The Effect of "Open Access" on Citation Impact: An Analysis of ArXiv's Condensed Matter Section

Henk F. Moed
Centre for Science and Technology Studies (CWTS), Leiden University, P.O. Box 9555, 2300 RB Leiden, The Netherlands. E-mail: moed@cwts.leidenuniv.nl

This article statistically analyzes how the citation impact of articles deposited in the Condensed Matter section of the preprint server ArXiv (hosted by Cornell University), and subsequently published in a scientific journal, compares to that of articles in the same journal that were *not* deposited in the archive. Its principal aim is to further illustrate and roughly estimate the effect of two factors, "early view" and "quality bias," on differences in citation impact between these two sets of papers, using citation data from Thomson Scientific's *Web of Science*. It presents estimates for a number of journals in the field of condensed matter physics. To discriminate between an "open access" effect and an early view effect, longitudinal citation data were analyzed covering a time period as long as 7 years. Quality bias was measured by calculating ArXiv citation impact differentials at the level of individual authors publishing in a journal, taking into account coauthorship. The analysis provided evidence of a strong quality bias and early view effect. Correcting for these effects, there is in a sample of six condensed matter physics journals studied in detail *no* sign of a general "open access advantage" of papers deposited in ArXiv. The study does provide evidence that ArXiv *accelerates* citation due to the fact that ArXiv makes papers available *earlier* rather than makes them *freely* available.

Introduction

The debate on costs and benefits of "open access" compared to other forms of scientific literature publishing has a political, an economical, and an information–scientific dimension. In this debate, the term "open access" has different meanings. It is used to indicate a particular business model of scientific publishing, in which essentially the authors of articles published in a journal pay the costs of the publication and their full texts are freely accessible once they are published. But the term "open access" also is used to indicate open or free accessibility of scientific documents in general, regardless of whether these are published in a journal running under an open access model or published in a journal applying other business models but also (often after several months) deposited in a freely accessible archive such as a personal Web site or an institutional depository, or as preprints in a freely accessible preprint server.

From an information–scientific perspective, the key issue in this debate is how scientific–scholarly communication, and particularly its publication processes, can optimally profit from the new developments in information and communication technologies. From this perspective, it is highly relevant to analyze and evaluate the feasibility of the various publication models and their effects, both at a short and at a longer term. It is no wonder that citation analysis constitutes one of the principal tools in this research. In 1972, more than 3 decades ago, Eugene Garfield, the founder of the *Science Citation Index*, showed how citation analysis can be used to study the scientific–scholarly communication system, and to contribute to its better functioning and hence to a better science. He and his followers illustrated this in numerous studies.

During the past years, several case studies applied citation analysis to examine the effects of open access business models or openly accessible publication archives upon the "visibility" or "impact" of published articles (e.g., Davis & Fromerth, 2006; Eysenbach, 2006; Harnad & Brody, 2004; Kurtz et al., 2005). These studies explored statistical relationships among variables of interest, in case studies examining particular data samples, variables, and access modalities. The study presented in this article also is a case study, primarily of a methodological nature. It relates to papers deposited in the Condensed Matter section of ArXiv, a preprint server founded by Ginsparg and currently hosted by Cornell University. The key questions this article addresses are:

- How does the citation impact of articles deposited in ArXiv and subsequently published in a scientific journal compare to that of articles in the same journal that were *not* deposited in that archive?
- How should the differences in citation impact among the two sets of articles be explained? Is it only or mainly the open accessibility of ArXiv that accounts for these differences,

Received November 16, 2006; revised January 22, 2007; accepted January 22, 2007

 • Published online 30 August 2007 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/asi.20663

or are there *other* factors responsible as well, and how strong are their effects?

This article builds upon the work by Harnad and Brody (2004). It calculates an ArXiv Citation Impact Differential (CID), a measure that is similar to Harnad and Brody's "open access (OA) to non-OA impact ratios (IRs)" but more appropriate for application at the level of individual authors. Results were compared to those presented by Harnad and Brody. A strict replication of their findings could not be carried out since their data related to other ArXiv sections.

Following the work by Kurtz et al. (2005) and Davis and Fromerth (2006), three effects were distinguished. The first is the genuine open access effect, in the sense that ArXiv increases access to research papers. Note that none of the journals analyzed in this article have adopted an open access business model. The second effect is termed the *early view effect*: Articles appear earlier in ArXiv than they do in the (electronic or printed) journal. The aspect of accessibility at stake here is "earlier" versus "later," distinct from "open (or free)" versus "not open," as in the open access effect. Finally, there is a *self-selection* effect or *quality bias*. Kurtz et al. distinguished two dimensions. The first is that prominent authors may tend to deposit their papers in ArXiv more often than do less prominent scientists. In other words, prominent authors may be overrepresented in ArXiv. The second is that authors—prominent or less so—may tend to deposit their "better" papers in ArXiv.

