Do Download Reports Reliably Measure Journal Usage? Trusting the Fox to Count Your Hens?

Alex Wood-Doughty, Ted Bergstrom, and Douglas G. Steigerwald

Download rates of academic journals have joined citation counts as commonly used indicators of the value of journal subscriptions. While citations reflect worldwide influence, the value of a journal subscription to a single library is more reliably measured by the rate at which it is downloaded by local users. If reported download rates accurately measure local usage, there is a strong case for using them to compare the cost-effectiveness of journal subscriptions. We examine data for nearly 8,000 journals downloaded at the ten universities in the University of California system during a period of six years. We find that controlling for number of articles, publisher, and year of download, the ratio of downloads to citations differs substantially among academic disciplines. After adding academic disciplines to the control variables, there remain substantial "publisher effects", with some publishers reporting significantly more downloads than would be predicted by the characteristics of their journals. These cross-publisher differences suggest that the currently available download statistics, which are supplied by publishers, are not sufficiently reliable to allow libraries to make subscription decisions based on price and reported downloads, at least without making an adjustment for publisher effects in download reports.

Introduction

Measures of the influence of academic research are valuable to many decision makers. University librarians use them to make purchasing and renewal decisions.[1] Academic departments use them in their hiring, tenure, and salary decisions.[2] Funding agencies use them to assess grant applicants. They are also used in determining the public rankings of journals, academic departments, and universities.[3]

Citation counts have long been the most common measure of research influence. Eugene Garfield's Institute for Scientific Information introduced the systematic use of citation data with the Science Citation Index in 1964 and Journal Citation Reports (JCR) in 1975.[4] The advent of electronic publishing has given rise to a new measure of research influence: download counts.[5]

Alex Wood-Doughty is Research Scientist, Lyft, email: awooddoughty@gmail.com. Ted Bergstrom, and Douglas G. Steigerwald are Professors in the Department of Economics at the University of California Santa Barbara; email: abwood@umail.ucsb.edu, tedb@econ.ucsb.edu, doug@ucsb.edu. *The authors thank Chan Li and Nga Ong of the California Digital Library for helping them obtain download data. They also thank Carl Bergstrom of the University of Washington and Kristin Antelman of the Caltech Library for helpful suggestions and discussions.*

For library evaluations, accurate download counts could offer important advantages over citation counts. Only a minority of those who download a journal article will cite it. Citation counts reflect the activities of scholars worldwide. Subscribing libraries can observe the number of downloads from their own institutions, which reflect their own patterns of research interest.

For academic departments and granting agencies, the use of download data in addition to citation records yields an enriched profile of the influence of individual researchers' work.[6] Download data have the advantage of being much more immediate than citation data, a valuable feature for tenure committees or grant review panels tasked with evaluating the work of younger academics.

Several previous articles have explored correlations between citations and recorded downloads.[7] Brody, Harnad, and Carr[8] examine the extent to which downloads from the physics e-print archive, arXiv.org, predict later citations of an article. McDonald[9] explores the ability of prior downloads at the California Institute of Technology (Caltech) to predict article citations by authors from Caltech. Previous studies of download behavior have been limited to a small number of journals within a few specialized disciplines. Our download data include recorded downloads at the ten University of California campuses from nearly 8,000 academic journals in a wide variety of academic disciplines.

The number of downloads from a volume of an academic journal is highly correlated with the number of times it is cited. A simple linear regression of downloads on citations finds that 78 percent of the variation in downloads can be "explained" by variation in citations. Despite this strong correlation, there are important systematic differences between download rates and citation rates. For example, the ratio of downloads to citations in the arts and humanities is significantly higher than in other disciplines while that in the physical sciences is significantly lower.

Given that a library's own download rates reflect the demands of its users more closely than citation rates, there appears to be a strong case for using download rates rather than citation rates to evaluate journal subscriptions. The case for using download rates depends, however, on the assumption that these rates are accurately measured.

Libraries do not, in general, maintain their own download counts. This information is collected and supplied by publishers in summary form to subscribing libraries. Davis and Price[10] suggest that

> The number of full-text downloads may be artificially inflated when publishers require users to view HTML versions before accessing PDF versions or when linking mechanisms, such as CrossRef, direct users to the full text rather than the abstract of each article. The publishers, who control the raw data on downloads, have a strong incentive to release statistics that may overstate the number of actual users.

Subscribers are not given access to the publishers' web server log files from which the reports they receive are compiled; thus, they have no independent way of verifying the artificial inflation of download counts. Publishers are well aware that their download reports will influence librarians' subscription decisions. Davis and Price quote Sir Crispin Davis, who as CEO of Reid-Elsevier in 2004 testified to the British House of Commons as follows:

> The biggest single factor is usage. That is what librarians look at more than anything else and it is what they [use to] determine whether they renew, do not

> renew and so on. We have usage going up by an average of 75 per cent each year. In other words, the cost per article download is coming down by around 70 per cent each year. That is fantastic value for money in terms of the institution, so I would say that [usage] is the single biggest factor.

Because download statistics are not managed in a transparent way by impartial arbiters, it is reasonable to ask whether publisher-supplied data on downloads can be reliably compared across publishers. The University of California has "Big Deal" subscriptions for all of the journals published by each of the seven publishers treated here ("Big Deal" refers to an agreement to purchase nearly the entire portfolio of journals from a publisher). If the relation between recorded downloads and actual usage is the same across publishers, we would expect that, after controlling for journal characteristics such as citations, number of articles, and academic discipline, the identity of the publisher should have little or no effect on the number of downloads at the University of California. We find, however, strong and statistically significant publisher effects that are persistent under a variety of specifications of variables.

Data

We have obtained download records from the California Digital Library (CDL), which handles subscriptions for all ten campuses of the University of California system. These records represent about 4.25 million downloads from 7,724 journals published by seven publishers during the years 2010–2016.

Reports on the number of downloads from each article are supplied by the publishers, who prepare this data according to guidelines set by COUNTER (Counting Online Usage of Networked Electronic Resources), a nonprofit organization established by libraries, data vendors, and publishers. Most publishers provide journal download reports at COUNTER level JR1, which records the monthly number of downloads to all articles that have ever been published in that journal, but they do not report the year in which the downloaded articles were published. A few publishers offer more detailed reports at COUNTER level JR5. The JR5 reports record the number of downloads in the current year, while specifying the year in which each downloaded article was published. For example, the JR5 data for 2015 reports the number of articles that were published in each year since 2000 and downloaded in 2015. While many publishers include clauses in their contracts that forbid public access to this information, the CDL contracts do not include such restrictive clauses. The seven publishers used in our analysis are those who supplied the CDL with download data at the JR5 level.

The JR5 reports released by publishers include downloads, not only from their subscription journals, but also from the open access journals that they publish. When a library uses download counts to evaluate a "Big Deal" package that allows access to its subscription journals, it would not be appropriate to include downloads from open access journals, since these are accessible whether or not the library subscribes. For this reason, we confine our analysis to journals that require paid subscriptions for access.

