Do download reports reliably measure journal usage? Trusting the fox to count your hens?

June 4, 2018

Abstract

Download rates of academic journals have joined citation rates as commonly-used indicators of the value of journal subscriptions. While citation rates reflect worldwide influence, the value that a single library places on access to a journal is probably more accurately measured by the rate at which it is downloaded by local users. If local download rates accurately measure local usage, there is a strong case for employing download rates to compare the cost-effectiveness of journals. We examine download data for more than five thousand journals subscribed to by the ten universities in the University of California system. We find that controlling for measured journal characteristics - citation rates, number of articles, and year of download - download rates, the ratio of downloads to citations, differs substantially between academic disciplines. Even after adding academic disciplines to the variables we control for, we find that there remain substantial "publisher effects", with some publishers recording significantly more downloads than would be predicted by the characteristics of their journals. These cross-publisher differences suggest that the download statistics currently supplied by publishers may not be sufficiently reliable to allow libraries to make subscription decisions based on price and reported downloads, without making an adjustment for these publisher effects.

1 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] 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 interests.

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.

Articles by Davis and Price and by Davis[10] have suggested that the download statistics released to libraries often exaggerate actual usage. According to Davis and Price,

> "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."

Davis and Price argue that differences between publishers' online platforms result in significant differences between the extent to which their downloads are double-counted. As evidence of this claim, they find that the ratio of PDF to HTML downloads differs substantially, even after they control for differences in publisher content. Davis and Price suggest that: "One solution may be to modify publisher numbers with adjustment factors deemed to be representative of the benefit or disadvantage due to its interface."

Most previous studies of download behavior have been limited to a small number of journals within a few narrowly defined disciplines. Our download data include recorded downloads at the ten University of California campuses from more than 5,000 academic journals in a wide variety of academic disciplines. We use this data to explore the relation between reported downloads and recorded citations. Among the questions we seek to answer are:

1) How do download rates differ across disciplines for journals with similar citation rates?

2) Does a journal's ratio of downloads to citations depend on its impact factor (citations per article)?

3) Do publisher-reported downloads measure library usage accurately and consistently across publishers?

2 Data

Our data include numbers of downloads, citations and articles per year for more than 5,000 scholarly and scientific journals. The citations data come from the website SCImago Journal & Country Rank, which records, for each journal, the number of citations in each year to articles published in that journal in the preceding three years. This website also reports annual numbers of "documents" and "citable documents". "Citable documents" refers to regular articles, while "documents" also includes book reviews, letters to the editor, and opinion pieces.[11]

Key to our analysis are the data on successful online full-text article requests (downloads) obtained from the ten-campus University of California library system. These were obtained form the California Digital Libraries (CDL) which handles subscriptions for all ten campuses. While most publishers supply their subscribing libraries with institution-specific data on downloads, restrictive clauses in publishing contracts typically forbid public access to this information. The CDL contracts do not include such restrictive clauses.

Publishers prepare download data according to guidelines set by COUNTER (Counting Online Usage of Networked Electronic Resources), a nonprofit organization set up by libraries, data vendors and publishers to ensure that online usage statistics are comparable. Most publishers provide journal download data to their institutional subscribers at COUNTER level Journal Report 1 (JR1), which reports the monthly number of downloads to all articles that have ever been published in that journal. A smaller number of publishers also provide data at the Journal Report 5 (JR5) level, which reports 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 would report the number of articles that were published in each year since 2000 and downloaded in 2015.

In this paper, we analyze University of California downloads from four large commercial publishers–Elsevier, Springer, Taylor & Francis, and Wiley–that publish across many

disciplines, one commercial publisher that specializes in life and physical sciences–Nature Publishing Group (NPG), and two professional society publishers–American Chemical Society (ACS) and Institute of Electrical and Electronics Engineers (IEEE). For each of these publishers, we have annual JR5 data on downloads, occurring in each of the years 2013 to 2016, of articles that were published in each year from 2000 to 2016. For three of the publishers we have additional data: downloads for 2012 for Elsevier, and downloads for 2011 and 2012 for Springer and Taylor & Francis. For each journal offered by these publishers, we have download data from four to six years, giving us a total of 26,793 journal-year observations.