The principal aim of the work described in this article is to further illustrate and roughly estimate the early view effect and quality bias of ArXiv on citation impact. Estimates for a number of journals in the field of condensed matter physics are presented. To discriminate between the open access effect and the early view effect, longitudinal citation data are analyzed covering a time period of 7 years, which is much longer than the time period of 18 months considered in a recent article by Eysenbach (2006) on the "citation advantage" of "OA" papers published in the *Proceedings of the National Academy of Sciences*.

The article is structured as follows. The next section describes data collection and elementary data handling as well as methodological issues. The empirical results are then presented. Finally, we draw conclusions and make suggestions for further research.

Data and Methods

A database was created of all 74,521 papers deposited in the Condensed Matter section of ArXiv (denoted as ArXiv-CM) during the time period 1992–2005. ArXiv-CM papers were linked to articles in journals processed by Thomson Scientific for the *Web of Science* (*WoS*), on the basis of first-author names, significant words from the papers' titles, and the information available in the "journal reference field" of an ArXiv publication record. The latter field, designed to give the source (e.g., journal, proceedings, book) in which a final version of the paper was published, was filled in only about 40% of the papers. Therefore, it was necessary to search for matches also on the basis of author names and title words. A base assumption underlying this approach is that documents from the two databases linked in this way represent one and the same article, and that ArXiv papers not found in the *WoS* were *not* (yet) published in *WoS* journals.

About 75% of ArXiv-CM papers were linked in the aforementioned way to a *WoS* source article. In this set of papers, the median time period between the date a paper was deposited in ArXiv and the date it was published in a journal was found to be about 6 months. The papers were published in several hundreds of journals, revealing a skewed distribution of publications among journals. Three journals—*Physical Review B*, *Physical Review Letters*, and *Physical Review E*—accounted for 50% of all linked articles. Of the 68 journals assigned in the *WoS* to the journal category "Physics, Condensed Matter," 24 had published at least 10 articles, or more than 1% of its articles, linked to an ArXiv-CM paper. It is on this set of 24 condensed matter physics journals that the analyses presented later is based. The six journals with the largest number of articles linked to an ArXiv-CM paper are listed in Table 1.

Citations to ArXiv versions of papers were traced in the *WoS* database by applying a citation match-key that included

TABLE 1. ArXiv Citation Impact Differential (CID) per journal and per citation window.

Journal	Total Publications 1992–2005	%Publications in ArXiv	%Share of ArXiv publications	%ArXiv CID	
				1–3 years after publication date	4–6 years after publication date
Physical Review B	13,285	19.7	70.8	43	27
European Physical Journal B	1,195	35.4	6.4	87	68
Journal Physics–Condensed Matter	1,143	7.2	6.1	88	68
Physica B–Condensed Matter	523	3.0	2.8	83	68
Solid State Communications	432	4.8	2.3	95	81
Intl. Journal of Modern Physics B	426	8.6	2.2	102	72
All selected journals ($n = 24$)	18,757	10.2	100.0	80	64

Cites to ArXiv versions are included. As can be seen, *Physical Review B* dominates the set of articles published in *WoS Condensed Matter* physics journals and deposited in ArXiv-CM, with a share as high as 70.8%. Note that the percentage of the *WoS* journal articles deposited in ArXiv-CM is an aggregate statistic for the total time period 1992–2005. Calculated on an annual basis, it increased from 0.2% in 1992 to 12.3% in 2000 and to 19.8% in 2005. All journals show a substantial increase over the years.

DOI: 10.1002/asi

parts of the first-author name and the ArXiv paper number. Citations to *WoS* articles were collected using an advanced matching algorithm that takes into account numerous variations, errors, or discrepancies between cited reference and intended target article (Moed, 2005). Author self-citations were *not* included in the citation counts. The study described in this article did *not* analyze citations *within* the ArXiv, from one ArXiv article to another.

The average citation impact of a journal's papers deposited in ArXiv-CM was compared to that of its articles *not* deposited in that archive. *CPP* denotes the number of received citations per article, and the subscripts *a* and *na* denote whether the cited paper was deposited in ArXiv. Harnad and Brody (2004) defined their OA versus non-OA IR as

$$IR = 100 \times \frac{CPPa}{CPPna}.$$

Apart from the fact that this ratio obtains a value of 100% if there is no "open access advantage" at all (In that case, the numerator and denominator have the same value.), the main problem of using this ratio is that it may reach extremely high values if CPP_{na} is much smaller than 1, and especially that it is undefined if this denominator equals zero. As long as the ratio is calculated for a journal as a whole or for a set of journals, this normally is not a problem. But the study of quality bias (this article analyzes IRs at the level of *individual authors*) for which numbers of articles and citations are generally much lower, a CPP_{na} value of zero is no exception. Therefore, in this article an ArXiv CID is calculated, defined as:[1]

$$CID = 200 \times \frac{CPPa - CPPna}{CPPa + CPPna}.$$

Its values range between −200 (if $CPP_a = 0$) and +200 (if $CPP_{na} = 0$). If both CPP_a and CPP_{na} are zero, its value is defined as 0. CID values are generally lower than those obtained by Harnad and Brody's (2004) OA to non-OA IR. Elementary calculus shows that

