ACCESS, READERSHIP, CITATIONS: A RANDOMIZED CONTROLLED TRIAL OF SCIENTIFIC JOURNAL PUBLISHING

by Philip Meir Davis

ACCESS, READERSHIP, CITATIONS: A RANDOMIZED CONTROLLED TRIAL OF SCIENTIFIC JOURNAL PUBLISHING

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

In Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Philip Meir Davis

August 2010

ACCESS, READERSHIP, CITATIONS: A RANDOMIZED CONTROLLED TRIAL OF OPEN ACCESS PUBLISHING IN SCIENTIFIC JOURNALS

Philip Meir Davis, Ph. D.

Cornell University 2010

This dissertation explores the relationship of Open Access publishing with subsequent readership and citations. It reports the findings of a randomized controlled trial involving 36 academic journals produced by seven publishers in the sciences, social sciences and humanities.

Between January, 2007 and February, 2008, 712 articles were randomly assigned free access status upon publication from the publisher's websites (the treatment), leaving 2,533 control articles that were accessible by subscription (the control). Article usage data was gathered from the publishers' websites and article citations were gathered from ISI's Web of Knowledge ™. At the time of this writing, all articles have aged at least two years.

Articles receiving the Open Access treatment received significantly more readership (as measured by article downloads) and reached a broader audience (as measured by unique visitors), yet were cited no more frequently, nor earlier, than subscription-access control articles.

A pronounced increase in article downloads with no commensurate increase in citations to Open Access treatment articles may be explained through social stratification, a process which concentrates scientific authors at elite, resource-rich institutions with excellent access to the scientific literature. For this community,

access is essentially a non-issue. The real beneficiaries of Open Access are the communities that consume, but do not contribute to, the scientific literature.

The focus on information consumers requires us to advance the theory of the attention economy. The linear transmission model, where information flows from the sender to the receiver is rejected for a two-sided market model, with authors on one side, readers on the other and journals fulfilling the role of the intermediary agent. The primary purpose of the journal-agent is to transmit quality signals to potential readers. I argue that this model is able to explain both author and reader behaviors as well as the persistent role of journals in an information environment that decouples certification from dissemination.

BIOGRAPHICAL SKETCH

Philip Davis was born in Toronto, Canada in 1968. He received his Bachelors of Science (Honors in Biology) in 1992 from the University of Guelph, and his Masters of Library and Information Science in 1994 from the University of Western Ontario. From 1995 through 2006, he worked for the Cornell University Library, first as a reference and instruction librarian and then as a collections librarian. In 2006, he left the library to pursue his doctorate in science communication at Cornell University.

To Suzanne, Anna and Isabelle

ACKNOWLEDGMENTS

While I am the sole author of this dissertation, I cannot claim sole credit for the work described therein. This study reflects a long-term collaboration with over two dozen individuals across the United States, without whom, this work would not have been possible. First, I would like to thank Mike Keller and Andrew Herkovic (Stanford University Library) for involving me in initial talks that would ultimately culminate in the creation of the research proposal. Next, I would like to thank my advisor, Bruce Lewenstein (Cornell University) for helping to write and submit a research proposal and funding request to support the research, and ultimately to Donald Waters (Andrew W. Mellon Foundation) for supporting this study financially. The study would not have been possible without the participation of seven society and commercial publishers: Marty Frank, Margaret Reich, Mark Goodwin and Mike Gentry (American Physiological Society); Beth Rosner, Monica Bradford, Emilie David, Betsy Harman and Stewart Wills (American Association for the Advancement of Science) and Don Kennedy (Stanford University); Heather Goodell (American Heart Association); Donna Blagdan and Rob Dilworth (Duke University Press); Jennifer Pesanelli (Federation of American Societies for Experimental Biology); Tracey DePellegrin Connelly (Genetics Society of America); Bob Howard, John Shaw and Peter Binfield (Sage Publications). Connecting me to these individuals would not have been possible without a central maven, John Sack (HighWire Press). I would also like to thank Michael Puff, Fran Steck and Elisabeth Ten Brink (HighWire Press) for their technical support regarding journal-level usage statistics, to Matthew Connolly (Cornell University Library) for developing software that allowed for the harvesting of these statistics, and to James Booth (Department of Biological Statistics and Computation Biology) for his help in statistical analysis. I would also like to

thank my special committee, Bruce Lewenstein, Tarleton Gillespie, Jeremy Birnholtz (Department of Communication), and Dan Simon (Department of Applied Economics and Management) for their mentoring throughout graduate school. Finally, I must show my greatest gratitude to my family who supported me through four difficult, yet rewarding, years of my life.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

AAAS	American Association for the Advancement of Science
AHA	American Heart Association
APS	American Physiological Society
CME	Continuing Medical Education
FASEB	Federation of American Societies for Experimental Biology
FRPAA	Federal Research Public Access Act
GSA	Genetics Society of America
HTML	Hypertext Markup Language
IP	Internet Protocol
ISI	Institute for Scientific Information
MLR	Multi-linear Regression
NBR	Negative Binomial Regression
NIH	National Institutes of Health
OA	Open Access
OR	Odds Ratio
PDF	Portable Document Format
PEER	Publishing and the Ecology of European Research
PPV	Pay per View
RCT	Randomized Controlled Trial
STM	Scientific, Technical and Medical
URL	Uniform Resource Locator
VIF	Variance Inflation Factor
WoS	Web of Science

INTRODUCTION

The marriage of digital publishing and the Internet has created both opportunities and challenges for science publishers. It has created economic forces that both favor the incumbents and encourage new entrants; it has allowed for alternative publishing business models; put pressure on the traditional relationships between authors, funders, publishers, and libraries; and created new legal responsibilities for those funding science with public monies. This marriage has been so profound because it has affected science publishing simultaneously from economic, organizational, legal, and social dimensions. At the same time, scientific publishing reflects the intensely conservative nature of academic values and practices (Harley, Acord, Earl-Novell, Lawrence, & King, 2010; Merton, 1973; Polanyi, 1962). Scientists are still driven chiefly by their desire to obtain public recognition from their colleagues through the published record (Hagstrom, 1965).

Since the launch of the *Philosophical Transactions of the Royal Society* in 1665, the modern journal has fulfilled four principle functions in science (Zuckerman & Merton, 1971):

1) *Registration*: the process of date-stamping received manuscripts, thereby establishing the priority of new discoveries and resolving disputes in cases of simultaneous publication (Merton, 1957).
2) *Certification*: conveying validity to a truth claim through editorial and peer-review.
3) *Dissemination*: distributing publicly the results of scientific discovery, and
4) *Archiving*: the preservation of the scientific record

A digital, networked publishing landscape allows these four functions to become decoupled from the printed journal and provides the opportunity for new services to be developed around one or more function (Crow, 2002). For example, the e-print repository *arXiv.org* fulfills the functions of registration, dissemination, and (as much as digital services can) archiving, but does little to offer certification services[1]. *Faculty of 1000* [2], a post-publication review service, purports to engage over 5,000 scientists worldwide to review new article publications, thus providing a secondary layer of certification that follows pre-publication editorial and peer-review.

Networked technologies have also reduced the cost of broadly disseminating scholarly information, thus lowering one of the barriers to entry that existed in the print world, and allowing new players into the digital publishing marketplace. At the same time, shifting the cost from *distributing* to *producing* information has favored economies of scale, permitting the growth of a few exceedingly large publishers through mergers and acquisitions (McCabe, 1998, 1999, 2002). The de-emphasis of distribution costs has allowed for new publishing business models that focus on the production of scientific information rather than its distribution. One of these new business inventions is the author-pays (or producer-pays) Open Access publication model, a model that funds the cost of publication on author (or funder) payments in lieu of reader revenue. While variations on this model abound (Willinsky, 2003), they all provide access to the information free of charge and can be categorized as some form of what is called "Open Access."

---

[1] arXiv submissions are checked briefly by a reviewer to ensure that they have some semblance to academic literature and are properly designated into at least one prescribed arXiv subject category. There is no vetting with regard to the accuracy and validity of the document content.

[2] http://f1000.com/

Open Access advocates perceive a fundamental problem in the ability of the incumbent subscription-access model to adequately accomplish the dissemination function of scientific publishing. The evidence for this claim is based on early research purporting to claim a large and positive citation advantage for freely-accessible (Open Access) scientific articles. These claims, however, are based upon weak methodology.

In this dissertation, we test the robustness of the Open Access citation claim under natural experimental conditions, using a much stronger methodology. Our study refutes the existence of a citation advantage to Open Access articles: within two years after publication, freely-accessible scientific articles were cited no more frequently, nor earlier, than subscription-access control articles. Open Access treatment articles did receive increased readership (as measured by article downloads) by a larger audience (as measured by unique visitors) suggesting that the real beneficiaries of Open Access are the communities that consume – but do not contribute to – the scientific literature. For the research community, access to the scientific literature (the "dissemination" function of publishing) is essentially a non-issue.

This dissertation begins with an introduction to the historical and political nature of the Open Access debate, and is followed by a literature review of studies of access to the scientific literature. The bulk of this dissertation focuses on the design, execution, and reporting of a large randomized controlled trial of Open Access publishing across seven publishers. In the discussion section, I argue that focusing on the dissemination function of scientific publishing is inadequate for explaining the current information market. In its place, I advance the theory of the *attention economy* by proposing a two-sided model with authors on one side, readers on the other, and the

journal functioning as the intermediary agent between the two communities. I argue that this model is able to explain both author and reader behaviors as well as the persistent role of journals in an information environment that decouples certification from dissemination.

Defining "Open Access"

The phrase "open access" has a long historical record. Its general meaning of unrestricted admission or access is documented by the Oxford English Dictionary as far back as 1602 (Oxford English Dictionary). In library science, the phrase can be traced to 1894 with reference to patrons' unrestricted access to the publications kept on library shelves. The current meaning of open access in the library and publishing communities is based on the widely cited declaration of the Budapest Initiative:

> By "open access" to this literature, we mean its free availability on the public internet, permitting any users to read, download, copy, distribute, print, search, or link to the full texts of these articles, crawl them for indexing, pass them as data to software, or use them for any other lawful purpose, without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. (Open Society Institute, 2001)

While unrestricted access forms the basis of the characterization of what is defined as "open access," subsequent attempts to narrow the definition have focused on how the information can be used. The Creative Commons, for example, has developed six specific licenses based upon whether a work can be used for commercial purposes or modified in any way (Creative Commons, 2001). Similar to the notion of giving credit to scientific authors for making their work public (Biagioli, 2003; Kaplan, 1965), all Creative Common licenses specify a public acknowledgement of the author when reusing a work.

In this study, "open access" will be defined in its broadest sense and equated exclusively to free online access of scientific articles, that is, access to an article that does not depend on a personal or institutional subscription, nor requires any form of monetary payment from the reader. Because the phrase "open access" is often used ambiguously and contains political and ideological baggage (Davis, 2009b), I will often defer to the phrases "free access" or "freely-accessible" in order to avoid confusion.

History of the Open Access Debate

The Open Access debate has changed over the years, taking on new terms and adopting different rationales. Its predecessor, *the serials crisis*, is an old concept concerned with the affordability of scientific journals.

Understanding the limited ability of college libraries to afford the entirety of chemistry literature before the Great Depression, Gross and Gross (1927) devised a method of creating a core list of chemistry journals by tabulating the list of citations from a volume of the *Journal of the American Chemical Society*. After World War II, indexing efforts to define core lists of journals became an issue again as Britain worked to rebuild its science libraries (Bradford, 1948). The inability of research libraries to collect comprehensively the world's literature called upon science as a way to prioritize what should be purchased. By analyzing article keywords, Samuel Bradford was able to derive a model of concentric "zones of decreasing productivity" of journals publishing papers on a particular topic (Bradford, 1948, pp. 110-111). Librarians could therefore begin with purchasing what was in the "nucleus" of these zones and move outward as their budgets permitted. It was never assumed that universal access to the corpus of scientific knowledge was possible or attainable. The

underlying assumption was that limited resources resulted in limited access. It was therefore incumbent upon the librarian to make decisions that best reflected the needs of their constituencies.

When the economies of Europe and North America rebounded, perspectives changed. The serials crisis became less of a matter of access and affordability and more an issue of overproduction. By the 1960s, two psychologists, William Garvey and Belver Griffith, began describing the "scientific information crisis" as a problem for scientists coping with the increasing publication of scientific articles (Garvey & Griffith, 1967, p. 1011). The solution to this crisis was to first understand scientific communication within a discipline before developing tools for improving the organization and retrieval of the literature. Their solution to the information crisis was not at the production end, but at the receiver end. New tools were needed by the researcher to deal with information overload. They write:

> It would appear, then, that the major information problem facing psychology is not so much that psychology is producing more information than its total manpower can assimilate, but rather that the individual scientist is being overloaded with scientific information. Perhaps the alarm over an "information crisis" arose because sometime in the last information doubling period, the individual psychologist became overburdened and could no longer keep up with and assimilate all the information being produced that was related to his primary specialty. (Garvey & Griffith, 1971, p. 350)

Understanding that publications and citations were concentrated among a small core of the scientific literature, Eugene Garfield extended Bradford's notion of dispersion to citation networks (Garfield, 1972). Considering that the vast majority of citations point to only a small core group of journals, scientists could use citation frequencies to help inform them what to read, and librarians could use citation data to assist with subscription decisions.

The historian of science Derek de Solla Price used networks of scientific papers to understand the nature of science communication. A very small core group of journals, he concluded, were active in what he described as the "active research front," (D.J.S Price, 1965, p. 512), publishing articles that advanced a field. Papers published in these journals continued to be cited over time while the vast majority of published articles were subsequently ignored in the citation record. In developing the notion of *obsolescence* in the literature, library scientist, B. C. Brookes (1971) developed a method for calculating an optimal sized library based on the life expectancy of journals. What is notable again is the argument that libraries (and librarians) were considered both part of the problem and the solution to the information crisis. Brookes writes:

> The "information explosion" becomes less alarming if it is appreciated that the "explosion" of new literature is almost exactly counterbalanced by the discarding of the old. But the librarian is conservative; he has yet to learn how to be as ruthless as his scientific colleagues in the art of discarding the obsolete. (Brookes, 1971, p. 461)

While librarians have always been practicing the art of collecting content, the development of the journal bundle, coined "The Big Deal" by Ken Frazier (2001), changed the relationships between librarians, publishers and readers. By the late 1990s, most large scientific publishers were producing electronic versions of their journals. Using a strategy similar to bundling of cable television programs, publishers began bundling electronic access to their entire journal title list. At first, this new business model presented itself as a win-win-win scenario: Libraries could subscribe to many more titles at prices only marginally higher than what they were currently paying for title-by-title access; researchers (especially at smaller institutions) would have access to collections unprecedented in most world-class libraries; and publishers

could realize a slight increase in revenue while increasing access to many more titles in their portfolio (Sanville, 1999). A stable relationship between these three players, however, is ultimately based on adequate financial resources. In difficult budget years, librarians expressed frustrations in their declining ability to cancel marginal journals in their collections. Big Deals were consuming larger percentages of their budgets and the remaining journals available for cancellation were often the important low-priced journals offered by non-profit societies and associations (Knight, 2003). Moreover, with many journal titles now being packaged and sold in publisher bundles, there was little need for so many collection experts in academic libraries. Collection decisions were increasingly being made by small teams of librarians as collection budgets became more centralized in libraries (L. L. Phillips & Williams, 2004). In effect, the Big Deal not only changed the power relationships between publishers, librarians and readers, it also had the effect of redistributing and centralizing power within the library.

The formation of library consortia helped solidify the centralization of power within the library. Many library consortia were formed as buying clubs for large journal packages. By representing a large group of potential buyers, libraries could negotiate preferential terms on prices, annual cost increases and stipulations such as long-term access (Hirshon et al., 1998). Yet, decisions on what major university libraries were subscribing were made almost exclusively by a small group of powerful individuals [3] and their decisions were long-lasting, often in the form of three to five-year contracts. With little option to return to a title-by-title purchasing model based on print subscriptions, librarians were left with little bargaining power.

Rising prices of scientific journals without commensurate increases in collection budgets put many academic libraries in the position of cancelling what was

[3] For example, the Associate University Librarian (AUL) for Collections

left outside of bundled packages and limiting the purchase of books and other non-journal materials. Historical graphs, such as the one published annually by the Association of Research Libraries (2004b) charted serial and monograph purchases since 1986. Until 2000, the narrative from these graphs was exceptionally clear: increasing total costs and rising unit costs results in a reduction in the number of serials and monographs purchased by the library.

Beginning in 2001, the narrative from these graphs became less clear. While total costs were still rising, unit costs underwent a sharp reduction and libraries were purchasing more journals than they had at any time in history – this was the effect of the Big Deal. By 2006, the effect of bundled deals on the ability of libraries to acquire more journals showed a persistent upward trend. Monograph purchases also reversed their downward trend and also returned to pre-Big Deal numbers (Association of Research Libraries, 2006).

After 2008, ARL changed the visual presentation of their data. No longer were total and unit costs of serials plotted – only total expenditures (Association of Research Libraries, 2009). Indeed, the ARL narrative shifted from a focus on declining access *as a result* of increasing costs, to an argument consisting entirely of increasing costs. The serials crisis has not gone away; it has merely changed scope, taken on new language and been pitched in new cognitive frames.

Changing Frameworks

By the mid 1990s, both publishers and librarians began using the much broader phrase "crisis in scholarly communication" to refer to issues beyond library affordability problems. Sandy Thatcher focused on the systemic problems confronted by university presses in publishing works in the humanities and argued that "it is the

entire system for distributing scholarship, not just one corner of it, that requires overhaul" (Thatcher, 1995, p. B2).

The Public Library of Science (PLoS) was formed in 2000 when a group of leading biologists, including the Nobel laureate Harold Varmus, circulated an open letter calling for biomedical journal publishers to make their articles freely available through PubMed Central within six months of initial publication (Public Library of Science, 2001). Their campaign addressed the needs of scientists but not the needs of the general public. A corpus of fully searchable, interlinked articles, they argued, would increase the utility of the scientific record, enhance productivity, and join disparate fields of knowledge within the biomedical sciences. In order to encourage publishers to change their strategy, scientists signing the letter pledged that they would only subscribe to, publish in, edit or review for, journals that abided by their stated conditions (Public Library of Science, 2001). Their letter attracted more than 30,000 signatories, yet few followed through on their pledge (Butler, 2003). Shortly thereafter, PLoS would change their approach and begin their role as a publisher of open access journals.

In 1997, the Association of Research Libraries launched an advocacy arm called "SPARC" (Scholarly Publishing and Resources Coalition) to work on various solutions to the crisis in scholarly communication. Taking a multi-faceted approach to a systemic problem, one of SPARC's first campaigns focused on *financial sustainability*. Monopoly pricing practices were believed to be one of the causes of serials price inflation. By introducing competition in the marketplace, SPARC hoped to put moderating forces on subscription prices. In the next few years, SPARC assisted society and other non-profit publishers to launch lower-priced journals and related projects that would compete with higher-priced alternatives.

Other SPARC campaigns appealed to *social responsibility*. Their 2002

campaign, called "Create Change," focused on educating college faculty on how their behaviors ultimately affected the college library (Association of Research Libraries, 2002). College faculty, they argued, should be sensitive to the prices of the journals in which they publish their work and provide editorial and peer-review services; those serving on journal boards should consider the effect of pricing on library budgets. Faculty authors should negotiate to retain their copyright, publish their work in an open access journal or deposit their work in an institutional repository. In addition, academic departments should consider electronic publication on par with traditional forms of publishing. ARL continued expanding their advocacy work into other frameworks of social responsibility. In 2004, SPARC launched their Open Access campaign, focusing on the "unprecedented public good" that results from the free access to scientific and scholarly journal articles (Association of Research Libraries, 2004c). The Open Access campaign continued to appeal to the social responsibility of the faculty author, only this time by highlighting *personal incentives*. Featuring the research of Steve Lawrence (2001) that online computer science conference papers were cited more frequently than their print-access counterparts (displayed under the heading "Open access *increases* research impact"), SPARC's Open Access campaign attempted to convince faculty that making one's work freely-accessible benefited both the author and the public. In their brochure, open access was described as being superior to subscription-access publishing in every way:

> Think about what this kind of distribution will mean for the enlargement of your audience, the widespread sharing of knowledge, and the acceleration of research. Open-access archives and journals are both practical and lawful. Implementations around the world are proving that they surpass traditional subscription-based journals in their cost-effectiveness and service to science and scholarship. (Association of Research Libraries, 2004c)

In 2004, ARL also launched the Alliance for Taxpayer Access (ATA), whose purpose was to use the legal system to push for universal access to the publications resulting from federally-funded research (Association of Research Libraries, 2004a). The ATA used a different approach from other social action campaigns. Using the deeply entrenched values of *transparency* and *accountability* in government affairs, ATA argued that taxpayers have a right to access the results of studies conducted with public monies. Instead of motivating scientists to change their behavior, ATA's approach has been to change the legal system, making public access a requirement for receiving federal research funds.

Several other prominent social responsibility arguments in favor of publicly accessible scientific information have been employed over the years by various agencies and individuals. The National Institutes of Health's Public Access Policy is based on the notion that increasing access to federally funded biomedical research "will speed discoveries, resulting in the prevention of death and disability" (Zerhouni, 2008), or, as stated on the NIH Public Access Policy website, will "help advance science and improve human health" (National Institutes of Health, 2009). The education researcher, John Willinsky, has made broader claims about the nature of scientific information as a public good, arguing that the open circulation of knowledge is beneficial to the "well-being of humanity" (Willinsky, 2006, p. 207), and that universal access to scientific information should be viewed as "a new civil rights issue" (Willinsky, 2009, p. 22). Heather Joseph, Executive Director of SPARC has framed access to scientific information as an issue of *social justice*, using the example that free medical literature on PubMed Central has helped her seek new treatments for her son (Joseph, 2008).

From this brief history of the open access movement in the United States, it appears that the argument for increased access to the scientific record has changed immensely over the years. Originally framed as an issue of financial sustainability for libraries (as the "serials crisis"), the access debate has vastly broadened, taking on new cognitive frameworks such as social responsibility, transparency and accountability.

Open Access is now much more than about economics or business models. It is about public good, fairness and social justice. It is about who controls the record of science, or alternatively, how to usurp that power. It is about personal empowerment and civil rights. While this makes "Open Access" ambiguous as a concept, it also makes it a powerful metaphor, allowing individuals to draw boundaries and craft messages depending on the purpose at hand (Gieryn, 1983). Ambiguity makes open access ultimately an issue of framing and language (Davis, 2009b).

LITERATURE REVIEW

This literature review will focus on Open Access solely as an issue of *access* – avoiding OA related topics such as economics, public good and social justice – and review what evidence there is to support a crisis of access to the scientific literature. It will begin by summarizing the literature on scientists' perceptions on access, followed by use of the scientific literature by the lay public. It will then review studies using unobtrusive methods for measuring access to the literature focusing on article downloads and citations.

Access Studies

Research on the accessibility of the scientific literature follows two main methodological approaches: the first is based on surveying researchers on their perceptions and desires of the journal publishing system; the second is based on unobtrusive studies of what scientists read and cite. Both approaches have their strengths and weaknesses.

Surveys can gather the responses of thousands of individuals and allow a researcher to generalize the results over a target population. In-depth interviews, while limited in their generalizability, can explore a topic in more detail and draw out values and motivations from a respondent. Poorly constructed questionnaires, however, can mislead respondents and result in biased results. Similarly, interviewees may be prompted to provide what researchers want to hear or what scientists ought to believe, leading to response bias (Spector, 2003). For example, since one of the central values of science is *openness* (Merton, 1942), scientists may be supportive of the phrase "open access" in spite of the ambiguity in how the term is used.

Furthermore, researchers have different – and often competing – interests when responding as *authors* or as *readers*. Authors want to publish more: readers want to read less (Mabe & Amin, 2002). This poses a problem for understanding the needs of researchers and makes the context of the study immensely important.

Unobtrusive measures (such as counting article downloads or measuring citations) are a more direct approach to measuring what scientists *actually do* and not what they say they do. We assume that these two measures are somewhat concordant. Unobtrusive studies, however, are unable to answer questions such as *why* an article was downloaded or cited. Clearly, both types of studies are required to develop a more complete picture of the state of access to the scientific literature.

In reviewing the literature on access (presented below in detail under two sections (*Survey Studies on Access* and *Article Download and Citation Studies on Access*), there is surprising consistency in the conclusions of these studies: *access to the published literature is improving, and those who generate knowledge view access issues as largely unimportant*.

The phrase "those who generate knowledge" cannot be overemphasized since there has been very little work on the dissemination of scientific information to those who use – but do not contribute to – the literature (i.e. teachers, medical practitioners, industrial researchers, and the lay public).

Survey Studies on Access

Since 1977, periodic surveys of the reading and information-seeking patterns of U.S.-based scientists have been performed allowing for longitudinal trends to be reported, e.g. (D. W. King & Tenopir, 1999; Tenopir & King, 2000, 2002, 2008; Tenopir et al., 2003; Tenopir, King, Edwards, & Wu, 2009). Over the previous three

decades, Tenopir et al. report, the average number of readings per scientist has been rising while the time spent finding and reading an article has been steadily decreasing (Tenopir et al., 2009). Their studies have also indicated that scientists are reading from a broader group of journals and extending their readership into the older literature – a trend that Tenopir and King attribute to the digitization of the journal literature and the creation of electronic archives (Tenopir et al., 2009). Scientists in the United States are relying primarily on institutional (library) access to journal collections although they do rely on informal sources (such as preprint serves or colleagues) for some of their literature needs (Tenopir et al., 2009). A recent survey of researchers in India illustrates the importance of informal sources of scientific literature in countries where institutional and library access is more restricted (Gaulé, 2009). In the previous three months, 84% of survey respondents reported either contacting an author or a colleague for a copy of an article when formal routes of access were unavailable.

A large, international survey of senior authors of scientific papers in 2005 revealed much about the values of researchers (Rowlands & Nicholas, 2005). In selecting journals to submit their work, factors such as reputation of the journal, readership, Impact Factor, and speed of publication were ranked as the top concerns of authors. Conversely, permission to post a copy of the article or retaining copyright were ranked last. At the time of this study, there seemed to be little knowledge of what Open Access meant – some authors claiming to have published in Open Access journals when in fact they had not. While the results of this survey reflected the views of over 5,000 authors, we should understand that the survey population consisted of a group of *corresponding authors* who were selected from the Institute for Scientific Information (ISI) author database. As a consequence, the results of this survey are biased toward senior authors who publish in higher impact journals. We should also

be aware that the response rate of the survey was just over 7% and may reflect a more motivated, and thus opinionated, group of respondents.

A later report, focusing on a subset of researchers in immunology and microbiology (Rowlands & Olivieri, 2006), indicated that two-thirds (67%) of respondents indicated they either had "good" or "excellent" access to the literature, and that nearly 84% claimed that access is much better than it was five years ago. Nearly all (97%) of respondents reported that they were "very up-to-date with the current literature in their area." In comparison with other impediments to conduct science, "access to the literature" ranked 12 out of 16, just above a desire for more conferences and networking opportunities, better management and training and clearer ethical guidelines. Surveying a similar author population (and using the same access questions as Rowlands), Mark Ware reported that some 69% of respondents claimed having either "good" or "excellent" access to the literature, although this figure varied by region of the world (Ware, 2007). The United States and Canada subgroup reported the highest satisfaction (85% "good or excellent" access versus 3% "poor"), with the "Rest of the world" subgroup reporting significantly less satisfaction (53% versus 15% respectively).

In a more recent survey of small and medium-sized businesses in the United Kingdom, over 70% of respondents claimed that they had reasonably good access to the journal literature, with 60% further reporting that access was easier than it was five years ago (Ware, 2009). The study was based upon a convenience sample of businesses known to be users of the academic literature and reports a response rate of only 4%.

The results of these large, broad surveys are confirmed by smaller, more focused studies of author preferences. Authors submitting manuscripts to the *British Medical Journal* reported that qualities such as Impact Factor, reputation, readership,

speed of publication, and the quality of peer review played an important role in their decisions to submit a manuscript (Sara Schroter, Tite, & Smith, 2005), Schroter and Tite (2006). Consistent with Rowlands (2005, 2006), authors placed little if any priority on the access policy of the journal.

The perceptions of faculty are inconsistent with those of librarians. A series of in-depth interviews of faculty, librarians and administrators at the University of California, Berkeley revealed a disjoint between the views of librarians and faculty. "Unlike many faculty," they write, "librarians who were interviewed strongly perceive a crisis in scholarly communication" (C. J. King et al., 2006, p. 8). For the most part, faculty were focused on quality concerns in academic publishing and were insulated from the consideration of costs in the publication process. The final report, released in January, 2010, summarized the values, motivations and behaviors of 160 interviewees located at 45 mostly elite research institutions across the United States and provided case studies of seven academic fields: archaeology, astrophysics, biology, economics, history, music, and political science (Harley et al., 2010). The reoccurring theme in this report is that academia is a highly conservative system, largely structured by disciplinary norms and organized around external peer-review and assessment. There is little room for experimentation in new forms of publication especially for new academics. The author-pays Open Access publishing model was viewed with some suspicion as it was perceived by several faculty to have a conflict of interest with unbiased and rigorous peer-review. Consistent with their earlier report, faculty did not perceive an access "crisis" in scholarly communication; indeed, their main concern was about access to publication outlets for their own work:

> We heard little about a crisis in scholarly communication from our interviewees, with a few exceptions […] Among humanists, there were quite a few rejections of the idea that there is a publishing crisis […] Good scholars doing good work at top-tier institutions seem to be able to get their books

> published with premier publishers […] For those who did see a publication crisis, this crisis was located in the fact that scholars producing good work could not get published—not because of the quality of their work—but because certain university presses had simply "stopped publishing" books in a number of areas and/or the costs of permissions were prohibitive. The oversupply of Ph.D.'s in some fields of the humanities and its effect on the monograph publication crisis (described as scandalous by some) cannot be ignored. (Harley et al., 2010, p. 10)

Several months after the Harley report was released, Ithaka S+R released its own report on the perceptions and behaviors of faculty with regards to scholarly communication (Schonfeld & Housewright, 2010). Reporting on a longitudinal survey conducted every three years since 2000, their 2009 survey of faculty in colleges and universities across the United States produced findings surprisingly consistent with the Harley report: With regard to publishing, faculty attitudes are fundamentally conservative and are guided by career advancement. Not surprisingly, faculty expressed little interest in transforming the scholarly communication system. Across all disciplines, free accessibility to journal content was consistently ranked last for scholars in their selection of a journal for publication. Moreover, faculty authors prioritized paying nothing to publish their own journal articles over free access, suggesting that the author-pays Open Access publication model may be at odds with the attitudes of many faculty. Consistent with previous surveys of author preferences e.g. (Rowlands & Nicholas, 2005; Rowlands & Olivieri, 2006; S. Schroter & Tite, 2006; Sara Schroter et al., 2005), publishing in a journal that is well-read among one's peers was the most important characteristic in the selection of a publication outlet. If transforming the scholarly publishing system is a concern for faculty, it is eclipsed by concerns of career advancement.