Journals are placed into field classifications according to Scimago's designation, which uses a classification system developed by Elsevier's Scopus to partition journals into disciplinary categories at three distinct levels of detail.[11] At the broadest level of classification, there are five major categories: life sciences, physical sciences, health sciences, social sciences, and arts and

humanities.[12] At an intermediate level of detail, Scopus specifies 27 "major fields." At the most detailed level, Scopus assigns each journal to one or more of 334 "minor fields." Many journals are classified as belonging to more than one category. Where a journal is assigned to k different categories, we treat it as if one kth of its articles are in each of the k categories to which Scopus has assigned it. For example, if Scopus designates a journal as belonging to three "major" fields (such as mathematics, computer science, and economics), we would assign it an indicator value of 1/3 for each of these three fields. Our sample includes four large commercial publishers—Elsevier, Springer, Taylor & Francis, and Wiley—that publish across many disciplines. We also include the Nature Publishing group, which specializes in life and physical sciences, and two professional disciplinary society publishers: the American Chemical Society (ACS) and the Institute of Electrical and Electronics Engineers (IEEE). For each of these publishers, we have four to six years of reports on the annual number of downloads occurring in years from 2011 to 2016, where the downloaded articles were published between 2000 and 2016.

Table 1 shows the distribution of subscription-based journals by broad research field across publishers.[13] As the table shows, each of the four large commercial publishers has a significant presence in all five research fields, while the other publishers have more limited scope. The Nature Publishing Group journals are sorted into *Nature*-branded, the 30 journals under the imprimatur *Nature (subject)* (such as *Nature Astronomy*), and Other, the remaining 42 NPG journals that do not include *Nature* in their title. As table 3 will show, articles in the *Nature*-branded journals are much more cited and even more frequently downloaded than the other NPG journals.[14]

TABLE 1
Number of Subscription Journals by Research Field and Publisher

	Arts and Humanities	Health Sciences	Life Sciences	Physical Sciences	Social Sciences	Number of Journals
American Chemical Society	0	7	9	31	1	47
Elsevier	28	808	405	681	314	2,235
IEEE	3	2	5	192	10	212
NPG: Nature-branded	0	11	16	7	1	34
NPG: Other	0	19	17	1	0	36
Springer	63	370	314	809	310	1,865
Taylor & Francis	283	259	189	377	947	2,054
Wiley	81	320	235	278	324	1,238
Total	457	1,796	1,190	2,375	1,906	7,724

Note: Journals classified as belonging to multiple disciplines are assigned fractionally to these disciplines. Totals are rounded to nearest integer.

The analysis will also require information on citations and the number of articles published, both by journal. The citations measure is obtained from the website *SCImago Journal & Country Rank*[15] that records, for each journal, and for each year, the number of citations to articles published in that journal in the preceding three years. The number of articles used in our calculation is the annual number of "documents" reported by Scimago for each journal.[16]

Downloads and Citation Patterns by Field and Publisher

Journal articles in the life and health sciences tend to be more frequently cited than those in the physical and social sciences, while journals in arts and humanities are significantly less frequently cited than those in all other disciplinary areas. Differences in downloads per recent article also differ by discipline, but less drastically.

In table 2, recent UC downloads are measured as the number of downloads from a journal in the first three years after publication (the year of publication and the following two years). The reported ratio is constructed by dividing this download count by the number of articles from the journal in the year of publication. The corresponding measure for citations, more commonly known as the *impact factor*, simply replaces the number of downloads with the number of citations. This table also shows the ratio of these two measures, the size of which depends on the fact that downloads are counted only from UC campuses, while citations are counted worldwide. For each measure, summary statistics are reported.

According to table 2, the ratio of downloads to citations in arts and humanities is significantly higher than for other disciplines. While journals in the arts and humanities tend to have fewer citations per article than other disciplines, the number of downloads per article is nearly as large as that for the physical and social sciences. This suggests that the use of citation rates rather than download rates is likely to undervalue journals in arts and humanities relative to other fields.

TABLE 2
Recent Downloads and Citations per Article, by Broad Research Area

	Mean	Median	75th Percentile	90th Percentile
Arts and Humanities				
Recent UC downloads per article	5.3	3.1	6.7	12.5
Recent citations per article (Impact factor)	2.7	1.7	3.7	7.9
Ratio: UC downloads/citations	1.8	1.7	1.8	1.8
Health Sciences				
Recent UC downloads per article	9.7	5.7	11.1	19.6
Recent citations per article (Impact factor)	7.4	5.9	9.4	13.7
Ratio: UC downloads/citations	1.3	0.9	1.1	1.4
Life Sciences				
Recent UC downloads per article	13.9	6.3	12.4	24.0
Recent citations per article (Impact factor)	9.4	7.3	11.0	16.6
Ratio: UC downloads/citations	1.4	0.8	1.4	1.1
Physical Sciences				
Recent UC downloads per article	5.5	2.6	5.8	10.7
Recent citations per article (Impact factor)	6.8	5.0	8.4	13.0
Ratio: UC downloads/citations	0.8	0.5	0.7	0.8
Social Sciences				
Recent UC downloads per article	5.6	3.0	6.9	13.2
Recent citations per article (Impact factor)	4.4	3.2	5.7	9.2
Ratio: UC downloads/citations	1.3	0.9	1.2	1.4

Table 3 shows the distribution of the three measures of downloads and citations for each of the seven publishers in our sample. This table shows that the *Nature*-branded journals have a far higher ratio of downloads to citations than any of the other publisher groups. The ratio for NPG's other journals is lower than for the *Nature*-branded journals, but it remains high relative to most other publishers. NPG's *Nature*-branded journals have a special feature that at least partially explains their high download-to-citation ratios: typically more than half of the

TABLE 3
Recent Downloads and Citations per Article, by Publisher

	Mean	Median	75th Percentile	95th Percentile
ACS				
Recent UC downloads per article	19.0	12.3	18.8	37.2
Recent citations per article (Impact factor)	21.1	15.3	19.2	41.2
Ratio: UC downloads/citations	0.9	0.8	1.0	1.0
Elsevier				
Recent UC downloads per article	12.8	7.4	13.9	25.5
Recent citations per article (Impact factor)	9.1	7.5	11.1	16.0
Ratio: UC downloads/citations	1.4	1.0	1.3	1.6
IEEE				
Recent UC downloads per article	5.6	4.2	7.0	11.3
Recent citations per article (Impact factor)	10.8	8.9	13.8	20.6
Ratio: UC downloads/citations	0.5	0.5	0.5	0.5
NPG: Nature-branded				
Recent UC downloads per article	196.0	198.4	252.7	350.9
Recent citations per article (Impact factor)	58.9	52.8	81.1	95.8
Ratio: UC downloads/citations	3.3	3.8	3.1	3.7
NPG: Other				
Recent UC recent downloads per article	26.6	19.9	32.2	55.5
Recent citations per article (Impact factor)	15.2	12.8	18.9	28.1
Ratio: UC downloads/citations	1.8	1.6	1.6	2.0
Springer				
Recent UC downloads per article	4.6	2.4	6.0	10.5
Recent citations per article (Impact factor)	4.9	4.0	6.8	9.9
Ratio: UC downloads/citations	0.9	0.6	0.9	1.1
Taylor Francis				
Recent UC downloads per article	3.0	1.4	3.6	6.9
Recent citations per article (Impact factor)	0.9	0.5	0.8	1.0
Ratio: UC downloads/citations	0.9	0.5	0.8	1.0
Wiley				
Recent UC downloads per article	7.4	5.2	9.0	15.4
Recent citations per article (Impact factor)	7.3	5.8	9.2	14.1
Ratio: UC downloads/citations	1.0	0.9	1.0	1.1

articles appear in a *News and Views* section. These articles are brief reports on recent research, targeted at nonspecialists. The *News and Views* reports are often commissioned to prestigious scholars and closely edited by professional staff. Because these articles are generally not the first to report new results, they are not often cited in the specialist literature. However, they are extremely popular and widely read because they are of high quality and easily absorbed by a wide audience. Among the publishers other than NPG, Elsevier has the highest ratio of reported downloads to citations.