$$CID = 2 \times \frac{IR - 100}{IR + 100}.$$

To analyze the *early view effect*, citation impact and ArXiv CID of a set of papers were analyzed in relation to their age, defined as the time period between publication date and date of citation. ArXiv CID was calculated for two *fixed* citation time windows: one for the time period involving the first 3 years after publication (the publication year included) and a second for the 4th until the 6th year of publication date.[2] In an in-depth analysis of the early view effect, citation counts per year were not sufficiently informative, and citations were therefore counted on a *monthly* basis as well. Following Harnad and Brody (2004), ArXiv CID also was measured for a *variable* citation time window, starting with the year of publication of an article up until 2005.

A methodological issue of interest is whether citations (given in *WoS* articles) to ArXiv versions of journal articles should be taken into account. One could argue that it is "unfair" to include these citations in a comparative analysis of journal papers deposited in ArXiv and articles not deposited in this archive because the latter do not have earlier versions that can be cited. Conversely, a base assumption underlying the analysis is that the two versions are different representations of the same paper. Therefore, it was decided to consider and count citations to ArXiv versions as well. When *fixed* citation windows were applied and the time interval between the date of a citation to a paper and the paper's publication date was determined, the publication date of an ArXiv paper receiving a citation was defined as its deposit date in ArXiv.

Results

Overall Results

Figure 1 presents ArXiv CIDs for the collection of all 24 condensed matter physics journals included in the study, applying a *variable* citation time window.

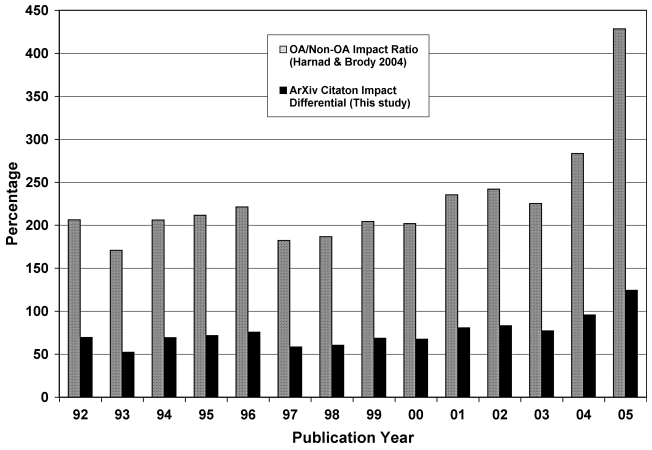
For the publication years between 1992 and 2003, the CID fluctuated between 50 and 75%. In the 2004 and 2005, it increased substantially and reached values of 96 and 124%, respectively. Figure 1 also shows the values of the "OA/non-OA IR" as defined by Harnad and Brody (2004). For publication years between 1992 and 2003, this ratio

[1]In interpreting the differences between ArXiv CID and "OA/non-OA IRs," one should first keep in mind that if there is *no* difference in average citation impact between papers deposited in ArXiv and nondeposited papers, ArXiv CID obtains a value of zero whereas Harnad and Brody's (2004) OA/non-OA IR is 100%. Second, rather than dividing the average citation impact of papers deposited in ArXiv to that of articles not deposited in that ArXiv, CID values are obtained by calculating in the denominator the mean value of the average citation impact of the two types of papers [$(CPP_{na} + CPP_a)/2$]. The table gives the IR and CID values for some of the ratio CPP_a/CPP_{na} values:

CPP_a/CPP_{na}	IR (Harnad & Brody, 2004)	CID (this article)
0.1	10%	−164%
0.2	20%	−133%
0.5	50%	−66%
1.0	100%	0%
2.0	200%	66%
5.0	500%	133%
10.0	1000%	164%

[2]Note that papers published in the beginning of a year may be followed during a longer time period than papers published in the end of that year; however, in aggregating papers from a particular year and calculating average citation rates of the aggregate during a fixed time window, such differences among individual papers tend to cancel out. Assuming that the publication date of papers published in a year is uniformly distributed across months, the average time period for citation is actually 2.5 years for both fixed windows.

DOI: 10.1002/asi



ArXiv CID values on the vertical axis are percentages, not absolute numbers. Citations to ArXiv versions of papers are included in the counts. In this figure, however, the horizontal axis gives the publication year of the journal articles and *not* that of their corresponding ArXiv versions (if there is one). The publication year (i.e., deposit date) of the ArXiv versions is on average about half a year earlier. A variable citation window is applied. This means that, for instance, for papers published in 1992, citations are counted during a 14-year time period (1992–2005) whereas for papers published in 2005, only citations are counted that were received in 2005. In fact, the citation per publication ratio for journal papers deposited in ArXiv gradually decreased from 28.6 for papers published in 1992 to 0.79 for those published in 2005. For journal papers not deposited in ArXiv, these ratios are 13.9 and 0.18, respectively.