> If faculty members have concerns about the established scholarly communications paradigm, their responses do not indicate a willingness to reshape their behaviors in response to those concerns. For most faculty members, our data seem to be consistent with other research indicating that faculty interest in revamping the scholarly publishing system is secondary to concern about career advancement, and that activities that will not be positively recognized in tenure and promotion processes are generally not a priority. (Schonfeld & Housewright, 2010, p. 26)

This is not to suggest that faculty were entirely satisfied with the current publication system. About one-third of respondents agreed that tenure and promotion guidelines "unnecessarily constrain" their publication choices and this belief was stronger in the humanities and social sciences than in the sciences.

Access and Clinical Decision-Making

To date, only one study on the clinical implications of access to the medical literature could be located (Hardisty & Haaga, 2008). In a pair of related experiments, researchers were interested in whether increased access would change the use of articles in clinical psychotherapy. Participating mental health professionals were provided with one of four access conditions: 1) a reference with no citation (the control); 2) a normal reference with citation; 3) a reference with an online linked citation; or 4) a reference with a linked citation to a free-access article. After one week, participants read a vignette on the same topic covered by the article and were asked for recommendations for a medical intervention. In both studies, those participants in the free-access linked citation were more likely to report having read the article; however, in only one of the two studies did reading the article translate into making a recommendation consistent with the read article. The researchers concluded that open access may increase the consumption of treatment research articles, but that

it may not necessarily influence clinical practice.

Scientific Literature and the Lay Public

While much is known about how researchers seek for, and make use of, the scientific literature, much less is known about the consumption of scientific literature by the general public. Previous studies have focused on the use of online medical and health-related information but do not distinguish the *type* of information found on the Internet. Other than anecdotal descriptions of patients bringing medical literature they found online into the doctor's office, little is known about the how the lay public uses the primary literature (e.g. scholarly journal articles) compared to public-focused websites.

Periodic telephone surveys of American adults conducted by the Pew Research Center report that the percentage of adults who look for health information online has increased between 2002 and 2008 (Pew Internet & American Life Project, 2009). In their 2006 survey, 80% of American Internet users have searched for information on at least one health topic (Pew Internet & American Life Project, 2006b). For those who were living with a disability or chronic disease, the percentage is even higher (about 86%). This group was more likely to report that online searching affected their treatment decisions including interactions with doctors (Pew Internet & American Life Project, 2006a). Respondents who had experienced a health crisis in the past year were also more likely to get a second opinion or ask their doctor new questions based on their research (Pew Internet & American Life Project, 2008). Not surprisingly, individuals with home broadband access were more than twice as likely to conduct online health research than dial-up users.

According to the Pew telephone surveys, most Internet users begin their

research with a general search engine such as *Google* when seeking information on a health topic, whereas a minority begin their inquiry at a health-related website (Pew Internet & American Life Project, 2006b). These results are confirmed by naturalistic observational studies of how laypersons search for online health information in a laboratory environment. In an early study of 21 users in Germany, all of the participants employed general search engines and simple keyword searches in order to find relevant web pages and explored only the first few links on the first page of their search results. None of the participants in this study used medical society or health library websites as starting points (Eysenbach & Kohler, 2002). The information seeking strategies described nearly ten years ago are validated by more recent studies. In a 2009 observational study of 41 lay persons seeking information on chronic diseases, search was also the preferred strategy. Browsing for information by using hyperlinks was perceived as "chaotic, misleading and time-consuming" (Mager, 2009, p. 1134). A similar observational study of 48 lay persons included a log analysis of participant search terms revealed significant difficulties in how users formulated their keyword search (Toms & Latter, 2007).

Most medical and health-related web pages suffer from significant problems dealing with accuracy, bias and completeness, according to an early review of the literature (Eysenbach, Powell, Kuss, & Sa, 2002). While laypersons claimed that they use a number of criteria in evaluating the credibility of a medical website, in practice, few of them checked the credentials of the source or were unable to later recall the sources of their information (Eysenbach & Kohler, 2002). Indeed, just 15% of telephone survey respondents claimed that they "always" check the source and date of the information, and 10% claimed they did "most of the time" (Pew Internet & American Life Project, 2006b).

In a study of the incidence and average position of professional websites in a

series of searches of medical terms, the user-generated online encyclopedia, *Wikipedia*, ranked higher than sites such as *MedlinePlus* (maintained by the National Library of Medicine and the National Institutes of Health). For the overwhelming majority of medical keyword searches in four different search engines, *Wikipedia* showed up in the first 10 results, the first results page (Laurent & Vickers, 2009). In spite of the absence of source attribution, the authors of the study maintain that the English language Wikipedia is a prominent source of online health information.

The Pew telephone surveys list many sources of medical information including websites, blogs, commentary and podcasts, but does not make specific mention of whether the sources were scholarly or professional in nature and makes no specific mention of journals or scientific articles as a source of medical information (Pew Internet & American Life Project, 2009). Likewise, the Health Information National Trends Survey (HINTS) supported and maintained by the National Cancer Institute asks several questions about the source of health information, but confuses sources, media format, and location (National Cancer Institute, 2007). For example, question HC02 asks "The most recent time you looked for information about heath or medical topics, where did you go first?" and reports: Internet (61.0%); Doctor or health care provider (13.9%); Books (8.4%); Brochures, pamphlets, etc. (3.8%); Magazines (3.4%); among others. Based on how this question is phrased, it is difficult to discern what the researcher is implying – or indeed what the survey respondent is *thinking* – when asked about Internet use. Technically, the Internet is a system of transfer protocols for moving data over a network, although the question most likely implies whether an individual is seeking information online. Still, we don't know *what* is being sought, *where* that information resides, and in what *format* that information is presented other than the obvious fact that it is online. Many of the traditionally printed information sources listed in the survey results (e.g. magazines, brochures,

pamphlets) are often found on the Internet, making categories of responses indistinguishable from this question. More importantly, from the standpoint of our study, there is no way to distinguish popular "magazines" from professional or scholarly magazines and journals.

Article Download and Citation Studies on Access

Article readership (as measured by publisher-reported fulltext downloads) has been rising steadily and publisher journal packages have opened up access to huge numbers of journals that were previously inaccessible to college communities (Research Information Network, 2009). Publishers who offer these package deals view these data as an indication that they are providing increasing value to academic communities. Ease of access to a greater range of published literature is supported by surveys of scientists as mentioned above (Tenopir et al., 2009).

There is some dispute, however, on whether increased access has broadened the scope of cited material. Using a complex statistical model, Evans (2008) suggests that online access to the literature is concentrating citations on a narrower group of more recent literature. Using a much simpler descriptive model, Larivière, Gingras, & Archambault (2008) report just the opposite.

Reporting on the first randomized controlled trial of Open Access publishing, Davis et al (2008) reported that freely-accessible articles received no more citations than subscription-access articles within the first year of publication, although the freely-accessible treatment cohort did receive significantly more article downloads from a larger group of visitors. The lack of a citation differential implies that the traditional subscription model is efficient in disseminating published results to the research community and is consistent with the surveys of authors as reported above.

The existence of a download advantage for freely-accessible articles may indicate a peripheral demand for scientific articles outside of the research community, although more research is required on illustrating *who* is accessing these articles, *where* they are being accessed, and for what *purpose*.

Access in Developing Nations

The high cost of western scientific journals has made much of the scientific literature inaccessible to researchers in developing nations. Collaborative projects such as HighWire's Free Access to Developing Economies (HighWire Press), and multi-publisher programs focusing on broad disciplines such as Agriculture (AGORA), health and medicine (HINARI), and the environment (OARE) have attempted to bridge the access gap and provide free (or highly-subsidized) access to institutions in the world's poorest regions (Research4Life). To date, there have been several studies designed to ascertain whether researchers in developing countries have benefited from free access as evidenced through their authorship and citation behavior.

Ross (2008) analyzed the citations to journals participating in two of the above programs (HINARI and AGORA) *before* and *after* the programs were initiated in an attempt to ascertain the effectiveness of the programs. Her results were mixed: in some regions, citations to the participating journals increased, in others they decreased. No generalizable pattern was reported.

An analysis of the citation patterns in 150 biology journals indicated that authors in developing countries were no more likely to cite, or publish, Open Access articles (Frandsen, 2009). While not statistically significant, Frandsen's Open Access regression coefficient was negative (-4.51, $p=0.16$) suggesting that authors in developing countries demonstrated an aversion to Open Access sources. If access to

the published literature were a dire concern for researchers in developing countries, we would expect that Open Access journals would play a significant role in the citation behavior of researchers. A recent study of eight conservation biology journals and chapters also revealed that authors in developing countries were no more likely to cite freely-available articles (Calver & Bradley, 2010). The sample size of the Fransden and Calver studies were small, limiting the detection of only large Open Access effects.

A much larger comparative study between Swiss and Indian researchers revealed that articles written by Indian researchers had shorter reference lists and were more likely to cite articles from Open Access journals (Gaulé, 2009). While statistically significant, the effect sizes reported by Gaulé were small. On average, Indian reference lists were 6% shorter (less than 2 references) and contained 0.16 (about one-eighth of one citation) more citations to Open Access articles. Considering that Indian research institutions have far poorer access to the published literature than their Swiss counterparts, Indian researchers reported that they routinely requested copies of articles from authors and their peers at better-endowed institutions to supplement their literature needs. Some researchers admitted asking former students who moved to North American or European institutions for access to the literature.

A similar large-scale analysis of citation patterns by international authors revealed that free access to the published literature had a small but statistically-significant effect on citation behavior. Freely-accessible articles received about 8% more citations on average and twice that for poorer countries (Evans & Reimer, 2009). In comparison, commercial access to the literature[4] could explain a 40% increase in citations. It should be noted that Evans & Reimer were measuring the effect of

[4] Access via a paid subscription or through a journal aggregator such as ProQuest, EBSCO Host or Lexis-Nexis

delayed free access (when publishers make older articles freely available) and not the effect of self-archiving or author-pays Open Access publishing. A report released by Research4Life, an organization coordinating three programs (HINARI, AGORA and OARE) designed to provide free and highly-subsidized access to health, agricultural and environmental literature to the poorest of the world's nations, claimed that article production has increased in participating countries (Research4Life, 2009). We should be aware that this study did not employ statistical controls for confounding variables such as GDP or national expenditures on research and development and should consider the link between access and article production to be associative, waiting for evidence to help make a causal claim.

In conclusion, the literature on access to the scientific literature indicates that access is 1) a low-priority concern for authors, and 2) access to the journal literature is steadily improving. There is mixed evidence on whether free access is making a difference in developing nations. Very little work has been conducted on whether free access to the scientific literature is making a difference in non-research contexts, such as in teaching, medical practice, industry and government policy making. Moreover, more work needs to be done on the dissemination of scientific papers through non-formal models such as peer-to-peer sharing networks.

Table 1. Summary of Key Access Studies

Author	Survey type	Survey Population	Response rate	Key findings
Rowlands and Nicholas (2005)	Web-based survey	International sample of corresponding authors extracted by ISI author database. Survey conducted in 2005.	5,513 of 76,790 invitations (7.2%)	In selecting a journal in which to publish, top concerns for authors were: Reputation of the journal, Readership, and Impact factor. Permission to post a copy of one's article and holding on to copyright were ranked last.
Rowlands and Olivieri (2006)	Web-based survey and interviews (phone, in-person)	Reanalysis of two prior author surveys undertaken in 2004 and 2005. Sample details not clear.	3,695 sample and subsample of 92 immunologists and microbiologists	67% of respondents (2004 survey) reported having either 'good' or 'excellent' access to the journal literature. 84% believe that access is improving.
King, C.J. et al (2006)	In-person interviews	49 interviewees (31 faculty, 5 librarians, 2 campus-level administrators, 11 steering committee members). Faculty selected from 5 departments. 22/31 faculty are/were editors of scholarly journals. Interviews conducted 2005-6.	n/a	Disciplinary norms, the review and reward structure defined faculty views and behavior. Faculty were largely focused on quality issues in publishing (e.g. peer review), and were insulated from affordability issues.

Table 1. (Continued)

Ware (2007)	Web-based survey	Recently published authors, reviewers, and editors of scientific journals. Data from ISI and journal websites. Survey conducted in 2007.	3,040 of 39,232 (7.7%)	69% reported having either 'good' or 'excellent' access to the journal literature; highest for USA and Canada (85%) and Australasia (84%), and lowest for rest of the world (53%)
Gaulé (2009)	Web-based survey	Corresponding India-based authors who had published in 2007 and extracted from ISI author database. Survey conducted in 2008.	348 of 2,212 invitations (16%)	Reports high incidence of article requests from informal sources (peers, authors). Most article requests were honored.
Ware (2009)	Web-based survey	Subscription lists to trade magazines, corporate authors of STM articles, purchasers of individual journal articles (PPV). Survey conducted in 2009	1,131 of 26,390 invitations (4%)	For those who claimed that the research literature was important, 71% described their access as "fairly easy" or "very easy." 60% reported that access was easier than 5 years ago, 20% claimed it was worse.

Table 2. Key Papers on the Citation Effects of Open Access

Author(s)	Study Design	Study Description	Main Results
Lawrence (2001)	Retrospective, Observational	111,924 conference papers in computer sciences published 1989-2000. Compared articles found freely on Internet with print-only access. Controls for venue. Online availability and citations from ResearchIndex.	Overall citation increase (mean=336%, median=158%). Greater citation effect reported for top 20 venues (mean=286%, median=284%)
Schwarz and Kennicutt (2004)	Retrospective, Observational	1,679 papers published in the *Astrophysics Journal* in 1999 and 2002. 484 (61%) and 608 (72%) OA respectively. OA defined as any version of the article appearing in the astro-ph section of the arXiv. Citations counts from ISI.	Papers posted to astro-ph cited more than twice as often. Reports demographic differences among those who post articles to the arXiv compared to those who do not.
Antelman (2004)	Retrospective, Observational	2,017 articles (802 OA (40%)) published in top 10 impact journals in philosophy, political science, engineering and mathematics 1999-2002. OA defined as any version of article freely-available on Web. Compares mean citations across disciplines. Citation counts from ISI.	Mean OA citation differences (45-91%) depending on discipline. Citation differential more exaggerated for highly-cited articles
Harnad and Brody (2004)	Retrospective, Observational	14 million articles published in physics between 1991 and 2001. OA defined as any version of article freely-available on the Web. Citation counts from ISI. Comparison methodology not defined.	Reports citation ratios between 2.5 and 5.8 in favor of OA.

Table 2. (Continued)

Metcalf (2005)	Retrospective, Observational	7,089 articles (4,156 OA (59%)) published in 13 journals. OA defined as any copy of the article found in the astro-ph section of the arXiv. Citation counts from ISI. Basic comparison without controls.	Citation increases between 1.6 and 3.5 in favor of OA. As high as 5 for articles appearing in *Science* and *Nature*.
Kurtz et al (2005)	Retrospective, Observational	Articles published in 7 core astrophysics journals. OA defined as any copy found in the arXiv. Citation data from ADS system. Various analytic techniques employed.	Strong evidence that citation effect caused by self-selection and early-view effects. No evidence of citation effect as a result of OA.
Eysenbach (2006)	Prospective, Observational	1,492 articles (212 OA (14%)) published in *PNAS* in 2004. Author-pays OA articles freely-available from journal website for first 6-mo, after which all articles become freely-available. Controls for article and author characteristics in a logistic regression model. Citation counts from ISI.	OA articles were more likely to be cited than subscription-access articles between 0-6 mo, 4-10 mo, and 10-16 mo after publication (Odds ratios: 1.7, 2.1, 2.9 respectively)
Davis and Fromerth (2007)	Retrospective, Observational	2,765 (511 OA (18.5%)) articles published in 4 math journals between 1997 and 2005. OA defined as any copy of the article present in the arXiv. Various analytic techniques with controls. Citation counts from MathSciNet.	OA articles received 35% more citations on average, more exaggerated for highly-cited articles. Self-selection argued as principle cause, not OA.

Table 2. (Continued)

Moed (2007)	Retrospective, Observational	18,757 articles published in 6 physics journals between 1992-2005 (1,913 OA (10.2%)). OA defined as any copy of the article found in the Condensed Matter section of the arXiv. Various analytic comparisons. Citation data from ISI.	No evidence of citation advantage as a result of access. Strong evidence that a quality differential between arXiv-deposited and non-deposited articles is responsible for citation effect. Evidence for earlier citation lifecycle for deposited articles.
Gaulé and Maystre (2008)	Retrospective, Observational	4,388 articles (17% OA) published in *PNAS* between 2004-2006. Author-pays OA articles freely-available from journal website for first 6-mo, after which all articles become freely-available. Linear regression model includes additional confounders over Eysenbach (2006).	When additional confounders (such as location of corresponding author and time of submission) were added to model, citation effect became insignificant.
Davis et al. (2008)	Randomized controlled trial	1,619 articles (247 OA (15%)) published in 11 physiology journals. Free access to articles from journal website. Controls for self-archiving. Logistic and negative-binomial regression analysis. Citation counts from ISI.	OA articles received more article downloads but were no more likely to be cited nor receive more citations within first year after publication
Norris et al (2008)	Retrospective, Observational	4,633 articles (2,280 OA (49%)) published in ecology, applied math, sociology and economics. OA defined as any freely-available copy of article on Web. Simple comparisons, no controls. Citation data from ISI.	Average citation advantage ranged between 44%-88% depending upon field.

Table 2. (Continued)

Evans and Reimer (2009)	Retrospective, Observational	26 million articles published in 8,000 journals 1998-2005. Measured effect of publisher-mediated free access with commercial online availability in Poisson regression model, controlling for journal volume effects. Citation counts from ISI.	Publisher-mediated free access increases citation rates by 8% on average (increasing for poorer countries), compared to 40% citation increase for commercial online access
Davis (2009a)	Retrospective, Observational	11,013 articles (OA=1,613) published in 11 biomedical journals from 2003-2007. Author-pays OA articles available from journal website. Linear regression models with article characteristics used as confounders. Citation counts from ISI.	Adjusted citation advantage of 17% for author-pays OA articles. Evidence of citation effect declining over time (from 32% in 2004 to 11% in 2007)
Frandsen (2009)	Retrospective, Observational	150 journals in biology (34 of which were OA). Measures share of articles published by authors in developing countries and citations to OA journals. Linear regression. Citations from ISI.	Authors in developing countries are no more likely to publish their articles in OA journals and are no more likely to cite OA journals. Some evidence that OA journals tend to cite OA journals more frequently.
Gaulé (2009)	Retrospective, Observational	43,150 articles in science and engineering published in 2007 by authors located in Switzerland and India. Linear regression with journal as fixed effect.	Indian reference lists were 6% shorter (2 fewer citations) and cited 50% more OA journals (0.16 more OA citations) than Swiss reference lists. Reference length differences were more pronounced in biology.

Table 2. (Continued)

Lansingh and Carter (2009)	Retrospective, Case-control study	895 articles published in 6 journals in ophthalmology (3 OA, 3 subscription paired by Impact Factor) published in 2003. Multiple linear regression controlling for article characteristics. Citations from Scopus and Google Scholar.	Access status was not a significant predictor of citations when article characteristics were added to the regression model.
Calver and Bradley (2010)	Retrospective, Observational	1,151 articles and book chapters published in 8 conservation biology journals in 2000. Distinguished OA publishing and OA from self-archiving. Linear regression controlling for article and author characteristics. Citation counts from Scopus.	When controlling for article and author characteristics, no citation benefit for OA articles, although OA article chapters received twice as many citations. No citation preference for authors in developing countries.

Predictors of Citations

Article-level prediction of citations has a history spanning several decades, beginning in 1983 (Stewart, 1983) and enjoying renewed interest recently in an attempt to explain the effect of free (or open) access on article citations (see Table 2). We assume that the underlying key motivations for citation are *relevance* and *quality*, and yet these two factors are fundamentally difficult to measure because they are abstract constructs.

Relevance is a characteristic of the article that persists only in the mind of the citer, determined as a relationship between the document being written and the object considered for referencing. While tools exist for matching documents based on semantic similarity or by co-occurrence of citations, these are *discovery tools* (tools for locating related documents) and not *authorship tools* (tools for deciding what to cite).

Quality is an abstract construct, which has both an individual component and a shared social component. Experienced readers may know quality when they see it; often however, we are guided by the opinions of others. As a result, we must seek indicator variables that attempt to operationalize these abstract constructs that underlie why articles are cited.

Prestige is also an abstract construct. A journal's impact factor – technically a measure of the average citation rate per published article – is often used as an empirical indicator for journal prestige. Traditionally, the prestigious journals are selective in what they accept and publish, thus filtering for high-quality material. They also tend to have high circulation rates, thus wide dissemination in the scientific community. We assume that journal effects work therefore in two complementary ways: as a system that stratifies articles into hierarchical levels of quality; and as a

method to disseminate relevant research to interested readers (S. Cole, 2000). Studies that follow the publication trajectory of rejected manuscripts confirm that the majority of rejected manuscripts are eventually published in journals with lower Impact Factors and more specialized (and thus more limited) readership (Cronin & McKenzie, 1992; Hall & Wilcox, 2007; Liesegang, Shaikh, & Crook, 2007; R. J. McDonald, Cloft, & Kallmes, 2009; Opthof, Furstner, van Geer, & Coronel, 2000; Wijnhoven & Dejong, 2010). In practice, the journal in which an article is published is a strong determinant of future citation (Baldi, 1998; Callaham, Wears, & Weber, 2002; Larivière & Gingras, 2010; van Dalen & Henkens, 2001).

Research on determining the predictors of citations has focused on several classes of indicators:

Article Effects (e.g. article length, number of authors, topic, article type, language). (Akre et al., 2009; Baldi, 1998; Callaham et al., 2002; Conen, Torres, & Ridker, 2008; Matthew E. Falagas & Kavvadia, 2006; Kostoff, 2007; Kulkarni, Busse, & Shams, 2007; Lokker, McKibbon, McKinlay, Wilczynski, & Haynes, 2008; Patsopoulos, Analatos, & Ioannidis, 2005; Piwowar, Day, & Fridsma, 2007; Stewart, 1983; van Dalen & Henkens, 2001).

Author Effects (e.g. university authorship, age, gender, rank and prestige of author, prior publications, social ties, geographical location, self-citation)(Akre et al., 2009; Baldi, 1998; Matthew E. Falagas & Kavvadia, 2006; Kostoff, 2007; Stewart, 1983).

Journal Effects (e.g. prestige, impact factor, indexing, and journal circulation)(Baldi, 1998; Callaham et al., 2002; Lokker et al., 2008; Piwowar et al., 2007; van Dalen & Henkens, 2001).

Media Effects (e.g. coverage in newspapers, newswire, network television)(Kiernan, 2003; Kulkarni et al., 2007; D. Phillips, Kanter, Bednarczyk, & Tastad, 1991).

Reader Effects (e.g. article downloads)(Brody, Harnad, & Carr, 2006; "Deciphering citation statistics," 2008; Perneger, 2004), and

Exogenous Effects (e.g. Funding source)(Conen et al., 2008; Kulkarni et al., 2007).

Details of these papers are provided in summary form in Table 3.

There are several explanations for why article characteristics are good indicators of future citations. Longer articles often contain more content and therefore have greater potential for at least some component to be cited. At the same time, journal editors may express preferential treatment to articles they perceive as higher quality, allowing these to be published in full length, while requiring that other articles be edited for brevity (van Dalen & Henkens, 2001). Articles with higher number of authors may benefit from collaborators with individual skill sets or may go through additional rounds of editing and revisions. As scientific information gets disseminated informally through peer-networks, articles with more authors have the potential to reach more colleagues through informal communication methods (M. E. J. Newman, 2001, 2004; van Dalen & Henkens, 2001). More authors per paper also increases the potential for self-citation by future articles (Matthew E. Falagas & Kavvadia, 2006; Fowler & Aksnes, 2007) and increases the discovery of these articles when the author names are indexed in article databases. The topic of an article may designate the size of a field or domain where other authors are working; the larger the field, the greater the potential to be cited. The type of article is also important; original research is often cited more frequently than case studies or opinion pieces. Review articles tend

to receive many more citations as authors use them as a citation shorthand instead of citing a full body of extant literature. Lastly, the dominant language of science is English, and as a result, research written in other languages may be overlooked by relevant potential audiences (van Dalen & Henkens, 2001).

Since articles are not anonymous but are branded with the authors' names, qualities of these authors (if known) may be related to article performance (Merton, 1968, 1988). Those authors with a history of high-quality output may receive disproportionate attention when they publish another article. Unknown authors, residing at little-known institutions may not receive the attention their article deserves.

In the evidence-based biomedical literature, citation patterns follow the relative importance placed on methods. Randomized controlled trials, controlled trials, and meta-analyses are valued over, and receive more citations than, uncontrolled trials, expert opinion and nonsystematic reviews (Patsopoulos et al., 2005). This is also true of studies with large sample sizes and those that include a control group (Callaham et al., 2002). These normative findings, however, do not hold up when it comes to industry funding. Kulkarni et al. (2007) report that even after controlling for article characteristics, studies with declared industry funding receive significantly more citations, but only if their results are supportive of the intervention. Conen et al. (2008) report that there is a consistent citation advantage to studies funded by commercial entities over all strata of comparison. For-profit funders have several advantages over non-profit governmental and non-governmental organizations in disseminating favorable research: For-profit entities have dense networks of sales representatives who are often in regular contact with academic researchers; industry-sponsored sessions are common at scientific meetings; industry has greater access to media sources and are likely to invest in secondary publications (often called

"throwaway journals" by medical researchers since they are distributed freely and are biased toward industry research); and lastly, financial ties between academic researchers and industry are more likely to result in a favorable conclusion and may lead to selectively citing other industry-favoring studies (Conen et al., 2008). In sum, the result of preferential treatment given to industry-sponsored studies may create unfounded authority through the citation record (Greenberg, 2009).

Better access to resources for the dissemination of industry-sponsored research results appears to affect Open Access publishing decisions as well. Authors declaring industry funding are more than twice as likely to pay the optional Open Access publication fees to make their work freely available in the *Annals of the Rheumatic Diseases*, a medical specialist journal published by the British Medical Journal (Jakobsen, Christensen, Persson, Bartels, & Kristensen, 2010).

Determining the role of mass media on the diffusion of scientific results has also been explored in several studies. Phillips et al. (1991) were concerned with the effect of the lay press, specifically the *New York Times*, on disseminating the results of medical research. Conducting a retrospective, cohort analysis of articles published in the *New England Journal of Medicine*, Phillips reported that research covered in the *Times* received many more citations, especially within the first few years. In comparison, this citation effect was not present in a cohort of articles covered in the *Times* during a three-month newspaper strike in 1978 where an "edition of record" was prepared and archived but not distributed. The absence of effect in the later control group provides strong evidence that the newspaper was amplifying the diffusion of scientific results. In further generalizing Phillips research to include 24 other leading newspapers and network television, Kiernan (2003) reported that coverage in newspapers was generally associated with greater citations although

network news coverage was not. These findings are corroborated by Chapman et al. (2007) who reported that press-released articles (sent by the publisher to the press) are associated with increased citations.

Characteristics of articles, their authors, the journals in which their articles are published, and the external network of media sources and peer-networks are all important predictors of future citations, not because of these features themselves, but in how they draw and support attention to an article. Several studies have investigated how readership (as measured by article downloads) are predictive of future citations. Using just the first week of fulltext article counts for articles published in the *British Medical Journal*, Perneger (2004) reported a weak, but significant correlation with citations five years after publication. Combined with article characteristics, he was able to explain about 33% of the citation variance. Working with articles deposited in the arXiv with article download statistics from the United Kingdom mirror site, Brody et al. (2006) describe moderate correlation strength between article downloads and citations, although their explanatory power with just download data is quite small, explaining less than 20% of citation variance. Like Perneger (2004), longer periods of observation time did not necessarily add to their ability to predict future citations; six months of download data was just as good as having two years of data.

Measuring download data as a predictor of future citations does have its caveats. When multiple versions of an article are residing in multiple locations (for example, an early draft of the paper in a subject repository such as arXiv, a final peer-reviewed manuscript residing in one's institutional repository, and a copy of the final publisher's version on the journal website), selecting usage data from only one source (such as in the case of Brody (2006)) may limit the extent of observation. Similarly, combining all of these sources together may obscure the results as different versions of

the article may serve unique functional purposes; for instance, alerting the reader on what research is being conducted in a field versus having an archival version of record for citation purposes (Henneken et al., 2007).

Key papers on predicting citations are summarized in Table 3. The rationale for these studies are several. Stewart (1983), Baldi (1998) and van Dalen (2001) were interested primarily in explaining whether citation practices followed a *normative process* of intellectual indebtedness and rewards, for example (S. Cole & Cole, 1967; Garfield, 1955; Hagstrom, 1965; Kaplan, 1965), or whether citation behavior was governed largely by a *rhetorical process*, whereby citations are a social construct used to bolster one's argument through appeals to established authorities (Gilbert, 1977; Latour, 1986, 1987).

If the citation process was chiefly normative, Stewart and others argue, then authors would primarily base their citation decisions on the characteristics and quality of the article and not on the characteristics and status of its author(s). Stuart (1983), Baldi (1998) and van Dalen (2001) show little support for the social constructive school of thought. In all three studies, author characteristics had little or no explanatory power in predicting citations after article characteristics had been explained.

The chief difficulty in determining the function of citations is that the citation process is ultimately a private act. As Blaise Cronin maintains, "Citation is a private process with a public face. Therefore, any attempt to understand the nature of citations is conjecture" (Cronin, 1984, p. 28). Attempts to decipher the meaning of the author have underscored the ambiguity in how citations are used in the authorship process. Many citations are *perfunctory*, that is, not necessary for a reader to understand the paper but merely acknowledging that similar work had been done in a

particular area (Chubin & Moitra, 1975; Moravcsik & Murugesan, 1975). Indeed, asking the authors themselves doesn't clarify the debate. Authors often attribute many motives for using particular citation with persuasion being a primary rationale (Brooks, 1985, 1986). It may suffice to state that citations reflect both normative *and* rhetorical processes and that these two views are not mutually exclusive but complementary (Cozzens, 1989; Luukkonen, 1997).