Possibly the differences between publishers' download-to-citation ratios could be explained by differences in the academic disciplines that they cover or by differences in the impact factors of their journals. Table 2 shows that the ratio of downloads to citations differs among academic disciplines and also differs with the impact factor of the journal, while table 1 shows that the publishers in our sample differ significantly in the distribution of academic disciplines that they cover. In the following sections, we apply statistical analysis to explore the extent to which these cross-publisher differences can be explained by observable characteristics of the journals that they publish.

Predicting Downloads from Journal Characteristics

Table 3 describes the relation between downloads per article and just two variables: impact factor and publishers. To account for the simultaneous effects on downloads of a longer list of characteristics, we estimate a function that predicts the number of downloads from a journal as a function of several variables describing that journal. Among the explanatory variables to be considered are the number of articles in a journal, the average number of citations per article (impact factor), the date of download, and the academic discipline to which the journal is devoted. We consider three specifications, which vary by the level of detail for the field classification of the journals: the five broad categories shown in table 1, the 27 major fields defined by Scopus, and the full set of 334 fields defined by Scopus.

Having controlled for a journal's citations, impact factor, academic discipline, and year of download, we might expect that the identity of the journal's publisher would have little or no effect on the predicted number of downloads. To determine whether this is the case, we fit an equation that includes all of the above-mentioned variables as well as an indicator variable for the publisher.

The Function to Be Estimated

We estimate an equation defined as follows. Let D_{jy} represent the number of times in year y that University of California libraries have downloaded articles that were published in journal j in year y and in the three years prior to year y. Let A_{jy} be the number of articles published in journal j in the three years previous to year y. Let C_{jy} be the number of times that articles published in journal j in the previous three years were cited in year y. We assign indicator variables for the academic discipline to which a journal is assigned, the year in which downloads are recorded, and the journal's publisher. (An indicator variable is either 0 or 1 and *indicates* a characteristic. If a journal is published by Elsevier, then, for all observations corresponding to this journal, the indicator for Elseveir equals 1 and the indicator for all other publishers equals 0.) We then employ maximum likelihood procedures to estimate a function that predicts downloads and takes the form

$$\mathbb{E}(D_{jy}) = A_{jy}^{\alpha} C_{jy}^{\beta} F_j Y_y P_j \qquad (1)$$

where F_j, P_j, and Y_y are multiplicative factors corresponding respectively to the journal's discipline, its publisher, and the year of download for the observed downloads. (Appendix A presents formal details of our estimation procedure.)

We can rewrite Equation 1 to explicitly show separate effects of citations per article (aka impact factor) and of number of articles (size of journal) on the number of downloads. Equation 1 is equivalent to

$$\mathbb{E}(D_{jy}) = A_{jy}^{\alpha+\beta}\left(\frac{C_{jy}}{A_{jy}}\right)^{\beta} F_j Y_y P_j. \qquad (2)$$

We estimate the parameters $\alpha + \beta$, β and the coefficients Y_y, F_j, and P_j, corresponding to indicator variables for year of download, journal discipline, and journal publisher. For each of the 7,724 journals in the sample, there are between four and six annual observations, corresponding to downloads in different years. We estimate standard errors using cluster- robust methods to account for within-journal correlation.[17]

Results

We estimate the joint effects of the variables impact factor, number of articles, year of download, journal discipline, and journal publisher by fitting Equation 2. We fit separate estimates in which fields are specified at each of Scopus's three levels of detail. We also consider a specification in which separate equations are fit for each of the five broad disciplinary areas, thus allowing the elasticities of downloads with respect to impact factor and to number of articles to differ between broad disciplines. Our discussion reports the effects of each group of explanatory variables, while controlling for the effects of all of the other variables.

The Effects of Impact Factor and Number of Articles

Table 4 shows the estimated elasticities of the number of downloads to the impact factor and number of articles. The estimates in the first column are best estimates when the elasticities are constrained to be the same for all categories. (These estimates are nearly the same for field specifications at all three levels of detail.) The remaining columns show separate estimates when the elasticities are allowed to differ among categories.

The coefficient β captures the responsiveness, in percentage terms, of downloads with respect to impact factor, holding constant the number of articles. Because the impact factor is the ratio of the number of citations to the number of articles, a 1 percent increase in the impact factor, holding articles constant, is equivalent to a 1 percent increase in citations. Thus we can also interpret β as an estimate of the elasticity of downloads with respect to citations.

TABLE 4
Elasticity of Downloads with Respect to Impact Factor and Number of Articles

	All Categories	Arts and Humanities	Health Sciences	Life Sciences	Physical Sciences	Social Sciences
Impact factor (β)	1.11	0.49	0.90	1.36	0.97	0.68
	(0.09)	(0.05)	(0.05)	(0.08)	(0.03)	(0.03)
Articles ($\alpha + \beta$)	0.91	1.03	0.87	0.95	0.91	0.94
	(0.02)	(0.05)	(0.02)	(0.04)	(0.03)	(0.03)

Note: Standard errors of coefficient estimates are reported in parentheses.

The coefficient 1.11 in the first column indicates that, for a given journal, a 1 percent increase in the number of citations, holding the number of articles fixed, would result in slightly more than a 1 percent increase in downloads.

The coefficient $\alpha + \beta$ measures the elasticity of downloads with respect to the number of articles, holding impact factor constant. The coefficient of 0.91 for articles in the first column indicates that holding constant a journal's impact factor, a 1 percent increase in the number of articles would result in slightly less than a 1 percent increase in the number of downloads.

The estimated elasticity of downloads with respect to the number of articles is close to 1 for all five disciplinary categories. For arts and humanities and for social sciences, the elasticity of downloads with respect to impact factor is significantly less than 1. (The statistical significance stems from the result that the estimated coefficient is more than 2 standard errors below 1.) For the health sciences and physical sciences, this elasticity is close to 1, and for the life sciences it is significantly greater than 1.

The Effect of Download Year

TABLE 5
Effect of Download Year on Downloads

Download Year	Arts and Humanities	Health Sciences	Life Sciences	Physical Sciences	Social Sciences
2011	0.92 (0.05)	0.99 (0.02)	0.92 (0.04)	0.82 (0.04)	0.89 (0.03)
2012	1.06 (0.04)	1.000 (0.02)	0.98 (0.02)	0.88 (0.02)	1.14 (0.02)
2013	1.10 (0.04)	1.29 (0.02)	1.16 (0.03)	1.24 (0.04)	1.16 (0.02)
2014	1.00 (.)	1.00 (.)	1.00 (.)	1.00 (.)	1.00 (.)
2015	0.93 (0.03)	1.03 (0.01)	1.01 (0.01)	0.93 (0.02)	1.01 (0.01)
2016	1.10 (0.04)	1.28 (0.03)	1.18 (0.05)	1.00 (0.02)	1.20 (0.02)
Average Annual	1.006	1.031	1.024	0.985	1.016
Growth Rate	(0.009)	(0.006)	(0.012)	(0.010)	(0.005)

Coefficients for download year are normalized relative to 2014. Robust standard errors appear in parentheses.

Table 5 shows the coefficients of year-of-download from the estimating equations for each of the four broad disciplinary categories. The rows for each download year report the multiplicative factor for that year. The year 2014 is selected as the base year because this is the first year for which we have data for all seven publishers.[18]

To estimate the average annual growth rate, we fit a linear time trend to annual data, controlling for the number of citations, number of articles, and publisher effects. This growth rate was about 3 percent in the health sciences and roughly 2 percent in the life sciences and social sciences. This rate was not significantly different from zero in the arts and humanities and in the physical sciences.