We use the California Digital Library's classification system to associate each journal with a broad research area and with a specialized discipline. Our data set consists of 5,423 journals classified into one of four broad research fields: Arts and Humanities, Life and Health Sciences, Physical Sciences and Engineering, and Social Sciences. Within these broad areas, journals are partitioned into 163 specialized research fields.[12] Table 1 shows the distribution of journals by broad research field across publishers. As the table shows, each of the four large commercial publishers has a significant presence in all four research fields, while the other publishers have more limited scope. Nature Publishing Group publishes 30 of its 72 journals under the imprimatur, *Nature Something*, (e.g. *Nature Astronomy*, *Nature Biomedical Engineering*,). As Table 3 suggests, articles in the *Nature*-branded journals are much more cited and even more frequently downloaded than the other NPG journals.[13]

Table 1: Number of Journals by Research Field and Publisher

	Arts and Humanities	Life and Health Sciences	Physical Sciences and Engineering	Social Sciences	Number of Journals
Elsevier	18	875	613	267	1773
Springer	30	517	476	178	1201
Taylor & Francis	94	107	171	546	918
Wiley	59	541	247	422	1269
ACS	0	9	35	0	44
IEEE	0	2	140	3	145
NPG: Nature-branded	0	22	8	0	30
NPG: Other	0	42	0	0	42
All Publishers	201	2115	1690	1416	5422

Note: Statistics for the universe of unique journals in our dataset

3 Patterns of Downloads and Citations by Field and Publisher

Because our download and citation data are compiled at the journal level, we account for differences in the number of articles per journal. For each journal in our dataset and for each year in which we have JR5 download data, we find the total number of University of California downloads of articles published in the current year and the previous two years. We divide the number of reported downloads by the number of articles published in that journal during this period. We call this ratio the *number of UC downloads per recent article* for the year in which the downloads take place. The number of citations per recent article is commonly known as the journal's *impact factor*.[14] Specifically, we estimate the impact factor as the number of citations to articles published in the previous three years, divided by the number of articles published in that period.[15]

Table 2: UC Downloads per Recent Article and Impact Factor by Broad Research Area

	Mean	Median	75th Percentile	90th Percentile
Arts and Humanities				
UC Downloads per Article	4.8	3.3	6.2	10.4
Impact Factor	1.8	1.2	2.4	4.1
Ratio	2.7	2.8	2.6	2.5
Life Sciences				
UC Downloads per Article	12.8	6.0	11.5	21.5
Impact Factor	8.6	6.5	10.2	15.6
Ratio	1.5	0.9	1.1	1.4
Physical Sciences				
UC Downloads per Article	5.3	2.6	5.3	9.7
Impact Factor	6.9	5.0	8.3	12.8
Ratio	0.8	0.5	0.6	0.8
Social Sciences				
UC Downloads per Article	5.6	3.3	7.0	13.3
Impact Factor	4.3	3.1	5.5	8.8
Ratio	1.3	1.1	1.3	1.5

Table 2 reports the mean, median, 75th percentile, and 90th percentile of the number of UC downloads per recent article, the number of citations per recent article, and the ratio of the impact factor to the number of downloads per article at each of these percentile ranks.[16] We see that although journals in the arts and humanities tend to have fewer citations per article than those in other disciplines, the ratio of downloads per recent article to impact factor is significantly higher. For the life sciences, physical sciences, and social sciences, the ratio of downloads per recent article

to citations per recent article is, on average, higher for journals with higher impact factors, while for journals in the arts and humanities, the ratio between downloads and citations is roughly the same for high and low impact factor journals. This suggests that in evaluating library subscriptions, the use of citation rates alone is likely to undervalue journals in arts and humanities relative to other fields.

Table 3 shows the distribution of reported downloads per article and reported citations per article (impact factor) for each of the seven publishers in our sample. The Nature Publishing Group (NPG) publishes 30 "Nature-branded" journals, with titles such as *Nature Cell Biology* or *Nature Chemistry* and 42 "other" journals that don't include "Nature" in their titles. We see that the ratio of reported downloads to recent citations is largest for the Nature Publishing Group's Nature-branded journals. NPG's Nature-branded journals have a special feature that at least partially explains their high download-to-citation ratios. Typically in the Nature-branded journals, more than half of the articles appear in a *News and Views* section. These articles are brief reports on recent research, targeted at non-specialists. The *News and Views* reports are often commissioned to prestigious scholars and closely edited by professional staff. Since 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. NPG's other journals have lower download-to-citation ratios than their Nature-branded journals, but these ratios are still high relative to those for journals from other publishers.