Table 3. Key Papers on Predicting Citation.

Author	Sample	Predictors	Methodology	Key findings
Stewart (1983)	133 articles published in geophysics and physical geography in 1968. Citations gathered from 1969-1974	**Article effects** (article length, topical relevance, publication delay, preprint, empirical studies, #references, recent references); **author effects** (university authorship, age, author rank, prior author publications)	linear and stepwise regression	After accounting for article characteristics (67% of citation variance), author characteristics only explain an additional 8%
Phillips et al. (1991)	25 articles covered by the *NY Times* vs. 33 not covered published in 1979; NY Times strike: 9 articles covered vs. 16 not covered published in 1978. 10 years of annual citations.	**Media effects:** coverage in *The New York Times*	paired, non-parametric comparisons	Articles covered in the *NY Times* received 73% more citations than control articles. Effect not present for articles published during the strike.
Baldi (1998)	100 articles on celestial masers (astrophysics) published 1965 - 1980 and citation links between them	**Article effects** (number of authors, length, content type, recency, article quality, years elapsed between citing and cited articles; **Journal effects:** (journal visibility and quality); **Author effects** (author gender, rank, university, institutional prestige, social ties)	logistic regression testing probability a citation exists between two papers in a network	Authors are likely to cite other articles based on relevancy, subject, recency, theoretical orientation (control and normative variables) with little concern for author characteristics (social constructivist variables)

Table 3. (Continued)

van Dalen and Henkens (2001)	1,371 articles in demography published between 1990-1992. Citations gathered 5 years after publication	**Article effects** (number of authors, US affiliation, article type, publication order, geography, language); **Journal effects** (journal impact, reputation of editorial board, journal circulation)	Negative binomial regression on citation count on accrued citations	Journal characteristics have highest explanatory power followed by paper characteristics. Contribution of author characteristics is small.
Callaham et al. (2002)	219 articles in emergency medicine published in early 1990s. Citations gathered after 3.5 years	**Journal effects** (Impact factor of journal); **Article effects** (study size, quality score, newsworthiness, study design (control group, hypothesis, retrospective/prospective, blinded, randomized, positive results))	Regression tree, calculating relative contribution to explanatory power	Impact factor of the journal is strongest predictor, followed by newsworthiness, sample size, and presence of control group.
Kiernan (2003)	2,655 articles published in *JAMA*, *NEJM*, *Science* and *Nature* published June 1997-May 1988	**Media effects**: Coverage in New York Times, coverage in 24 other leading newspapers, coverage on network television. Citations gathered in 2002	Hierarchical linear regression to explore relationship between news coverage and citations	Coverage by newspapers was generally associated with greater citations; network news coverage was not.
Perneger (2004)	154 articles published in *BMJ* in 1999. Citations gathered after 5 years	**Reader effects**: Fulltext views (HTML downloads) during week after publication with citations counted 5 years later.	Pearson correlation; linear regression	Early reading counts can predict citations 5 years later. Page length, study design also predictors.

Table 3. (Continued)

Patsopoulos et al (2005)	2,646 medical articles published in 1991 and 2001. Citations gathered after 2 years	**Article effects:** Study type (meta-analysis, randomized controlled trial, review articles, epidemiological studies, decision and cost-effectiveness study, case reports)	Non-parametric comparisons; logistic regression on highly-cited articles	Citations reflect relative importance of papers as established by evidence-based medicine
Brody et al. (2006)	14,917 articles deposited into the arXiv. Article downloads from UK mirror site only.	**Reader effects**: Article downloads during first 2 years after article deposit with citations from Citebase.	Pearson correlation	Moderate strength correlation between article downloads and citations (r=0.46 for physics)
Falagas and Kavvadia (2006)	340 papers published in 6 leading biomedical journals in 2005	**Article effects**: Number of authors; **Author effects:** self-citation	Comparison of means	Number of authors of a paper is associated with higher rates of self-citation
Kostoff (2007)	102 articles published in Lancet between 1997-1999 that received the top and bottom 5% of citations	**Article effects**: Number of authors, references, abstract words, page length study design; **Author effects**: organization type and geographical location	Comparison of characteristics of articles in the top and lower 5%	Author, article, study, and location differences reported

Table 3. (Continued)

Kulkarni et al. (2007)	328 articles published in *The Lancet*, *JAMA* and *NEJM* 1999-2000. Citations gathered 5 years after publication	**Exogenous effects:** Industry funding, industry favoring result, location of study; **Article effects**: topic, group authorship, sample size, study design, journal; **Media effects**: coverage in news media	Forward stepwise regression analysis	After controlling for independent variables, studies with declared industry funding received more citations only if their results were industry-supportive
Piwowar et al. (2007)	85 clinical cancer trials articles dealing based on microarray data	**Article effects:** Publicly-available dataset, Journal impact (high/low); **Author effects**: US author	Linear and logistic regression	Articles with publicly-available data was associated with 69% increase in citations.
Conen (2008)	303 cardiovascular articles published in *JAMA*, *The Lancet* and *NEJM* published 2000-2005. Citations per year to 2006	**Exogenous effects:** Funding source (profit vs. not-for profit); trial outcome (favorable/unfavorable results); **Article effects:** sample size, end point, single vs. multi-center study, intervention type	Non-parametric comparisons	Citation advantage to studies funded by for-profit entities, consistent over all strata of comparison
Lokker (2008)	1,274 medical articles published in 2005. Citations gathered at 2 years	**Article effects:** 20 article and journal features including indexing and abstracting in journals and databases; article quality rating	Linear regression	Variables measured at 3 weeks can predict citations at 2 years

Table 3. (Continued)

Akre (2009)	4,724 articles published between 1998 and 2002 in *BMJ*, *The Lancet*, *JAMA*, and *NEJM*. Citations taken in 2008.	**Author effects:** Geographic location of corresponding author; **Article effects**: type of study	Logistic regression predicting likelihood of being highly-cited or poorly-cited (top/bottom quartile)	Authors from low-mid income countries less likely to be highly-cited and more likely to be poorly-cited

INTRODUCTION TO THE EXPERIMENT

There are many reasons why a rational scientist would attempt to seek publication outlets that maximize the chances of his or her work being cited. Citations are an indicator of the dissemination of an article in the scientific community (Garfield, 1955). They provide stable links to cited documents and make a public statement of intellectual recognition for the cited authors (Biagioli, 1998, 2003; Franck, 1999; Kaplan, 1965; Merton, 1988). Reified, citations provide a quantitative system for the public recognition of one's work by qualified peers (Cronin, 1984; Merton, 1988), and in many institutions, citations form the basis for the evaluation of scientists.

In 2001, Steve Lawrence, a computer scientist working at the NEC Research Institute, first reported that freely-accessible online computer science proceedings garnered more than three-times the average number of citations received by articles found only in print (Lawrence, 2001). This "Open Access citation advantage" has since been validated in other subject disciplines, such as astrophysics (Metcalfe, 2005, 2006; Schwarz & Kennicutt, 2004), physics (Harnad & Brody, 2004), mathematics (Antelman, 2004; Davis & Fromerth, 2007), philosophy (Antelman, 2004), political science (Antelman, 2004), engineering (Antelman, 2004), and multi-disciplinary sciences (Eysenbach, 2006). Craig et al. provide a critical review of the literature (Craig, Plume, McVeigh, Pringle, & Amin, 2007).

The primary explanation offered for the Open Access citation advantage is that freely available articles are more accessible, and thus read more frequently, than their subscription-only counterparts. While often unstated, the theoretical proposition for the citation advantage may be written:

THEORETICAL PROPOSITION:

Free access → Increased readership → Increased citations

The basis of this proposition has been made on inferential evidence. All of the studies references above make a logical leap between access status and citations without including readership as an intermediary causal variable.

Studies of single journals have described weak, but statistically significant, correlations between article downloads from a publisher's website and future citations ("Deciphering citation statistics," 2008; Moed, 2005; Perneger, 2004), between downloads from a subject-based repository and future citations (Brody et al., 2006), and between downloads from a repository and downloads from a publisher's website (Davis & Fromerth, 2007). While these studies provide a connection between readership and citations, the validity of the theoretical model is based upon making causal connections between these three parts.

In recent years, a growing number of studies have failed to provide evidence supporting the citation advantage, leading researchers to consider alternative theoretical models (Calver & Bradley, 2010; Davis & Fromerth, 2007; Davis et al., 2008; Kurtz et al., 2005; Kurtz & Henneken, 2007; Moed, 2007). In addition to the Open Access explanation, Kurtz et al. (2005) proposed two other non-exclusive theoretical explanations for the citation effect: the *Self-Selection postulate*, and the *Early View postulate*. Self-selection postulates that authors tend to preferentially promote their best (and thus most citable articles) by making them freely-available. The early access postulate suggests that manuscripts that have been posted freely on the Internet benefit from additional time to be read and cited.

Several studies using the arXiv as a free source of journal articles show support

for these alternative postulates. For astronomy articles, Kurtz et al. found strong support for the Self-Selection and Early View postulates, but not for the Open Access postulate (Kurtz et al., 2005; Kurtz & Henneken, 2007). For physics articles, Moed (2007) found strong support for the Early View postulate, but not for Open Access. For mathematics articles, Davis and Fromerth (2007) found support for the Self-Selection postulate, but not for the Open Access nor Early View postulates.

Studies of the prevalence of self-archiving one's article on the public Internet also support the existence of alternate theoretical explanations for a citation effect. For the economics literature, self-archiving is much more prevalent for the most-cited journals than for less-cited journals (Bergstrom & Lavaty, 2007), and for the medical literature, articles from higher-impact journals were more likely to be found on non-publisher websites (Wren, 2005).

In sum, the literature is inconclusive on whether Open Access is a cause of increased citations. Part of the difficulty in discerning the relationship between access and citations may be explained by the methodology employed in these studies. All previous studies on the impact of access on article citations were based on uncontrolled experimental methods, meaning that the researcher could observe but have no control over the access status of the articles. While uncontrolled observational studies are more realistic in nature compared to controlled experiments, they suffer from three major deficiencies:

1) *Discerning the direction of causality.* Uncontrolled observational studies are unable to discern the direction of causality. Free access may lead to increased citations; however, we are unable to rule out reverse causality that highly-cited papers may be more likely to be made freely-accessible.

2) *Unobserved intermediary cause*. The association between accessibility and citations assumes an intermediary variable (readership), which has been ignored in previous studies connecting access with citations. Access itself is not a sufficient explanation for a citation effect. We assume that free access leads to more *readership*, and that readership is responsible for increased citations. Thus in order to build an explanatory model, it is critical to include readership as an intermediary cause.

3) *Confounding*. Uncontrolled observational studies may suffer from the presence of unobserved and/or unmeasured variables that are associated with freely-accessible articles. When these confounding variables are unknown, the researcher may attribute a citation effect to access, when in fact other variables were responsible. Statistical controls may be used to help mitigate against bias; yet, the use of statistical controls assumes that all confounding variables are known, observable and thus measurable. While it is simple to code for article characteristics such as the number of authors, page length, or funding source of an article, more abstract characteristics of an article, such as novelty, readability, and significance of the results are more difficult to observe and thus include in a statistical analysis.

While still useful, studies based on uncontrolled observational methods should be examined with some caution before strong causal claims are inferred from them. The inconclusive results on the association between access and citation performance may be based, at least partially, on the methods employed for gathering and analyzing the data. A more robust methodology is necessary to understand the relationship

between access, readership, and citations. This dissertation will describe the results of an experimental approach.

Randomized Controlled Trials

Randomized Controlled Trials (RCTs) are a type of methodology used to isolate the effect of the treatment under investigation. Through the randomization process, subjects are allocated either to the treatment arm or the control arm of the study. The allocation process ensures that confounding variables are equally distributed between the two groups and minimizes any bias between the two groups present at the beginning of the experiment (Friedman, Furberg, & DeMets, 1985; Greenhalgh, 1997; Jadad, 1998; Koepsell & Weiss, 2003; Stanley, 2007). If differences are detected *after* the start of the trial, they are likely to be the result of the treatment alone. In this study, I applied such a methodology in order to isolate the effect of access from other potential confounding effects.

Research Questions

The two research questions posed in this study are:

RQ1: How does free and immediate access to the scientific literature affect readership?

RQ2: How does free and immediate access to the scientific literature affect citation behavior?

Operational Variables

Readership is measured using four different, although related, indicators: abstract downloads; full text (HTML) downloads; PDF downloads; and unique Internet Protocol (IP) addresses (an indicator for the number of visitors). Although these are imperfect measures of readership, because an article download may not result in that article actually being read, they provide good proxies for readership. Data for each article in this study were made available to the researcher from participating publishers' websites. All publishers agreed to provide the researcher with direct access to their administrative reporting systems. In measuring readership, article downloads and unique visitor counts are known to be affected by non-human software robots, which are designed to crawl the web for the purpose of indexing freely-accessible content. These data were excluded from the analysis (when possible) in order to arrive at a tighter relationship between downloads and readership.

Hypotheses

This experiment will test two null hypotheses:

$H_0$1: Free access to scientific articles does not increase article downloads

$H_0$2: Free access to scientific articles does not increase article citations

Ethical Issues

This study is a manipulation of the *output* of authors (their articles), and not of authors *themselves*. From a legal standpoint, the publisher owns the copyright of the research articles, so permission of the author is not technically required. One may

argue that there is no need to consider the ethical implications of this study. Still, if authors are evaluated and rewarded based on the performance of their output, then manipulation of their articles should be taken very seriously. We considered two important ethical questions in designing this study:

1) Is there potential for harming authors?
2) Is consent necessary to participate in the study?

Potential for Harm

Our research is unidirectional. We randomly provide free access to some articles, but do not close access to others. Previous research has indicated that freely accessible articles receive more citations than subscription-based articles, and while some studies fail to confirm that access is the cause, there is no evidence that Open Access is associated with negative effects. As a medical analogy, the treatment may result in a benefit or it may be a placebo, however there is no evidence that it may lead to harm. In sum, we see no potential in this study to do harm to authors.

Author Consent

While harm to authors was not preconceived as a consequence of our study, we felt it prudent to confirm our speculation with the authors themselves. During the feasibility study with the American Physiological Society, corresponding authors of accepted manuscripts were sent letters alerting them of the study and providing an option to *opt-out*. Those authors who wished to opt-out of the study would have their article removed from the pool of articles from which we randomly assigned Open

Access status. The letters were sent by electronic mail by the publisher (rather than the researcher) to minimize confusion and to maximize the chance that the letter would be read. In the feasibility study, none of the approximately 1,500 corresponding authors wished to opt-out of the study. What was curious, and somewhat amusing, was that several authors wanted us to randomly select *their* article for the treatment (this does change the definition of random selection!).

Institutional Review Board

As a final precaution, we submitted our research proposal to the Institutional Review Board on Human Subjects (IRB) at Cornell University before the commencement of our experiment. The IRB responded that our study did not involve human subjects and therefore did not require formal approval from the committee.

Since we could not identify potential for harm and because authors were not opposed to us manipulating the access status of their papers in this experiment, we discontinued the practice of alerting authors of the study. We believe that ceasing this practice would also remove the potential of ascertainment bias in our study (see *Statistical Bias : ascertainment bias* below for more details).

METHODS

The nature of our experiment required the participation of several journal publishers. We wanted to create as natural an experiment as possible while maintaining autonomy and objectivity in the process. This meant having control over the selection and treatment of experimental articles, and secondly, having access to the publisher's statistics system for gathering usage data. Our autonomous and independent role in this study was critical to maintain credibility of the study to those outside the publishing community. The Open Access debate is intensely political and emotional; it was important therefore to avoid the impression that the study was being directed – or unduly influenced by – industry insiders. Maintaining that autonomy meant that participating publishers consent to an outsider having access to, and control of, their online publishing and data gathering systems and trust that these powers would not be abused.

Participation also required that the publisher recognize the editorial independence of the author and not attempt to influence the reporting of the results. Publishers would be given an opportunity to check for errors or comment on a draft manuscript, but not reserve the right to edit the manuscript in any way. Stated another way, participation did not imply authorship. In early negotiations, one publisher demanded line editing rights to any and all future manuscripts. Accepting this agreement would have violated academic independence, and as a result, we could not accept this publisher into the study.

Publisher Recruitment

The journal publishing community is relatively small, stable, and highly

connected. This creates an environment where personal reputation, trust, and favorable recommendation largely define the dealings of individuals. The inclusion of publishers into this study would not have been possible without the initial involvement of several highly-influential people who connected the researcher with interested publishers and vouched for the importance of the study. The most central of these individuals was John Sack, the publisher at HighWire Press, who was invaluable for providing contacts of potential participants in the publishing community. The first contact was Martin Frank, the Executive Director of the American Physiological Society, without whom the feasibility study would not have been possible. Success with the APS subsequently made it easier to recruit other publishers into the full experiment. The process of negotiation became easier as each successive commitment legitimized and augmented the importance of the study. Negotiations with publishers took place by phone and email: none of them required formal, written agreements – a testament to the relationship of trust between the journal publishing community and the researcher.

Journal Recruitment

The goal of the selection process was to recruit participation from a diverse group of journals representing several disciplines in the sciences, social sciences, and humanities. While our goal was to aim for a representative sample of journals in our study, this was not possible for several reasons: First, we were limited to a single publisher hosting platform (HighWire Press); second, several journals had particular publishing practices that made inclusion in the study difficult, if not impossible; [5]

[5] For example, one journal made all articles written by members of its society freely accessible, but charged non-society members a fee for this service. Another publisher made all subscription-access articles freely available at the beginning of the next

third, in order to conduct a statistical analysis, we required journals that received sufficient article downloads and citations over the period of study; and finally, the decision to participate was out of the researcher's control. Whereas our sample consists of journals from various academic fields (biological sciences, medical sciences, multi-disciplinary sciences, social sciences, and the humanities), we should not consider our selection to represent a representative sample of all academic fields. They do however represent a diversity of different disciplines, and it is possible with such a diverse group to investigate similarities and differences within and between journals. Descriptions of the specific publishers and journals included in our study are found under *Exploratory Study* and *Full Study* subsections below. A discussion of the methodological limitations of our journal set is provided under *Methodological Limitations*.

Randomized Controlled Trial

Following the general methodology of the randomized controlled trial, articles were randomly assigned into two groups: a *treatment group* and a *control group*. The treatment in this study is immediate free access from the publisher's website. The articles in the control group follow their natural trajectory of publication. For some articles, this means available by subscription; for other articles, this means they are available by subscription for the first part of their publication life after which they become freely available.[6] We will refer to this as the *delayed access model*. The

---

calendar year (January 1st), rather than basing access on the date that the article was published. For example, articles published in December waited only one month before becoming freely available, whereas articles published in January needed to wait 12 months until they became freely available.

[6] A list of HighWire-hosted journals that provide free access to backfiles can be found at: http://highwire.stanford.edu/lists/freeart.dtl (accessed April 28, 2010).

American Physiological Society (11 journals), the American Heart Association (5 journals), the Federation of American Societies for Experimental Biology (1 journal), the Genetics Society of America (1 journal), and the American Association for the Advancement of Science (1 journal) all employed a delayed access model, the details of which can be found in Table 4. Duke University Press (7 journals) and Sage Publishers (10 journals) both employed a full subscription access model.

Articles were randomly assigned to either the treatment group or the control group upon online publication. We ruled out other potential forms of experimental allocation (such as retrospectively making published articles freely available), as the interpretation of the data becomes difficult when articles are moved into treatment groups after they have been published.

Exploratory Study

Since our experimental design was novel, it was important to test the methodology on a single publisher before expanding the experiment into a full study. The American Physiological Society (APS) was willing to partner with us in working out the methodological details for the full study. The American Physiological Society is a non-profit membership society located in Bethesda, Maryland. Started in 1887, the society has over ten thousand members and currently publishes 12 research journals, an education journal and an online newsletter, in addition to several books and book series.[7] APS research journals are ranked as some of the best in their field. Of their 12 research journals, we were able to work with 11. At the time of our experiment, one journal, *Physiological Genomics*, had been providing a service where authors could pay the publisher a small fee ($750) to make their article freely available

[7] The American Physiological Society http://www.the-aps.org/about/index.htm (accessed April 28, 2010).

immediately upon publication.[8] While the uptake for this service was small, the publisher wanted to avoid a potential public relations conflict with these authors if other authors were given the same service for free. As a result, *Physiological Genomics* was used as an observation journal, that is, we would observe how articles in this journal performed, but would not manipulate articles in an experimental setting.

The choice of treatment articles was made using a random sequence generator ("Random.org,"). Only research articles and reviews were included in the randomization. Editorials, letters to the editor, corrections, retractions, announcements, etc., were ignored in the sampling. For those journals that employed structured categories for their articles, a stratified random sampling technique was employed to ensure that certain categories of articles were adequately represented in the sample.

The allocation of articles into the treatment arm or control arm of the study was made entirely by the researcher. Direct access to the online journals' administration system was made possible by the production staff at the APS. Through the journals' administration system, the access status of articles could be manipulated by the researcher and usage data for each article could be tracked.

From January through April 2007, 247 research articles were randomly assigned to the immediate free access (treatment) group. The remaining 1,372 articles formed the control group and were available to readers by subscription for the first 12 months after which they became freely available. Details on allocation numbers per journal are presented in Table 4.

[8] This service, called "AUTHORCHOICE" is available now for all 13 APS journals. In 2009, the fees were $2,000 for research articles and $3,000 for reviews.

Full Study

After a successful start to the experiment, six additional publishers were recruited into the experiment: one publisher of a prestigious multidisciplinary science journal (AAAS), a medical publisher (American Heart Association), two biology societies (FASEB and the Genetics Society of America), a social sciences publisher (Sage), and a publisher of the humanities literature (Duke University Press). The addition of these publishers provided us with greater breadth over the scholarly journal landscape, and allow us to make more general statements about scholarly publishing beyond the specific field of physiology (Godlee, 2008).

Randomization of treatment articles began in June 2007 and proceeded through January 2008. An additional 465 articles were made freely available upon publication, bringing total treatment articles to 712. The control group of articles increased from 1,372 to 2,533 (see Table 4 for details). Some variations on the sampling methodology should be noted: For the *FASEB Journal* and *Genetics*, a balanced sampling technique was followed, allocating half of the articles in each issue to the treatment arm and the other half to the control arm of the study. Every other article in each issue was chosen for the treatment group with a random start (i.e. start on article 1 or 2) to avoid any possibility of bias if there was anything meaningful about the first article listed in the journal issue. The publication boards of both of these journals wished for accurate point estimates for their journals, and not simply to have their data aggregated with other journals in the study. Based on power estimates (see section on *Sample Size Calculation* below), adequate sample sizes were met within five issues for the *FASEB Journal* and four issues for *Genetics*. For Duke University Press, nearly half of the articles in each issue were also allocated into the treatment arm of the study. All of the Duke journals in our study with the exception of one (*Journal of Health Politics, Policy and Law*, which publishes six times per year) publish on a

quarterly basis. It was therefore necessary to allocate more articles to the treatment arm per issue in order to achieve an adequate sample size. With considerably fewer articles published in each journal, it was necessary to analyze all participating Duke journals as a group. An aggressive sampling regime was also used for Sage journals, although the publisher was willing to allow a maximum treatment of one-third of articles per issue. In working with Sage, we ran into an early complication: during the summer of 2007, Sage announced free access to many of the journals in our study. We delayed allocating treatment during these promotional periods for fear that they would compromise our study. The AAAS was more cautious of providing free access to large numbers of articles and limited our treatment allocation to two original articles per issue, or about 1 in 8 articles. Access control for the AAAS was controlled by publishing staff. A day or two before publication, I was sent the table of contents for what would appear in the next issue. From the table of contents, I randomly selected two articles and relayed these selections back to the AAAS contact, who made these articles freely available upon publication. For the five journals published by the American Heart Association, a stratified random sampling methodology was followed for those journals that published journal sections and a simple random sample for those that did not. Treatment articles were indicated as freely available with an open green lock on the table of contents page on the journal website. In addition to the random treatment allocation, AHA journals employed an Editor's Pick section. In each issue, an editor could choose one article to highlight as the "Editor's Pick." These articles were displayed prominently on the journal websites and were also made freely available. Thus in each issue there were two kinds of free access articles: an article chosen by the editor and several articles chosen randomly by the researcher. The existence of these Editor's Pick articles provided us with a unique opportunity to compare the effect of expert *selection* independently from the effect of *access*.

Table 4. Allocation of Random Open Access Articles by Publisher and Journal

Publisher (Journal)	**No. Treatment articles**	**No. Control articles**	**% (Treatment / Total)**
American Physiological Society, (Jan - Apr, 2007), Delayed access: 12mo			
American Journal of Physiology-Regulatory, Integrative and Comparative Physiology	34	161	17%
American Journal of Physiology-Endocrinology and Metabolism	21	126	14%
American Journal of Physiology-Renal Physiology	18	122	13%
American Journal of Physiology-Heart and Circulatory Physiology	32	201	14%
American Journal of Physiology-Lung Cellular and Molecular Physiology	14	95	13%
American Journal of Physiology-Gastrointestinal and Liver Physiology	22	112	16%
American Journal of Physiology-Cell Physiology	36	119	23%
Journal of Applied Physiology	27	174	13%
Journal of Neurophysiology	39	239	14%
Physiological Reviews	2	14	13%
Physiology	2	9	18%
Total	247	1372	15%

Table 4. (Continued)

Publisher (Journal)	No. Treatment articles	No. Control articles	% (Treatment/ Total)
American Heart Association, (Jun - Sep, 2007), Delayed access: 12mo			
Arteriosclerosis, Thrombosis, and Vascular Biology	20	85	19%
Circulation	20	76	21%
Circulation Research	19	41	32%
Hypertension	20	75	21%
Stroke	22	110	17%
Total	101	387	21%
Duke University Press, (Jun - Dec, 2007), Delayed access: never			
Journal of Health Politics, Policy and Law	7	10	41%
American Speech	3	5	38%
Neuro-Oncology	12	15	44%
Public Culture	6	5	55%
Ethnohistory	5	6	45%
GLQ: A Journal of Lesbian and Gay Studies	6	7	46%
Social Science History	4	6	40%
Total	45	54	45%

Table 4. (Continued)

Publisher (Journal)	**No. Treatment articles**	**No. Control articles**	**% (Treatment/ Total)**
SAGE Publishers, (Jun 2007 - Feb 2008), Delayed access: never			
Comparative Political Studies	11	17	39%
Communication Research	8	11	42%
New Media & Society	8	22	27%
Social Studies of Science	8	17	32%
American Behavioral Scientist	10	31	24%
Progress in Human Geography	8	18	31%
Administration & Society	9	15	38%
Theory & Psychology	10	23	30%
Applied Psychological Measurement	7	13	35%
Organization	8	16	33%
Total	87	183	32%
Federation of American Societies for Experimental Biology, (Jun - Oct, 2007), Delayed access: 12mo			
FASEB Journal	81	84	49%

Table 4. (Continued)

Genetics Society of America, (Jun - Sep, 2007), Delayed access: 6mo			
Genetics	103	108	49%
American Association for the Advancement of Science, (Jun - Nov, 2007), Delayed access: 12mo			
Science	48	345	12%
Grand Total	**712**	**2533**	**22%**

Data Gathering

Beginning the first month after publication and extending throughout this study, usage statistics for each article were harvested, by permission, from the publisher's website. For the first few months, data lookup and recording was performed manually. This proved to be an unsustainable work model. In the feasibility study alone, there were 1,619 articles under study, and for each of them, 4 attributes were recorded (abstract downloads, fulltext downloads, PDF downloads and unique visitors) per month, resulting in recording nearly 6,500 data points per month. Later in 2007, the publisher began providing additional columns in their usage report, reporting the data after being filtered for known robot activity. This doubled the number of potential data points. Without automation, the study could not take on additional journals, let alone keep up with the monthly pace of data collection for four years. A programmer within the Cornell University Library was contracted to create a Perl script that could do this data harvesting semi-automatically, thus greatly reducing the amount of manual labor spent gathering usage data.

While we assume that readers procure their copy of an article directly from the publisher's website, we do not exclude the possibility that free copies of research articles can be found elsewhere on the Internet. An article may be found freely on the Internet because someone (often the author) decided to post a copy on a personal, laboratory, class, or departmental website; deposit a copy in a disciplinary or institutional repository; or make a copy available through a peer-to-peer sharing network such as *Facebook* or *Mendeley*. To arrive at an estimate of the prevalence of these free copies, as well as calculating their effect on our readership and citation data, our programmer wrote a second Perl script to search for PDF copies of articles anywhere on the public Internet ignoring the publisher's website. The Perl script

submitted article title searches to the Yahoo search engine and processed the returned results, including the Uniform Resource Locator (URL) of a PDF with a similar match. Care was taken in determining the sensitivity of the match: Too little sensitivity would retrieve too many false-positives; too much sensitivity would miss real cases of self-archiving (false-negatives). We decided to err on retrieving too many false-positives and then proceed to manually verify each case to determine if the document identified was indeed a match. "Self-archiving" was considered when either a publisher's PDF version or final peer-reviewed copy of an article was found on a freely-available website, irrespective of who was responsible for making the article freely available and for what reason.

Lastly, the number of citations to each article involved in this study was collected monthly from ISI's *Web of Science*. This routine collection of citation data allowed us to observe the longitudinal performance of these articles over the length of the full study period (4 years). The flow of study data is presented in Figure 1.

Sample Size Calculation

Because randomized controlled trials involve intervention, they should be designed in such a way that there is a reasonable likelihood that they will provide a clear and definitive answer (Stanley, 2007). Sample sizes that are too small often yield inconclusive results. In planning the experiment, I conducted power analyses in order to determine adequate article sample sizes required from each participating journal and publisher.