The Effect of Academic Disciplines

TABLE 6
Download Rates Relative to Social Science

Broad Discipline Category	Simple Ratio Downloads to Citations	Category Coefficient
Arts and Humanities	1.91	2.53 (0.31)
Health Sciences	1.13	0.76 (0.05)
Life Sciences	0.70	1.09 (0.07)
Physical Sciences	0.51	0.44 (0.03)
Social Science	1	1

Table 6 compares the download rates across broadly defined disciplines. To facilitate comparison across disciplines, we express these rates relative to those for social sciences. The second column shows a simple ratio of the numbers of downloads to citations without controlling for other variables. The column *Category Coefficient* reports the coefficient in our fitted equation or an indicator variable for a journal's disciplinary category. This measures the effect of discipline when controlling for our other variables: impact factor, number of articles, year of download, and publisher. This table shows that, controlling for these other factors, articles in arts and humanities are more than twice as likely to be downloaded as those in the social sciences, while articles in the physical sciences are less than half as likely to be downloaded.

A possible explanation for the relatively low download rates for articles in the physical sciences is that, in many of the physical science fields, a large proportion of published articles also appear on the freely available source arXiv. For example, roughly two-thirds of articles published in astronomy and astrophysics and in nuclear and particle physics, and roughly one-third of articles published in mathematics and in general physics, are available in arXiv.[19] A study by Davis and Fromerth[20] concluded that arXiv-deposited articles in mathematics received more than 20 percent fewer downloads from the publisher's website.

Download Rates by Major Field

Table 7 shows the results of fitting download rates to a more finely drawn division of disciplines consisting of the 27 "major fields" assigned to journals by Scopus. The second column compares a simple ratio of downloads to citations in each major field relative to that ratio for all articles in the social sciences. The third column shows the coefficient of each discipline when we fit an equation accounting for impact factor, number of articles, year of download, and publisher. (Table 12, which is found in the appendix, reports the coefficients of 334 "minor fields" as classified by Scopus.) This table shows substantial differences in download behavior between major fields, even within the same broad discipline.

TABLE 7
Download Rates by Major Field Categories

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Arts and Humanities	1.91	2.53
Arts and Humanities	1.91	3.26
Health Sciences	1.13	0.76
Dentistry	0.90	1.18
Health Professions	0.76	0.90
Medicine	1.10	1.00
Nursing	1.15	1.44
Veterinary	2.18	2.30
Life Sciences	0.70	1.09
Agricultural and Biological Sciences	0.56	0.92
Biochemistry, Genetics and Molecular Biology	0.78	1.84
Immunology and Microbiology	0.58	1.26
Neuroscience	0.92	1.79
Pharmacology, Toxicology and Pharmaceutics	0.53	0.81
Physical Sciences	0.51	0.44
Chemical Engineering	0.51	0.63
Chemistry	0.51	0.59
Computer Science	0.38	0.42
Earth and Planetary Sciences	0.49	0.78
Energy	0.90	0.44
Engineering	0.52	0.65
Environmental Science	0.72	0.63
Materials Science	0.45	0.55
Mathematics	0.42	0.62
Physics and Astronomy	0.35	0.72
Social Sciences	1.00	1.00
Business, Management and Accounting	0.43	0.41
Decision Sciences	0.40	0.47
Economics, Econometrics and Finance	0.80	1.33
Psychology	0.69	1.49
Social Sciences (other)	1.19	2.03

The Effect of Journal Publisher

Table 8 shows the effect of an indicator variable for each publisher on reported download rates in an estimated equation that controls for each journal's number of articles, impact factor, major field, and the year in which downloads occurred. To facilitate comparison among publishers, these coefficients are expressed as their ratio to the publisher effect of Elsevier.

Table 8 presents three alternative specifications, which differ in fineness of detail by which fields are distinguished. This table demonstrates that, after controlling for discipline, impact factor, and number of articles, there remain dramatic publisher effects. These effects are little changed by changes in the granularity with which fields are defined.

TABLE 8
Estimated Publisher Effects with Alternative Specifications of Fields

	5 Broad Categories	27 Major Fields	334 Minor Fields
ACS	0.88 (0.11)	0.92 (0.10)	0.88 (0.08)
Elsevier	1 (.)	1 (.)	1 (.)
IEEE	0.54 (0.05)	0.57 (0.06)	0.49 (0.04)
NPG: *Nature*-branded	2.24 (0.41)	2.02 (0.36)	1.95 (0.21)
NPG: Other	0.98 (0.10)	0.91 (0.11)	0.97 (0.07)
Springer	0.67 (0.03)	0.68 (0.03)	0.66 (0.02)
Taylor & Francis	0.46 (0.03)	0.43 (0.03)	0.40 (0.01)
Wiley	0.72 (0.06)	0.72 (0.05)	0.68 (0.03)
R^2	0.88	0.89	0.91
Number of Observations	35,722	35,722	35,722
Number of Journals	7,728	7,728	7,728
Coefficients for publisher are normalized relative to Elsevier. Robust standard errors appear in parentheses.			

To explore the robustness of our estimated publisher effects to alternative specifications, we estimated publisher effects based on a model in which we fit separate equations for each of the five broad disciplinary categories. This specification allows the effects of impact factor, number of articles, and year of download to differ across broad categories. These results are shown in table 9.

TABLE 9
Publisher Effects by Broad Category

	Arts and Humanities	Health Sciences	Life Sciences	Physical Sciences	Social Sciences
ACS		1.18 (0.14)	0.77 (0.10)	1.26 (0.11)	2.55 (0.81)
Elsevier	1 (.)	1 (.)	1 (.)	1 (.)	1 (.)
IEEE	0.33 (0.15)	0.52 (0.09)	0.17 (0.07)	0.67 (0.06)	0.56 (0.09)
NPG: *Nature*		2.24 (0.29)	1.37 (0.21)	4.54 (0.47)	2.71 (0.19)
NPG: Other	0.78 (0.11)	0.98 (0.09)	0.85 (0.09)	1.51 (0.44)	
Springer	0.57 (0.05)	0.58 (0.02)	0.69 (0.04)	0.84 (0.06)	0.66 (0.04)
Taylor & Francis	0.40 (0.04)	0.35 (0.02)	0.419 (0.03)	0.46 (0.04)	0.37 (0.02)
Wiley	0.80 (0.09)	0.61 (0.03)	0.59 (0.03)	1.20 (0.14)	0.74 (0.04)
R^2	0.84	0.86	0.92	0.90	0.83
Num. of Obs.	3586	10299	8137	13900	12410
Num. of Journals	773	2377	1798	2919	2575
Coefficients for publisher are normalized relative to Elsevier. Robust standard errors appear in parentheses.					

Tables 8 and 9 show that the strongest publisher effect by far is for *Nature*-branded journals. This effect probably is due to the fact that about half of the articles in *Nature*- branded journals are commissioned summaries of recent research called *News and Views*, which are written by prominent scholars and intended for nonspecialists. These papers do not present original research that is likely to be cited, but they are frequently downloaded and read by scientists who wish to learn about research that is not directly related to their own work.