Elsevier comes next in the ratio of reported downloads to citations. Elsevier reports about 50% more downloads per citation than the other three large commercial publishers who publish across many disciplines. As can be seen from Table 1, the publishers in our sample differ significantly in the distribution of academic disciplines that they cover. Table 2 shows that the ratio of downloads to citations differs among academic disciplines and also differs with the impact factor of the journal. Thus it could be that the differences between publishers' download-to-citation ratios are explained by differences in the academic disciplines that they emphasize and/or by differences in the impact factors of the journals that they publish. The next section of this paper explores the extent to which these differences can be explained by observable characteristics of the journals that they publish.

Table 3: Downloads and Citations per Recent Article, by Publisher

	Mean	Median	P75	P90
NPG: Nature-branded				
UC Downloads per Article	221.0	215.9	287.0	422.5
Citations	61.4	54.2	79.2	113.2
Ratio	3.6	4.0	3.6	3.7
NPG: Other				
UC Downloads per Article	28.1	23.8	36.0	56.6
Citations	15.4	13.4	21.3	27.8
Ratio	1.8	1.8	1.7	2.0
Elsevier				
UC Downloads per Article	12.6	6.8	13.0	23.3
Impact Factor	8.7	6.8	10.4	15.4
Ratio	1.4	1.0	1.3	1.5
Springer				
UC Downloads per Article	4.8	3.0	6.2	10.4
Impact Factor	4.9	3.9	6.8	9.9
Ratio	1.0	0.8	0.9	1.1
Taylor Francis				
UC Downloads per Article	2.8	1.8	3.7	6.6
Impact Factor	3.1	2.4	3.8	5.7
Ratio	0.9	0.8	1.0	1.2
ACS				
UC Downloads Per Article	14.8	11.4	16.8	29.2
Impact Factor	18.7	14.2	18.9	36.9
Ratio	0.8	0.8	0.9	0.8
Wiley				
UC Downloads per Article	5.9	3.6	7.4	13.1
Impact Factor	7.2	5.7	8.9	13.7
Ratio	0.8	0.6	0.8	1.0
IEEE				
UC Downloads per Article	5.1	4.0	6.6	9.5.
Impact Factor	10.5	9.0	13.1	18.8
Ratio	0.5	0.4	0.5	0.5
Entire Sample				
UC Download per Articles	8.3	3.9	8.1	15.4
Impact Factor	6.7	4.8	8.3	12.8
Ratio	1.2	0.8	1.0	1.2

4 Estimating a function to predict downloads

Table 2 describes the behavior of downloads as a function of a single explanatory variable, citations, for each of four broad disciplinary categories. In order to investigate the relation of downloads to several variables simultaneously, it will be useful to estimate a function that predicts the number of downloads as a function of these variables. We see from Table 2 that the ratio of downloads to citations tends to be higher for more prestigious journals with relatively high ratios of citations to articles (impact factors). This suggests that the number of downloads from a journal can be better predicted if one accounts for the number of articles in the journal as well as the number of citations. From Table 2 it is apparent that the number of downloads from a journal depends not only on its number of citations and number of articles, but also on the academic discipline to which it is devoted. Since for each journal we have download data taken from each of several years, it is also appropriate to control for the year of download.

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. In order to determine whether this is the case, we fit an equation that includes an indicator variable for each publisher.

The equation that we estimate includes the following variables. 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. We then employ maximum likelihood procedures to estimate a function that predicts downloads and takes the form

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

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 1 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

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

We use maximum likelihood methods, as described in the Appendix of this paper, to 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 journal we have between four and six observations, corresponding to downloads in different years. We estimate standard errors using cluster-robust methods to account for within-journal correlation.[17]

5 Results

Table 4 reports estimates of some of the parameters of Equation 2 which predicts a journal's reported download rate as a function of its download rate, number of articles, academic discipline,

year of download, and publisher. The second column of Table 4 reports coefficient estimates when academic discipline is represented by one of the four mentioned broad categories. (These coefficients are normalized to express their ratio to that of social science.) The third column of Table 4 reports estimates when indicator variables are used for each of 163 narrowly defined academic disciplines. Listings of these 163 fields and coefficients of indicator variables for each field appear in Tables 10-13 of the Appendix.

The coefficient $\alpha + \beta$ measures the elasticity of downloads with respect to number of articles, holding impact factor constant. The coefficient β measures responsiveness of downloads to impact factor, holding the number of articles constant.