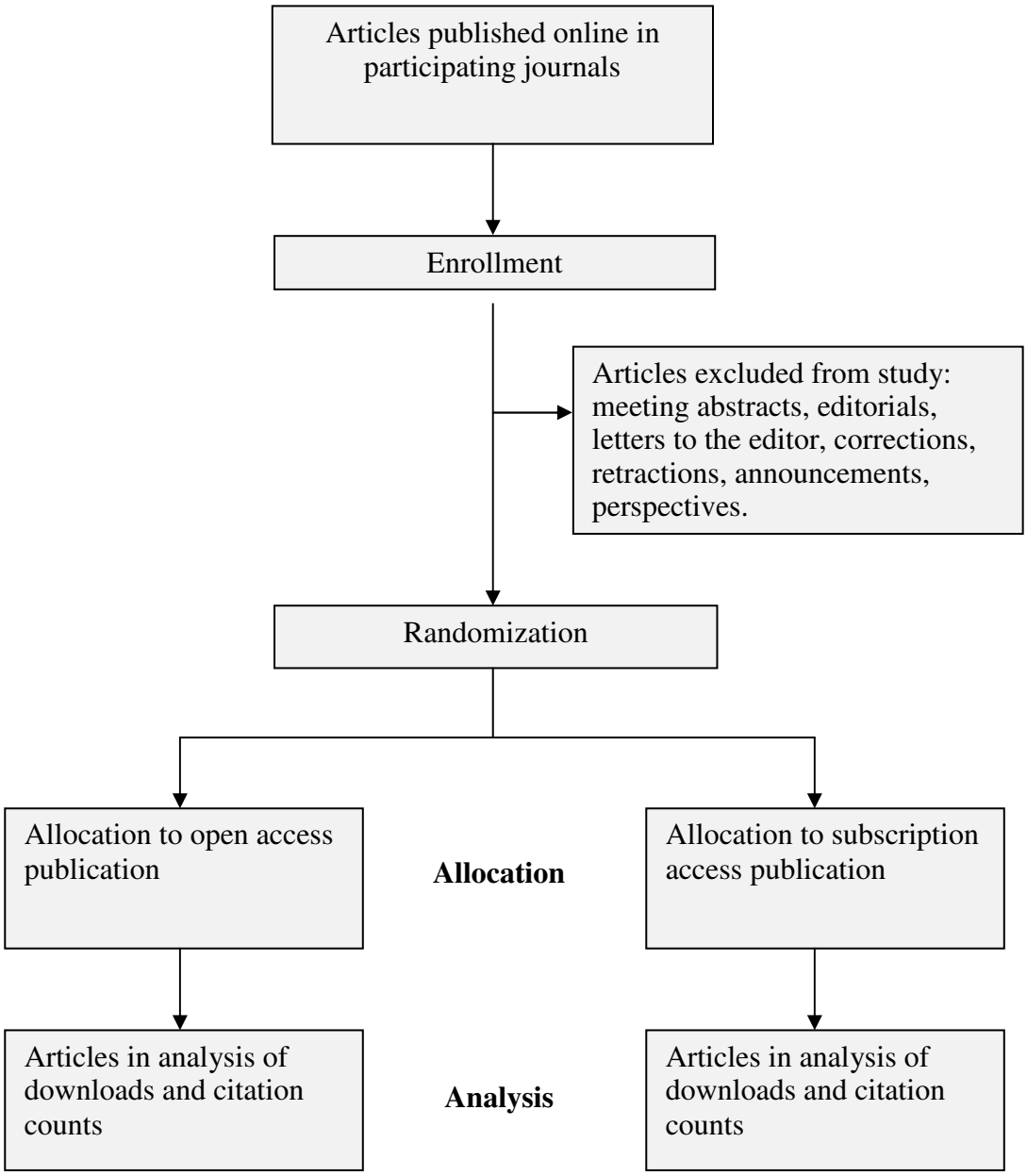


Figure 1. Flow of study data.

The statistical power of a study is based on the probability of making a Type II error, or in plain words, not detecting a significant difference between the treatment and control group when one really exists. When the probability of making such an error is set at 20% (i.e. $\beta = 0.2$), the statistical power of the study is therefore 80% (1-β) or 0.8.

Before calculating the sample size for each journal, the citation variance for each journal needed to be estimated. This was achieved by analyzing the citations to articles after two years in science journals and four years in social science and humanities journals.[9] The proposed number of treatment articles necessary for each journal (or groups of journals) was calculated in order to detect a 25% difference in citations between the treatment and the controlled cohorts.

The retrospective nature of previous studies did not help us to predict our expected difference in citations, although we assumed that it would be smaller than the 200-700% difference routinely reported in the literature (see Table 2). Considering that the literature has reported much greater differences in the citation advantage, we feel that our sample sizes are more than ample to detect differences if they exist. Sample sizes are presented in Table 4.

Our first participating publisher, the American Physiological Society, agreed to make 1 in 8 (15%) articles freely available. Based on a 0.7 standard deviation in *log* citations in the society's journals and a power of 0.8 to detect a significant difference (two-sided, P=0.05), 247 Open Access articles allowed us to detect significant differences of about 20%. These calculations were based on equal sample sizes and a two-sided test. Given that our subscription sample was much larger (n=1,372) and that we did not anticipate a negative effect as a result of the Open Access treatment,

[9] While the citation half-life of many social science and humanities can exceed ten years, our study is limited to four years.

these calculations are conservative.[10] Sample sizes for other publishers were based on similar calculations. In several cases, the publisher urged us to increase the treatment size in order to increase the statistical power of the tests – this was the case for *Genetics* and the *FASEB Journal*. In the case of the journal *Science*, the number of treatment articles was reduced due to concerns expressed by the publisher. It was necessary to aggregate several of the journals in the individual studies in order to achieve an adequate sample size. For example, we analyze all of the social science journals together (10 journals) and the humanities journals together (6 journals) because we simply do not have enough treatment subjects in each journal in order to make point estimates.

The intent of the analysis is not to make point estimates for individual journals, but to view these journals as samples from the larger literature. Considerable effort was made to include as many journals and publishers as possible, rather than focus on detailed analyses of individual titles.

Statistical Analysis

The analysis of the data involved several statistical techniques. Within the first year after publication, when the frequency of citations is low for most articles and many articles have yet to receive any citations, logistic regression is used to estimate how the Open Access treatment changes the *likelihood* to be cited. As the articles age and accrued more citations, the effect of the treatment on the frequency of citations is measured using multi-linear regression (MLR) and Negative Binomial Regression (NBR). MLR is a widely-known and understood technique; however, citation and usage data rarely conform to linear models and it is necessary to transform the data in

[10] Assuming a unidirectional (positive) effect would allow us to use a one-sided test, and thus move α from 0.05 to 0.10.

order to conform to the model assumptions. MLR also assumes independence of the data, an assumption which is technically violated in our dataset. Article downloads are correlated with each other, and the act of citing an article increases the probability that the cited article will be cited again in the future (Burrell, 2002; D.J.S. Price, 1976). While linear models may be able to approximate the data using transformation (often a simple log transformation is sufficient), they are often a bad fit when citation rate is low and the dataset includes a high proportion of zeros (Bensman, 1996; Burrell, 2003; Glanzel & Schubert, 1993; Leydesdorff & Bensman, 2006; Nadarajah & Kotz, 2007; Simon, 1955). Since the logarithm of zero is a logical impossibility, it is common to add 1 to every observation in the dataset prior to log transformation. Observations with 0's now become 1's and the entire dataset is right-shifted. This type of transformation is not a problem for the type of questions posed in our study: We are not interested in the value of the intercept of the regression line, but in the slope of the regression lines and the additional effects of each of the variables in our model. Right-shifting before transformation is thus a justifiable data treatment.

The Negative Binomial Regression (NBR) is based on a non-linear model, and as a result, no transformation is necessary for the dependent variables (citations and article downloads). The NBR model is ideally suited for count data (like the Poisson model), and can accommodate datasets with high-proportion of zeros (Hilbe, 2007). The Poisson model is based on the assumption that the mean is equal to the variance – an assumption that is rarely met in datasets. When the variance exceeds the mean (called "overdispersion") errors in point estimates and standard errors may occur. In addition, overdispersion may make a variable appear to be significant when it is not (Hilbe, 2007). The NBR model incorporates an overdispersion variable which makes it ideal for bibliometric analyses e.g. (Burrell, 1985, 2003, 2005; Davis et al., 2008; Evans, 2008; Evans & Reimer, 2009; Fowler & Aksnes, 2007; J. McDonald, 2007;

van Dalen & Henkens, 2001). To ensure that the statistical analysis provides consistent results, the dataset will be analyzed using both linear and negative binomial model techniques.

While we test for evidence of an access effect on article downloads and citations, evidence of non-uniform effects in the data will also be explored. For example, Open Access publishing may amplify the effects of higher-impact journals more than lower-impact journals or may have stronger effects in particular fields like medicine than fields like the humanities. Free access may increase the citations to review articles more than original research articles, or the effect of providing free access may be amplified by editorial highlights or by media press releases. In sum, one should not assume that the effect of Open Access (if one exists independently of other explanatory variables) acts uniformly on the literature. Relaxing this assumption will allow us to make qualified and conditional statements about the effects of Open Access as opposed to the generalized and unconditional hyperbole most frequently seen in the Open Access debate.

Methodological Limitations

While a randomized controlled trial removes many sources of possible bias, RCTs are not immune from biases that may affect the validity, reliability and generalizability of experiments. Below are described the limitations of our study, their potential effects on the results, and how we attempted to mitigate against their effects.

Journal Selection

The journals participating in this study (Table 4) represent some of the most prestigious titles in their respective fields, and in the case of one journal, *Science*, of

the entire science publishing domain. It is important to consider how these journals were selected, whether this suite of journals is representative of scientific publishing, and consequently, whether the results of this study are therefore generalizable.

The distribution of citations among papers is highly skewed, favoring a small percentage of articles published in an even smaller percentage of journals (Garfield, 1996; Hamilton, 1990; Ioannidis, 2006). Even within a journal, a small number of articles can account for the vast majority of citations to that journal (Seglen, 1992), with many papers (especially in the arts and humanities) remaining uncited after 5 years (Hamilton, 1991; Larivière et al., 2008). Larivière et al. (2008) illustrate that the percentage of uncited papers has been declining over time for the medical, natural, and social sciences, but not for the humanities. It is important, therefore, that the journals included in the study publish articles that have a chance of being cited within the time-frame of the study. Moreover, studying articles that attract few (if any) citations severely limits the kind of statistical analysis that can be done, especially when citation events are rare and unpredictable.

Secondly, previous studies studying the effect of access on citations have focused on prestigious journals. Antelman (2004) selected the ten most highly-ranked journals (as measured by their Impact Factor) for each discipline in her study. Eysenbach focused his study on one journal, the *Proceedings of the National Academy of Sciences of the United States* (PNAS), a multidisciplinary science journal that is comparable in scientific prestige with our top-tier journal, *Science*. While not the same suite of journals, our set of participating journals is not dissimilar to journals involved in other access-citation studies, increasing the comparability of our conclusions.

Moreover, our study is unique in that it is the first to study the effect of access on *readership* as well as on citations. Previous work by Davis and Price (2006) has

illustrated that publisher platforms can have a significant effect on user behavior and therefore on readership statistics. In order to isolate and measure the effect of access on readership, it was important to keep the publishing platform constant; hence, our study was limited to journals that publish on the HighWire Press platform.

Lastly, while most attention on the Open Access publication model has focused on the biomedical sciences, we were determined to include social sciences and humanities journals into the study in order to be able to generalize our findings beyond the biomedical literature to academic journal publishing in general. While the social sciences and humanities literatures show longer citation patterns, title choices were made under the believe that sufficient citation data would be available by the end of the study for meaningful statistical analyses.

Ascertainment Bias

Ascertainment bias can result when subjects know which treatment they are receiving (Jadad, 1998). In our case, we are conscious that labeling treatment articles may have an effect on author and reader behavior. The presence of a treatment indicator, such as the standard open padlock icon or the words "OPEN ACCESS" beside the article on a journal's table of contents may send a cue to the reader that the article should be downloaded and read. In addition, an author knowing that his article received the Open Access treatment might promote it differently to his community and to the general public. Both the American Physiological Society and American Heart Association employed the open padlock icon beside Open Access articles: the rest of the publishers in our study did not. In this study, we can look for evidence that the presence or absence of an Open Access indicator may result in differential behaviors in readership. Participating publishers in this study told us that the vast majority of

article requests do *not* originate from the Table of Contents page (the page that displays the padlocks), but from outside services such as Google, Medline, or directly from citation linking from other online articles. As a result, few readers ever see the padlock icon. For the remaining publishers who have not implemented the padlock icon (AAAS, Genetics, FASEB, Sage, Duke), there was no physical indication on the Table of Contents which articles received the Open Access treatment. Potential readers would either gain access to the freely accessible articles when they attempted to read them or would be prevented from accessing the subscription-based articles. At the time of the experiment, there was no indication in the literature indexing and referral services (such as Google or PubMed) which articles had received the Open Access treatment.

As previously described in the feasibility study section with the American Physiological Society, we sent out email notices to all corresponding authors of forthcoming articles alerting them of the study and providing an opportunity to opt-out. None of the respondents wished to opt-out. Moreover, the Institutional Review Board for Human Subjects at Cornell University determined that it was unnecessary to alert authors of the study. Consequently, we discontinued the practice of alerting authors in the full study in the hope that this would reduce possible ascertainment bias.

Expectation Bias

While readers may not be cognisant that a particular *article* may be freely-available – due to the lack of visible cues (such as the open green lock icon) – there may be access expectations for particular *journals*. For instance, a potential reader may not attempt to click through to an article in a journal to which her institution does not subscribe nor attempt to visit the journal in the first place. The result of this prior

access expectation may compromise the internal validity of our experiment, by setting up an *artificial environment* that does not mimic the true information seeking environment of the reader.

While all true experiments are artificial, that is, they never can replicate the exact conditions of the environment, I believe that our randomized controlled trial closely approximates the existing information landscape on several levels:

1) Articles were manipulated on the publisher's own website. This eliminates the need to create a separate publishing platform on which the experiment is conducted.
2) Many scientific journals already provide free access to older issues, so there are pre-existing expectations – even for non-subscribers – that some degree of free article access exists within subscription-access journals.
3) Conducting a reverse experiment in full Open Access journals (i.e. making a random selection of articles available only by subscription) is impossible since full Open Access journals do not employ subscription-access controls. Indeed, such an experiment (were it to be even possible to conduct) would be even more artificial than our present experiment.

In sum, our experiment closely approximates the true scientific publishing environment, and as a result, there is little concern that the design of the study should raise issues of internal validity.

Scope of Citation Data

Citation counts to the articles under observation were derived from ISI's *Web of Science (WoS)*. While this resource is considered a standard tool for collecting citations, it does not index all academic journals. The *Web of Science* focuses on indexing core titles in each discipline. For this reason, our citation counts should not be considered complete. The *Scopus* database (produced by Reed-Elsevier) is an alternate source of citation data and indexes approximately 18,000 journals compared to 10,000 by WoS. Unfortunately, Cornell University does not have a subscription to *Scopus*. Recent studies also report that these two datasets provide reliable and comparative citation data (Archambault, Campbell, Gingras, & Larivière, 2009; M. E. Falagas, Pitsouni, Malietzis, & Pappas, 2008). *Google Scholar* was also considered a source for citation data, although it has been criticised for its inadequate and often poorly updated data (M. E. Falagas et al., 2008). A possible future extension to this study may determine whether freely-accessible scientific articles attract more informal links from web pages, blog posts, and e-mail lists.

Access as a Precondition of Citation

Our study assumes that access to an article is a precondition of citation. In other words, we assume that authors have read what they cite. In reality, an author may cite directly from an abstract or copy a citation directly from another article without having read the referenced article. Studies of propagated errors in citations provide some evidence that some citations are merely copied from one reference list to another e.g. (Broadus, 1983; M.V. Simkin & Roychowdhury, 2005; M. V. Simkin & Roychowdhury, 2007). The result of this kind of citation behavior may attenuate any

observed access effect on citations.

Circumventing Formal Access Routes

By counting article downloads directly from the publisher's website, we ignore other avenues of access to the journal literature. A survey of Indian scientific authors revealed a high frequency of peer-to-peer sharing of published articles (Gaulé, 2009). Even the best Indian research library has sub-optimal access to the scientific journal literature. The Indian Institute of Science, for example, lacks access to one-third of the top biology journals. Gaulé reports that Indian researchers routinely request copies of articles directly from the corresponding author of a paper, or ask colleagues or former students located at institutions in Europe or the United States for copies. Most requests for copies were honored, and the strong sharing ethos in science (Merton, 1973) may help attenuate the effects of subscription access barriers.

Data Granularity

Our source of article download data does not allow us to make statements about *who* is requesting a copy of an article. The usage data has been digested and tabulated for us – we have no access to the raw usage logs. Consequently, if we find a difference in readership, we are unable to discern the source of the difference. Open Access may have discriminating effects based on location, country, or other socio-economic profiles. To date, only a few studies have explored the differential effect of Open Access on country location (Calver & Bradley, 2010; Evans & Reimer, 2009; Frandsen, 2009).

Changing Publishing Landscape

The landscape of information access is changing rapidly. Institutional self-archiving mandates, such as those implemented at Harvard and MIT are requiring authors to deposit versions of their articles in public repositories. Government policies, such as the NIH Public Access Policy, as well as two Federal Acts presently active at the time of writing (FRPAA and COMPETES) may require public deposit of research findings as a condition of federal funding. More publishers are offering Open Access publishing options for their authors and several research libraries have made publishing funds available for their researchers. At the same time, increased bandwidth and new peer-to-peer and collaborative technologies are making articles easier to share.

The results of this study will ultimately reflect the conditions of access during the time period the experiment was conducted and the specific findings may not remain reliable over time. However, it is my hope that this dissertation adds to the literature in two other ways: First, that it provides a new and rigorous methodological approach for evaluating changes in publishing practices, especially in cases where the system is complex, the causes are numerous, and the direction of causality is uncertain. Second, that the results of the dissertation direct us to a theory of information that helps us understand the underlying and dominant processes that drive readership and citation.

RESULTS

At the writing of this dissertation all the articles made freely available upon publication in the 36 journals under investigation have aged at least two years, allowing us to observe common trends.

The initial results of our prototype study focusing on 11 research journals published by the American Physiological Society and reported in 2008 (Davis et al., 2008) appear to be both robust and generalizable across journals and disciplines: *open access appears to result in more article downloads from a larger group of readers but no more citations*.

Readership

Figure 2 presents the differences (in percent) between the Open Access treatment articles and subscription-access articles for the 11 journals published by the American Physiological Society for the first year after article publication. Download and unique visitor counts were calculated with, and excluding, known robot activity. While total downloads for fulltext version of articles was more than 200% greater for those articles made freely accessible, nearly half of that increase could be explained by automated robot activity. Removing known robot activity from our analysis reduced the number of unique visitors by only one percent (from 37% to 36%), confirming that a small number of visitors to the journal were responsible for a remarkably large number of fulltext downloads. Because we wish to measure the effect of free access on human behavior, all future download analyses will be reported without robot activity.

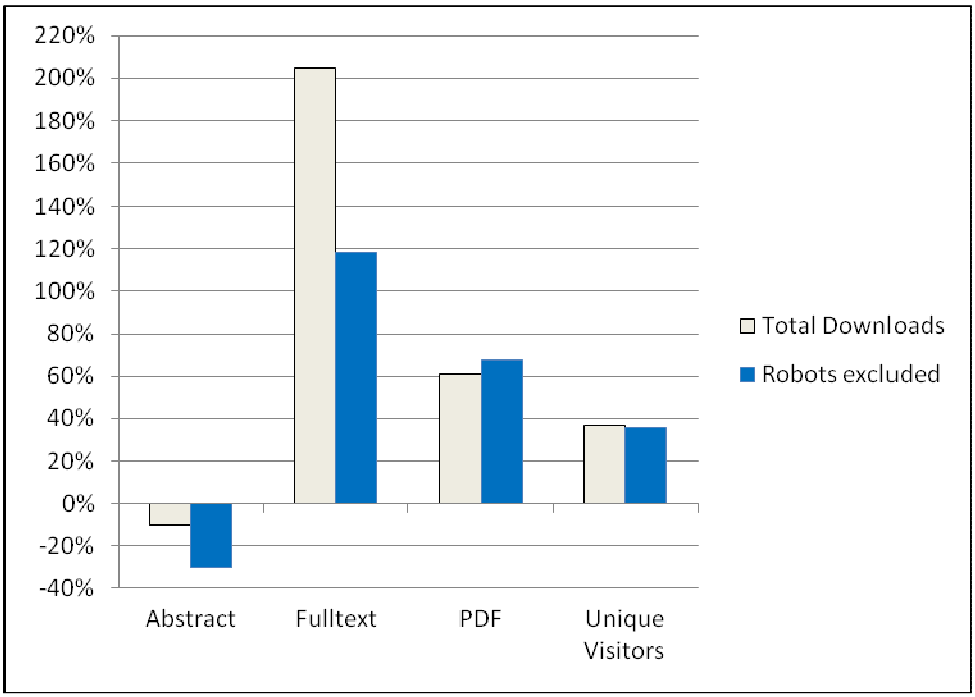


Figure 2. Percent increase in article downloads and unique visits to Open Access treatment articles compared to subscription-access articles published in 11 journals by the American Physiological Society. Data with and without known indexing robots are presented.

Figure 3 illustrates the increase in article downloads and visitors for the first year after publication for the 20 science journals included in our study. Social sciences and humanities journals were left out of this analysis since many of the articles do not include abstract or fulltext versions, and in some cases, "reference view" (a document with just the list of references) was added to fulltext counts. This made the interpretation of these data problematic and incomparable with the science journals.

Removing article downloads from known indexing robots and focusing on what may be best consider the results of human-based activity during the first year after publication, providing free access to the treatment articles resulted in a doubling

of fulltext (HTML) downloads on average (112%, Standard Error = ±3.0%), and to a lesser degree, a significant increase in the number of full image (PDF) downloads (59%, S.E. = ±2.7%). As measured by I.P. addresses, freely accessible articles received about a quarter more unique visitors (25%, S.E. = ±2.2%). Abstract views decreased for free articles (-23%, S.E. = ±2.5%), suggesting a reader preference for the full document, when available.

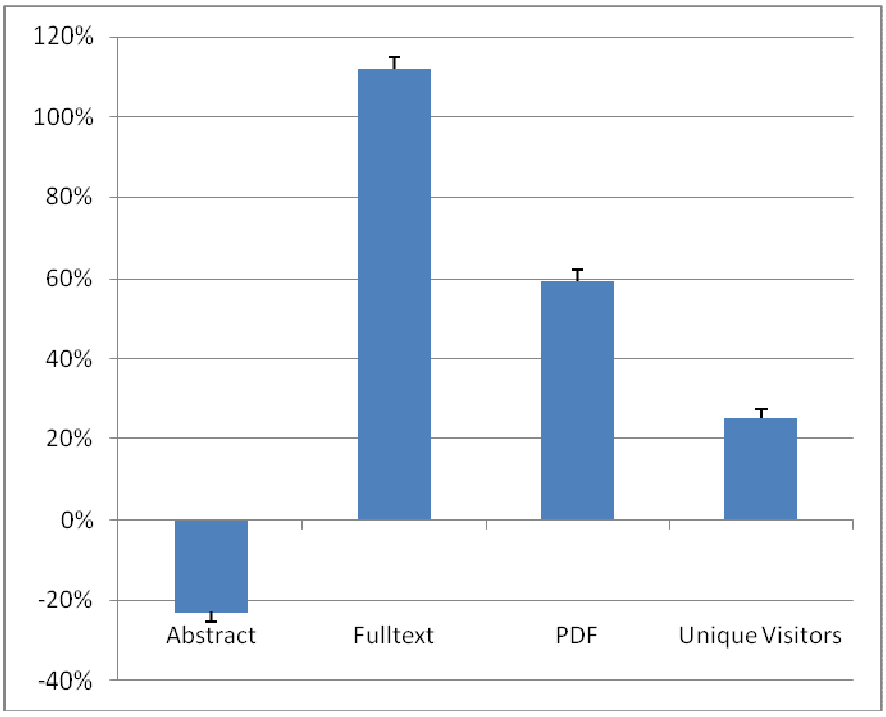


Figure 3. Percent increase in article downloads and unique visits after 1 year (± Standard Error) to articles made freely accessible upon publication. The figure represents the mean difference across 20 science journals controlling for journal as a fixed effect. Article downloads from known indexing robots were removed prior to analysis.

American Physiological Society

Figure 4 plots the monthly performance of American Physiological Society articles during the first two years. During the first 12 months after publication, articles in the Random Open Access treatment cohort consistently received more article downloads than their subscription-access cohort. Random OA articles outperformed subscription-access articles by a median of 20% in the first month after publication, increasing to a 50% by the 12th month, for a 12-month median difference of 39%. In the 13th month of publication, when all subscription-access articles became freely available, this difference disappeared, after which with both lines converged on a similar trajectory through to the 24th month after publication. If we succeeded in performing an unbiased random allocation of articles into both arms of the study, we should expect that articles would follow a similar trajectory when the access treatment becomes normalized over both cohorts.

In contrast, articles observed in *Physiological Genomics* (a journal that we reserved for observation purposes) displayed several noteworthy differences (Figure 5). In this journal, authors could purchase immediate free access for their articles and we label this group "Author Choice OA," a distinction from the previous figure in which the selection of articles receiving free access was made randomly.

Author-selected OA articles received a median of 55% more PDF downloads in their first month, increasing steadily to 72% by month 12, for a median difference of 70%. After month 12, however, when all subscription-access articles were made freely-accessible, Author Choice OA articles still outperformed Random OA articles for the next 12 months, maintaining nearly a 30% difference over time.

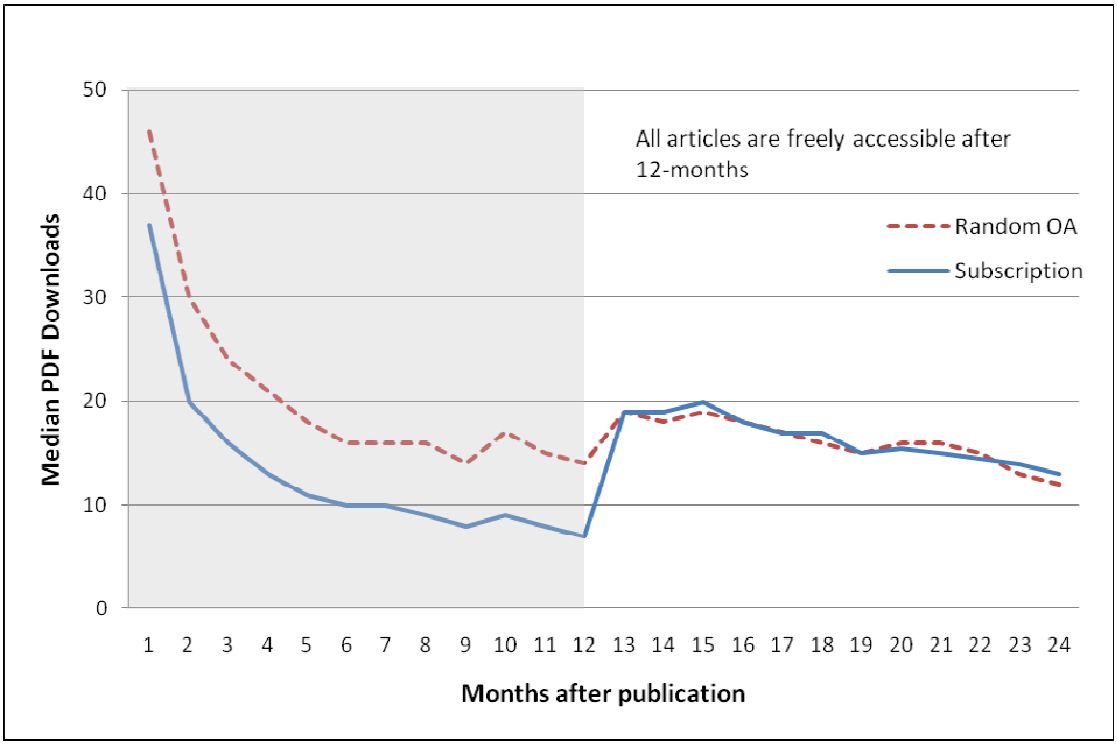


Figure 4. Median PDF downloads for Random Open Access articles (n=247) and Subscription access articles (n=1372) by month after publication for 11 journals published by the American Physiological Society.

The protracted difference in article downloads for Author-Choice articles published in *Physiological Genomics* suggests that there is something different about these articles from their access status. Indeed, Author-Choice articles are 1.5 pages longer on average than subscription-access articles (11.2 versus 9.8 respectively) and also list one more author on average (7.6 versus 6.6). Since payment to the publisher is required for immediate free access status, we may imagine that willingness (or ability) to pay the publishing fees establishes a barrier to entry that may differentiate these two groups of articles. I explore this issue in more detail in the citation analysis section of this dissertation.

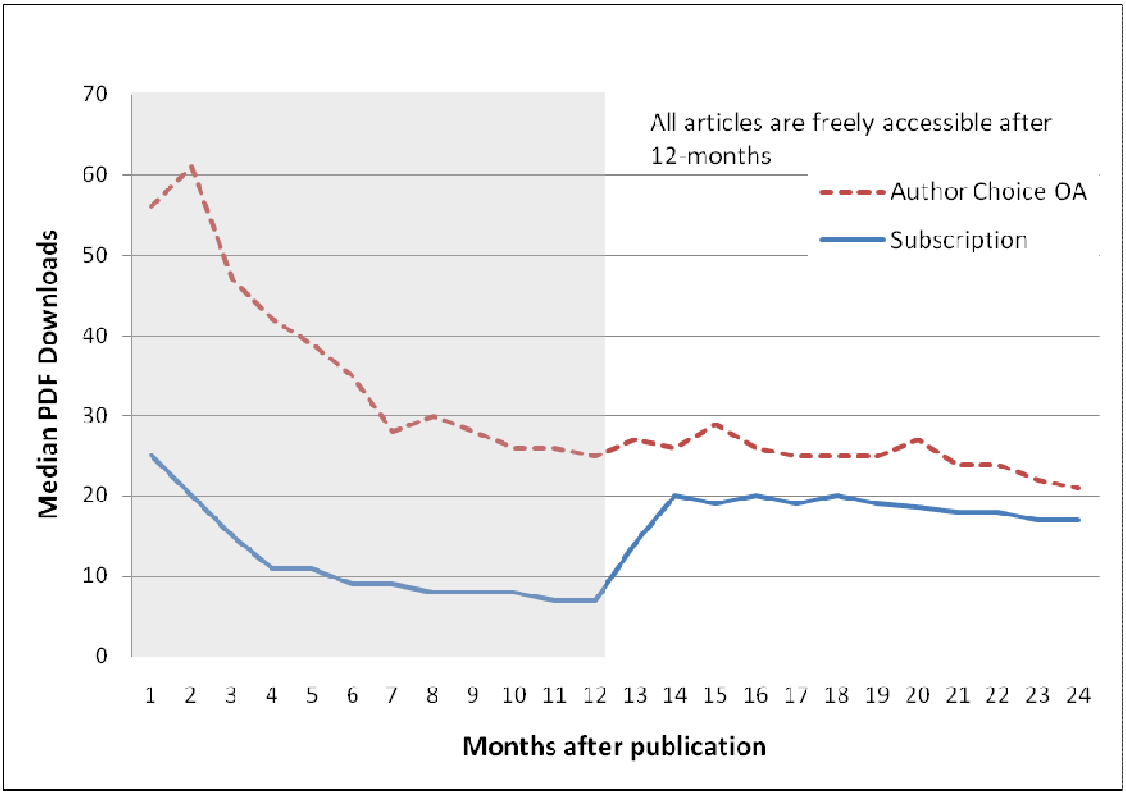


Figure 5. Median PDF downloads by month after publication for *Physiological Genomics* comparing the performances of Author Choice Open Access (n=94) articles with subscription-access articles (n=627).