The remaining publisher effects fall roughly into two groups. Journals published by Elsevier, American Chemical Society, and by the Nature Publishing Group without the *Nature* brand consistently show higher publisher effects than those published by IEEE, Springer, Wiley, and Taylor & Francis. These coefficients indicate that Elsevier reports more than twice as many downloads as Taylor & Francis from journals that are in the same discipline and have similar impact factors and numbers of articles. Elsevier reports about 50 percent more downloads than Springer and about 40 percent more than Wiley from journals of similar quality and disciplinary specialization.

Double-counting and the Ratio of PDF to Total Downloads

Davis and Price[21] and Li and Wilson[22] have suggested that differences in publisher platforms are likely to result in large differences in the number of downloads recorded in a single usage. Some platforms may make it more likely that a user who wants to read an article will download both a PDF copy and an HTML copy, thus counting two downloads for a single usage. In a study of records of downloads from about 800 journals at the Cornell University library in 2004, Davis and Price found wide divergence in the ratio of PDF to HTML downloads among the six publishers they studied.

To explore this hypothesis, we performed a similar exercise with our sample of nearly 8,000 journals from seven publishers at the ten University of California campuses.[23] The first column of table 10 displays the ratio of total downloads to PDF downloads for each publisher. The second and third columns provide comparisons of all publishers with Elsevier. In the second column, the ratio of the first column entries appears, thus 0.46 for ACS indicates that the total/PDF ratio for ACS is only 46 percent of the total/PDF ratio for Elsevier. The third column repeats results presented earlier (column 2 of table 8), which allows comparison of the download ratios with the publisher effects reported above.

TABLE 10
Estimated Publisher Effects and Ratios of PDF to Total Downloads

	Ratio of Total to PDF Downloads	Ratio Total/PDF Relative to Elsevier	Publisher Effect Relative to Elsevier
ACS	1.19	0.46	0.92
Elsevier	2.59	1.00	1.00
IEEE	1.03	0.40	0.57
NPG: *Nature*	2.81	1.08	2.02
NPG: Other	2.94	1.14	0.91
Springer	1.40	0.54	0.68
Taylor & Francis	1.38	0.53	0.43
Wiley	1.46	0.56	0.72

This table shows that the journals published by Nature and by Elsevier, which have the largest publisher effect, also have much higher ratios of total to PDF downloads than the journals published by Springer, IEEE, Wiley, and Taylor & Francis. This tends to confirm the view of Davis and Price, Wilson and Li, and Wiersma[24] that the extent to which download statistics double-counts downloads varies widely among publishers. (The correlation we find is not universal, however. The American Chemical Society's publisher effect is nearly as large as Elsevier's, yet its ratio of PDF to total downloads is closer to that of the group of publishers with low estimated publisher effect.)

Because libraries frequently use download statistics to evaluate journal subscriptions, publishers have an incentive to induce users to download the same article multiple times. Some publishers seem to have been more successful in this endeavor than others. The links that appear if one begins a search at the journal's table of contents appear to be quite similar among the seven publishers in our study.[25] The platform that one encounters when accessing an article through a search engine or through Crossref seems much more variable. For some journals, the first link that the search engine points to will open an HTML copy immediately, while offering the option to also download a PDF. For other journals, it opens a page that offers an option to download a PDF before it opens an HTML. Sometimes the first link will take one directly to a PDF file.

The statistics that publishers release to libraries appear in a summary form that conceals much of the information that would be necessary for libraries to estimate the usage that a subscription pays for. When a user is shown an HTML version of a paper and file and then downloads the PDF version, this is counted as two downloads. If the same user downloads the same paper a few hours later, this is counted as an additional download.[26] The JR1 and JR5 download reports are compiled from log files that record the exact time of each download, the IP address of the user, and, in some cases, whether the download is an HTML or a PDF download. As Bergstrom, Uhrig, and Antelman[27] demonstrate, these log files can be used to estimate the extent of double-counting by publishers.

Conclusion

This paper originated as an exploration of the relation between journal downloads and journal citations. Our study indicates that there is substantial correlation between citations and reported downloads, with an R^2 of about .78 in a simple regression. It also shows that the ratio of downloads to citations differs sharply among disciplines and that this ratio tends to be higher for journals with higher impact factors. This suggests that, if download reports accurately measure usage, there is a compelling case that libraries should use download data in addition to or perhaps instead of citation data in deciding how to allocate their subscription expenditures among journals.

Our study uncovered a disconcerting dependence of reported journal downloads on the identity of the journal's publisher. This dependence persists when we control for academic discipline, impact factor, number of articles, and year of download. When we fit an estimating function that controls for these variables, the numbers of recorded downloads from journals published by Elsevier, the American Chemical Society, and Nature Publishing Group are roughly twice as high as those for journals published by Springer, Wiley, Taylor & Francis, and IEEE.

Large differences in the ratio of reported PDF downloads to reported total downloads provide circumstantial evidence that a) actual usage is exaggerated because users who download

both a PDF copy and one or more additional HTML copies are counted as making multiple downloads; and b) this exaggeration differs substantially among publishers.

If the amount of double-counting was relatively constant across disciplines and across publishers, then reported downloads would remain useful for comparing the relative cost-effectiveness of competing journals. But our estimates suggest that this is not the case. Differences among publishers' ratios of reported downloads to actual usage would mean that download statistics cannot be used to compare the value of similar journals published by different publishers, at least without an adjustment factor to account for publisher effect.

If we use the publisher fixed effects reported in the third column of table 10 to estimate the amount of double-counting by each publisher, then to compare relative numbers of downloads across publishers we must deflate the numbers reported by publishers with large publisher effects. Table 11 reports these deflation factors. Because Taylor & Francis has the smallest publisher effect, their reported downloads are not deflated. For Elsevier, which has a large publisher effect, deflated downloads are only 43 percent of reported downloads.

TABLE 11
Deflating by Estimated Publisher Fixed Effect

Publisher	Publisher-effect Deflator
ACS	0.47
Elsevier	0.43
IEEE	0.75
NPG: *Nature*	0.21
NPG: Other	0.47
Springer	0.63
Taylor & Francis	1.00
Wiley	0.60

Our study suggests that the organization COUNTER has not achieved the objective stated on their website:[28]

> "COUNTER provides the Code of Practice that enables publishers and vendors to report usage of their electronic resources in a consistent way. This enables libraries to compare data received from different publishers and vendors."

Our results strongly indicate that the COUNTER data currently available to university libraries does not enable libraries to reliably compare the value of journals received from different publishers. This is unfortunate because accurate reports of downloads would be a better measure of local usage and hence of the value of a subscription than are citation counts.

We suggest that, if librarians wish to use download statistics to compare the cost- effectiveness of journals offered by different publishers, they should consider adjusting the reported download statistics for each publisher to account for the platform effects shown in table 11.

Download data is currently collected by publishers and reported to subscribing libraries in summary form, often subject to a confidentiality clause that prevents them from sharing download information with researchers or with other libraries. We suggest that, when negotiating contracts with publishers, libraries insist on the right to share this information with researchers and with other libraries.

In the long run, if download statistics are to be a credible and reliable tool for estimating usage, it seems that it would be advisable for libraries to develop a uniform interface for downloading articles from all publishers and to maintain their own records of journal downloads, which they would share as public information.

APPENDIX A. Statistical Methods

The number of downloads is a count variable taking nonnegative integer values. Because count data are not continuous, the traditional approach of specifying the conditional mean of the variable of interest together with a normal error is not always the best approach. For the problem at hand, $D_{j,y}$ has many small integer values, a large number of zeros, and a small number of very large counts (the source of the positive skewness in the downloads distribution), all of which suggest the normal distribution is not appropriate. One common alternative is to convert the integer values to noninteger values (by using the log of the variable of interest) that are then well approximated by a normal distribution. Such an approach is not appealing here, because the log is not defined for the many observations that equal zero.