To help interpret the values in Table 4, consider the estimate of β when journals are grouped into broad categories. An increase in the impact factor of 10% would result in an increase in downloads of 11% (10*1.146). The standard errors account for statistical uncertainty and allow one to compute the range of values we are confident that β falls in. The range is obtained by adding, and subtracting, 1.96*0.109 from the estimate, yielding (1.04, 1.26). In response to a 10% increase in the impact factor we are 95% confident that downloads will increase by no less than 10% and by no more than 13%, so accounting for statistical uncertainty does not alter the conclusion that downloads increase in proportion to the impact factor.

The estimates for field and publisher are multiplicative constants and must have a base value (we select Social Sciences and Elsevier as the base values). For the estimates of publisher effects in Table 4, consider comparison of two journals - one from Elsevier and one from Wiley - that are in the same broad field, are of the same size, and have similar impact factors. The journal for Wiley has reported downloads that are only 54% (100*.535) of those reported by Elsevier. Again the estimated standard error is quite small, so accounting for statistical uncertainty does not alter the conclusion that reported downloads are much smaller for Wiley than for Elsevier.

The estimates shown in Table 4 are constructed under the assumption that the coefficients β, $\alpha + \beta$, which measure the effects of impact factor and scale of a journal, and the coefficients P_j, which measure the publisher effect, are the same across all disciplines. Table 5 shows results when we relax this assumption by fitting separate equations for each of the four broad disciplinary categories. The first subsections describe how journal size, impact factor, research category, and download year affect downloads. Control for these factors is needed to allow us to estimate a publisher effect, which is contained in the last subsection.

Table 4: Effect of Journal Characteristics on Downloads

	Broad Cat.	Fine Cat.
Impact Factor (β)	1.146	1.053
	(0.109)	(0.058)
Articles ($\alpha + \beta$)	0.879	0.902
	(0.030)	(0.026)
Arts and Humanities	2.117	
	(0.354)	
Life and Health Sciences	0.975	
	(0.056)	
Physical Sciences and Engin.	0.520	
	(0.037)	
Social Sciences	1	
NPG: Nature	2.148	1.933
	(0.427)	(0.282)
NPG: Other	0.993	1.046
	(0.105)	(0.102)
Elsevier	1	1
ACS	1.012	0.888
	(0.155)	(0.097)
IEEE	0.509	0.578
	(0.053)	(0.049)
Springer	0.607	0.608
	(0.029)	(0.027)
Taylor & Francis	0.559	0.448
	(0.059)	(0.029)
Wiley	0.535	0.514
	(0.040)	(0.027)
R^2	0.838	0.878
Number of Observations	26793	26793

Note: Coefficients for broad discipline categories are normalized relative to Social Sciences. Coefficients on publishers are normalized relative to Elsevier. Standard errors, clustered at the journal level, are reported in parentheses and are measured around 1.

Table 5: Estimates Allowing Elasticities to Differ by Broad Category

	Arts and Humanities	Life and Health Sciences	Physical Sciences and Engineering	Social Sciences
Impact Factor (β)	0.327	1.171	0.929	0.655
	(0.049)	(0.068)	(0.052)	(0.058)
Articles ($\alpha + \beta$)	0.955	0.870	0.937	0.903
	(0.077)	(0.033)	(0.030)	(0.034)
NPG: Nature		1.663		
		(0.230)		
NPG: Other		0.981		
		(0.087)		
Elsevier	1	1	1	1
ACS			1.106	
			(0.083)	
IEEE			0.641	
			(0.050)	
Springer	0.824	0.509	0.845	0.755
	(0.146)	(0.030)	(0.060)	(0.056)
Taylor & Francis	0.474	0.444	0.480	0.363
	(0.068)	(0.052)	(0.047)	(0.025)
Wiley	0.628	0.403	0.851	0.527
	(0.102)	(0.018)	(0.076)	(0.031)
R^2	0.653	0.876	0.882	0.811
Number of Observations	1016	10337	8404	7036

Note: Coefficients are normalized relative to Elsevier whose coefficient is set to 1. Standard errors for other publishers are around 1. Standard error estimates, clustered at the journal level, are reported in parentheses.