FIG. 1. ArXiv CID values for 24 journals in condensed matter physics using a variable citation time window.

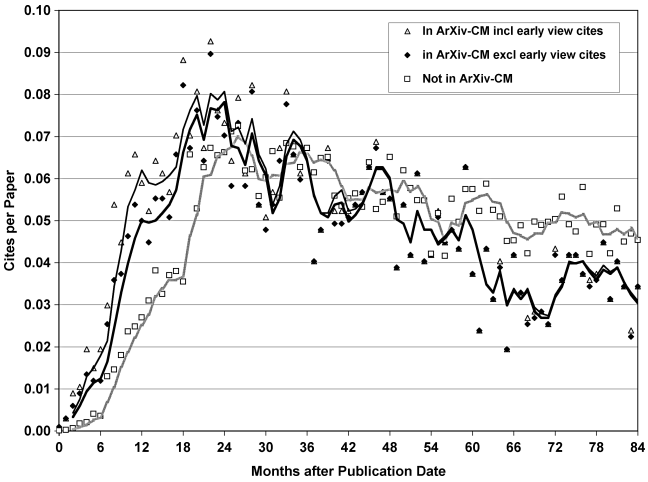
fluctuated between 170 and 225, and reached values of 283 and 428 for papers published in 2004 and 2005, respectively.

Early View Effect

The differences in ArXiv CIDs among publication years in Figure 1 are largely caused by the fact that when a variable citation time window is applied, the time period over which impact differentials are calculated shortens as the publication year becomes more recent. It is hypothesized that this pattern is mainly due to an early view effect. Figures 2 and 3 further corroborate this hypothesis. They show for articles published during 1996–1999 the number of citations per article on a *monthly* basis during the first 7 years after publication date, and compare the age distribution of citations for articles deposited in ArXiv-CM to that of papers *not* deposited in that archive.

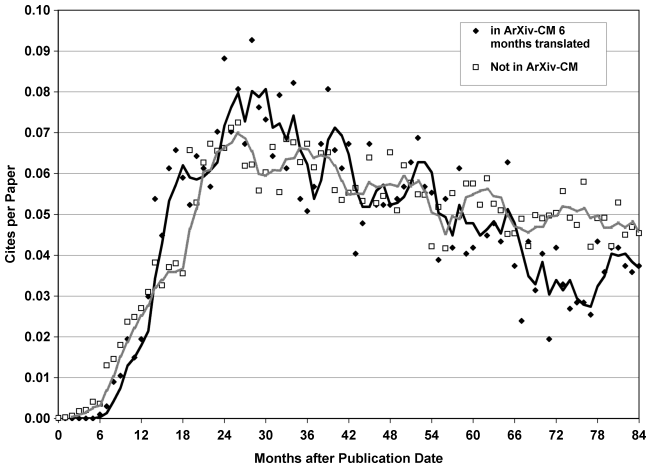
Age distributions of citations to some extent depend on the number of citations received: High-impact papers and poorly cited papers may show different aging patterns. Therefore, papers were categorized into classes on the basis of their citation frequency—1–2, 3–6, 7–18, and >18 citations—and age distributions were calculated per class. Figure 2 presents the results for articles cited between three and six times. Outcomes for the other citation classes were similar.

Note that the median time interval between the date a paper was deposited in ArXiv and the date it was published in a *WoS* journal is about 6 months. In Figure 3, the curve for ArXiv



Data relate to articles cited between three and six times. The curves represent 3 months' moving averages. Publication date of citing and cited journal articles was measured in this analysis as the date an article was included in the *WoS*. For cites to ArXiv versions of papers, the publication date of the cited article is its ArXiv deposit date.

FIG. 2. Age distribution of citations to papers deposited in ArXiv-CM and to nondeposited papers.



Data relate to articles cited between three and six times. The curves represent 3 months' moving averages. Publication date of citing and cited journal articles was measured in this analysis as the date an article was included in the *WoS*. The curve for papers deposited in ArXiv including cites to ArXiv versions in Figure 2 was translated with 6 months to the right, and the values of the number of cites per paper during Months 1 to 5 were set to zero.

FIG. 3. Age distributions of citations with the curve for citations to ArXiv-CM papers translated by 6 months.

deposited papers shown in Figure 2 is translated with 6 months along the time axis in a positive direction. During the first 24 months, this new curve roughly coincides with that for nondeposited papers. Around Month 24, both curves reach a maximum, followed by a decline. For papers deposited in ArXiv, the maximum is slightly higher and the decline afterward more rapid than that for articles not deposited in ArXiv.