Instead, we model the distribution of downloads, conditional on the covariates $x_{j,y}$, as a Poisson random variable with distribution defined by

$$\mathbb{P}[D_{j,y} = k|x_{j,y}] = \frac{e^{-\mu_{j,y}}(\mu_{j,y})^k}{k!} \qquad k = 0, 1, 2, \ldots \tag{3}$$

where $\mu_{j,y}$ depends on $x_{j,y}$. The Poisson approximation to the distribution of downloads is unlikely to work well for noninteger random variables, in particular for the ratio of downloads to citations.

The key is to specify the relationship between $\mu_{j,y}$ and the covariates, for which a natural specification would be $\mu_{j,y} = x_{j,y}^{\mathrm{T}}\beta$. One feature of the Poisson distribution is that $\mathrm{E}[D_{j,y}|x_{j,y}] = \mu_{j,y}$, hence $\mu_{j,y} > 0$ because downloads are restricted to be nonnegative. Unfortunately, the linear specification does not satisfy the restriction $\mu_{j,y} > 0$ for all values of $x_{j,y}^{\mathrm{T}}\beta$, so the common specification is $\mu_{j,y} = exp(x_{j,y}^{\mathrm{T}}\beta)$. Thus

$$\mathbb{E}[D_{j,y}|x_{j,y}] = exp(x_{j,y}^{\mathrm{T}}\beta). \tag{4}$$

The parameters are estimated via quasi-maximum likelihood. The density for an individual observation is

$$f(D_{j,y}|x_{j,y}) = \frac{e^{-exp(x_{j,y}^{\mathrm{T}}\beta)}e^{x_{j,y}^{\mathrm{T}}\beta \cdot D_{j,y}}}{D_{j,y}!} \tag{5}$$

If we let the full set of observations be denoted $(D, x) := \{D_i, x_i^{\mathrm{T}}\}_{i=1}^n$, the log likelihood is

$$L(\beta|d, x) = \sum_{i=1}^{n}[D_i \cdot x_i^{\mathrm{T}}\beta - e^{x_i^{\mathrm{T}}\beta} - log(D_i!)], \tag{6}$$

with first-order conditions

$$\sum_{i=1}^{n}[D_i - e^{x_i^{\mathrm{T}}\hat{\beta}}]x_i = 0, \tag{7}$$

where $\hat{\beta}$ is the maximum likelihood estimator of β. Although (7) does not have a closed-form solution, L is a concave function of β and standard numeric optimization methods can be employed.

Under the Poisson distribution the mean equals the variance, a restriction that is unrealistic for downloads. Yet $\hat{\beta}$ remains consistent for β even if this restriction is violated, as long as the conditional mean is correctly specified in (4). More care needs to be taken in estimating the standard error of $\hat{\beta}$. To produce consistent estimators of the standard errors we use the robust variance estimator where $\hat{\mu}_i = exp(x_i^{\mathrm{T}}\hat{\beta})$.

APPENDIX B. Discipline Effects by Minor Field

The coefficients in table 12 show the coefficients of an indicator for each minor field on our fitted estimate of the annual number of downloads. These are normalized to be expressed as ratios to the coefficient of social science. For example, the coefficient 0.61 for Accounting means that controlling for impact factor, number of articles, year of download, and publisher, accounting journals are downloaded about 61% as often as the average journal in social science.

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Accounting	0.61
Acoustics and Ultrasonics	0.22
Advanced and Specialized Nursing	1.55
Aerospace Engineering	0.3
Aging	0.34
Agricultural and Biological Sciences (miscellaneous)	1.27
Agronomy and Crop Science	0.41
Algebra and Number Theory	0.52
Analysis	0.17
Analytical Chemistry	0.4
Anatomy	0.84
Anesthesiology and Pain Medicine	0.92
Animal Science and Zoology	0.74
Anthropology	4.49
Applied Mathematics	0.43
Applied Microbiology and Biotechnology	0.84
Applied Psychology	0.56
Aquatic Science	0.56
Archeology	0.27
Archeology (arts and humanities)	1.94
Architecture	1.24
Artificial Intelligence	0.12
Arts and Humanities (miscellaneous)	1.42
Assessment and Diagnosis	1.38
Astronomy and Astrophysics	0.25
Atmospheric Science	0.56
Atomic and Molecular Physics, and Optics	0.39
Automotive Engineering	0.71
Behavioral Neuroscience	0.87
Biochemistry	0.74
Biochemistry (medical)	0.59
Biochemistry, Genetics and Molecular Biology (miscellaneous)	2.31

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Bioengineering	1.26
Biological Psychiatry	0.74
Biomaterials	0.27
Biomedical Engineering	1.03
Biophysics	1.3
Biotechnology	0.89
Building and Construction	0.27
Business and International Management	0.22
Business, Management and Accounting (miscellaneous)	0.27
Cancer Research	0.77
Cardiology and Cardiovascular Medicine	0.58
Catalysis	0.64
Cell Biology	1.23
Cellular and Molecular Neuroscience	0.79
Ceramics and Composites	0.23
Chemical Engineering (miscellaneous)	0.32
Chemical Health and Safety	165.2
Chemistry (miscellaneous)	0.38
Chiropractics	0.67
Civil and Structural Engineering	0.42
Classics	9.99
Clinical Biochemistry	0.47
Clinical Psychology	0.67
Cognitive Neuroscience	0.89
Colloid and Surface Chemistry	0.68
Communication	1.63
Community and Home Care	3.11
Complementary and Alternative Medicine	0.78
Complementary and Manual Therapy	1.01
Computational Mathematics	0.33
Computational Mechanics	2.18
Computational Theory and Mathematics	0.85
Computer Graphics and Computer-Aided Design	0.8
Computer Networks and Communications	0.18
Computer Science (miscellaneous)	0.36
Computer Science Applications	0.37
Computer Vision and Pattern Recognition	0.21
Computers in Earth Sciences	0.32

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Condensed Matter Physics	0.42
Conservation	0.025
Control and Optimization	0.94
Control and Systems Engineering	0.16
Critical Care Nursing	1.18
Critical Care and Intensive Care Medicine	0.61
Cultural Studies	2.93
Decision Sciences (miscellaneous)	0.23
Demography	1.54
Dentistry (miscellaneous)	0.62
Dermatology	1.14
Development	1.11
Developmental Biology	1.78
Developmental Neuroscience	1
Developmental and Educational Psychology	1.2
Discrete Mathematics and Combinatorics	0.58
Drug Discovery	0.51
Drug Guides	0.044
E-learning	0.17
Earth and Planetary Sciences (miscellaneous)	0.59
Earth-Surface Processes	0.46
Ecological Modeling	0.8
Ecology	0.62
Ecology, Evolution, Behavior and Systematics	0.9
Economic Geology	0.18
Economics and Econometrics	0.85
Economics, Econometrics and Finance (miscellaneous)	0.64
Education	1.03
Electrical and Electronic Engineering	0.54
Electrochemistry	0.32
Electronic, Optical and Magnetic Materials	0.56
Embryology	0.9
Emergency Medicine	1.58
Emergency Nursing	0.75
Endocrine and Autonomic Systems	0.92
Endocrinology	0.59
Endocrinology, Diabetes and Metabolism	0.45
Energy (miscellaneous)	0.32

