the authors showed an average 35% increase in citations (or 1.1 citations per article) of articles posted in arXiv versus those that were not posted. The Early View postulate was tested by regression analysis of citation counts for individual articles posted in arXiv and subsequently published in the same journal against the number of days before journal publication they were posted in arXiv. Although there was no significant correlation (indeed, many highly-cited articles were deposited in arXiv *after* publication), we might not expect one in a field such as mathematics, in which citation practices are such that the average age of cited references is relatively high, and the frequency and speed of publication are relatively low. By contrast, articles with a version posted in arXiv were over-represented among the most highly cited articles in this study, implying a Selection Bias effect.

[25] Davis, P.M. and Fromerth, M.J. (2007) Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics* 71: 203-2150.

What does Open Access mean for individual authors?

In an attempt to investigate the effect of Gold Open Access on citations, Eysenbach[26] undertook an analysis of articles published in the latter half of 2004 in a single hybrid Gold journal, the *Proceedings of the National Academy of Sciences* (*PNAS*). *PNAS* is a large multidisciplinary journal, publishing in areas as diverse as biochemistry, neuroscience, genetics, biophysics, chemistry, evolution, microbiology, and plant sciences. Articles whose cost of publication was borne by the authors of the article (or a sponsor) were termed 'OA', while the remainder were termed 'non-OA'. As all the articles were published without delay after peer review and typesetting, no Early View effect would have been present to confound the analysis.

The two groups of articles were subjected to logistic regression using several variables, including author, funding, subject, and other publication characteristics. While OA status was found to remain a significant predictor of the likelihood that an article would be cited at least once within 10-16 months after publication, so too were several other factors. Among these were the number of authors on the paper and funding from competitive grants, each of which can be construed as an independent marker for scientific rigour and thus article 'quality'. Indeed, an earlier study showed a clear correlation between grant funding and citation counts for articles published in *PNAS*.[27] Given the presence of these potentially confounding factors, further analysis would be needed to determine whether OA status is indeed the primary 'driver' for citation, as Eysenbach maintains. Moreover, the first authors of OA papers also tended to be more senior (more lifetime publications), and the combined average citations per paper for the first and last authors of OA papers tended to be higher too. Logistic regression analyses were not presented for either of these potential confounders. Taken together, these data suggested a strong influence from Selection Bias effects, which were not explored in the study. Differences between subject areas are alluded to, but sample sizes were variable and reduced the robustness of the findings. The main study was not stratified by subject area and so may have suffered from subject-specific citation biases. Eysenbach extended his study to assess the effect on citations of online availability of the full-text articles at any location other than the *PNAS* website or PubMed Central. This analysis suggested that additional online availability of OA articles (which are of course already freely available) did not significantly enhance citability further, arguing against any general Open Access effect.

Eysenbach's work highlighted the importance of author characteristics (reputation, prior citation history, lifetime publication count, country, funding organization, etc.) as confounding variables in the analysis of Open Access and citations, but he did not fully explore the effect of author prestige when looking at the Selection Bias effect. Moed[20] undertook a study methodologically similar to that of Harnad and Brody,[10] matching almost 75,000 pre-print manuscripts deposited in the condensed matter physics section of arXiv with the final published journal article indexed in Thomson Scientific's Web of Science (appearing mostly in one of 24 physics journals). Citation counts (excluding self-citations) to both the arXiv version *and* the final published version of these 'OA' articles were then compared with those for all other 'non-OA' articles published in the same journal and a ratio of the two values derived: the Citation Impact Differential (CID). This analysis confirmed that articles with a corresponding pre-print version deposited in arXiv had higher citation counts than those that did not, and that the size of increase varied by year of publication and by journal. It should be noted that Moed's CID values were systematically lower than the 'OA'/'non-OA' ratios of Harnad and Brody. However, Moed did not ascribe the CID he observed to an OA effect but went on to explore the observed difference further.

[26] Eysenbach, G. (2006) Citation advantage of Open Access articles. *PLoS Biology* 4, e157.