American Heart Association

In addition to the 101 randomly-chosen articles made freely available in the five AHA journals under investigation, editors simultaneously made 32 Editor's Pick articles freely available. Four of these Editor's Pick articles overlapped our treatment articles and were classified as Editor's Pick for univariate analysis. In our regression analysis articles were classified as both Editor's Pick and Random OA. Performed this way, Editor's Pick variable represents the effect of editorial selection on top of free access allowing us to separate the effect of *access* from the effect of *editorial selection*. Article downloads (Abstract, HTML, PDF) and unique visits for Random

OA, Editor's Pick OA and subscription-access cohorts are presented in Figure 6.

For each of the reporting variables, Editor's Pick articles (dotted line) outperformed both Random Open Access articles (dashed line) and subscription-access articles (solid line) throughout the first two years after publication. In Figure 6 we see a familiar pattern: the effect of randomly-selected OA articles disappears (relative to subscription-access articles) after the first year when all articles become freely accessible. These results are consistent with the APS study. The advantage of Editor's Pick articles, however, continues to persist in year two. Even controlling for article characteristics, editorial selection continues to have a positive effect on article views in year two, increasing Abstract views by 22%, Fulltext views by 25%, PDF views by 23% and unique visitors by 23% (Table 5).

Table 5. Effect of free access on article downloads and visitors in Year 1 and Year 2 after Publication. Effects (± 95% Confidence Interval) are estimated controlling for journal and article characteristics (page length, number of authors, review, press release, and CME component) and are measured against subscription-access articles. All articles become freely-accessible after 12 months.

	Random Free Access	**Editor's Pick (in addition to free access effect)**
Abstract		
1-12mo	-14% (-22% to -5%)	40% (16% to 60%)
13-24mo	1% (-10% to 14%)	22% (-1% to 51%)
Full text (HTML)		
1-12mo	74% (56% to 94%)	131% (89% to 181%)
13-24mo	0% (-10% to 11%)	25% (3% to 51%)
PDF		
1-12mo	55% (39% to 74%)	88% (53% to 131%)
13-24mo	-4% (-15% to 8%)	23% (-1% to 52%)
Unique Visitors		
1-12mo	27% (15% to 39%)	79% (51% to 111%)
13-24mo	-3% (-12% to 7%)	23% (2% to 47%)

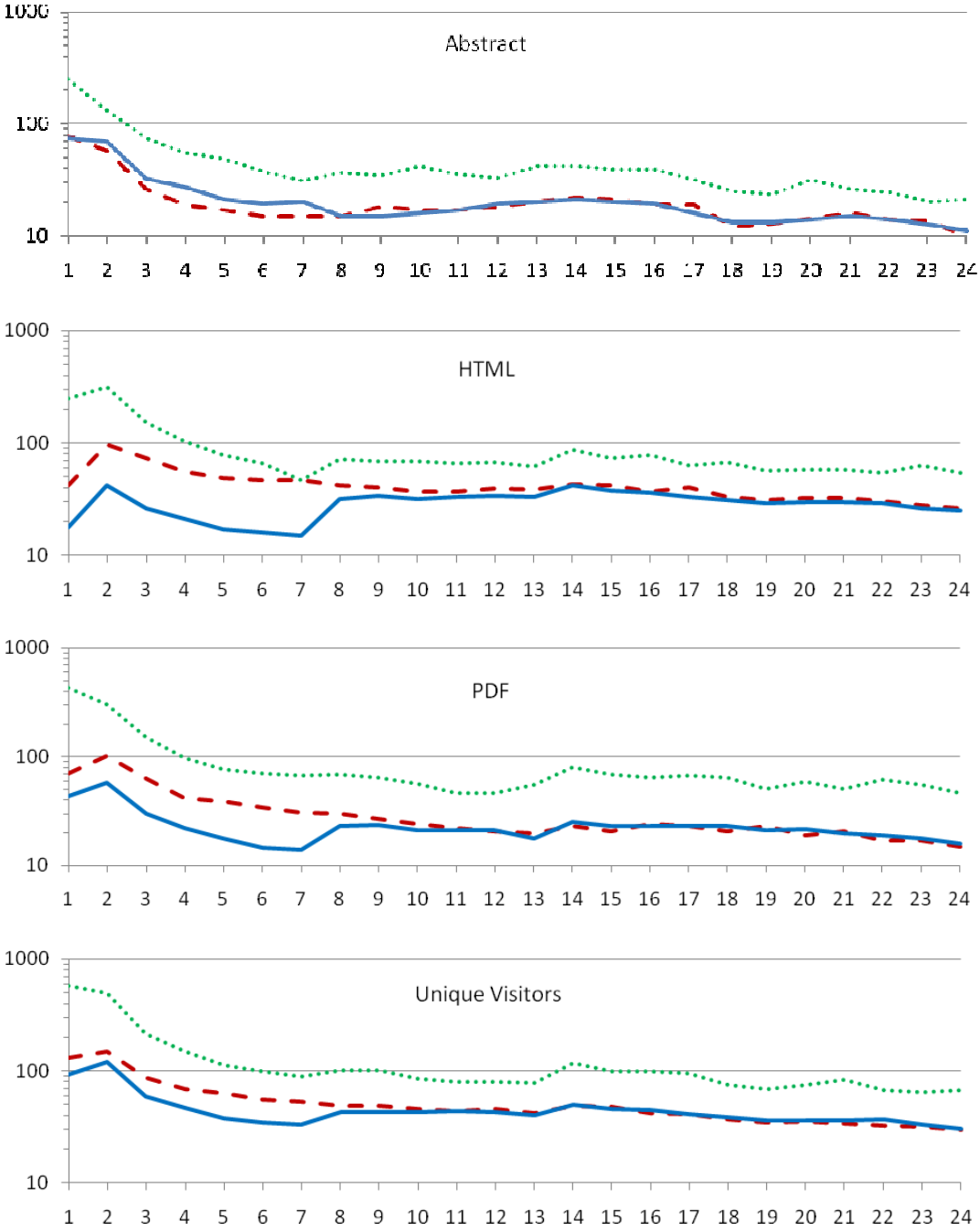


Figure 6. Median Abstract, HTML and PDF views and Unique Visitors for Editor's Pick articles (····), Random Open Access articles (- -), and Subscription-access articles (—) published by the American Heart Association during the first 24 months after publication. All articles become freely available 12 months after publication.

In our American Heart Association study, 12 articles (or 2.5% of 488) received press-releases by the AHA. While this number is rather small, it is worthwhile to investigate whether press releases are associated with greater readership (as measured by article downloads) than other articles. Figure 7 illustrates that press-releases are indeed associated with a positive effect on readership within the first two years after publication. The data are indeed noisy due to the small sample size, although two general patterns are apparent from the graph. First, the effect becomes apparent in the second month of publication, suggesting some latency of effect. Secondly, the effect does not appear to be ephemeral in nature, but continues to be relatively consistent throughout the first two years representing about a 60% increase for press-released articles. This consistency suggests that the press-release alone is not responsible for increased readership but that press-releases may be associated with articles of greater scientific interest or importance.

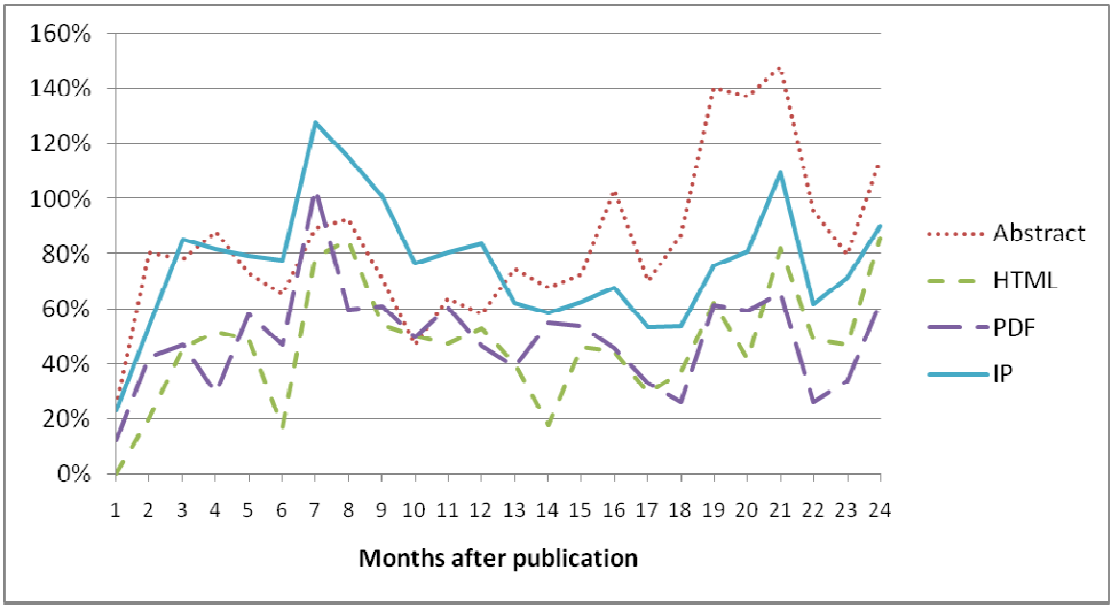


Figure 7. Effect of press releases (n=12) on article downloads and unique visits for articles published in five journals by the American Heart Association. Regression controls for Journal, OA, Editor's Pick, Review, and the interaction between Press Release and OA.

In addition to articles receiving press-releases, 16 were designated as Continuing Medical Education (CME) articles. These articles must be read by physicians participating in continuing education programs and are associated with a short lesson and quiz. AHA Continuing Medical Education lessons are active for just one year, after which they expire.

Table 6 reports a regression analysis of the effects of press releases and CMEs on article downloads during the first and second year after publication. Consistent with Figure 7 (presented above), press release appears to have a positive, significant and sustained effect on article downloads and visits in both years one and two. Likewise, articles with CME components show a positive association with article downloads and visits, but only in year 1; the estimate of effect (while still positive) becomes non-significant in year two.

Table 6. Effect (± 95% Confidence Interval) of Press Release and Continuing Medical Education (CME) on Article Downloads. Regression controls for Journal, OA, Editor's Pick, Review, and the interaction between Press Release and OA.

	Press Release (n=12)	**CME** (n=16)
Abstract		
1-12mo	54% (19% to 100%)	66% (30% to 111%)
13-24mo	91% (40% to 163%)	13% (-16% to 52%)
Full text (HTML)		
1-12mo	38% (1% to 89%)	23% (-6% to 62%)
13-24mo	48% (10% to 99%)	9% (-16% to 41%)
PDF		
1-12mo	41% (2% to 94%)	42% (8% to 88%)
13-24mo	45% (3% to 105%)	28% (-5% to 73%)
Unique Visitors		
1-12mo	66% (28% to 116%)	34% (7% to 68%)
13-24mo	69% (28% to 123%)	13% (-12% to 44%)

Science Magazine

There are many variables apart from *access* that affect the frequency of article downloads and *Science Magazine* provides us with an interesting case. *Science* is a multidisciplinary sciences journal that covers a wide variety of topics in the life and medical sciences, physical sciences, and to a lesser degree, the social sciences and mathematics (Table 7 lists the subject classifications and frequencies in our dataset). In each issue of the journal, editors publish short summaries highlighting noteworthy articles in the current issue,[11] and important time-sensitive articles are released early (before print) in the form called "Science Express".[12] These article characteristics, among others such as article page length, number of authors and type of article, can be analyzed simultaneously in a single regression model.

In our dataset of 393 *Science* articles, editors employed 44 different subject classifications, many only once (Table 7). There was clearly not enough data in each category to calculate reliable point estimates for each subject in a fixed effects model. Moreover, it was difficult to construct a high-level classification structure (i.e. physical sciences, life sciences, social sciences) to the articles as many papers fit into two or more categories. For example, an article in behavioral economics may be classified as social sciences because its application to group decision-making behavior but also as life sciences because its researchers used functional Magnetic Resonance Imaging (fMRI) to scan the brains of its subjects. Furthermore, the readership and citation patterns within a broader rubric may be as diverse as the differences between them. For example, ecology papers show more resemblance to the empirical social

[11] Highlighted articles are reviewed in an editorial called "Editor's Choice." Unlike the American Heart Association, these articles are not made freely available. There is no indication on the article nor on the table of contents page than an article was selected as an "Editor's Choice."

[12] See Science Express http://www.sciencemag.org/sciencexpress/

sciences than to biological chemistry. As a result, the subject category was used as a *random effect* in the regression model. As a random effect, we assume that our group of subject categories represents a *sample* from the larger population of subject categories; the goal of the regression is to estimate the effect of *subject* and not the effect of *individual subjects*. In statistical parlance, considering subject category to be a random effect, we can estimate the total variance component for subject category as a whole rather than estimating the variance component for each subject. This approach allowed us to control for known, large subjects effects in our analysis. Combining a random effect with fixed effects in the same regression equation is called a *mixed-effects model*.

Table 7. Subject classification of study articles in *Science Magazine*.

Subject Category	**No.**	**Subject Category**	**No.**
AIDS	1	GENOMICS	1
ANTHROPOLOGY	3	GEOCHEMISTRY	8
APPLIED PHYSICS	16	GEOPHYSICS	11
ARCHAEOLOGY	3	IMMUNOLOGY	15
ASTRONOMY	7	MATERIALS SCIENCE	15
ASTROPHYSICS	3	MATHEMATICS	1
ATMOSPHERIC SCIENCE	6	MEDICINE	12
BEHAVIOR	2	MICROBIOLOGY	7
BIOCHEMISTRY	17	MOLECULAR BIOLOGY	20
BIOPHYSICS	1	NEUROSCIENCE	23
CELL BIOLOGY	13	OCEAN SCIENCE	7
CELL SIGNALING	2	PALEOCLIMATE	1
CHEMISTRY	28	PALEONTOLOGY	9
CLIMATE CHANGE	9	PHYSICS	18
COMPUTER SCIENCE	2	PHYSIOLOGY	1
DECISION-MAKING	4	PLANETARY SCIENCE	10
DEVELOPMENTAL BIOLOGY	7	PLANT SCIENCE	14
ECOLOGY	19	PSYCHOLOGY	11
ECONOMICS	1	SOCIOLOGY	2
EPIDEMIOLOGY	1	STRUCTURAL BIOLOGY	10
EVOLUTION	22	SYSTEMS BIOLOGY	1
GENETICS	26	VIROLOGY	3
GENOMICS	1		
		Total	393

Table 8 presents the results of the regression model. Since the dependent variable (fulltext downloads) required logarithmic transformation to address the requirement of normality, the effect estimates are reported as a multiplicative effect; thus for example, 1.50 represents a 50% increase in article downloads. Our regression model could explain 58% of the variance in fulltext article downloads (RSq=0.58) with nearly one-half (49%) of this variance explained by article subject classification alone, underscoring the importance of subject classification in the model. The mean response of the dependent variable was 7.48. Exponentiating this response to arrive at mean fulltext downloads ($e^{7.48}$) we get an average of 1,772 fulltext article downloads during the first year after publication. There was some correlation in the dataset, for example, with review articles tending to be longer than original articles, although a VIF (Variance Inflation Factor) analysis revealed that there was little correlation in the dataset and that we should not be worried about multicollinearity.[13]

Holding all other model variables constant, the Open Access treatment had the largest single effect on fulltext downloads during the first year following publication, increasing fulltext downloads by 69% on average (95% C.I. 42% to 200%, $p<0.001$). *Ceteris paribus*, being a review article increased expected fulltext downloads by 51% (95% C.I. 15% to 98%, $p=0.003$). Self-archived articles had no significant effect on fulltext downloads: Articles that were found freely-available on the Internet received 2% fewer fulltext downloads (point estimate=0.98), although the confidence interval for this estimate ranged between -20% and +20%. We should note that we were able to find only 36 self-archived articles, two of which also received the Open Access treatment. Due to the severely limited sample size of self-archived articles, we should

[13] Multicollinearity can occur when there is strong correlation between two or more variables in the data. The result is an unstable regression model, with small changes in the data often resulting in wild and unpredictable changes in the estimates of the coefficients.

be cautious with over generalizing these results.

Articles listing more authors and longer articles (as measured in pages) received more fulltext downloads. For each additional (log) author, article downloads increased by about 12% (95% C.I. 3% to 21%, p=0.006), and for each additional (log) page, fulltext downloads increased by 15%, although this estimate was not statistically significant (95% C.I. -1% to 35%, p=0.072). Being a Science Express article did not result in more article downloads, although our dataset only counts the usage of the article *after* it was formally published – we miss initial usage of the article when it first appeared in Science Express. Finally, being highlighted by an editor ("Issue Highlights") appears to have little effect on fulltext article downloads (1%, 95% C.I -12% to 16%, p=0.861).

Table 8. Multiplicative effect on fulltext (HTML) downloads for first 12 months after publication in *Science Magazine*.

Fixed Effects	**Estimate**	**Lower 95% C.I.**	**Upper 95% C.I.**	**P-value**
Intercept	-	-	-	
Open Access	1.69	1.42	2.00	<.0001
Authors†	1.12	1.03	1.21	0.006
Review	1.51	1.15	1.98	0.003
Pages†	1.15	0.99	1.35	0.072
Self-archived	0.98	0.80	1.20	0.840
Science Express	1.07	0.93	1.23	0.341
Issue Highlights	1.01	0.88	1.16	0.861
Random Effects	Variance Component Estimate	Lower 95% C.I.	Upper 95% C.I.	Percent of Total Variance
Subject	1.30	1.13	1.49	49%

Notes:
† Log transformed continuous variable. Other variables are categorical and take the value of 1 or 0
R-sq model= 0.58; Mean of (log) response= 7.48; n=387

In light of reporting a strong, positive editorial effect on article downloads in American Heart Association journals, we did not find a similar effect in *Science*. In explaining the discrepancies, we should consider several differences: First, AHA's "Editor's Pick" articles are displayed prominently on the home page of the journal, are given priority order in the listing of articles, and are accompanied by an icon stating that the article is free.[14] An interested reader need only click on the title of the article and be connected with the fulltext. In comparison, *Science Magazine* includes a contextual summary of the highlighted articles in a separate section of the magazine called "Issue Highlights." Articles listed in Issue Highlights are not free and are not indicated elsewhere in the journal that they received an editorial highlight. In sum, the effect of editorial decision making may be more to do with signaling editorial choice to readers than the editorial choice itself.

Summary of Readership Analysis

In summarizing the effect of free access on article readership (as measured by downloads and unique visitors), we may make the following statement: Free access to the scientific literature results in an increase of article downloads from a larger group of visitors compared to a similar group of subscription-access articles. As a result of these general findings, we may reject our first null hypothesis that free access to scientific articles does not increase article downloads. In addition, we may make several additional observations:

1) Free access to the scientific literature has differential effects on format preference, with fulltext views showing the largest effect and abstract views showing a negative effect.

[14] For example, see Hypertension http://hyper.ahajournals.org/

2) The greatest contribution to fulltext downloads was attributed to indexing robots and not human action.

3) These download effects disappear when access conditions are equal.

4) There is a large article download advantage for author-sponsored Open Access articles over subscription-access articles and this differential persists when access conditions between the two cohorts are made equal.

5) There is a large article download advantage for Editor-selected Open Access articles over subscription-access and author-sponsored Open Access articles and this differential persists when access conditions between these cohorts are made equal.

6) Article characteristics (apart from access status) explain article download patterns.

Citations

Articles accrue citations at different rates. Some of the articles in our study have received several hundred citations by the end of their second year while others (especially in the humanities) have remained uncited. Whereas various authors have claimed evidence that free access speeds up the citation process – either through Open Access publishing (Eysenbach, 2006; ISI, 2004) or through self-arching (Henneken et al., 2006; Kurtz et al., 2005; Kurtz & Henneken, 2007; Moed, 2007) – these claims have not been verified in a randomized controlled trial, leaving open the possibility that the early citation effect may be caused by unobserved variables or through confounding.

Likelihood of Being Cited

Table 9 presents the frequency and likelihood of being cited 12, 18 and 24 months after publication. As an entire group 74.1%, 86.9% and 92.0% of articles were cited at 12, 18 and 24 months respectively. These figures however differ between subfields. Grouping our journals into five subgroups (Medical[15], Life Sciences[16], Multidisciplinary Sciences[17], Social Sciences[18], and Humanities[19]) we observe that the frequency of first citation differs dramatically amongst these groups. Nearly all of the articles in our Medical subgroup (99.4%) were cited by the end of 2 years. This figure was similar for the Life Sciences (95.5%) and the Multidisciplinary Sciences (99.7%). For the Social Sciences, however, just over half (59.4%) were cited by the end of the second year; for the Humanities, this figure was just over one-third (36.8%).

A logistic regression was performed to determine whether the Open Access treatment increased the odds of being cited over time. The analysis was run for the entire group of journals and then for each subset, controlling for journal as a fixed effect in each regression model. It was not necessary to run the analysis on journal sets in which nearly all articles were cited at least once during the specified timeframe.

Table 9 presents the Odds Ratio (O.R.), which is a ratio of the estimated probability of being cited between the Open Access cohort and Subscription cohort. An odds ratio of 1.0 is interpreted as no difference between the two groups. The Odds Ratio for Medical articles at 12 months was 1.21, meaning a 21% increase in the odds

[15] Medical included the five journals published by the American Heart Association plus *Neuro-Oncology*, published by Duke University Press.
[16] Life Sciences included the *FASEB Journal*, *Genetics* and the 11 journals published by the American Physiological Society.
[17] Multidisciplinary Sciences included just one journal, *Science Magazine*.
[18] Social Sciences included the 10 journals published by Sage Publications.
[19] Humanities included 6 journals published by Duke University Press.

of treatment articles being cited, although this estimate was not statistically significant (p=0.55).

Table 9. Frequency and likelihood of being cited 12, 18 and 24 months after publication.

	% Articles Cited	Odds Ratio (Open Access /Subscription)	ChiSq	P>ChiSq
All journals				
12 mo	74.1	0.96	0.11	0.74
18 mo	86.9	1.23	1.95	0.16
24 mo	92.0	0.98	0.02	0.90
Medical				
12 mo	85.2	1.21	0.35	0.55
18 mo	95.3	-	-	-
24 mo	99.4	-	-	-
Life Sciences				
12 mo	74.5	0.87	1.10	0.29
18 mo	89.7	-	-	-
24 mo	94.5	-	-	-
Multidisc. Sciences				
12 mo	97.5	1.26	0.05	0.83
18 mo	99.5	-	-	-
24 mo	99.7	-	-	-
Social Sciences				
12 mo	35.6	1.36	1.19	0.28
18 mo	51.5	1.21	0.48	0.49
24 mo	59.4	0.85	0.48	0.49
Humanities				
12 mo	10.0	0.43	0.96	0.33
18 mo	21.3	1.09	0.02	0.88
24 mo	36.8	1.02	0.00	0.97

The Odds Ratios for the entire journal group, as well as each subgroup, displayed no general pattern; some were above 1.0 while others were below 1.0. All Chi Square tests for statistical significance failed to reject the null hypothesis of no difference. In

sum, there is a lack of evidence in our study that free access to scientific articles leads to earlier citations.

Frequency of Citations

If our Open Access treatment articles do not appear to accrue their first citations earlier than subscription-access articles, do they accrue more citations over time? To answer this question, we need to apply a model that takes into consideration the *frequency* of citation. Statistically, we need to analyze citations as a continuous variable instead of a categorical one. For simplicity and for easier interpretation of results, I employed several linear models for the analysis. The dependent variables are the number of citations accrued by each article at 12, 18 and 24 months after publication. For each journal (or group of journals), I built a regression model estimates the effect of the Open Access treatment independently from other potential predictive variables (Table 10). For groups of journals, the journal variable was used as a *random effect*, which, as explained earlier, is used for estimating its contribution to the overall variance of the model and not for providing point estimates for each journal. In the case of *Science Magazine*, Subject category was used as a random effect.

In a randomized controlled trial, it is not necessary to control for other explanatory variables. If the randomization process was successful, each of the cohorts under investigation should be similar with each other in all respects at the commencement of the experiment, with the treatment being the only difference between the two. Elaborate statistical models controlling for possible bias between the two arms of the study are not necessary – the setup of the experiment allows for a simple comparison between the two groups to be made by the researcher. While such

a simple analysis is tempting, there are reasons for building a more elaborate model:

Controlling for exogenous processes. While we have been manipulating access conditions directly on the publisher's journals websites, there are some access conditions that cannot be controlled by the researcher. For example, an author may make a copy of an article freely available from an institutional repository, departmental, lab or personal website. This is referred as "self-archiving." The presence of these free copies may result in *attenuating* the effect of the free access treatment made by the researcher leading to an underestimate of its true effect. Moreover, if self-archiving were not made uniformly between the two arms of the study, experimental bias may be present.

Increasing experimental precision. Controlling statistically for known predictors of readership and citations increases the precision of the estimate of the treatment effect. It does this by explaining some of the variance left over in the regression model that cannot be explained by the treatment alone – the *residual error* or ε. In a simple model with only the dependent variable (number of citations) and one independent variable (Open Access), any other contributor that helps to predict citations is folded into the model's residual error. When other independent variables are added into the model, the residual error is reduced and the result is a narrower confidence interval surrounding the treatment estimate. For example, review articles tend to receive many more downloads and citations than original articles. By controlling for known differences between these two types of articles, we explain some of the residual error in the model that cannot be explained by the treatment alone. Adding explanatory variables to a regression model should not change the point estimate of the treatment effect, since we assume that the randomization process

leads to similar, unbiased distributions of variables across both arms of the study. Nevertheless, it should reduce the confidence interval associated with that point estimate, leading to greater precision in reporting.

Giving context to the treatment effect. Lastly, the addition of known explanatory variables to the regression models provides context to the interpretation of the treatment effect. Given a large enough dataset, even small differences between two groups may be considered statistically significant but have little practical significance in context of other explanatory variables. For example, the Open Access treatment may have a statistically observable effect, although its effect may be small compared to other contributing effects.

Model Building

In building the regression models, several variables required logarithmic transformation to adhere to the assumption of normality – a necessary condition for linear regression models. These independent variables were the number of authors, number of article pages, and number of references. The dependent variables (number of citations) were also log transformed to adhere to the normality assumption. Since some articles had received zero citations (and the logarithm of zero is a logical impossibility), 1 was added to each value before transformation. Thus 0's became 1's and the logarithm of 1 is zero. The effect of adding a constant to the dependent variables raises the intercept of the regression line; however, we are not interested in the intercept but in the contribution of each explanatory variable in the overall model. As a result, the transformation technique should not pose a concern for the interpretation of each variable.

There are several categorical variables in the regression models: Self-archived (indicated as a 1 when a PDF copy of the final manuscript or publisher's version of the article was found on a non-publisher website and zero when otherwise); Review article (1=review, 0=other); Cover (1=article was featured on the cover of the journal issue, 0=not); Press release (1=article was featured in a press-release by the publisher or society, 0=not); Editor's Pick (1=article was made freely available by editor, 0=not, AHA journals only); Data supplement (1=article is accompanied with a data supplement, 0=not, AHA journals only); Continuing Medical Education (1=article has a CME component, 0=not, AHA journals only); Editor's Choice (1=article is highlighted by editor, 0=not, *Science* only); Issue Highlights (1=article highlighted by editor, 0=not, *Science* only); Science Express (1=article was released ahead of print, 0=not, *Science* only). A detailed comparison of each regression model may be found in Table 10.

For each regression model, I looked for evidence of poor fit. Lack of fit tests were performed, and errant data points that may have high leverage on the model were investigated. The analyses were performed on all observations in the dataset. While several data points were unusually large, often 3 standard deviations beyond the mean, none of the observations were the result of recording error. For this reason, there were no compelling reasons to omit any of the observations from the analysis.[20] Variance Inflation Factors (VIF) revealed that there was little correlation in our dataset and that we should not be worried about multicollinearity. Residual error plots were screened for evidence of a poor model fit and for violation of model assumptions.

For the humanities group of journals, citation rates were generally low and there was a high frequency of zeros in the dataset. To validate the linear model

[20] Omitting extreme outliers is a convenient way to reduce the variance associated with point estimates.

results, I also performed a Generalized Linear Regression using a Poisson distribution and log link function, testing for overdispersion and estimating its size.[21] Since the results were similar to the linear model, and because my goal is to make general statements about the article dataset as a whole without excluding the humanities, I employed a single linear model for the entire dataset. All analyses were performed using JMP 8.0, a statistical software produced by SAS.

Regression Results

Articles randomly selected for immediate free access showed no significant citation effect (positive or negative) 12, 18 and 24 months after publication. Table 11 displays the point estimate for the Open Access treatment along with 95% Confidence Intervals. The data are also presented in a Forest plot (Figure 8), which gives a visual display of the data in a single graphic. Used routinely for reporting meta-analyses in the medical sciences, the purpose of a forest plot is to visually describe the variation between similar studies and an estimate of the overall effect (Bax et al., 2009; Lewis & Clarke, 2001; Schriger, Altman, Vetter, Heafner, & Moher). Forest plots should include the effect estimate for each study along with a corresponding confidence interval (Liberati et al., 2009). The vertical axis in our forest plot represents the multiplicative effect of the treatment on article citations taken at 12, 18 and 24 months after publication; consequently, estimates above the line represent positive effects and estimates below the line represent negative effects. The associated 95% confidence interval associated with each estimate tests whether the effect is significantly different than zero. If the confidence interval includes 1.0 the estimate is non-significant at the α=0.05 level.

[21] JMP software does not perform Negative Binomial Regression (NBR). This technique makes the regression model similar to the NBR model.

Table 10. Regression model used for citation analysis. The dependent variables were total (log) citations at 12, 18 and 24 months.