Figure 3 provides evidence that there is an early view effect at stake. It was therefore decided to apply *fixed* rather than *variable* citation time windows, and calculate the ArXiv CID

DOI: 10.1002/asi

during (a) the first *3* years after publication date (with the year of publication included) and (b) during the *4th to 6th* years after publication. The final row in Table 1 gives the CID for these two fixed citation time windows, calculated for the total collection of 24 journals included in the study. It shows that during the first 3 years after publication, the ArXiv CID is 80%, and during the 4th to 6th years, the ArXiv is 64%. The *absolute* decline is 16% (80−64), and the *relative* decline, calculated as [(80−64)/80], amounts to 20%.

Table 1 also presents the ArXiv CID for the six *WoS* journals with the largest number of papers deposited in ArXiv-CM. It reveals large differences in CID values among journals, but each journal shows the same pattern as does the total collection of 24 journals: a decline in the CID rate during the 4th to 6th years after publication, compared to that calculated for the first 3 years after publication date. The mean relative decline rate over these six journals is about 24%. *Physical Review B* shows the highest relative decline rate, and *Solid State Communications* shows the lowest.

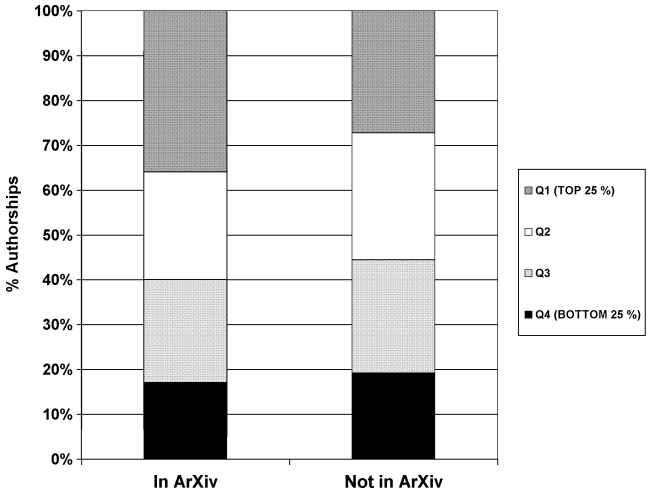


FIG. 4. Distribution for deposited versus nondeposited papers of authorships in *Physical Review B* among author citation impact quartiles.

Quality Bias

A first analysis addresses whether prominent authors are overrepresented in the bylines of papers deposited in ArXiv-CM. A methodological problem is how to measure author prominence *independently* of a possible ArXiv advantage effect. Therefore, it was decided to measure author prominence by calculating, on a journal-by-journal basis, the average number of citations received by an author's papers that were *not* deposited in ArXiv. Two indicators were calculated: one based on citations received during the first 3 years after publication date and a second one based on those received during the 4th to 6th years after publication. Per journal, and for each indicator, authors were categorized into four quartiles containing the top 25%, the 25% above the median but not in the top 25, the 25% below the median but not in the bottom 25%, and the bottom 25%, respectively.

The average number of authors per paper is about four. There is no one-to-one correspondence between an author and a paper. In this analysis, one paper contributes as many times to the counts as the number of authors it has in its byline, so that papers of large author teams have a greater weight than those of small teams. Since coauthorship often reflects collaboration between a senior researcher and one or more junior researchers, this bias can to some extent be reduced by taking into account only "senior" authors whose publication output exceeds a certain threshold. Therefore, the analysis was carried out per journal for all authors publishing at least one paper not deposited in ArXiv and for authors with at least 5 nondeposited articles.

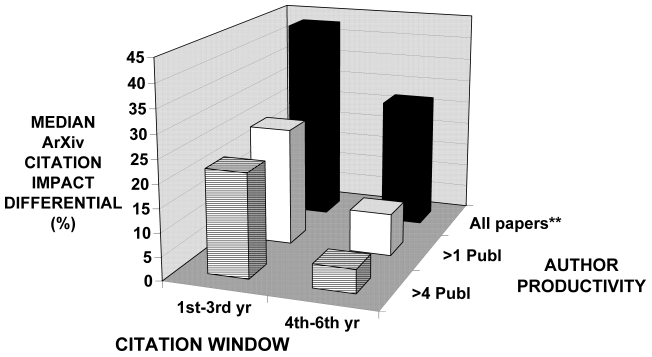
As an illustration, Figure 4 presents the outcome of this analysis for the journal *Physical Review B*, and for authors publishing in that journal at least one paper that was not deposited in ArXiv. It shows that authors in the highest citation impact quartile ("top" authors) account for 36% of authorships in the set of papers deposited in ArXiv-CM, and for 27% in the set of nondeposited papers. The fact that both percentages are above 25% reflects that top authors tend to publish more papers than do less prominent authors. For authors in the bottom 25% of the citation impact distribution, the percentage of authorships in deposited and nondeposited papers are 17 and 19, respectively. Authors in this quartile tend to publish less papers than do more prominent authors. All journals listed in Table 1 show for both author productivity levels the same pattern: Top authors in terms of average citation impact per nondeposited paper are overrepresented in ArXiv.[3]

To correct for this phenomenon, ArXiv CIDs were calculated at the level of an *individual* author publishing in a journal, and the *median* ArXiv CID was determined over publishing authors. This can be done for authors publishing at least one paper deposited in ArXiv-CM and at least one paper not deposited in that archive. Figure 5 gives the results for *Physical Review B*.