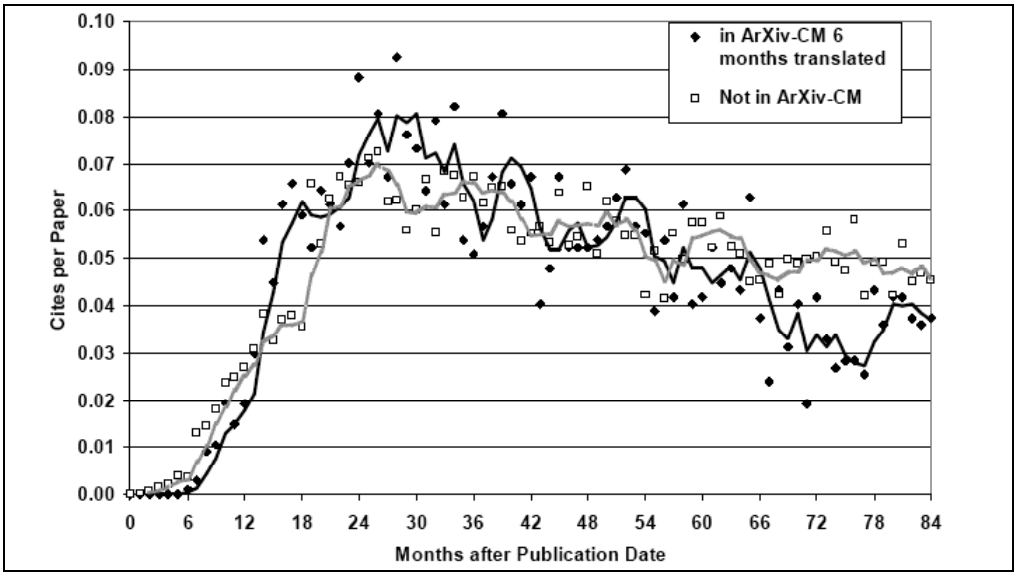
[27] Boyack, K.W. (2004) Mapping knowledge domains: Characterizing PNAS. *Proceedings of the National Academy of Sciences USA* 101: 5192-5199.

Moed's analysis of the underlying reasons for this difference introduced innovative approaches and methods to the study of Open Access status and citations for the first time, drawing heavily on the standard practices of citation analysis practiced by research bibliometricians. This was the first study to impose fixed windows of time for counting citations to each article analysed. As noted earlier, this is imperative for fair comparison between articles published at different times. Moed analysed the Early View effect by imposing two fixed citation periods for each article analysed, either the first three years after publication or the fourth to the sixth year after publication. Citations were also counted on a monthly basis after publication to give a granular view of early citation counts, thus minimizing the effects of differences in publication frequency of the cognate journals. Both methods showed a clear Early View effect, and the latter approach produced the most visually striking results: when monthly citation curves for articles with or without a version posted in arXiv were plotted on the same chart, translation of the arXiv curves by the average length of deposit in arXiv (6 months, the average time between deposit in arXiv and formal publication of the refereed, final journal article) resulted in almost indistinguishable curves (Figure 3; cf. Schwarz and Kennicutt[21] and Henneken *et al*.[19]). Moed also found a strong quality bias effect, reflected in a significant over-representation of prominent authors in arXiv-deposited articles, and showed that articles with higher proportions of prominent authors are more likely to be cited than those with a higher proportion of less prominent authors. When both the Early View effect and the Selection Bias effect were controlled for, the magnitude of the CID across all 24 major physics journals was variable but averaged just 7% (for authors with more than four lifetime articles).

Moed concluded that, based on these results, there is no *general* Open Access citation advantage for individual authors, especially the more prolific ones. Like all of the other studies discussed above, the findings of this study are only directly applicable to the subject discipline it is based on; generality cannot be assumed across all areas of research, as citation behaviour and author attitudes to Open Access are cultural factors that differ across different fields.[3,28] However, this study has set a benchmark for a rigorous method that can be applied to further studies of these effects in other subject areas.

[28] Swan, A. and Brown, S. (2005) Open access self-archiving – An author study. Available at http://eprints.ecs.soton.ac.uk/10999 (link verified 23rd April 2007)

Figure 3



Citation time-course for published articles with ('in ArXiv-CM') and without ('not in ArXiv-CM') a self-archived version in the Condensed Matter section of arXiv. The curves represent three-month moving averages. The open squares and grey line represent articles published in a journal only, while the black diamonds and black line represent articles published in a journal but also deposited in arXiv an average of 6 months prior to their publication in a journal. To account for the Early View effect the latter have been translated to the right by 6 months. Reprinted from Moed[19] with permission.

Conclusions

We posed two questions at the beginning of this article: firstly 'Do Open Access research articles have a greater citation impact?', and secondly 'Will my research paper(s), and therefore will I, get a citation benefit from the Gold and Green Open Access models?' These two questions in turn represent the main stages in the development of the research literature on this subject. While early work was simply concerned with seeking a positive correlation between Open Access and citation counts, more recent work has begun methodically to dissect the factors that drive the observed correlation and to discover what this might mean for individual authors using the Green or Gold Open Access models.