APS	AHA	FASEB
Open Access (treatment)	Open Access (treatment)	Open Access (treatment)
Self-archived ‡	Editor's Pick	Self-archived ‡
Authors†	Self-archived ‡	Authors†
Pages†	Authors†	Pages†
References†	Pages†	References†
Review article	References†	Review article
Cover	Review article	Press release
Press release	Cover	
Journal*	Press release	
	Continuing Medical Education (CME)	
	Data supplement	
	Journal*	
Genetics	**Science**	**Neuro-Oncology**
Open Access (treatment)	Open Access (treatment)	Open Access (treatment)
Self-archived ‡	Self-archived ‡	Self-archived ‡
Authors†	Authors†	Authors†
Pages†	Pages†	Pages†
References†	References†	References†
Review article	Review article	Review article
	Subject category*	
	Editor's Choice	
	Issue Highlights	
	Science Express	
Sage Publications	**Duke University Press**	**All journals**
Open Access (treatment)	Open Access (treatment)	Open Access (treatment)
Self-archived ‡	Self-archived ‡	Self-archived ‡
Authors†	Authors†	Authors†
Pages†	Pages†	Pages†
References†	References†	References†
Review article	Review article	Review article
Journal*	Journal*	Journal*

Notes:
† The natural logarithm of these variables was used.
‡ A self-archived article indicates when a free copy of the article or final manuscript could be found on the Internet.
Journal and Subject category (*Science*) was used as a random effect

For example, the treatment articles published in American Physiological Society journals received 1% fewer citations, 2% more citations, and 3% more citations, on average, 12, 18 and 24 months after publication. In all three instances, the 95% confidence intervals contained 1.0, meaning that these estimates were not significantly different than zero.

Generally, point estimates from each of the journals (or journal groups) were small and consistent over the observation period. While some of the groups displayed positive citation effects others displayed negative effects and there was no discernable pattern to the data. *Neuro-Oncology*, a medical journal published by Duke University Press was reported outside of the rest of the Duke package since the rest of the Duke journals were humanities titles.

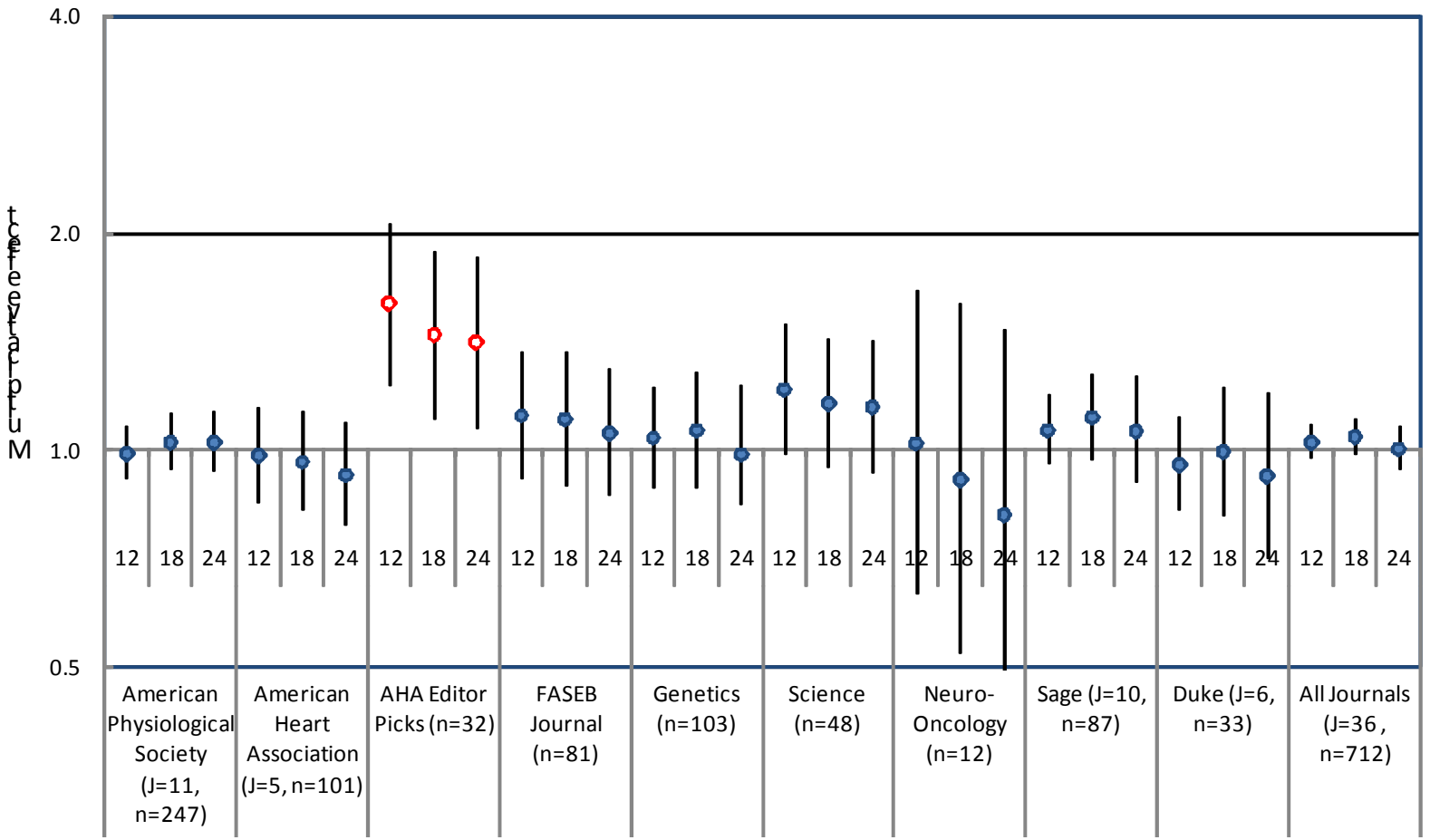
The group of articles selected by the editors of the American Heart Association journals and made freely available upon publication (labeled as "Editor Picks" in Table 11 and Figure 8.) did show positive and significant citation effects. On average, Editor Picks articles were cited about 1.4 times more frequently after 24 months. In comparison, AHA articles *randomly* selected for free access displayed no significant increase (they received, on average, 8% *fewer* citations than the subscription-access control group). These results suggest that factors besides access are responsible for the citation advantage for Editor Picks articles. Editors may be selecting more citable articles (based on their novelty, quality or importance to science), or the editors are signaling which articles should be read and cited to their community of readers. Both factors may also be in play simultaneously as evidenced by article download patterns. What is clear, however, is that free access does not appear to be driving article citation behavior in any of the journals in this study.

Table 11. Multiplicative effect of the Open Access treatment 12, 18 and 24 months after publication.

Journal/Group	Month after publication	Point Estimate	Lower 95% C.I.	Upper 95% C.I.
American Physiological Society (J=11, n=247)	12	0.99	0.91	1.08
	18	1.02	0.94	1.12
	24	1.03	0.93	1.13
American Heart Association (J=5, n=101)	12	0.98	0.84	1.14
	18	0.96	0.82	1.13
	24	0.92	0.78	1.09
AHA Editor Picks (n=32)	12	1.59*	1.23	2.06
	18	1.43*	1.10	1.87
	24	1.40*	1.07	1.85
FASEB Journal (n=81)	12	1.12	0.91	1.36
	18	1.10	0.89	1.37
	24	1.06	0.86	1.29
Genetics (n=103)	12	1.04	0.88	1.22
	18	1.06	0.89	1.28
	24	0.99	0.84	1.22
Science (n=48)	12	1.21	0.98	1.49
	18	1.16	0.94	1.42
	24	1.15	0.93	1.41
Neuro-Oncology (n=12)	12	1.02	0.63	1.66
	18	0.91	0.52	1.59
	24	0.81	0.45	1.46
Sage (J=10, n=87)	12	1.06	0.95	1.19
	18	1.11	0.97	1.27
	24	1.09	0.92	1.28
Duke (J=6, n=33)	12	0.95	0.82	1.11
	18	1.00	0.81	1.22
	24	0.94	0.73	1.20
All Journals (J=36 , n=712)	12	1.03	0.97	1.08
	18	1.04	0.98	1.10
	24	1.00	0.94	1.08

Notes: *Statistically significant at α=0.05 (two-sided test)

Figure 8. The effect of the Open Access treatment on article citations 12, 18 and 24 months after publication. Circles represent point estimates (P.E.) with vertical lines conveying their 95% Confidence Intervals (C.I.). The only article cohorts illustrating a significant and positive citation effect are those articles selected by the editors of the AHA and made freely available ("AHA Editor Picks"). Analyzed collectively, 24 months after publication, articles selected for immediate free online access show no citation advantage (P.E.=1.01, 95% C.I.=0.95 to 1.07). J=number of journals involved in the study; n=number of articles made freely available.


4.0
2.0
1.0
0.5
12 18 24 12 18 24 12 18 24 12 18 24 12 18 24 12 18 24 12 18 24 12 18 24 12 18 24 12 18 24
American Physiological Society (J=11, n=247)
American Heart Association (J=5, n=101)
AHA Editor Picks (n=32)
FASEB Journal (n=81)
Genetics (n=103)
Science (n=48)
Neuro-Oncology (n=12)
Sage (J=10, n=87)
Duke (J=6, n=33)
All Journals (J=36 , n=712)

AHA's Editor Picks

While we have failed to observe a citation differential for freely-accessible scientific articles within the first two years after publication under controlled experimental conditions, we clearly observe a positive citation effect for those articles selected and highlighted by Editors and made freely-available in AHA journals.

Are editors simply picking better articles to highlight or are editors signaling what should be cited? The real strength of the randomized controlled trial is the ability to account for systematic differences between the treatment and control group through the randomization process. The articles selected by Editors ("Editor Picks") do not represent a random selection of articles but the deliberate choice of experienced individuals. As such, we may consider three, non-exclusive, explanations for the performance of these articles:

1) Article characteristics
2) Editorial signaling
3) Accessibility

We may rule out the Accessibility postulate since our AHA experiment included a cohort of randomly-selected Open Access articles; hence, we have already controlled for accessibility. Regarding postulate 1 (Article characteristics), there is evidence that Editor's Pick articles are different from subscription and treatment articles. For instance, Editor's Pick articles are much longer on average (11.3 pages versus 7.4 and 7.9 respectively). More importantly, half (50%) of the Editor Picks were *review articles* (16 of 32) compared to 6% (23 of 359) and 13% (13 of 101) for subscription and treatment cohorts respectively. When we control for article characteristics (review article, number of authors, page length, press release,

continuing medical education component (CME) and data supplement (Table 14), the citation effect becomes statistically insignificant. In other words, it appears that we can explain the citation effect of Editor Picks articles by the fact that editors are generally selecting more citable articles.

While the sample size of Editor Picks articles is small, and thus our statistical power is limited to detecting large differences in our data, we did report large and long-term effects of Editor Picks on article downloads (as reported in Table 12), even when controlling for article characteristics. What we may be observing are two different user behaviors in play: Editors may be highly effective in directing readers to download an article, but play little (if any) role in the citation process. I will explore the role of the editor in more detail in the *Discussion* section.

Table 12. Unadjusted versus adjusted estimates of the citation effect due to editorial selection in AHA Journals.

Month	Unadjusted Estimate	P-value	Adjusted Estimate†	P-value
12	1.59 (1.23 - 2.06)	<0.001	1.27 (0.97 - 1.66)	0.085
18	1.43 (1.10 - 1.87)	0.006	1.11 (0.84 - 1.46)	0.480
24	1.40 (1.07 - 1.85)	0.012	1.04 (0.78 - 1.37)	0.800

† Controlling for review article, number of authors, page length, press release, CME, and data supplement

Summary of Citation Analysis

In summarizing the effect of free access on article citations in a controlled experiment, we may make the following statements:

1) There is no evidence that free access speeds up the citation process. Article receiving the open access treatment were no more likely to be cited earlier than their subscription-access control group.
2) There is no evidence that free access increases the frequency of citations within the first two years after publication.
3) Articles that were selected by journal editors and made freely available received significantly more citations than both randomly-selected articles receiving the same access treatment and subscription-access articles.
    a. This citation advantage for editor-selected articles appears to be explained more by article characteristics (i.e. more citable articles being chosen for free access), than by the editorial signaling process itself.

We may therefore accept our second null hypothesis (H2), that free access to scientific articles does not increase article citations.

DISCUSSION

Our experiment suggests that free access to the scientific literature may increase readership, as measured by article downloads and unique visitors, but have no effect on article citations. Open Access articles were not cited earlier than subscription-access articles, nor did they receive more citations than subscription-access articles. These results were consistent within the first two years after publication and generalizable across all journals in our study. Whereas the point estimates for a treatment effect was positive for some journals, it was negative for others, and none of the differences were significantly different than zero. When analyzing the entire dataset, the overall effect of Open Access publishing on citations was precisely zero.

While there is still ample time for articles to accrue citations, our time-frame is sufficient to detect a citation advantage for open access treatment articles, if one indeed exists. Prior uncontrolled studies were able to detect large and significant differences within as little as six months after publication e.g. (Eysenbach, 2006). Even for disciplines that follow a longer citation cycle (i.e. humanities and descriptive social sciences) our analysis shows no evidence that open access reduces the time until first citation.

Our finding that Open Access publishing does not result in earlier or more article citations challenges established dogma and suggests that the citation advantage associated with Open Access publishing may be the product of other explanations such as *self-selection* as first postulated by Kurtz et al. (2005), leading us to revise the original theory with a proposed alternative:

a. ORIGINAL THEORY:

   Free access → Increased readership → Increased citations

b. PROPOSED REVISED THEORY:

   Higher quality articles → More likely to be made freely-accessible AND more likely to be cited.

Reconciling a Readership Effect with no Citation Effect

If increased accessibility improves readership, and readership is associated with citation, then it is necessary to reconcile a pronounced readership effect (a doubling of full-text downloads) with no citation effect. This is possible to do if we replace the notion of a single reader community with *multiple reader communities*.

The universe of readers (those individuals who successfully downloaded an article from the journals within our study) may be represented as the largest concentric circle in Figure 9. Within this universe of readers is a much smaller group of readers who also function as scientific *authors*. This group adds new knowledge to the corpus of scientific literature and connects their work to the rest of the corpus through the process of citation. This author community is further distinguished by a smaller community of authors who have their work published in a journal that is indexed by Thompson/ISI's Web of Science – the source from which citation data were extracted. As mentioned above in the Methods section (see *Methodological Limitations: Scope of Citation Data*), Web of Science indexes only a fraction of the extant scientific literature. While this small corpus of journals publishes the majority of scientific articles and attracts the vast majority of all citations (Garfield, 1996), it is far from a comprehensive literature index. It is therefore necessary to separate the author community into two distinct groups based on whether their work is indexed by the

Web of Science.

While these multiple reader communities may be considered distinct, they are not mutually exclusive: We assume that authors are also readers although the reverse is not necessarily true. Moreover, while these communities are largely fixed, they are not completely static: We assume that there is some temporal movement of individuals between these concentric groups. Some readers may become authors at some time in their career; and some authors may have only partial representation of their work in the ISI-indexed community. In spite of the limitations for representing readers into three functional groups, this simple nested representation is sufficient for explaining the results of our study.

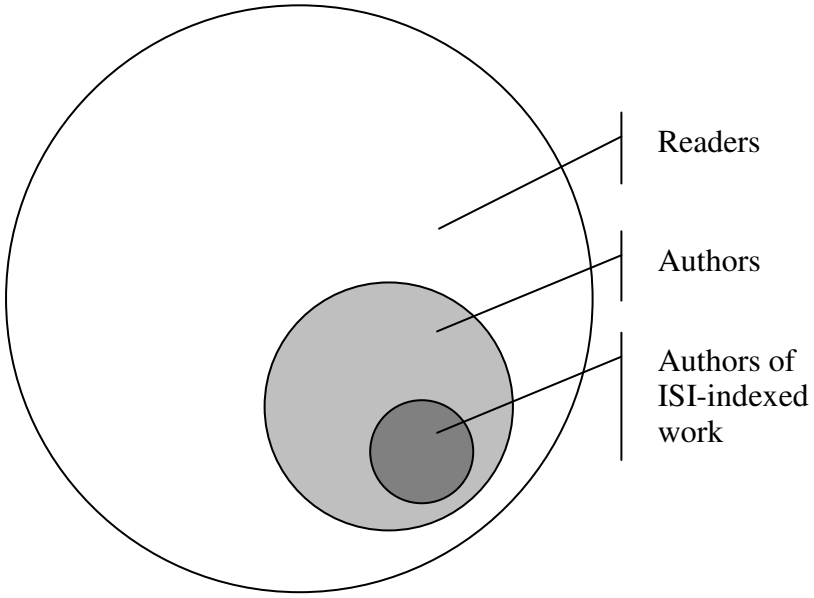


Figure 9. Concentric reader communities.

Authors of ISI-indexed work, for the most part, are stratified into elite institutions with excellent access to the scientific literature. Understanding this stratification is crucial for reconciling the readership effect reported in our study with no corresponding citation effect.

In order to contribute meaningfully to the scientific literature, in most cases, one must have access to resources (equipment, trained individuals, and money), as well as to the literature of one's discipline. These two requirements result in the concentration (or "stratification") of researchers at elite institutions around the world (J. R. Cole & Cole, 1973; Crane, 1972; D.J.S Price, 1986). Elite institutions are also known for their extensive library collections and online access to the research literature. Providing free access to journal articles may make little difference to researchers located at elite institutions: from their standpoint, researchers already have "free access" to the literature. Indeed, the journals selected for our study are core research journals – not obscure titles – and should be available in nearly all research libraries.[22]

The fact that we observe an increase in readership and visitors for Open Access articles but no citation advantage suggests that the increase in readership is taking place outside the core author community. The real beneficiaries of Open Access may not be the author community, who traditionally have excellent access to the research literature, but communities of practice that consume, but rarely contribute to, the corpus of literature. These individuals may include students, educators, physicians, patients, and researchers employed by private companies (such as the pharmaceutical industry) who depend on the publication of scientific literature.

The increase in fulltext downloads for Open Access articles during their first

[22] Selecting obscure titles for a citation analysis provides its own problems. Articles published in these journals receive few citations, making it difficult to conduct an empirical analysis.

year after publication (Figures 2 and 3) suggests that the primary benefit to the non-subscriber community is in *browsing*, as opposed to printing or saving, which would have been indicated by a commensurate increase in PDF downloads. The fact that Internet robots were responsible for almost half of the fulltext downloads, yet represented only about 1% of unique visitors to the journal sites, may also imply that Internet search engines are helping to direct non-subscribers to free journal content. We do not have access to the transaction logs of the publishers – only the aggregated summary statistics of individual actions – and therefore can only infer on how users are discovering research articles and their resulting behavior when these individuals are brought to a research article.

Lastly, while we need to be careful not to equate article downloads with readership (we have no idea whether downloaded articles are actually read), measuring success by only counting citations may miss the broader impact of the free dissemination of scientific results beyond the research community. Our study suggests that free access to the scientific literature may have little impact on readership within the scientific authorship community, although it may have impact outside this small core of readers.

Implications for Scientific Authors

Early studies linking Open Access status to a citation advantage suggested the benefit for making one's work freely available was immense. The citation advantage for Open Access articles was routinely reported within the range of 200-700% (see Table 2), implying that the subscription model was largely ineffective for disseminating scientific articles to scientific authors. As citation to one's work provides evidence of peer recognition – the basis of the reward system in science

(Hagstrom, 1965; Merton, 1988) – a large citation return was used as an incentive for faculty to change their publication behavior and formed the central message of several campaigns, e.g. (Association of Research Libraries, 2004c). The lack of a citation effect for Open Access articles in our study suggests that claims of the inefficiency of the subscription-model to disseminate scientific results to the research community – and conversely, the payoff to faculty to adopt alternative dissemination models – may be greatly exaggerated. While we cannot make claims for researchers who read but do not cite the research literature, our results are consistent with repeated surveys and interviews that suggest that access is not a primary concern for scientific authors (see *Survey Studies on Access*).

Advancing the Theory of the Attention Economy

Given the conclusions of our study, it is necessary to revisit and reconsider the four functions of the journal (registration, certification, dissemination, and archiving). Open Access presupposes a fundamental problem in the dissemination function of journals, a problem which is not supported by our study nor generally by the extant literature on this topic (see *Literature Review*). New opportunities created by digital scholarly publishing have implications for our understanding of journals that go far beyond the function of dissemination.

Worldwide, there are approximately 25,000 active, peer-reviewed journals (ProQuest, 2009) producing an estimated 0.75 million articles (National Science Foundation, 2010) to 1.5 million articles annually (Björk, Roos, & Lauri, 2009). Between 1995 and 2007, world annual output of science and engineering articles grew at an average annual rate of 2.5%, generally tracking growth in the global R&D

workforce (National Science Foundation, 2010). Estimates over the last two centuries estimate article growth at about 3% annually (Ware & Mabe, 2009).

What has not kept pace with the inflation of published knowledge is human attention. The amount of time we have to read and process the work of others is fixed, and remains fixed while the number of potentially relevant articles continues to grow. Specialization of journals has been a common solution for editors and publishers wishing to create separate channels or "space of attention" for reader communities (Klamer & van Dalen, 2002, p. 302). While this is but one solution to dealing with the plethora of articles on a related topic, it offers only a partial solution for readers. Scientists themselves see the vast amount of literature produced as a fundamental problem for their ability to assimilate the literature produced by their field (Garvey & Griffith, 1967, 1971). There is simply too much to read and not enough time. As economist Herbert Simon wrote:

> In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention and a need to allocate that attention efficiently among the overabundance of information sources that might consume it. (Simon, 1971, pp. 40-41)

The traditional approach to the problem of information overabundance is at the *receiver side* of the equation. While one could equally argue for solutions that put restrictions on the production of new knowledge,[23] most solutions have focused largely on tools and strategies for dealing with a crisis of attention. Echoing Simon, Pirolli and Card (1999) write:

[23] For example, promotion and tenure boards could put more emphasis on quality rather than quantity of publication, requiring candidates to submit only their top 3 articles. Journal editorial boards could also decide to accept and publish fewer articles, and research foundations could decide to cease subsidizing article publication.

> Providing people with access to more information is not the problem. Rather, the problem is one of maximizing the allocation of human attention to information that might be useful. (p.645)

Pirolli & Card's *information foraging model* (1999) is an adaptation of how animals forage for food and other resources. Their model assumes that valuable and relevant information resources are not distributed homogenously in the environment but clustered into patches; as a result, individuals develop strategies that maximize their gathering of valuable information while minimizing expended effort. Their model also assumes that the receiver is more than a passive receiver of information, but an active agent in the information seeking process. In the information foraging model, the agent is constantly evaluating the information resources being discovered, its own expenditures in terms of time and effort, uses this information to modify its behavior in order to adapt to its environment. Through this iterative process, the agent develops heuristics for seeking and evaluating the quality and relevance of information.

Like the simple, linear *Transmission Model of Communication*, first proposed by Shannon and Weaver (1949), Pirolli & Card's information foraging model assumes that information flows from the sender to the receiver along a simple channel with no feedback to the sender. While the foraging model is useful for developing information seeking tools at the *receiver* end, it does little to explain the behaviors of the message *sender* (the author), nor does it explain the *conduit* of the message (the journal). More importantly, viewing scientific publishing as a linear transmission model fails to explain several phenomena in science communication, specifically:

1) *Certification*. Why do academics continue to seek out the certification process in the publication process when certification is no longer a

precondition for the dissemination of one's work to one's peers and the rest of the academic community?

2) *Branding*. Why is the journal brand so important when academics can create their own brand through self-publishing and distribution?

In this section, I will argue that we need to abandon the linear sender-receiver model (in any form) for a more complicated two-sided market model in which both groups, authors and readers, are actively sending and receiving value signals through an intermediary, the publisher, who is responsible for organizing the communication between these two groups.

Using the classic example of used cars, this chapter begins with a description of the problem of *information asymmetry* in markets and how certification is used by sellers to signal quality to potential buyers. I then apply this theory to the market of scholarly articles where authors use the peer-review certification process to send quality signals to potential readers. I continue with how readers now construct their own market signals through the article download and citation process and how these signals influence peer behavior. I then conclude with a discussion of how increased transparency of the collective behavior of readers may lead to greater concentration of attention placed on fewer scholarly articles with implications for science.

Information Asymmetry and the Market for Used Cars

In his seminal paper, *The Market for "Lemons,"* George Akerlof (1970) defined the market for used cars as a market in which the sellers have more information about the quality of an automobile than buyers. Because of this

information asymmetry, the potential buyer can only assume that a vehicle is of average quality, and should therefore offer a price reflective of the average quality of all cars in the market.

This is not a problem for a seller possessing a car of below-average quality. This seller is in the position of profiting from the transaction, being offered a price that is higher than the true value of the car he or she possesses. It *is* a problem for a seller possessing a car of superior quality since the seller will be offered a price which is lower than the true value of the automobile. In this case, the seller may decide to remove the car from the market and simply not sell it. Alternatively, the seller may also pursue a different strategy.

A seller may decide to *certify* a used vehicle prior to putting it on the market. Certification, in this case, is a signal to the potential buyer that the vehicle meets certain quality standards. The certification process is not free for the seller – he must pay a disinterested evaluator to inspect the vehicle and provide some type of quality guarantee that is transparent, recognized, and acceptable to the potential buyer. In addition, the time and resources that go into certifying a vehicle represents lost opportunity costs for the seller, who could have sold the car faster at a reduced (average) price.

The stamp of certification does not indicate a quality estimate of a used car – it only specifies that the car meets *minimum* quality standards necessary to receive the stamp of certification. The assumption here is that this minimal level of certification is higher than the average level of quality in the used car market; otherwise, there is no incentive for the seller to certify a car. The result is that a seller may charge more – and a buyer may pay more – for a certified vehicle than an uncertified one.

The Market for Academic Articles

The market for scientific articles has similar properties to the used car market: a great heterogeneity in the quality of scientific articles and an asymmetry of information between the author and reader on the value of each article. There are several key differences in the nature of information from physical resources that should be noted before we continue.

First, traditional economies are governed by a limited supply (often a scarcity) of physical resources, and the interaction between supply and demand results in determining market prices. Information, on the other hand, is overly abundant, and unlike physical resources, using information does not deplete its source nor leave others information-poor. Information, especially in its digital form, can be easily replicated, and the marginal costs of producing and distributing another copy of an article are close enough to zero to be considered irrelevant (Kingma, 2001).

Second, readers do not spend their own money on purchasing articles, but they do spend their limited attention on reading articles. We assume that most academics are located at research institutions with access to libraries that purchase journal subscriptions on behalf of their constituents. While academics are essentially taxed to support the overall infrastructure of the universities, they rarely pay directly for their access to the literature; for them, access to the literature is essentially free. Free however, does not mean that the reader is engaged in a costless economic transaction – what is exchanged for information is reader attention. Like money in financial transactions, attention is the limited resource being traded for the content embedded in academic articles. Being conscious that attention is a limited resource, readers must calculate *in advance* whether an article is worth their attention. The lost opportunity costs of reading the wrong paper means that another higher-quality article may be ignored.

Last, *quality* is a multi-dimensional construct. The details of what makes a "high-quality" paper will not be explored in this chapter. It will suffice to state that quality is essentially a *private measure*, defined by each reader only *after* some information has been consumed. In other words, the potential reader does not know the value of an article until the article has been read. Later in this paper, I will discuss how quality indicators are created and transmitted in groups; yet at this stage, we assume that quality is private, fixed, and unknown by the consumer in advance of the transaction.

Evaluating Articles is like Evaluating Used Cars

In the case of a market where a reader does not know the value of a paper, one will first assume, as in the case of used cars, that a new article is of *average quality*. Since a reader spends the same amount of attention reading a low-quality paper as a high-quality paper (there is no cost difference on the side of the reader), there is a strong incentive for readers to seek out articles of high-quality.

Because an author is rewarded by having one's work widely read and recognized in one's discipline (Biagioli, 2003; Hagstrom, 1965), there is a strong incentive in a market of heterogeneous quality for authors to seek out forms of quality certification that will send signals to potential readers that one's article is of high-quality and worthy of readers' attention. These quality signals become more important as the size of the market grows (Rosen, 1981) as more articles compete for the limited attention of readers.

Modeled in this way, we may view scientific publishing as a two-sided market, with authors on one side, readers on the other, and journals fulfilling the role of intermediary agent. Authors use the journals to send out quality signals, which

compete with each other for the limited attention of the reader community. Readers, in turn, seek out the quality signals in the market as indicators to what they should devote their limited attention. Next, I will argue that this two-sided market approach (in contrast to the linear transmission approach) is able to explain certification and market branding in science publishing.

Certification in Science

Journal publishers traditionally perform four fundamental roles in scientific communication: registration, certification, dissemination, and archiving (Zuckerman & Merton, 1971). In an electronic, networked environment, these four functions can be disaggregated from the printed journal. Preprint servers can function to establish priority claims (registration) by date-stamping submissions; digital repositories can function to disseminate articles to a wide networked community of readers; and digital libraries and archives can function to preserve the scientific copy of record. Often each of these services provides more than one function.

Like the certification of used cars, the certification process for academic articles [24] performs the role of guaranteeing that an article meets minimum quality standards established by the editorial board of a journal. The method of certification often involves refereeing by an author's peers in the academic community. There are many variations of peer-review (single-blinded, double-blinded, editorial, open, post-publication, etc.) I will not go into the details of these variations, only to generalize that certification is a community-defined process that establishes whether a submitted article is worthy of being given the journal's stamp of quality.

[24] Because the certification process in science ultimately reflects a dichotomous decision (accept or reject), it has been referred to as "gatekeeping" (Crane, 1967).

Although some journals set very high quality standards for what is accepted for publication, rejecting 19 out of 20 submitted manuscripts for some top-tier journals, the gatekeeping function of journals does not appear to stem the tide of an increasing number of manuscripts being published each year. Rejected articles are often submitted to lower-status journals until publication is secured (Cronin & McKenzie, 1992). [25] The function of the peer-review system, therefore, should not be thought of as a mechanism of *preventing publication*, but as a system of stratifying articles into tiers of quality, making it easier for readers to select what to read (S. Cole, 2000).

While it is easy to look for examples of some of the most prestigious articles in science rejected by top journals (for example, those articles becoming the basis of Nobel prizes for their authors) e.g. (Campanario, 1996), the system works fairly efficiently. The ethical norms of science prevent authors from submitting their manuscript simultaneously to multiple journals except in limited and well-defined cases (Fulda, 1998; International Committee of Medical Journal Editors, 2009). To avoid wasting their time (in addition to wasting the time of editors and reviewers), authors evaluate potential publication outlets and typically select venues that are commensurate with the perceived value of their manuscripts. Thus, while journals select which manuscripts are worth publishing, authors have already *pre-selected* the journals in which their manuscripts have a chance of being accepted (S. Cole, 2000).