Figure 5 reveals the following patterns. First, CID rates calculated for the 4th to 6th years after publication date tend to be lower than those based on the 1st to 3rd years after publication. This phenomenon was observed earlier in this article, where it was ascribed to an early view effect. Second, compared to the CID rate for the journal as a whole, the median CID values over authors are lower, and they decrease with increasing author productivity. For authors with more than four publications, applying the 4th- to 6th-year citation window, the median CID is 5%. In this case, there is hardly any "impact advantage" of papers deposited in ArXiv.

[3]Authors publishing papers that were deposited only in ArXiv are not included in this analysis. The percentage of these authors strongly depends on the publication productivity threshold applied. Considering all authors with at least one paper in a journal, it was 23% for *European Physical Journal B*, 6% for *Physical Review B*, and less than 4% for the other four journals listed in Table 1. For authors with greater than four publications, it was 18% for *European Physical Journal B*, 1.2% for *Physical Review B*, and almost zero for the other five journals.

DOI: 10.1002/asi



**The percentages for all papers are not median values over authors, but the overall CID for the journal as a whole, presented in Table 1.

FIG. 5. ArXiv CIDs over authors publishing in *Physical Review B*.

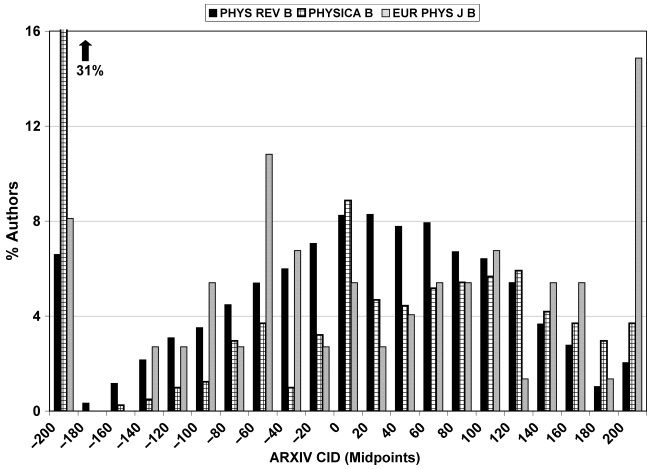


FIG. 6. Distribution of median CID values among authors for three journals.

Table 2 presents the outcomes for each of the six journals with the largest number of papers deposited in ArXiv-CM. The table shows a large variability among journals. Note that the number of authors over which median CID values are calculated varies substantially among journals, citation windows, and author-productivity thresholds, and it is rather low in several cases. But the general pattern is similar to that of *Physical Review B*: Calculating median values over authors leads in most cases to a substantial reduction of the ArXiv CID compared to the overall rates presented in Table 1, and the more so if citations are counted during the 4th to 6th years after publication.

Comparing the median CID values over productive authors (i.e., authors with more than four papers) to the overall values for a journal as a whole, the average relative reduction rate over the six journals in Table 2 is 56% if citations are counted during the first 3 years after publication date, and 60% if they are counted during the 4th to 6th years. *European Physical Journal B* obtained the lowest reduction rate (23%), and *Physica B* the highest (93%).

Figure 6 shows for the latter two journals and for *Physical Review B* the distribution of median CID among authors. The distribution of the first journal is almost symmetrical around the class with midpoint 0. *Physica B* has a large share of authors' papers deposited in ArXiv with a zero average impact while their papers not deposited in that archive have impact above zero. These authors are included in the class with midpoint −200. For *European Physical Journal B*, it is the opposite: There are relatively many authors for which the impact of their papers deposited in ArXiv is above zero while that of their nondeposited articles is zero. These are included in the class with midpoint +200.

CID values for all authors with at least one paper deposited in ArXiv and one nondeposited paper are generally higher than those calculated for more productive authors (i.e., publishing more than four papers). To explain these differences, note that a substantial fraction of papers with at least one less productive author was coauthored by at least one

TABLE 2. Median ArXiv Citation Impact Differential (CID) over authors.

	1–3 years after publication date				4–6 years after publication date			
	>1 publication		>4 publications		>1 publication		>4 publications	
Journal	*n*	CID	*n*	CID	*n*	CID	*n*	CID
Physical Review B	7,741	**25%**	5,158	**22%**	4,424	**9%**	2,813	**5%**
European Physical Journal B	394	**67%**	74	29%	163	**57%**	18	68%
Journal of Physics–Condensed Matter	874	**40%**	332	**37%**	441	**29%**	146	29%
Physica B–Condensed Matter	766	**6%**	406	11%	465	**2%**	213	0%
Solid State Communications	445	**40%**	153	29%	230	**33%**	72	13%
Intl. Journal of Modern Physics B	259	**38%**	54	**67%**	121	**40%**	14	−14%
All 24 journals	11,937	**29%**	6,747	**24%**	6,495	**14%**	3,528	**7%**

Values in the cells are median values of the ArXiv CID over authors publishing in a journal. Values in bold and italics are significantly different from 0 at $p = .01$ according to the Sign Test. The second row indicates the publication thresholds for author productivity (>1 or >4 publications). The last row gives outcomes for all 24 journals aggregated. These median values are calculated on a journal-by-journal basis. In other words, the CID was calculated for each author's publication oeuvre in a particular journal, and the median was calculated over all author–journal pairs in which the number of papers exceeded the publication threshold.