5.1 The effects of impact factor and number of articles

The coefficient $\alpha + \beta$ represents the elasticity of the number of reported downloads from a journal with respect to the number of articles it contains, holding constant the journal's impact factor. Thus a 1% increase in the number of articles, holding impact factor constant, is predicted to result in an $(\alpha + \beta)$% increase in the number of downloads from that journal. Tables 4 and 5 both show estimates of $\alpha + \beta$ that are slightly less than one for all broad disciplinary categories. This indicates that if a journal expands its number of articles by 1%, while holding its impact factor constant, its predicted number of downloads would increase by slightly less than 1%.

The coefficient β represents our estimate of the elasticity of the number of downloads of a journal with respect to its impact factor, while holding the number of articles in the journal constant. Thus, holding the number of articles constant, a 1% increase in impact factor would result in a β% increase in the downloads. Since the impact factor is the ratio of the number of citations to the number of articles, a 1% increase in the impact factor, holding articles constant, is equivalent to a 1% increase in citations. Thus we can also interpret β as an estimate of the elasticity of downloads with respect to citations.

In Table 4, we see that in both of the regressions with broad and fine categories, the estimates for β are slightly greater than, but not statistically significantly different from, unity. This suggests that if a journal holds its number of articles constant, but experiences a 1% increase in citations, then its expected number of downloads would also increase by about 1%.

In Table 5, where the parameters β and $\alpha + \beta$ are allowed to differ among broad categories, a slightly different picture emerges. The elasticity, β, of downloads with respect to impact factor is approximately unity for the physical sciences and engineering, but this elasticity is much smaller for the arts and humanities and for the social sciences and significantly greater than unity for the life and health sciences.

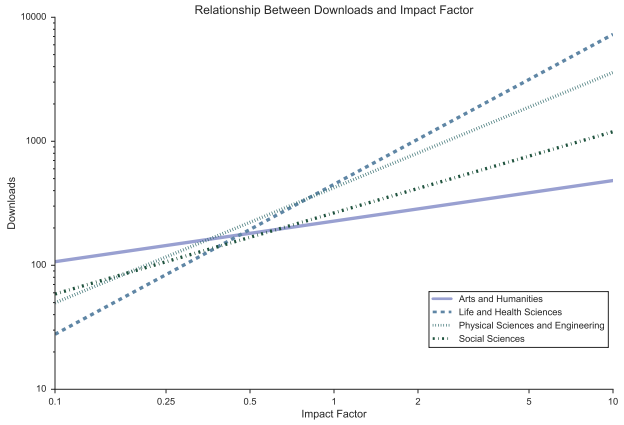

Figure 1: Relationship between Downloads and Impact Factor

Figure 1 plots the predicted relation between impact factor and downloads for each of the four broad disciplinary categories, controlling for the number of articles, the publisher and the date of download. We see that for journals with relatively low impact factors, journals in the

arts and humanities and the social sciences have more downloads per citation than journals in life and health sciences and in physical science and engineering, while for journals with relatively high impact factors, this relation is reversed.

5.2 The effect of download year

Table 6 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 2013 is selected as the base year because we do not have data for all publishers in 2011 and 2012.[18] Thus changes for the first two years reflect not only trends in downloading, but also the changing composition of our sample. We also replace the multiplicative factors with a linear time trend; the estimated trend coefficient is reported in the last row. There appears to have been a substantial increase in downloading for journals in Life and Health Sciences. For the other categories there appears to have been modest growth, except in the case of physical sciences and engineering from 2013 to 2016, where the download coefficient has remained roughly constant from 2013-2016.

Table 6: Effect of Download Year

Download Year	Arts and Humanities	Life and Health Sciences	Physical Sciences and Engineering	Social Sciences
2011	0.978	0.810	0.753	0.914
	(0.102)	(0.031)	(0.052)	(0.035)
2012	1.122	0.901	0.800	1.259
	(0.107)	(0.015)	(0.042)	(0.039)
2013	1	1	1	1
	(.)	(.)	(.)	(.)
2014	1.308	0.985	0.948	1.192
	(0.134)	(0.015)	(0.062)	(0.044)
2015	1.035	1.002	0.869	1.182
	(0.095)	(0.016)	(0.053)	(0.043)
2016	1.245	1.267	0.977	1.393
	(0.123)	(0.041)	(0.066)	(0.052)
Average annual growth rate	3.33%	7.5%	1.6%	5.5%
	(1.62)	(0.88)	(1.44)	(0.81)

5.3 The effect of academic discipline

Table 7 records discipline effects for our four broad disciplinary categories and for a sample of narrowly defined academic disciplines. The second column of this table shows a simple ratio of