Unlike getting a used car inspected and certified by a mechanic, the

[25] Studies of the fate of articles rejected for publication reveal that high percentages eventually are accepted in lower impact journals although eventual publication is not guaranteed: *New England Journal of Medicine*: 86-89% (Groves, 2009); *Epidemiology*: 75% (Hall & Wilcox, 2007); *British Journal of Surgery*: 66% (Wijnhoven & Dejong, 2010); *American Journal of Neuroradiology*: 56% (R. J. McDonald et al., 2009); *American Journal of Ophthalmology*: 50% (Liesegang et al., 2007); *Cardiovascular Research* (47%) (Opthof et al., 2000). These studies suffer from two major weaknesses: First, a paper may be published much later after first rejection. Second, a rejected paper may eventually be published in a journal, which is not included in a literature index.

certification of scholarly articles is far from a timely process. In fields such as economics, the delay from submission to publication may take several years, involving several iterations of manuscript resubmission and review (Mason, Steagall, & Fabritius, 1992) and the process has been growing longer, not shorter (Ellison, 2002).

Considering that much of the transfer of information among colleagues within a discipline occurs *before* journal publication – through conference presentations, working papers, and the informal network of researchers in one's field (Garvey & Griffith, 1971) – why do academics continue to seek out the slow and expensive certification process when certification is no longer a precondition for the dissemination of one's work? Or posed another way, why do academics choose to put much of their time and resources into having their work certified when these same resources could be put into publishing more articles?

Certification as Market Signaling

Quality signaling is important in large markets where most participants are not in the market frequently enough to develop their own reputation signals (Spence, 1973). In the market for academic articles, most authors publish very few papers (D.J.S Price, 1986), and less than 20% are repeat authors (Ware & Mabe, 2009). Even for those who do publish regularly, it may take years to develop a reputation for quality as an author builds a portfolio of publications.

In a large market of academic papers, where quality is heterogeneous and information is asymmetric, readers will rely on various signals to identify what is worth their attention. In seeking the attention of readers, authors, in turn, will seek out certification for their articles in order to send out high-quality signals even when the certification process is slow, costly and detracts the author from publishing more

papers. In sum, this two-sided market, formed with readers on one side, authors on the other and journals mediating the transaction, explains reader *and* author behavior, as well as the persistence of journals in an information environment where dissemination may be decoupled from the certification process.

Types of Quality Signals

There are many types of signals in the market for scholarly information. These signals can be universal, institutional, discriminatory, or communal. Potential readers will often evaluate whether an article is worth attention according to multiple signals, often using them in combination.

a) *Universal*. The most basic certification for scholarly articles is whether the article has been peer-reviewed at all, that is, certified by a group of other individuals within a community of practice. With nearly 25,000 active, peer-reviewed, academic journals, peer-review does not signal very much to the reader except that the article belongs to this large set of academic literature and passes the gatekeeping function of a small number of one's qualified peers. Being able to distinguish levels of quality within this massive set of publications is left to more salient indicators of article quality.

b) *Institutional*. The past performances of an aggregate class of localized peers (such as one's department or institution) can create a strong signal to potential readers. As a reader associates high-quality articles with members of a certain group, an expectation is created that new articles emanating from this group are of significant quality. A granting agency can create a type of institutional

signal to the quality of the article as well. Research projects go through similar expert evaluation, similar to gatekeeping by journals, where only those projects with prospects of generating high-quality research findings are furnished with funds. Granting agencies thus create quality certifications and market signals to potential readers.

In a controversial experiment, twelve published articles in top psychology journals were resubmitted back to the journals that accepted them with one modification: the author's names and affiliations were changed to reflect unknown authors and fictitious institutions (Peters & Ceci, 1982). At the time, these journals practiced single blind review, meaning that author details were known to the editor and reviewers. Only three of the twelve articles were detected as duplicates. Eight of the remaining nine articles were rejected, mostly on methodological grounds and only one of the twelve resubmitted articles was accepted. While the results of this research may be interpreted in many ways, the strongest conclusion is that editorial and gatekeeping decisions may be influenced by the prestige of the author and his or her institutional affiliation.

c) *Journal*. The name of a journal conveys a type of brand in the marketplace – a signal to potential readers of some expected level of quality. This association of quality with brand name is very important for a journal since a loss of readership is possible when the level of quality does not meet customer expectations. Given that the journal market represents a gift economy, with authors freely providing their manuscripts to a publisher in exchange for peer-recognition (Biagioli, 2003; Hagstrom, 1965), the reputation for quality is also critical for attracting future manuscripts. While it may take years to develop a

reputation for quality, individual events such as a forced retraction of a paper due to falsification of data (for example the Hwang South Korean stem cell controversy (Couzin, 2006)), can shake the community's sense of confidence in quality assurance. This is why journals are so careful in preserving the integrity of the peer review and editorial process and spend so much time defending their reputation under such circumstances.

In addition to adhering its stamp of certification to an article, a journal can create additional quality signals for articles. An editor can signal which articles are of exceptional quality or newsworthiness by highlighting them in an editorial, by establishing the order of article publication or by affixing an "Editor's Pick" mark for exceptional articles. Moreover, many leading scientific journals generate press-releases that are picked up by the lay-press and other outlets intended to translate and interpret the results of scientific research. Several studies indicate that press-releases not only help disseminate research to the lay public, but also lead to greater dissemination of research within the scientific community, as evidenced by increased citations to articles (Chapman et al., 2007; Kiernan, 2003; D. Phillips et al., 1991).

d) *Individual*. Individual authors who have a history of publication and have gained reputation from one's peers can build a personal signal in the marketplace. With the exception of a few glorified academics, such as Nobel Laureates or others who have achieved status across disciplines, the signal that an author creates is likely only interpreted by readers in one's own field.

There is evidence from the field of economics that high-profile authors are increasingly bypassing the journal certification market to disseminate their own work (Ellison, 2007). This phenomenon has been made possible by the

Internet, which allows the certification and dissemination function of journals to be decoupled. As authors are no longer required to go through the lengthy and costly process of journal certification in order to have their work disseminated, some may decide to market their own work using discipline-based repositories, institutional repositories, or one's departmental, laboratory, or personal website (Davis & Connolly, 2007).

e) *Community constructed signals*. So far, we have focused entirely on quality market indicators that are fixed and are created by the producer side of the market, that is, by the authors and the certifications they seek. I will now discuss the influence of citations and more recently article downloads as community constructed signals that may determine reader choice.

Through the citation process, authors create signals that convey status on other authors (Merton, 1988). Citations also create functional links between documents, helping to guide readers to related material (Cronin, 1984). When citations are aggregated to form a single count, they transmit a type of communal quality indicator (an impact factor) that alerts the reader to the influence of an article on the scientific literature (Garfield, 1955). In addition, many journal websites now provide frequency of article downloads (providing either raw counts, or the ranked order of the highest downloaded papers) as guides for their readers.

The notion that the collective behaviors of other readers can provide a useful heuristic on what is worth attending to has been described in many situations as the *wisdom of the crowds* (Surowiecki, 2004). These signals are created, however, only *after* an article has been published. The first readers of an article cannot rely on signals from previous readers to help them guide their

choice but must depend solely on producer-side quality indicators, such as the reputation of the journal, author, or institutional affiliation. Whereas articles published in prestigious journals attract sufficient downloads early enough after publication to create the beginnings of a quality signal, most articles take months to generate significant readership and years to generate significant citation signals, if any at all. As a result, community constructed signals may disproportionately benefit those who already attract early and significant attention.

Implications of Social Influence on Scholarly Communication

Objectivity in science is the ability to separate the contribution of a piece of work from its context (author, place of publication, etc.) This is what Robert K. Merton described as the ethos of *universalism* in science (Merton, 1973). And yet, science operates as a social institution, an obvious fact that does not go unnoticed by Merton, who acknowledges that "universalism is deviously affirmed in theory and suppressed in practice"(p. 273).

We cannot ignore that scientists, like everyone else, are highly sensitive to what their peers are doing. In an environment of too much information and a scarcity of attention, readers actively seek out signals of article quality designed to guide them to what is worth reading. More importantly, we can expect that signaling becomes more important as the market of academic articles continues to grow (Rosen, 1981).

As a response to new technologies that send additional quality signals to potential readers (e.g. citation indexes, download counts, search engines that rank based on the behaviors of others), we would expect that these communal signals would reinforce the disproportionate attention given to a small number of authors and

their work. In a massive, longitudinal study of citation patterns since 1965, Evans (2008) documented that scholars are indeed showing less diversity in their citation practices, citing fewer journals and unique articles. This type of social amplification has been described for similar phenomena such as why eminent scientists often receive credit for discoveries when priority claims are ambiguous ("Matthew Effect" (Merton, 1968, 1988)), why highly-cited articles are cited more often ("Cumulative Advantage" (D.J.S. Price, 1976)), why highly-linked websites attract more in-links ("Preferential attachment" (Barabasi & Albert, 1999)), or why famous people get paid so much ("Economics of Superstars" (Rosen, 1981)).

Environments in which actors can see each other's decisions may result in early advantages that are amplified over time. Because of this amplification, early entrants in a market have an advantage over those who arrive later. A study analyzing citation rates to physics articles suggests that early papers published on a particular topic are cited at a rate much higher than subsequent papers (M.E.J. Newman, 2009). Lastly, increasing the strength of social influence can increase both inequality and unpredictability of success. In a study of artificial music markets, success was determined only partly by quality. While high-quality songs rarely did poorly and low-quality songs rarely performed well, any other outcome was possible (Salganik, Dodds, & Watts, 2006).

Increasing the transparency of peer behavior may amplify the social influence of actors who participate in scientific publishing. It may also change the institution of science as a whole. Evans warns that signaling, as it affects reading and citation behaviors of authors, may hasten the process of consensus-building in science, such that unpopular ideas that do not find their way to consensus early in the community may be quickly forgotten (Evans, 2008, p. 398).

Conclusion

In a large market of academic papers, where quality is heterogeneous, information about quality is asymmetric, and the vast majority of authors appear only once, readers will use various signals to identify what is worth attending to. Authors, in attempting to maximize the attention given to their papers, will seek out certification for their articles in order to send out high-quality signals to potential readers. Authors and readers interact with each other in this two-sided market with the journal forming an intermediary between the two. Unlike the linear transmission model of communication, the two-sided market model is able to explain reader behavior as well as author behavior, and adequately explains why authors continue to use the journal as a mechanism to certify their work when other cheaper and more timely distribution channels are available to the author.

Future Research

In this study, we considered the effects of publisher-mediated access on readership and citations. Access conditions were controlled at the journal websites and the usage data represented activity at those websites. As a result, the setup of this experiment assumes a directional and hierarchical flow of information from publisher to reader and ignores other avenues of access to the scientific literature. Indeed, most of the extant research on information usage assumes a traditional and hierarchical flow of information from the publisher to the reader.

Alternative Sources of Scientific Literature

Very little has been done investigating alternative access routes to the scientific literature. If consumers of the scientific literature operate anything like consumers of cultural media, such as music and film, we may miss much of the flow of scientific media that does not emanate from the publisher. Some of this alternative flow of documents may be mediated through academic libraries in the form of interlibrary lending; however, the largest flow of documents may move *between* peers – between the author and reader or between readers themselves. Gaulé's 2009 study of access to scientific information in India suggests that informal peer-to-peer sharing is very common in countries with a history of poor access to the scientific research literature. For authors, the practice of ordering reprints of one's article for the purpose of fulfilling reader requests by physical post has largely been replaced by sending copies by electronic mail or by directing a reader to a copy placed in a publicly accessible electronic archive. More recently, productivity software, like Mendeley,[26] a bibliographic database for managing academic literature, include functions for sharing articles among small social networks. Although such networks are currently limited to 10 individuals, information can diffuse very quickly when members of one social network overlap with members of another. In sum, by measuring only the distribution of articles from the publisher's website, we miss all other alternative forms of distribution that may be taking place within informal networks.

While we acknowledge that these alternate access venues exist, little is known about the extent and magnitude of informal sharing of the scientific literature. Part of the problem is one of tracking and reporting: individuals do not keep count of the articles they share with others and subsequently report these figures to a publisher.

[26] Mendeley research networks. http://www.mendeley.com. Accessed 15 Feb, 2010.

Similarly, most academic communication networks maintain privacy when dealing with interpersonal communications, and as a matter of policy, do not track what is being sent over their networks. Finally, established systems that enable peer-to-peer sharing of documents do not talk to each other and thus cannot aggregate usage data. In our experiment, the best we could do was to acknowledge that alternative sources of access to the scholarly literature were available and that their combined effects were to attenuate the effect of our access treatment. Future research should attempt to estimate the extent of access to alternate sources of scientific literature.

Self-archiving as an Alternative to Publisher Access

Self-archiving, as a form of open access, has become much more prevalent since the start of our study, mostly as a result of new institutional and funding agency requirements. Many universities, colleges or departments now mandate [27] that authors deposit a copy of their final, peer-reviewed manuscript in their institutional repository within 12-months of publication. [28] The National Institutes of Health requires the deposit of an author's final manuscript into PubMed Central within 12 months as a condition of funding (National Institutes of Health, 2009).

Introduced into the U.S. Senate on June 25th, 2009 by Senators Joseph Lieberman (Connecticut) and John Cornyn (Texas), The Federal Research Public Access Act of 2009 ("Federal Research Public Access Act (S.1373),"), also known by its acronym, "FRPAA," would require all Federal agencies that dispenses over $100 million dollars in extramural research to require self-archiving of final, peer-reviewed

[27] Mandates do not always guarantee whether authors will comply.
[28] See ROARMAP (Registry of Open Access Repository Material Archiving Policies) for a current list of policies and repositories http://www.eprints.org/openaccess/policysignup/ Accessed 29 April, 2010.

manuscripts into a digital archive no longer than six-months after publication. The bill was referred to the Committee on Homeland Security and Governmental Affairs. On April 15th, 2010, the same bill ("Federal Research Public Access Act (HR. 5037),") was introduced into the U.S. House of Representatives by Michael F. Doyle (Pennsylvania-14) with five co-sponsors. It was referred to the House Committee on Oversight and Government Reform.

The effects of self-archiving are not entirely clear. To date, there has been a dearth of rigorous controlled trials on the effects of self-archiving and previous research has indicated that multiple factors are at play simultaneously, making it difficult to determine and disambiguate causes and effects (see *Literature Review*). In 2008, a group of publishers and universities began collaborating on a large study to investigate the effect of systematic archiving of manuscripts on article and journal visibility ("PEER: Publishing and the Ecology of European Research,"), with results expected beginning in 2011.

In our study, self-archiving was not frequent enough for us to estimate its general effect on readership and citations. While we had enough data to conduct an analysis on *Science Magazine* (see Appendix), we should be hesitant to generalize the results to the rest of scientific publishing. Wren (2005) reported that there is a tendency for articles published in higher impact journals to be found freely on the Internet, an association also reported for the economics literature (Bergstrom & Lavaty, 2007). Moreover, the citation advantage attributed to free access has been reported to have a disproportionate effect on highly-cited articles (Antelman, 2004; Davis & Fromerth, 2007; Gargouri et al., 2010; Lawrence, 2001). More recently, Gargouri and others (2010) have made a strong and declarative causal link between self-archiving and increased citation performance. Their claim should be considered tentative until it can be confirmed with more rigorous studies.

CONCLUSIONS

The results of this experiment provides strong evidence that free access to the academic journal literature increases readership (as measured by article downloads) and reaches a broader audience (as measured by unique IP addresses), yet may not have any effect on article citations at least within the first two years after publication. The lack of a citation effect suggests that traditional models of disseminating scientific knowledge work efficiently for the research community, or more specifically, for those who generate new knowledge.

Free access to scientific articles may speed up the transfer of knowledge to industry, improve health care, empower the general public, and reach individuals at institutions with limited access to the subscription-access literature. While there are many proposed benefits to the free access of scientific information, the results of this study suggests that a citation advantage is not one of them.

The dissemination function of the journal appears to be inadequate for explaining scholarly behavior as well as the persistence of journals in an information environment that decouples the four functions of the journal. We should reject this limited view for a more expansive theory that views scholarly publishing as part of a larger attention economy.

APPENDIX

Self-archiving

Before reporting the readership and citation results, it is important to describe the extent of self-archiving and estimate its effects on our experiment. As mentioned earlier, high prevalence of self-archiving may attenuate any Open Access effects we may observe.

Self-archiving rates were generally low for the articles under investigation (Table 13). Most of the journals in our study reported zero or few cases of self-archiving. The journal *Science* showed the highest number and percentage of self-archived articles (36/393 or 9.2% of articles in our study), followed by the *Journal of Neurophysiology* (12/278 or 4.3%). The overall detection rate was about 2%.

Table 13. Self-archiving rates by journal.

Journal (abbreviation)	**N, self-archived**	**N, total**	**%**
Adm. Soc.	0	24	0.0%
AJP-C	1	155	0.6%
AJP-E	0	147	0.0%
AJP-F	0	140	0.0%
AJP-G	0	134	0.0%
AJP-H	1	233	0.4%
AJP-L	0	109	0.0%
AJP-R	2	195	1.0%
Am. Behav. Sci.	0	41	0.0%
Am. Speech	0	8	0.0%
Appl. Psychol. Meas.	0	20	0.0%
Arterioscler. Thromb. Vasc.	0	105	0.0%
Circ.Res.	0	60	0.0%
Circulation	1	96	1.0%
Commun. Res.	0	19	0.0%

Table 13. (Continued)

Comp. Polit. Stud.	2	28	7.1%
Ethnohistory	0	11	0.0%
Faseb J.	3	165	1.8%
Genetics	3	211	1.4%
GLQ-J. Lesbian Gay Stud.	0	13	0.0%
Hypertension	0	95	0.0%
J. Appl. Physiol.	1	201	0.5%
J. Health Polit. Policy Law	0	17	0.0%
J. Neurophysiol.	12	278	4.3%
Neuro-Oncology	0	27	0.0%
New Media Soc.	0	30	0.0%
Organization	1	24	4.2%
Physiol. Rev.	1	16	6.3%
Physiology	0	11	0.0%
Prog. Hum. Geogr.	0	26	0.0%
Public Cult.	1	21	4.8%
Science	36	393	9.2%
Soc. Sci. Hist.	0	10	0.0%
Soc. Stud. Sci.	0	25	0.0%
Stroke	0	132	0.0%
Theory Psychol.	0	33	0.0%
Total	65	3253	2.0%

The rate of self-archiving does not appear to differ significantly between subject disciplines as reported in Table 14. With the exception of the Multidisciplinary group (a class that includes just *Science Magazine*), all other categories report low rates of self-archiving. Because of such low frequencies of self-archiving, this variable was dropped in several of the inferential statistical analyses.

Table 14. Self-archiving by journal category.

Journal Category	N, self-archived	N, total	%
Medical	1	515	0.2%
Life Sciences	24	1995	1.2%
Multidisciplinary Sciences	36	393	9.2%
Social Sciences	3	270	1.1%
Humanities	1	80	1.3%
Total	65	3253	2.0%

Notes:
Medical included the five journals published by the American Heart Association plus *Neuro-Oncology*, published by Duke University Press.
Life Sciences included the *FASEB Journal*, *Genetics* and the 11 journals published by the American Physiological Society.
Multidisciplinary Sciences included just one journal, *Science Magazine*.
Social Sciences included the 10 journals published by Sage Publications.
Humanities included 6 journals published by Duke University Press.

Case Study: Science Magazine

As reported earlier, there was low frequency of self-archiving in this study, with most journal cohorts in this study showing few (if any) instances of self-archived articles. *Science Magazine* was an exception and we were able to detect 36 examples of articles found on publically-available websites – nearly 10% of the 393 articles involved in this study, compared to an average of about 2%. As illustrated in Table 15, self-archived articles appear to receive about 30% more citations at 24 months (1.30, 95% C.I. 1.01 to 1.67, p=0.038) than their subscription-access cohort.

Self-archived and non-self-archived articles published in *Science* appear to be similar to each other in many ways. They include similar mean numbers of authors per paper (8.6 versus 8.9), similar number of article pages (3.5 versus 4.1), and similar number of review articles (8% versus 6%) respectively. Self-archived articles, however, appeared more frequently in Science Express (published online before print)

(36% versus 21%) respectively. Conceiving that the association between self-archiving and appearing in Science Express may be responsible for the self-archiving citation effect reported in Table 15, rerunning the regression model *without* the Science Express variable resulted in a similar self-archiving effect. It appears that self-archiving, at least in *Science Magazine*, is associated with an independent and positive effect with article citations. Because self-archiving behavior could not be controlled in this experiment, we should be cautious with attributing a causal link between self-archiving and article citation performance. It is equally possible that more citable articles are made freely-available through self-archiving. While it was impossible to measure the number of article downloads when an author places a copy on a public website, there was no evidence in this study that self-archived articles received more article readership from the journal website. Indeed, self-archived articles demonstrated no more article downloads than their subscription-access cohort (Table 8). The premise that free access through self-archiving may increase readership and article citations deserves further investigation.

Table 15. The effect of article and access characteristics on article citations in *Science Magazine*, 24 months after publication. Estimates are reported as multiplicative effects.

	Estimate	Lower 95% C.I.	Upper 95% C.I.	P-value
Open Access	1.15	0.93	1.41	0.209
Self-archived	1.30	1.01	1.67	0.038
Number of Authors†	1.26	1.14	1.39	<.0001
Review	1.31	0.93	1.84	0.124
Length in Pages†	1.73	1.42	2.10	<.0001
Science Express article	1.35	1.13	1.60	0.001
Issue Highlights	0.98	0.83	1.17	0.833
Cover article	1.17	0.83	1.64	0.363

Notes:
† Log transformed variable
Mean response=3.31; N=393; RSq=0.47;
Model includes Section as random effect

REFERENCES

Akerlof, G. A. 1970. The Market for "Lemons": Quality Uncertainty and the Market Mechanism. *The Quarterly Journal of Economics* 84: 488-500.

Akre, O., Barone-Adesi, F., Pettersson, A., Pearce, N., Merletti, F., & Richiardi, L. 2009. Differences in citation rates by country of origin for papers published in top-ranked medical journals: do they reflect inequalities in access to publication? *Journal of Epidemiology and Community Health*. http://dx.doi.org/10.1136/jech.2009.088690

Antelman, K. 2004. Do Open-Access Articles Have a Greater Research Impact? *College & Research Libraries* 65: 372-382. http://eprints.rclis.org/archive/00002309/

Archambault, E., Campbell, D., Gingras, Y., & Larivière, V. 2009. Comparing of Science Bibliometric Statistics Obtained From the Web and Scopus. *Journal of the American Society for Information Science and Technology* 60: 1320-1326. http://dx.doi.org/10.1002/asi.21062

Association of Research Libraries. (2002). Create Change. Retrieved Sept 21, 2006, http://www.arl.org/createchange/bm~doc/createchange2003.pdf

Association of Research Libraries. (2004a). Alliance for Taxpayer Access. http://www.taxpayeraccess.org/

Association of Research Libraries. (2004b). Monograph and Serial Costs in ARL Libraries, 1986-2004. http://www.arl.org/bm~doc/monser04.pdf

Association of Research Libraries. (2004c). Open Access. Retrieved April 11, 2008, http://www.arl.org/sparc/bm~doc/openaccess.pdf

Association of Research Libraries. (2006). Monograph and Serial Costs in ARL Libraries, 1986-2006. http://www.arl.org/bm~doc/monser06.pdf

Association of Research Libraries. (2009). *ARL Statistics 2007-2008*. Washington, DC, http://www.arl.org/bm~doc/arlstat08.pdf

Baldi, S. 1998. Normative versus Social Constructivist Processes in the Allocation of Citations: A Network-Analytic Model. *American Sociological Review* 63: 829-846. http://www.jstor.org/stable/2657504

Barabasi, A.-L., & Albert, R. 1999. Emergence of Scaling in Random Networks. *Science* 286: 509-512. http://dx.doi.org/10.1126/science.286.5439.509

Bax, L., Ikeda, N., Fukui, N., Yaju, Y., Tsuruta, H., & Moons, K. G. M. 2009. More Than Numbers: The Power of Graphs in Meta-Analysis. *Am. J. Epidemiol.* 169: 249-255. http://dx.doi.org/10.1093/aje/kwn340

Bensman, S. J. 1996. The Structure of the Library Market for Scientific Journals: The Case of Chemistry. *Library Resources and Technical Services* 40: 145-170.

Bergstrom, T. C., & Lavaty, R. (2007). How often to economists self-archive? University of California, Santa Barbara.

Biagioli, M. 1998. The Instability of Authorship: Credit and Responsibility in Contemporary Biomedicine. *FASEB Journal* 12: 3-16.

Biagioli, M. (2003). Rights or Rewards? Changing Frameworks of Scientific Authorship. In M. Biagioli & P. Galison (Eds.), *Scientific Authorship: Credit and Intellectual Property in Science* (pp. 253-279). New York: Routledge.

Björk, B.-C., Roos, A., & Lauri, M. 2009. Scientific journal publishing: yearly volume and open access availability. *Information Research* 14. http://informationr.net/ir/14-1/paper391.html

Bradford, S. C. *Documentation*. London, Lockwood, 1948, pp.156.

Broadus, R. N. 1983. An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science* 34: 132-135. http://dx.doi.org/10.1002/asi.4630340206

Brody, T., Harnad, S., & Carr, L. 2006. Earlier Web Usage Statistics as Predictors of Later Citation Impact. *Journal of the American Society for Information Science and Technology* 57: 1060-1072. http://dx.doi.org/10.1002/asi.20373

Brookes, B. C. 1971. Optimum P% Library of Scientific Periodicals. *Nature* 232: 458-461. http://dx.doi.org/10.1038/232458a0

Brooks, T. A. 1985. Private Acts and Public Objects: An Investigation of Citer Motivations. *Journal of the American Society for Information Science* 36: 223-229. http://dx.doi.org/10.1002/asi.4630360402

Brooks, T. A. 1986. Evidence of complex citer motivations. *Journal of the American Society for Information Science* 37: 34-36. http://dx.doi.org/10.1002/asi.4630370106

Burrell, Q. L. 1985. The 80/20 Rule: Library Lore or Statistical Law? *Journal of Documentation* 41: 24-39.

Burrell, Q. L. 2002. Will this paper ever be cited? *Journal of the American Society for Information Science and Technology* 53: 232-235.

Burrell, Q. L. 2003. Predicting future citation behavior. *Journal of the American Society for Information Science and Technology* 54: 372-378. http://dx.doi.org/10.1002/asi.10207

Burrell, Q. L. 2005. Are "Sleeping Beauties" to be expected? *Scientometrics* 65: 381-389. http://dx.doi.org/10.1007/s11192-005-0280-5

Butler, D. 2003. Scientific publishing: Who will pay for open access? *Nature* 425: 554-555. http://dx.doi.org/10.1038/425554a

Callaham, M., Wears, R. L., & Weber, E. 2002. Journal prestige, publication bias, and other characteristics associated with citation of published studies in peer-reviewed journals. *JAMA* 287: 2847-2850. http://dx.doi.org/10.1001/jama.287.21.2847

Calver, M. C., & Bradley, J. S. 2010. Patterns of Citations of Open Access and Non-Open Access Conservation Biology Journal Papers and Book Chapters. *Conservation Biology* 24: 872-880. http://dx.doi.org/10.1111/j.1523-1739.2010.01509.x

Campanario, J. M. 1996. Have Referees Rejected Some of the Most-Cited Articles of All Times? *Journal of the American Society for Information Science* 47: 302-310. http://dx.doi.org/10.1002/(SICI)1097-4571(199604)47:4<302::AID-ASI6>3.0.CO;2-0

Chapman, S., Nguyen, T. N., & White, C. 2007. Press-released papers are more downloaded and cited. *Tobacco Control* 16: 71. http://dx.doi.org/10.1136/tc.2006.019034

Chubin, D. E., & Moitra, S. D. 1975. Content Analysis of References: Adjunct or Alternative to Citation Counting? *Social Studies of Science* 5: 423-441.

Cole, J. R., & Cole, S. *Social stratification in science*. Chicago, University of Chicago Press, 1973, pp.283.

Cole, S. (2000). The Role of Journals in the Growth of Scientific Knowledge. In B. Cronin & H. B. Atkins (Eds.), *The Web of Knowledge: A Festschrift in Honor of Eugene Garfield* (pp. 109-142). Medford, N.J.: Information Today.

Cole, S., & Cole, J. R. 1967. Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review* 32: 377-390.

Conen, D., Torres, J., & Ridker, P. M. 2008. Differential Citation Rates of Major Cardiovascular Clinical Trials According to Source of Funding: A Survey From 2000 to 2005. *Circulation* 118: 1321-1327. http://dx.doi.org/10.1161/circulationaha.108.794016

Couzin, J. 2006. And how the problems eluded peer reviewers and editors. *Science* 311: 23-24. http://dx.doi.org/10.1126/science.311.5757.23

Cozzens, S. E. 1989. What do citations count? The rhetoric-first model. *Scientometrics* 15: 437-447. http://dx.doi.org/10.1007/BF02017064

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. 2007. Do Open Access Articles Have Greater Citation Impact? A critical review of the literature. *Journal of Informetrics* 1: 239-248. http://www.publishingresearch.net/Citations-SummaryPaper3_000.pdf.pdf

Crane, D. 1967. The gatekeepers of science: Some factors affecting the selection of articles for scientific journals. *American Sociologist* 2: 195-201.

Crane, D. *Invisible colleges; diffusion of knowledge in scientific communities*. Chicago, U. Chicago Press, 1972, pp.213.

Creative Commons. (2001). Licenses. http://creativecommons.org/about/licenses/

Cronin, B. *The citation process: the role and significance of citations in scientific communication* London, Taylor Graham, 1984, pp.103.

Cronin, B., & McKenzie, G. 1992. Documentation note: the trajectory of rejection. *Journal of Documentation* 48: 310-317.