DOI: 10.1002/asi

productive author. For instance, for *Physical Review B*, 11% had only productive authors while 51% had only less productive authors, and 38% both a productive and a less productive author. The fraction of papers of less productive authors coauthored by a productive one is therefore 43%. A first relevant finding is that this fraction for papers deposited in ArXiv is much higher than that for nondeposited articles (39 vs. 82%, respectively). The other five journals studied in detail in this article showed a similar pattern.

A second finding was that the average citation impact of papers with only productive authors tends to be higher than that of articles with both productive and less productive authors, which in turn is higher than the impact of papers authored only by less productive authors. This pattern was in most cases found for all six journals, for both citation windows, and for both papers deposited in ArXiv and nondeposited papers. For instance, for papers published in *Physical Review B*, counting citations during the 4th to 6th years after publication, these three IRs amount to 15.1, 13.6, and 7.8, respectively, for papers deposited in ArXiv, and 13.1, 12.3, and 8.7, respectively, for articles not deposited in that archive.

Discussion and Conclusions

The analysis of ArXiv CID for the collection of all 24 condensed matter physics journals presented in Figure 1 showed that the OA/non-OA IRs, as defined by Harnad and Brody (2004), fluctuated for papers between 1992 and 2003 around a level of 200% and increased to 450 for papers published in 2005. These outcomes are on the same order of magnitude as those given by these two authors for "All Physics Fields" and for "Nuclear and Particle Physics," even though it is uncertain whether they included citations to ArXiv versions in their counts. Differences between the outcomes of the two studies may reflect differences among research fields.

The observation that the ArXiv CID calculated for a set of papers varies with the age of those papers is crucial. The differences between the citation-age distributions of deposited and nondeposited ArXiv-CM papers presented in Figure 2 can to a large extent—though not fully—be explained by the publication delay of about 6 months of nondeposited articles compared to papers deposited in ArXiv. This outcome provides evidence for an *early view effect* upon citation impact rates, and consequently upon ArXiv CIDs. The early view effect is caused by the fact that colleagues in the field start the process of reading a paper, processing its information, and citing it in their own articles *earlier* if a paper is deposited in ArXiv because of its earlier availability.

The early view effect explains why CID values for recent years (i.e., 2004 and 2005) are so much higher than those for earlier years. The observation that CIDs of journals calculated during the 4th to 6th years after publication are on average about 20% lower than those calculated for the first 3 years after publication date also should be attributed to an early view effect. The outcomes illustrate that a citation time window of 18 months, as applied in a recent article by Eysenbach (2006), may not be sufficiently long to adequately capture how "OA versus non-OA IRs" vary with the age of cited articles.

Figure 2 in the article by Kurtz et al. (2005, p. 1399) shows an increase in the very short term average citation impact (Citations were counted during the first 5 months after publication date.) of astronomy papers as a function of their publication date published after 1995. This reflects an increase in the share of astronomy papers deposited in ArXiv (and other preprint servers) over time. More and more papers had become available at the date of their submission to a journal rather than on the formal publication date. Therefore, their findings for astronomy are fully consistent with the outcomes presented in this article for journals in condensed matter physics.

Figure 4 of this article provides evidence that prominent authors—measured per journal by the average citation impact of their papers *not* deposited in ArXiv—are statistically overrepresented in the bylines of papers deposited in ArXiv. Therefore, it is appropriate to calculate CID rates at the level of individual authors even though such an analysis is to some extent hampered by the fact that the numbers of authors in the analysis on a journal-by-journal basis are in a number of cases rather low, especially if the publication productivity threshold is set to 4.

For the six journals presented in Table 1, the calculation of *median* ArXiv CIDs *over authors* leads on average to a reduction in the CID with 56 to 60% compared to the overall CID for a journal. This outcome suggests a strong quality bias in ArXiv CIDs or "OA versus non-OA IRs." This conclusion is in agreement with that in Kurtz et al. (2005); however, the evidence provided in this paper is based on an analysis at the level of individual authors, and therefore more direct and much stronger than that provided by Kurtz et al.

Considering more productive authors, and calculating citation impact during the 4th to 6th years after publication date, the median CID rates over authors in the six journals do not significantly differ from zero, except in the case of the 5% rate for *Physical Review B* authors. For two journals, the CID is zero or even negative, for two other journals it is between 13 and 29%, and for still another journal it is 68. Note that the two extreme values (i.e., −14% for *International Journal of Modern Physics B* and +68% for *European Physical Journal B*) are based on low numbers of authors (18 and 14, respectively).