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Energy Engineering and Power Technology	0.31
Engineering (miscellaneous)	0.27
Environmental Chemistry	0.32
Environmental Engineering	0.14
Environmental Science (miscellaneous)	0.55
Epidemiology	0.75
Equine	2.43
Experimental and Cognitive Psychology	0.83
Family Practice	1.29
Filtration and Separation	0.12
Finance	0.76
Fluid Flow and Transfer Processes	0.17
Food Animals	0.39
Food Science	0.23
Forestry	0.28
Fuel Technology	0.061
Fundamentals and Skills	205.5
Gastroenterology	0.59
Gender Studies	3.6
Genetics	0.76
Genetics (clinical)	0.72
Geochemistry and Petrology	0.4
Geography, Planning and Development	0.74
Geology	0.27
Geometry and Topology	0.37
Geophysics	0.89
Geotechnical Engineering and Engineering Geology	0.29
Geriatrics and Gerontology	0.52
Gerontology	0.84
Global and Planetary Change	0.53
Hardware and Architecture	0.36
Health (social science)	1.56
Health Informatics	0.83
Health Information Management	0.16
Health Policy	0.89
Health Professions (miscellaneous)	1.23
Health, Toxicology and Mutagenesis	0.25
Hematology	0.6

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Hepatology	0.42
Histology	1.07
History	3.78
History and Philosophy of Science	1.47
Horticulture	0.45
Human Factors and Ergonomics	0.35
Human-Computer Interaction	0.46
Immunology	0.87
Immunology and Allergy	0.69
Immunology and Microbiology (miscellaneous)	0.63
Industrial Relations	1.01
Industrial and Manufacturing Engineering	0.061
Infectious Diseases	0.81
Information Systems	0.71
Information Systems and Management	0.091
Inorganic Chemistry	0.49
Insect Science	0.53
Instrumentation	0.3
Internal Medicine	0.89
Issues, Ethics and Legal Aspects	1.52
LPN and LVN	0.76
Language and Linguistics	1.41
Law	1.18
Leadership and Management	1.1
Library and Information Sciences	0.94
Life-span and Life-course Studies	1.08
Linguistics and Language	1.18
Literature and Literary Theory	9.42
Logic	0.48
Management Information Systems	0.064
Management Science and Operations Research	0.37
Management of Technology and Innovation	0.29
Management, Monitoring, Policy and Law	0.63
Marketing	0.35
Materials Chemistry	0.34
Materials Science (miscellaneous)	0.41
Maternity and Midwifery	1.03
Mathematical Physics	0.33

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Mathematics (miscellaneous)	0.55
Mechanical Engineering	0.43
Mechanics of Materials	0.4
Media Technology	0.14
Medical Laboratory Technology	0.44
Medical and Surgical Nursing	3.61
Medicine (miscellaneous)	0.57
Metals and Alloys	0.22
Microbiology	0.75
Microbiology (medical)	0.36
Modeling and Simulation	0.2
Molecular Biology	1.23
Molecular Medicine	1.49
Multidisciplinary	0.55
Museology	22.3
Music	5.11
Nanoscience and Nanotechnology	0.55
Nature and Landscape Conservation	0.69
Nephrology	0.44
Neurology	0.87
Neurology (clinical)	0.67
Neuropsychology and Physiological Psychology	0.8
Neuroscience (miscellaneous)	1.47
Nuclear Energy and Engineering	0.5
Nuclear and High Energy Physics	0.29
Numerical Analysis	0.14
Nursing (miscellaneous)	0.79
Nutrition and Dietetics	0.72
Obstetrics and Gynecology	0.97
Occupational Therapy	0.26
Ocean Engineering	1.02
Oceanography	0.64
Oncology	0.48
Oncology (nursing)	1.22
Ophthalmology	1.01
Optometry	2.33
Oral Surgery	0.83
Organic Chemistry	0.65
Organizational Behavior and Human Resource Management	0.48

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Orthodontics	1.29
Orthopedics and Sports Medicine	0.53
Otorhinolaryngology	0.95
Paleontology	0.53
Parasitology	0.38
Pathology and Forensic Medicine	0.61
Pediatrics	2.13
Pediatrics, Perinatology and Child Health	1.11
Periodontics	0.58
Pharmaceutical Science	0.42
Pharmacology	0.57
Pharmacology (medical)	0.7
Pharmacology (nursing)	0.7
Pharmacology, Toxicology and Pharmaceutics (miscellaneous)	0.34
Pharmacy	3.14
Philosophy	2.44
Physical Therapy, Sports Therapy and Rehabilitation	0.56
Physical and Theoretical Chemistry	0.39
Physics and Astronomy (miscellaneous)	0.55
Physiology	0.68
Physiology (medical)	0.61
Plant Science	0.68
Podiatry	2.08
Political Science and International Relations	2.01
Pollution	0.52
Polymers and Plastics	0.15
Process Chemistry and Technology	0.066
Psychiatric Mental Health	1.03
Psychiatry and Mental Health	0.82
Psychology (miscellaneous)	1.32
Public Administration	0.5
Public Health, Environmental and Occupational Health	1.01
Pulmonary and Respiratory Medicine	0.47
Radiation	0.7
Radiological and Ultrasound Technology	0.37
Radiology, Nuclear Medicine and Imaging	0.68
Rehabilitation	0.67
Religious Studies	2.2
Renewable Energy, Sustainability and the Environment	0.36

TABLE 12
Discipline Effects by Minor Field

Discipline	Coefficient
Reproductive Medicine	0.52
Research and Theory	0.02
Rheumatology	0.5
Safety Research	0.34
Safety, Risk, Reliability and Quality	0.38
Sensory Systems	0.84
Signal Processing	0.23
Small Animals	3.62
Social Psychology	1.18
Social Work	1.94
Sociology and Political Science	1.52
Software	0.35
Soil Science	0.45
Space and Planetary Science	0.81
Spectroscopy	0.25
Speech and Hearing	0.49
Statistical and Nonlinear Physics	0.49
Statistics and Probability	0.7
Statistics, Probability and Uncertainty	1.12
Strategy and Management	0.33
Stratigraphy	0.34
Structural Biology	0.98
Surfaces and Interfaces	0.12
Surfaces, Coatings and Films	0.29
Surgery	0.67
Theoretical Computer Science	0.16
Tourism, Leisure and Hospitality Management	0.1
Toxicology	0.55
Transplantation	0.97
Transportation	1.25
Urban Studies	2.15
Urology	0.65
Veterinary (miscellaneous)	1.13
Virology	0.89
Visual Arts and Performing Arts	6.56
Waste Management and Disposal	0.23
Water Science and Technology	0.55

Notes

1. David Coughlin, Mark Cambell, and Bernard Jansen, "Measuring the Value of Library Content Collections," *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–13; John Gallagher, Kathleen Bauer, and Daniel Dollar, "Evidence-based Librarianship: Utilizing Data from All Available Sources to Make Judicious Print Cancellation Decisions," *Library Collections, Acquisitions, and Technical Services* 29, no. 2 (2005): 169–79.

2. John Gibson, David Anderson, and John Tressler, "Which Journal Rankings Best Explain Academic Salaries? Evidence from the University of California," *Economic Inquiry* 52, no. 4 (2014): 1322–40; Glenn Ellison, "How Does the Market Use Citation Data? The Hirsch Index in Economics," *American Economic Journal: Applied Economics* 5, no. 3 (2013): 63–90.