Crow, R. (2002). *The Case for Institutional Repositories: A SPARC Position Paper*. Washington, DC: SPARC, http://www.arl.org/sparc/bm%7Edoc/ir_final_release_102.pdf

Davis, P. M. 2009a. Author-choice open access publishing in the biological and medical literature: a citation analysis. *Journal of the American Society for Information Science and Technology* 60: 3-8. http://dx.doi.org/10.1002/asi.20965

Davis, P. M. 2009b. How the Media Frames 'Open Access'. *Journal of Electronic Publishing* 12: e. http://dx.doi.org/10.3998/3336451.0012.101

Davis, P. M., & Connolly, M. J. L. 2007. Institutional Repositories: Evaluating the Reasons for Non-use of Cornell University's Installation of DSpace. *D-Lib Magazine* 13. http://www.dlib.org/dlib/march07/davis/03davis.html

Davis, P. M., & Fromerth, M. J. 2007. Does the arXiv lead to higher citations and reduced publisher downloads for mathematics articles? *Scientometrics* 71: 203-215. http://dx.doi.org/10.1007/s11192-007-1661-8

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., & Connolly, M. J. L. 2008. Open access publishing, article downloads and citations: randomised trial. *BMJ* 337: a568. http://dx.doi.org/10.1136/bmj.a568

Davis, P. M., & Price, J. S. 2006. eJournal interface can influence usage statistics: implications for libraries, publishers, and Project COUNTER. *Journal of the American Society for Information Science and Technology* 57: 1243-1248. http://dx.doi.org/10.1002/asi.20405

Deciphering citation statistics. 2008. *Nature Neuroscience* 11: 619-619. http://dx.doi.org/10.1038/nn0608-619

Ellison, G. 2002. The Slowdown of the Economics Publishing Process. *Journal of Political Economy* 110: 947-993. http://dx.doi.org/10.1086/341868

Ellison, G. (2007). *Is Peer Review in Decline?* (NBER Working Paper W13272): MIT, Dept. of Economics, http://ssrn.com/abstract=1002051

Evans, J. A. 2008. Electronic Publication and the Narrowing of Science and Scholarship. *Science* 321: 395-399. http://dx.doi.org/10.1126/science.1150473

Evans, J. A., & Reimer, J. 2009. Open Access and Global Participation in Science. *Science* 323: 1025-. http://dx.doi.org/10.1126/science.1154562

Eysenbach, G. 2006. Citation Advantage of Open Access Articles. *PLoS Biology* 4. http://dx.doi.org/10.1371/journal.pbio.0040157

Eysenbach, G., & Kohler, C. 2002. How do consumers search for and appraise health information on the world wide web? Qualitative study using focus groups, usability tests, and in-depth interviews. *BMJ* 324: 573-577. http://dx.doi.org/10.1136/bmj.324.7337.573

Eysenbach, G., Powell, J., Kuss, O., & Sa, E.-R. 2002. Empirical Studies Assessing the Quality of Health Information for Consumers on the World Wide Web: A Systematic Review. *JAMA* 287: 2691-2700. http://dx.doi.org/10.1001/jama.287.20.2691

Falagas, M. E., & Kavvadia, P. 2006. "Eigenlob": self-citation in biomedical journals. *FASEB J.* 20: 1039-1042. http://dx.doi.org/10.1096/fj.06-0603ufm

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. 2008. Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. *FASEB Journal* 22: 338-342. http://dx.doi.org/10.1096/fj.07-9492LSF

Federal Research Public Access Act (HR. 5037)(2010).

Federal Research Public Access Act (S.1373)(2009).

Fowler, J. H., & Aksnes, D. W. 2007. Does self-citation pay? *Scientometrics* 72: 427-437. http://dx.doi.org/10.1007/s11192-007-1777-2

Franck, G. 1999. Scientific Communication--A Vanity Fair? *Science* 286: 53-55. http://dx.doi.org/10.1126/science.286.5437.53

Frandsen, T. F. 2009. Attracted to open access journals: a bibliometric author analysis in the field of biology. *Journal of Documentation* 65: 58-82. http://dx.doi.org/10.1108/00220410910926121

Frazier, K. 2001. The Librarians' Dilemma: Contemplating the Costs of the "Big Deal". *D-Lib Magazine* 7. http://www.dlib.org/dlib/march01/frazier/03frazier.html

Friedman, L. M., Furberg, C. D., & DeMets, D. L. *Fundamentals of clinical trials*. Littleton, MA, PSG Pub., 1985, pp.302.

Fulda, J. S. 1998. Multiple Publication Reconsidered. *Journal of Information Ethics* 7: 47-53.

Garfield, E. 1955. Citation Indexes for Science: A New Dimension in Documentation through Association of Ideas. *Science* 122: 108-111. http://dx.doi.org/10.1126/science.122.3159.108

Garfield, E. 1972. Citation analysis as a tool in journal evaluation. *Science* 178: 471-479. http://dx.doi.org/10.1126/science.178.4060.471

Garfield, E. 1996. The Significant Scientific Literature Appears in a Small Core of Journals. *The Scientist* 10: 13.

Gargouri, Y., Hajjem, C., Larivière, V., Gingras, Y., Carr, L., Brody, T., et al. (2010). *Self-Selected or Mandated, Open Access Increases Citation Impact for Higher Quality Research*, http://arxiv.org/abs/1001.0361v2

Garvey, W. D., & Griffith, B. C. 1967. Scientific Communication as a Social System. *Science* 157: 1011-1016. http://dx.doi.org/10.1126/science.157.3792.1011

Garvey, W. D., & Griffith, B. C. 1971. Scientific communication: Its role in the conduct of research and creation of knowledge. *American Psychologist* 26: 350-362.

Gaulé, P. 2009. Access to scientific literature in India. *Journal of the American Society for Information Science and Technology* 12: 2548-2553. http://dx.doi.org/10.1002/asi.21195

Gaulé, P., & Maystre, N. (2008). *Getting cited: does open access help?* : CEMI Working Paper 2008-007, http://ilemt.epfl.ch/repec/pdf/cemi-workingpaper-2008-007.pdf

Gieryn, T. F. 1983. Boundary-Work and the Demarcation of Science from Non-Science: Strains and Interests in Professional Ideologies of Scientists. *American Sociological Review* 48: 781-795. http://www.jstor.org/stable/2095325

Gilbert, G. N. 1977. Referencing as Persuasion. *Social Studies of Science* 7: 113-122.

Glanzel, W., & Schubert, A. 1993. A characterization of scientometric distributions based on harmonic means. *Scientometrics* 26: 81-96.

Godlee, F. 2008. Open access to research. *BMJ* 337: a1051-. http://dx.doi.org/10.1136/bmj.a1051

Greenberg, S. A. 2009. How citation distortions create unfounded authority: analysis of a citation network. *BMJ* 339: b2680-. http://dx.doi.org/10.1136/bmj.b2680

Greenhalgh, T. 1997. How to read a paper : getting your bearings (deciding what the paper is about). *BMJ* 315: 243-246. http://www.bmj.com/cgi/content/full/315/7102/243

Gross, P. L. K., & Gross, E. M. 1927. College Libraries and Chemical Education. *Science* 66: 385-389. http://dx.doi.org/10.1126/science.66.1713.385

Groves, T. 2009. Nine in 10 articles rejected by NEJM appear in another journal. *BMJ* 339: b3777. http://dx.doi.org/10.1136/bmj.b3777

Hagstrom, W. O. *The scientific community*. New York, Basic Books, 1965, pp.304.

Hall, S. A., & Wilcox, A. J. 2007. The Fate of Epidemiologic Manuscripts: A Study of Papers Submitted to Epidemiology. *Epidemiology* 18: 262-265. http://dx.doi.org/10.1097/01.ede.0000254668.63378.32

Hamilton, D. P. 1990. Publishing by-and for?-the Numbers. *Science* 250: 1331-1332. http://dx.doi.org/10.1126/science.2255902

Hamilton, D. P. 1991. Research Papers: Who's Uncited Now? *Science* 251: 25. http://dx.doi.org/10.1126/science.1986409

Hardisty, D. J., & Haaga, D. A. F. 2008. Diffusion of treatment research: does open access matter? *Journal of Clinical Psychology* 64: 821-839. http://dx.doi.org/10.1002/jclp.20492

Harley, D., Acord, S. K., Earl-Novell, S., Lawrence, S., & King, C. J. (2010). *Assessing the Future Landscape of Scholarly Communication: An Exploration of Faculty Values and Needs in Seven Disciplines*: Center for Studies in Higher Education, UC Berkeley, http://escholarship.org/uc/item/15x7385g

Harnad, S., & Brody, T. 2004. Comparing the Impact of Open Access (OA) vs. Non-OA Articles in the Same Journals. *D-Lib Magazine* 10. http://www.dlib.org/dlib/june04/harnad/06harnad.html

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Thompson, D., et al. 2006. Effect of E-printing on Citation Rates in Astronomy and Physics. *Journal of Electronic Publishing* 9. http://hdl.handle.net/2027/spo.3336451.0009.202

Henneken, E. A., Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C. S., Thompson, D., et al. 2007. E-prints and journal articles in astronomy: a productive co-existence. *Learned Publishing* 20: 16-22. http://dx.doi.org/10.1087/095315107779490661

Hilbe, J. M. *Negative binomial regression*. Cambridge, Cambridge University Press, 2007, pp.251.

Hirshon, A., Sanville, T., Okerson, A., Kohl, D., Friend, F., Mittler, E., et al. 1998. Statement of current perspective and preferred practices for the selection and purchase of electronic information. *Information Technology and Libraries* 17: 45-50.

International Committee of Medical Journal Editors. (2009). Uniform Requirements for Manuscripts Submitted to Biomedical Journals: Publishing and Editorial Issues Related to Publication in Biomedical Journals: Overlapping Publications. http://www.icmje.org/publishing_4overlap.html

Ioannidis, J. P. A. 2006. Concentration of the Most-Cited Papers in the Scientific Literature: Analysis of Journal Ecosystems. *PLoS ONE* 1: e5. http://dx.plos.org/10.1371/journal.pone.0000005

ISI. (2004). *The Impact of Open Access Journals: A Citation Study from Thomson ISI*: Thomson Corporation,

Jadad, A. *Randomised Controlled Trials: a user's guide*. London, BMJ Books, 1998, pp.123.

Jakobsen, A. K., Christensen, R., Persson, R., Bartels, E. M., & Kristensen, L. E. 2010. Open access publishing And now, e-publication bias. *BMJ* 340: c2243-. http://dx.doi.org/10.1136/bmj.c2243

Joseph, H. D. (2008). Testimony before the Subcommittee on Courts, the Internet, and Intellectual Property Committee on the Judiciary on: H.R. 6845, the "Fair Copyright in Research Works Act," September 11, 2008. Retrieved Feb 11, 2009, http://judiciary.house.gov/hearings/pdf/Joseph080911.pdf

Kaplan, N. 1965. The Norms of Citation Behavior: Prolegomena to the Footnote. *American Documentation* 16: 179-184. http://dx.doi.org/10.1002/asi.5090160305

Kiernan, V. 2003. Diffusion of news about research. *Science Communication* 25: 3-13. http://dx.doi.org/10.1177/1075547003255297

King, C. J., Harley, D., Earl-Novell, S., Arter, J., Lawrence, S., & Perciali, I. (2006). *Scholarly Communication: Academic Values and Sustainable Models*. Berkeley, CA: Center for Studies in Higher Education, University of California, Berkeley, http://cshe.berkeley.edu/publications/docs/scholarlycomm_report.pdf

King, D. W., & Tenopir, C. 1999. Using and Reading Scholarly Literature. *Annual Review of Information Science and Technology* 34: 423-477.

Kingma, B. R. *The Economics of Information* (2nd ed.). Englewood, Libraries Unlimited, 2001, pp.180.

Klamer, A., & van Dalen, H. P. 2002. Attention and the art of scientific publishing. *Journal of Economic Methodology* 3: 289-315. http://dx.doi.org/10.1080/1350178022000015104

Knight, J. 2003. Cornell axes Elsevier journals as prices rise. *Nature* 426: 217-217. http://dx.doi.org/10.1038/426217a

Koepsell, T. D., & Weiss, N. S. *Epidemiologic methods*. New York, Oxford University Press, 2003, pp.513.

Kostoff, R. N. 2007. The difference between highly and poorly cited medical articles in the journal Lancet. *Scientometrics* 72: 513-520. http://dx.doi.org/10.1007/s11192-007-1573-7

Kulkarni, A. V., Busse, J. W., & Shams, I. 2007. Characteristics Associated with Citation Rate of the Medical Literature. *PLoS ONE* 2: e403. http://dx.doi.org/10.1371/journal.pone.0000403

Kurtz, M. J., Eichhorn, G., Accomazzi, A., Grant, C., Demleitner, M., Henneken, E., et al. 2005. The effect of use and access on citations. *Information Processing and Management* 41: 1395-1402. http://dx.doi.org/10.1016/j.ipm.2005.03.010

Kurtz, M. J., & Henneken, E. A. (2007). *Open Access does not increase citations for research articles from The Astrophysical Journal*: Harvard-Smithsonian Center for Astrophysics, http://arxiv.org/abs/0709.0896

Lansingh, V. C., & Carter, M. J. 2009. Does Open Access in Ophthalmology Affect How Articles are Subsequently Cited in Research? *Ophthalmology* 116: 1425-1431. http://dx.doi.org/10.1016/j.ophtha.2008.12.052

Larivière, V., & Gingras, Y. 2010. The impact factor's Matthew effect: a natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology* 61: 424-427. http://dx.doi.org/10.1002/asi.21232

Larivière, V., Gingras, Y., & Archambault, É. 2008. The decline in the concentration of citations, 1900-2007. *Journal of the American Society for Information Science and Technology* 60: 858-862. http://dx.doi.org/10.1002/asi.21011

Latour, B. *Laboratory life : the construction of scientific facts*. Princeton, N.J., Princeton University Press, 1986, pp.294.

Latour, B. *Science in action : how to follow scientists and engineers through society*. Cambridge, Mass, Harvard University Press, 1987, pp.274.

Laurent, M. R., & Vickers, T. J. 2009. Seeking Health Information Online: Does Wikipedia Matter? *Journal of the American Medical Informatics Association* 16: 471-479. http://dx.doi.org/10.1197/jamia.M3059

Lawrence, S. 2001. Free online availability substantially increases a paper's impact. *Nature* 411: 521. http://dx.doi.org/10.1038/35079151

Lewis, S., & Clarke, M. 2001. Forest plots: trying to see the wood and the trees. *BMJ* 322: 1479-1480. http://dx.doi.org/10.1136/bmj.322.7300.1479

Leydesdorff, L., & Bensman, S. 2006. Classification and Powerlaws: The Logarithmic Transformation. *Journal of the American Society for Information Science and Technology* 57: 1470-1486. http://dx.doi.org/10.1002/asi.20467

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., GÃ¸tzsche, P. C., Ioannidis, J. P. A., et al. 2009. The PRISMA Statement for Reporting Systematic Reviews and Meta-Analyses of Studies That Evaluate Health Care Interventions: Explanation and Elaboration. *Annals of Internal Medicine* 151: W-65-W-94. http://dx.doi.org/10.1059/0003-4819-151-4-200908180-00136

Liesegang, T. J., Shaikh, M., & Crook, J. E. 2007. The outcome of manuscripts submitted to the American Journal of Ophthalmology between 2002 and 2003. *American Journal of Ophthalmology* 143: 551-560. http://dx.doi.org/10.1016/j.ajo.2006.12.004

Lokker, C., McKibbon, K. A., McKinlay, R. J., Wilczynski, N. L., & Haynes, R. B. 2008. Prediction of citation counts for clinical articles at two years using data available within three weeks of publication: retrospective cohort study. *BMJ* bmj.39482.526713.BE. http://dx.doi.org/10.1136/bmj.39482.526713.BE

Luukkonen, T. 1997. Why has Latour's theory of citations been ignored by the bibliometric community? Discussion of sociological interpretations of citation analysis. *Scientometrics* 38: 27-37. http://dx.doi.org/10.1007/BF02461121

Mabe, M. A., & Amin, M. 2002. Dr. Jekyll and Dr. Hyde: author-reader asymmetries in scholarly publishing. *ASLIB Proceedings* 54: 149-157. http://dx.doi.org/10.1108/00012530210441692

Mager, A. 2009. Mediated health: sociotechnical practices of providing and using online health information. *New Media Society* 11: 1123-1142. http://dx.doi.org/10.1177/1461444809341700

Mason, P. M., Steagall, J. W., & Fabritius, M. M. 1992. Publication delays in articles in economics: what to do about them. *Applied Economics* 24: 859-874. http://dx.doi.org/10.1080/00036849200000054

McCabe, M. J. 1998. The Impact of Publisher Mergers on Journal Prices: A Preliminary Report. *Bimonthly Newsletter of Research Library Issues and Actions*. http://www.arl.org/resources/pubs/br/br200/br200mccabe.shtml

McCabe, M. J. 1999. The Impact of Publisher Mergers on Journal Prices: An Update. *A Bimonthly Report on Research Library Issues and Actions from ARL, CNI, and SPARC* 207. http://www.arl.org/resources/pubs/br/br207/br207jrnlprices.shtml

McCabe, M. J. 2002. Journal Pricing and Mergers: A Portfolio Approach. *American Economic Review* 92: 259-269. http://dx.doi.org/10.1257/000282802760015702

McDonald, J. 2007. Understanding Journal Usage: A Statistical Analysis of Citation and Use. *Journal of the American Society for Information Science and Technology* 58: 39-50. http://dx.doi.org/10.1002/asi.20420

McDonald, R. J., Cloft, H. J., & Kallmes, D. F. 2009. Fate of Manuscripts Previously Rejected by the American Journal of Neuroradiology: A Follow-Up Analysis. *American Journal of Neuroradiology* 30: 253-256. http://dx.doi.org/10.3174/ajnr.A1366

Merton, R. K. 1942. Science and Technology in a Democratic Order. *Journal of Legal and Political Sociology* 1: 115-126.

Merton, R. K. 1957. Priorities in scientific discovery: a chapter in the sociology of science. *American Sociological Review* 22: 635-659.

Merton, R. K. 1968. The Matthew Effect in Science. *Science* 159: 56-63. http://dx.doi.org/10.1126/science.159.3810.56

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

Merton, R. K. 1988. The Matthew Effect in Science, II: Cumulative Advantage and the Symbolism of Intellectual Property. *Isis* 79: 606-623.

Metcalfe, T. S. 2005. The rise and citation impact of astro-ph in major journals. *Bulletin of the American Astronomical Society* 37: 555-557. http://arxiv.org/abs/astro-ph/0503519v1

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

Moed, H. F. 2005. Statistical relationships between downloads and citations at the level of individual documents within a single journal. *Journal of the American Society for Information Science and Technology* 56: 1088-1097. http://dx.doi.org/10.1002/asi.20200

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* 58: 2047-2054. http://dx.doi.org/10.1002/asi.20663

Moravcsik, M. J., & Murugesan, P. 1975. Some Results on the Function and Quality of Citations. *Social Studies of Science* 5: 86-92.

Nadarajah, S., & Kotz, S. 2007. Models for citation behavior. *Scientometrics* 72: 291-305. http://dx.doi.org/10.1007/s11192-007-1717-9

National Cancer Institute. (2007). Health Information National Trends Survey (HINTS). http://hints.cancer.gov/questions/index.jsp

National Institutes of Health. (2009). National Institutes of Health Public Access. Retrieved Mar 23, 2009, http://publicaccess.nih.gov/

National Science Foundation. (2010). *Science and Engineering Indicators 2010*. Arlington, VA, http://www.nsf.gov/statistics/seind10/

Newman, M. E. J. 2001. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America* 98: 404-409. http://dx.doi.org/10.1073/pnas.021544898

Newman, M. E. J. 2004. Coauthorship networks and patterns of scientific collaboration. *Proceedings of the National Academy of Sciences of the United States of America* 101: 5200-5205. http://dx.doi.org/10.1073/pnas.0307545100

Newman, M. E. J. 2009. The first-mover advantage in scientific publication. *EPL* 86: 68001-. http://dx.doi.org/10.1209/0295-5075/86/68001

Norris, M., Oppenheim, C., & Rowland, F. 2008. The citation advantage of open-access articles. *Journal of the American Society for Information Science and Technology* 59: 1963-1972. http://dx.doi.org/10.1002/asi.20898

Open Society Institute. (2001). Budapest Open Access Initiative. Retrieved April 23, 2003, http://www.soros.org/openaccess/

Opthof, T., Furstner, F., van Geer, M., & Coronel, R. 2000. Regrets or no regrets? No regrets! The fate of rejected manuscripts. *Cardiovascular Research* 45: 255-258. http://dx.doi.org/10.1016/s0008-6363(99)00339-9

Oxford English Dictionary. *open access* (2nd ed.). New York,

Patsopoulos, N. A., Analatos, A. A., & Ioannidis, J. P. A. 2005. Relative Citation Impact of Various Study Designs in the Health Sciences. *JAMA* 293: 2362-2366. http://dx.doi.org/10.1001/jama.293.19.2362

PEER: Publishing and the Ecology of European Research. (2010). http://www.peerproject.eu/

Perneger, T. V. 2004. Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ. *BMJ* 329: 546-547. http://dx.doi.org/10.1136/bmj.329.7465.546

Peters, D. P., & Ceci, S. J. 1982. Peer-review practices of psychological articles: the fate of accepted, published articles, submitted again. *Behavioral and Brain Sciences* 5: 187-195.

Pew Internet & American Life Project. (2006a). *E-patients With a Disability or Chronic Disease* http://www.pewinternet.org/~/media//Files/Reports/2007/EPatients_Chronic_Conditions_2007.pdf.pdf

Pew Internet & American Life Project. (2006b). *Online Health Search 2006*, http://www.pewinternet.org/~/media//Files/Reports/2006/PIP_Online_Health_2006.pdf.pdf

Pew Internet & American Life Project. (2008). *The Engaged E-patient Population*, http://www.pewinternet.org/~/media//Files/Reports/2008/PIP_Health_Aug08.pdf.pdf

Pew Internet & American Life Project. (2009). *The Social Life of Health Information*, http://www.pewinternet.org/~/media//Files/Reports/2009/PIP_Health_2009.pdf

Phillips, D., Kanter, E., Bednarczyk, B., & Tastad, P. 1991. Importance of the lay press in the transmission of medical knowledge to the scientific community. *New England Journal of Medicine* 325: 1180-1183.

Phillips, L. L., & Williams, S. R. 2004. Collection development embraces the digital age - A review of the literature, 1997-2003. *Library Resources & Technical Services* 48: 273-299.

Pirolli, P., & Card, S. 1999. Information Foraging. *Psychological Review* 106: 643-675. http://dx.doi.org/10.1037/0033-295X.106.4.643

Piwowar, H. A., Day, R. S., & Fridsma, D. B. 2007. Sharing Detailed Research Data Is Associated with Increased Citation Rate. *PLoS One* 2: e308. http://dx.doi.org/10.1371/journal.pone.0000308

Polanyi, M. 1962. The Republic of Science: Its Political and Economic Theory. *Minerva* I: 54-73.

Price, D. J. S. 1965. Networks of Scientific Papers. *Science* 149: 510-515. http://dx.doi.org/10.1126/science.149.3683.510

Price, D. J. S. 1976. A General Theory of Bibliometric and Other Cumulative Advantage Processes. *Journal of the American Society for Information Science* 27: 292-306. http://dx.doi.org/10.1002/asi.4630270505

Price, D. J. S. (1986). Collaboration in an Invisible College. In *Little science, big science...and beyond* (pp. 119-134). New York: Columbia University Press.

ProQuest. (2009). Ulrich's Periodicals Directory. Retrieved Feb 1, 2010, http://www.ulrichsweb.com/ulrichsweb/

Public Library of Science. (2001). Open Letter. http://www.plos.org/support/openletter.shtml

Random.org. http://www.random.org/sequences/

Research Information Network. (2009). *E-journals: their use, value and impact*, http://www.rin.ac.uk/files/E-journals_use_value_impact_Report_April2009.pdf

Rosen, S. 1981. The Economics of Superstars. *The American Economic Review* 71: 845-858.

Rowlands, I., & Nicholas, D. (2005). *New journal publishing models: an international survey of senior researchers*: CIBER, http://www.ucl.ac.uk/ciber/ciber_2005_survey_final.pdf

Rowlands, I., & Olivieri, R. (2006). *Journals and Scientific Productivity: A case study in immunology and microbiology*. London: Publishing Research Consortium, http://www.publishingresearch.net/documents/Rowland_Olivieri_summary_paper.pdf

Salganik, M. J., Dodds, P. S., & Watts, D. J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* 311: 854-856. http://dx.doi.org/10.1126/science.1121066

Sanville, T. 1999. A license to deal. *Library Journal* 124: 122-124.

Schonfeld, R. C., & Housewright, R. (2010). *Faculty Survey 2009: Key Strategic Insights for Libraries, Publishers, and Societies*. New York: Ithaka S+R, http://www.ithaka.org/ithaka-s-r/research/faculty-surveys-2000-2009/faculty-survey-2009

Schriger, D. L., Altman, D. G., Vetter, J. A., Heafner, T., & Moher, D. Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. *Int. J. Epidemiol.* online before print. http://dx.doi.org/10.1093/ije/dyp370

Schroter, S., & Tite, L. 2006. Open access publishing and author-pays business models: a survey of authors' knowledge and perceptions. *Journal of the Royal Society of Medicine* 99: 141-148. http://dx.doi/org/10.1258/jrsm.99.3.141

Schroter, S., Tite, L., & Smith, R. 2005. Perceptions of open access publishing: interviews with journal authors. *BMJ* 330: 756-. http://dx.doi.org/10.1136/bmj.38359.695220.82

Schwarz, G. J., & Kennicutt, R. C. J. 2004. Demographic and Citation Trends in Astrophysical Journal papers and Preprints. *Bulletin of the American Astronomical Society* 36: 1654-1663. http://arxiv.org/abs/astro-ph/0411275v1

Seglen, P. O. 1992. The Skewness of Science. *Journal of the American Society for Information Science* 43: 628-638. http://dx.doi.org/10.1002/(SICI)1097-4571(199210)43:9<628::AID-ASI5>3.0.CO;2-0

Shannon, C. E., & Weaver, W. *Mathematical Theory of Communication*. Urbana, University of Illinois Press, 1949, pp.125.

Simkin, M. V., & Roychowdhury, V. P. 2005. Stochastic modeling of citation slips. *Scientometrics* 62: 367-384. http://dx.doi.org/10.1007/s11192-005-0028-2

Simkin, M. V., & Roychowdhury, V. P. 2007. A mathematical theory of citing. *Journal of the American Society for Information Science and Technology* 58: 1661-1673. http://dx.doi.org/10.1002/asi.20653

Simon, H. A. 1955. On a Class of Skew Distribution Functions. *Biometrika* 42: 425-440.

Simon, H. A. (1971). Designing organizations for an information-rich world. In M. Greenberger (Ed.), *Computers, Communications, and the Public Interest* (pp. 37-72). Baltimore: Johns Hopkins Press.

Spector, P. E. (2003). Response Bias. In *Encyclopedia of Social Science Research Methods*: SAGE Publications.

Spence, M. 1973. Job Market Signaling. *The Quarterly Journal of Economics* 87: 355-374.

Stanley, K. 2007. Design of Randomized Controlled Trials. *Circulation* 115: 1164-1169. http://dx.doi.org/10.1161/CIRCULATIONAHA.105.594945

Stewart, J. A. 1983. Achievement and Ascriptive Processes in the Recognition of Scientific Articles. *Social Forces* 62: 166-189.

Surowiecki, J. *The wisdom of crowds*. New York, Doubleday, 2004, pp.296.

Tenopir, C., & King, D. W. (2000). Readership of Scientific Scholarly Journals. Chapter 7. In *Toward Electronic Journals: Realities for Scientists, Librarians, and Publishers.* Washington: Special Libraries Association.

Tenopir, C., & King, D. W. 2002. Reading behaviour and electronic journals. *Learned Publishing* 15: 259-265. http://dx.doi.org/10.1087/095315102760319215

Tenopir, C., & King, D. W. 2008. Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns. *D-Lib Magazine* 14. http://www.dlib.org/dlib/november08/tenopir/11tenopir.html

Tenopir, C., King, D. W., Boyce, P., Grayson, M., Zhang, Y., & Ebuen, M. 2003. Patterns of Journal Use by Scientists Through Three Evolutionary Phases. *D-Lib Magazine* 9. http://www.dlib.org/dlib/may03/king/05king.html

Tenopir, C., King, D. W., Edwards, S., & Wu, L. 2009. Electronic journals and changes in scholarly article seeking and reading patterns. *ASLIB Proceedings* 61: 5-32. http://dx.doi.org/10.1108/00012530910932267

Thatcher, S. G. 1995. The crisis in scholarly communication. *Chronicle of Higher Education* 41: B1.

Toms, E. G., & Latter, C. 2007. How consumers search for health information. *Health Informatics Journal* 13: 223-235. http://dx.doi.org/10.1177/1460458207079901

van Dalen, H. P., & Henkens, K. 2001. What makes a scientific article influential? The case of demographers. *Scientometrics* 50: 455-482. http://dx.doi.org/10.1023/A:1010510831718

Ware, M. (2007). *Peer review in scholarly journals: Perspectives of the scholarly community -- an international study*. Bristol, UK: Publishing Research Consortium,http://www.publishingresearch.net/documents/PeerReviewFullPRCReport-final.pdf

Ware, M. (2009). *Access by UK small and medium-sized enterprises to professional and academic information*. Bristol, UK: Publishing Research Consortium, http://www.publishingresearch.net/documents/SMEAccessResearchReport.pdf

Ware, M., & Mabe, M. (2009). *The stm report: An overview of scientific and scholarly journals publishing*. Oxford: International Association of Scientific, Technical and Medical Publishers, http://www.stm-assoc.org/2009_10_13_MWC_STM_Report.pdf

Wijnhoven, B. P. L., & Dejong, C. H. C. 2010. Fate of manuscripts declined by the British Journal of Surgery. *British Journal of Surgery* 97: 450-454. http://dx.doi.org/10.1002/bjs.6880

Willinsky, J. 2003. The Nine Flavours of Open Access Scholarly Publishing. *Journal of Postgraduate Medicine* 49: 263-267. http://www.jpgmonline.com/text.asp?2003/49/3/263/1146

Willinsky, J. *The access principle: the case for open access to research and scholarship*. Cambridge, MA, MIT Press, 2006, pp.287.

Willinsky, J. 2009. The Publishers' Pushback against NIH's Public Access and Scholarly Publishing Sustainability. *PLoS Biology* 7: e30. http://dx.doi.org/10.1371/journal.pbio.1000030

Wren, J. D. 2005. Open access and openly accessible: a study of scientific publications shared via the internet. *BMJ* 330. http://dx.doi.org/10.1136/bmj.38422.611736.E0

Zerhouni, E. A. (2008). Testimony before the Subcommittee on Courts, the Internet, and Intellectual Property Committee on the Judiciary on: H.R. 6845, the "Fair Copyright in Research Works Act," September 11, 2008. Retrieved Feb 11, 2009, http://judiciary.house.gov/hearings/pdf/Zerhouni080911.pdf

Zuckerman, H., & Merton, R. K. 1971. Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva* 9: 66-100. http://dx.doi.org/10.1007/BF01553188