Median CID values for *all* authors publishing at least one paper deposited in ArXiv and one nondeposited paper were found to be higher than those for productive authors, and are for all six journals significantly positive. This outcome can be explained by the finding that (a) if one selects from a set of papers deposited in ArXiv a paper authored by a junior (or less productive) scientist, the probability that this paper is coauthored by a senior (or more productive) author is higher than it is for a paper authored by a junior scientist but not deposited in ArXiv; and (b) papers coauthored by both productive *and* less productive authors—regardless of whether these papers were deposited in ArXiv—tend to have a higher citation impact than do articles authored solely by less productive authors. Since senior authors tend to be more

DOI: 10.1002/asi

productive in terms of numbers of published papers and tend to generate per paper a higher citation impact than do junior authors, the outcome can be interpreted as a *quality bias*: More productive, influential senior authors are overrepresented in the bylines of the papers deposited in ArXiv and authored by junior (or, in any case, less productive) researchers.

Controlling for quality bias and the early view effect, the conclusion is that in the sample of six journals analyzed in detail in this study, there is no sign of a general "open access advantage" of papers deposited in ArXiv-CM. In the citation analysis by Kurtz et al. (2005), both the citation and target universe contain a set of seven core journals in astronomy. They explained their finding of no apparent OA effect in their study of these journals by postulating that "essentially all astronomers have access to the core journals through existing channels" (p. 1401). In the study presented in this article, the target set consists of a limited number of core journals in condensed matter physics, but the citation universe is as large as the total *WoS* database, including a number of more peripherical journals in the field. Therefore, the result in this article is stronger than that obtained by Kurtz et al.: Even in this much wider citation universe, no evidence was found for an OA advantage effect.

The empirical findings presented in this article do provide evidence that ArXiv *accelerates* citation. This is actually a primary function of a preprint archive, and the outcomes reveal that ArXiv was successful in carrying out this function. Accelerating communication is definitely a positive effect of ArXiv. But the findings presented in this study suggest that this acceleration is due to the fact that ArXiv makes papers available *earlier* rather than makes papers *freely* available.

It would be illuminating to further analyze early view effects in other publication environments. For instance, in electronic journal collections of publishers adopting a subscription-based business model, journal issues may be available electronically several months before their formal publication. The time interval between electronic and formal publication date varies across journals and changes over time. Selecting appropriate groups of journals and control groups, the early view effect could be further studied and quantified.

The effect of delays in the publication process and their effect on age distributions of citations is an important topic of bibliometric and informetric research (e.g., Egghe & Rousseau, 2000). A further analysis and modeling of such phenomena, particularly in relation to preprint archiving, falls beyond the scope of this article and awaits further research. This research also should include citations within ArXiv (i.e., citations from one ArXiv paper to another). These data were not available for the current study.

The analysis of quality bias focused on the extent to which prominent authors are overrepresented in the bylines of papers deposited in ArXiv. A second dimension, the extent to which authors—be it prominent or less so—tend to deposit their "better" papers in ArXiv had not been examined and awaits further research. This can be done only by using measures of author prominence that are not based on citation impact.

Acknowledgments

The main lines of the work described in this article were presented at the 1st International Conference of the Association of Publishers in Europe on April 4–5, 2006 in Berlin, Germany, and at the Open Scholarship 2006 Conference on October 18–20, 2006 in Glasgow, United Kingdom. The author is grateful to numerous participants at these conferences for their comments, to two anonymous referees, and to his CWTS colleague Martijn Visser for technical assistance and stimulating discussions. The research described here was partly funded by Elsevier within the framework of a research project entitled "Reference behavior of scientific authors," carried out at CWTS. Several techniques applied in this article were developed in a research project funded by the Netherlands Organisation for Scientific Research on the development of bibliometric indicators for the assessment of research performance in the field of computer science. The author of this article emphasizes that both granting organizations respect the academic and entirely independent position of the researchers involved in these projects.

References

Davis, P.M., & Fromerth, J. (2006). Does the ArXiv lead to higher citations and reduced publisher downloads for mathematics articles? Scientometrics, 71(2), 2007. Available: http://arxiv.org/abs/cs.DL/0603056

Egghe, L., & Rousseau, R. (2000). The influence of publication delays on the observed aging distribution of scientific literature. Journal of the American Society for Information Science, 51, 158–165.

Eysenbach, G. (2006). Citation advantage of open access articles. PLOS Biology, 4, 692–698.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. Science, 178, 471–479.

Harnad, S., & Brody, T. (2004). Comparing the impact of open access (OA) vs. non-OA articles in the same journals. D-Lib Magazine, 10(6).

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

Moed, H.F. (2005). Citation analysis in research evaluation. Dordrecht, The Netherlands: Springer.

DOI: 10.1002/asi