3. Ellen Hazelkorn, *Rankings and the Reshaping of Higher Education* (London, UK: Palgrave Macmillan, 2015).

4. A brief history of the science citation index and the impact factor appears in Eugene Garfield, "The Evolution of the Science Citation Index," *International Microbiology* 10 (2007): 65–69.

5. A broad-ranging summary and history of the application of download information and other direct measures of journal usage is presented in Michael Kurtz and Johan Bollen, "Usage Bibliometrics," *Annual Review of Information Science and Technology* 44, no. 1 (2010): 1–64.

6. Michael Kurtz, Gunther Eichhorn, Alberto Accomazzi, and Stephen Murray, "The Effect of Use and Access on Citations," *Information Processing and Management* 41, no. 6 (2005): 1395–1402; Michael Kurtz and Edwin Henneken, "Measuring Metrics: A 40-year Longitudinal Cross-validation of Downloads and Peer Review in Astrophysics," *Journal of the Association for Information Science and Technology* 68, no. 3 (2017): 695–708.

7. Henk Moed, "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* 58, no. 1 (2005): 1088–97; Joanna Duy and Liwen Vaughan, "Can Electronic Journal Usage Data Replace Citation Data as a Measure of Journal Use? An Empirical Examination," *Journal of Academic Librarianship* 32, no. 5 (2006): 512-17; Jin-kun Wan, Ping-huan Hua, Ronald Rousseau, and Xiu-kun Sun, "The Journal Download Immediacy Index (DII): Experiences Using a Chinese Full-text Database," *Scientometrics* 82, no. 3 (2010): 555–66; Juan Gorraiz, Christian Gumpenberger, and Christian Schloegl, "Usage versus Citation Behaviours in Four Subject Areas," *Scientometrics* 101, no. 2 (2014): 1077–95; Daniel Coughlin, Mark Campbell, and Bernard Jansen, "Modeling Journal Bibliometrics to Predict Downloads and Inform Purchase Decisions at University Research Libraries," *Proceedings of the American Society for Information Science and Technology* 50, no. 1 (2013): 1–13; Henk Moed and Gali Halevi, "On Full Text Download and Citation Distributions in Scientific-Scholarly Journals," *Journal of the Association for Information Science and Technology* 67, no. 2 (2016): 412–31; Liwen Vaughan, Juan Tang, and Rongbin Yang, "Investigating Disciplinary Differences in the Relationships between Citations and Downloads," *Scientometrics* 111, no. 3 (2017): 1533–45.

8. Tim Brody, Stevan Harnad, and Leslie Carr, "Earlier Web Usage Statistics as Predictors of Later Citation Impact," *Journal of the American Society for Information Science and Technology* 57, no. 8 (2006): 1060–72.

9. John McDonald, "Understanding Journal Usage: A Statistical Analysis of Citation and Use," *Journal of the American Society for Information Science and Technology* 58, no. 1 (2007): 39–50.

10. Philip Davis and Jason Price, "eJournal Interface Can Influence Usage Statistics: Implications for Libraries, Publishers, and Project COUNTER," *Journal of the American Society for Information Science and Technology* 57, no. 9 (2006): 1243–48.

11. "Appendix D: Field Classification Systems" (2018), available online at https://www.csg.org/programs/knowledgeeconomy/images/appendixd.pdf [accessed 1 February 2019].

12. Scopus treats "Arts and Humanities" as one of the subject areas within Social Sciences. We have chosen to define Arts and Humanities as a separate broad subject area from Social Sciences. Scopus also has a category called Multidisciplinary. In our sample, the Multidisciplinary category includes only 15 minor journals. Therefore, this category is omitted from most of our discussion.

13. Scopus also includes a category titled "Multidisciplinary." Because our sample includes only about 15 minor journals classified as Multidisciplinary, we do not include this category in table 1.

14. For NPG we exclude the journal *Nature* due to its broader, general interest readership.

15. "SJR: Journal and Country Rank" (2018), available online at https://scimagojr.com/journalrank.php [accessed 1 February 2019].

16. Scimago reports both the number of "documents" and the number of "citable documents." "Citable documents" refers to regular articles, while "documents" also includes book reviews, letters to the editor, and opinion pieces. The "number of articles" used by Web of Science in calculating its impact factor is essentially the same

as SCImago's citable documents. The number of citations reported by SCImago (and also by Web of Science) includes citations to all documents, not only "citable documents." Elsevier's Scopus calculates an impact factor called *CiteScore* that is based on SCImago's reported number of documents rather than citable documents. See "Journal Metrics-FAQ" (2017), available online at https://journalmetrics.scopus.com/index.php/Faqs [accessed 1 February 2019].

17. When these results are compared with robust standard errors that only account for heteroskedasticity, we find that the cluster-robust standard errors are about twice the estimates found without accounting for within-journal correlation.

18. Our data for the year 2011 included only the publishers Springer and Taylor & Francis. For 2012 we have data from Springer, Taylor & Francis, and Elsevier. For 2013 we have data from all publishers except Wiley.

19. Vincent Lariviere, Cassidy Sugimoto, Benoit Macaouso, Stasa Milojevic, Blaise Cronin, and Mike Theiwall, "ArXiv E-prints and the Journal of Record: An Analysis of Roles and Relationships," *Journal of the American Society for Information Science and Technology* 65, no. 2 (2014): 1157–69.

20. Philip Davis and Michael Fromerth, "Does ArXiv Lead to Higher Citations and Reduced Publisher Downloads for Mathematics Articles?" *Scientometrics* 71, no. 2 (2007): 203–15.

21. Philip M. Davis and Jason S. Price, "eJournal Interface Can Influence Usage Statistics: Implications for Libraries, Publishers, and ProjectCounter," *Journal of the American Society for Information Science and Technology* 57, no. 9 (July 2006): 1243–48.

22. Chan Li and Jacqueline Wilson, "Inflated Journal Value Rankings: Pitfalls You Should Know about HTML and PDF Usage" (paper presented at the American Library Association Annual Conference in San Francisco, CA, June 25-30, 2015).

23. The JR5 data that we use does not separately report HTML and PDF downloads, but the JR1 data for recent years does report separate numbers of HTML and PDF downloads. For each of the seven publishers, we have JR1 data with separate reports for PDF and HTML downloads for only some of the years that our JR5 data covers. To estimate ratios of PDF to HTML downloads for our sample, for each publisher, we use the ratio of total PDF downloads to total HTML downloads in the years for which we have JR1 data.

24. Gabriella Wiersma, "Report of the ALCTS CMS Collection Evaluation and Assessment Interest Group Meeting, American Library Association Conference, San Francisco, June 2015," *Technical Services Quarterly* 33, no. 2 (2016): 183–92.

25. Typically one sees the article title along with a link for downloading a PDF copy and a link for viewing the abstract. If one clicks the article title, the article is opened as an HTML file and an option to also open it as a PDF appears.

26. According to Counter Project Release 4 (2017), available online at https://www.projectcounter.org/about [accessed 1 February 2019], the data presented in the Counter reports screens for double-clicking by impatient users in the following way. If a user clicks the link to an HTML copy twice within 10 seconds, or a PDF copy twice within 30 seconds, the two clicks count as only one access. It appears, however, that if one clicks a link to an HTML file and also a PDF file, within a short interval, both are counted.

27. Working paper, by Ted Bergstrom, Richard Uhrig, and Kristin Antelman, "Looking under the Counter for Overcounted Downloads," available online at https://escholarship.org/uc/item/0vf2k2p0 [accessed 1 February 2019].

28. "About Counter" (2017), available online at https://www.projectcounter.org/about [accessed 1 February 2019].