These questions have driven the development of increasingly robust research methods that account for potentially confounding factors and biases. Citation analysis is not a trivial undertaking, not least because it requires technical ability with data manipulation and analysis, and also requires an understanding of the underlying drivers for citation in scholarly publications. All but one of the studies discussed above failed to determine accurately the date of earliest dissemination of each article, and then to impose a defined citation window, which must be used if citation analysis of Open Access status is to yield definitive results. Even the most robust methods developed to date are unable to show causality unequivocally, nor can they generalize the observed effects at the author level or across a large number of varied disciplines.

Initially, the observed positive correlation between the OA status of a given article and higher citation counts was interpreted as causal, since there was both an intuitive and sociological appeal to this mechanism. However, closer analysis has not confirmed any causal relationship, and has actually shown a more complex set of contributors to the effect itself. Assuming that citation differences are due solely to the free availability of an article implies that many scholars working in a given discipline are currently totally unaware of important, relevant literature in their field and are unable to read and cite it. This further suggests that authors will limit their citations to those works that are readily available in favour of citations to works that are of the highest relevance. This view of citation behaviour dismisses any contributing role from long-established and robust means of scientific and scholarly communication – namely, all mechanisms of peer communication, the influence and availability of cited references, and the inherent value a given researcher will place on the content of a paper, independently of the mechanism by which it might have been retrieved.

Instead, the concept of a general and highly influential Open Access effect requires us to imagine a situation in which authors either refer to a particular article simply *because* it is Open Access (not for any inherent relevance to the topic at hand) or fail to cite a relevant paper which they are unable to read (and so unable to cite) because it *is not* Open Access. Familiarity with the citation history of influential or canonical works in a discipline suggests that the overriding determinant of lifetime citations of an article is the quality, importance, and relevance of the work reported in the article.

What does Open Access provide for an individual author? The most rigorous study available to date[20] suggests that any residual Open Access effect in condensed matter physics is negligible, after accounting for Selection Bias and Early View effects. This suggests that the benefits of self-archiving for an individual article or the work of an individual author are uncertain and could be as much affected by subject area, inherent variations in publication, and citation patterns generally, and the presence

and/or importance of a specialized online pre-print archive. Scientific citation is influenced, overwhelmingly, by the relevance and importance of a given scholarly work to other scholars in the field. While other factors might have moderate effects, the process of science is driven not by access, but by discovery.

With this in mind, we invite the bibliometrics and broader scientific community to contribute methodologically sound and well-interpreted studies of the relationships between OA and citation counts across diverse disciplines. Such studies need to ensure that the signature skewed distribution of citation patterns is not casually assigned on the basis of simple correlations and that artefacts such as Selection Bias and Early View are accounted for. True randomised studies (with articles randomly selected from the same journal for OA treatment) may offer one approach, provided these can be managed practically, and other factors (e.g. the citing window, seasonally) can be controlled for.

Acknowledgements

The authors would like to thank Jeffrey Aronson at the Department of Clinical Pharmacology, Oxford University, and Henry Small, Chief Scientist at Thomson Scientific, for helpful comments on the manuscript.

Glossary

Term	Definition
Gold Open Access	Publication of a peer-reviewed article in a journal where the final published article is made free to read to anyone wishing to do so; the cost of publication is typically borne by the author or a sponsor on behalf of the author. Gold OA comprises two distinct approaches: one where publication is contingent on the author or sponsor paying (often called 'author pays'), and another where publication is not dependent on payment but where an author or sponsoring organisation can opt to make the published article freely available through the payment of a fee (sometimes called the 'sponsored article' option)
Green Open Access	Posting into an institutional or subject-based electronic archive of a pre-print or post-print article. Some have used a narrower definition to include only post-print articles but in this case we use a broader definition as it applies best to the studies reviewed here
Pre-print article	Author's manuscript of an article as submitted to a journal for peer review
Post-print article	Author's peer-reviewed manuscript of an article accepted for publication in a journal
Citation window	Time-span over which citations to an article are counted
Author self-citation	Citation to an article by one or more of the same authors in a subsequent article
Citation index	A database containing information about articles and the citations between them
Open Access postulate	Suggests that authors are more likely to read, and thus cite, articles that are made available under an OA model
Early View postulate	Suggests that the period between the early posting of an article in a repository and the appearance of the cognate published journal article allows time for earlier citation; refers to the Green OA model where articles (pre-prints or post-prints) are posted earlier than the final published article
Selection Bias postulate	Suggests that the most prominent (and thus most citable) authors are more likely to make their articles available under an OA model, and that they are more likely to do so with their most important (and thus most citable) articles. In other words, authors tend to promote their best work and the best authors are more likely to do so.