Table 7: Coefficients for Selected Disciplines

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Arts and Humanities	2.69	2.12
Fine Arts	2.55	3.61
Literature	4.98	4.49
Philosophy	2.00	3.30
Religion	4.66	3.27
Life and Health Sciences	0.83	0.97
Biology	1.12	1.72
Medicine	1.04	1.20
Oncology	1.13	0.98
Pharmacy, Therapeutics, & Pharmacology	0.67	0.80
Physical Sciences and Engin.	0.42	0.52
Chemical Engineering	1.41	0.37
Chemistry	0.45	0.86
Computer Science	0.61	0.39
Electrical Engineering	0.59	0.60
Mathematics	0.82	0.81
Mechanical Engineering	0.80	0.41
Physics	0.49	0.59
Social Sciences	1.00	1.00
Economics	0.68	0.91
Education	0.85	1.11
History	4.42	3.86
Law	1.75	1.57
Library & Information Science	0.87	0.44
Political Science	2.44	2.11
Psychology	0.90	1.39

downloads to citations for each of these disciplines. The third column shows discipline effects when we control for the effects of impact factor, journal publisher, and date of download. [19]

5.4 The effect of journal publisher

Libraries do not, in general, maintain their own download counts. This information is collected and supplied by publishers in summary form to subscribing libraries. Libraries are typically forbidden from making this information public. Subscribers are not given access to the original web server

log files from which the reports they receive are compiled, and thus have no independent way of verifying the frequency of double-counting.

Davis and Price point out that publishers have strong incentives to manipulate their reported 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." (Sir Crispin Davis, House of Commons, 2004)

Since 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.

Table 8 summarizes our estimates of publisher effects, with alternative specifications of control variables as shown in Tables 4 and 5. All of these effects are expressed relative to the publisher effect for Elsevier. The second column reports simple ratios of the mean ratio of downloads to citations for journals published by each publisher. The third column reports the effect of an indicator variable for each publisher when we estimate Equation 2, which controls for impact factor, number of citations, year of download and broad disciplinary category. The fourth column shows these effects when controlling for each of 163 narrowly defined disciplinary categories as well as the other variables. The final four columns show the effects of publisher indicators relative to that for Elsevier when we allow the possibility of different effects of impact factor and number of articles in each of the four broad categories.

We see that after controlling for disciplinary concentration and impact factor, there remain dramatic publisher effects, reflecting differences in the number of reported downloads that we have not been able to explain as arising from differences in characteristics of their journals. From Table 8, it appears that the listed publishers fall into two groups, distinguished by very different publisher effects on the number of reported downloads. One group consists of Elsevier, the American Chemical Society, and the Nature Publishing Group, while the second group consists of three broad-based commercial publishers, Springer, Taylor & Francis, and Wiley, and the professional society, IEEE. This table indicates that after controlling for academic discipline, impact factor, and year of download, the number of downloads reported by publishers in the first group is on average about twice the number reported by publishers in the second group.

Davis and Price and Li and Wilson[20] have argued 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

Table 8: Estimated Publisher Effects Normalized Relative to Elsevier

	Simple Ratio	Broad Cat.	Fine Cat.	Arts & Hum.	Life & Health Sci	Physics & Engineering	Social Science
NPG: Nature	1.69	2.15	1.93	·	1.66	·	·
NPG: Other	0.92	0.99	1.05	·	0.98	·	·
Elsevier	1	1	1	1	1	1	1
ACS	0.49	1.01	0.89	·	·	1.11	·
IEEE	0.37	0.51	0.58	·	·	0.64	·
Springer	0.63	0.61	0.61	0.82	0.51	0.85	0.76
Taylor & Francis	1.11	0.56	0.45	047	0.44	0.48	0.36
Wiley	0.60	0.54	0.51	0.63	0.40	0.85	0.53

Price found wide divergence in the ratio of pdf to html downloads among the six publishers that they studied. We are able to perform a similar exercise for our sample of more than 5,000 journals from seven publishers at the ten University of California campuses.[21]

Table 9 relates the estimated publisher effects found in Table 8 to our estimate of the ratio of pdf downloads to total downloads. From this table, we see that the journals published by Nature and by Elsevier, which have more reported downloads than the predicted numbers of citations, also have much higher ratios of total downloads to PDF downloads than the journals published by Springer, IEEE, Wiley, and Taylor & Francis. This seems to confirm the view of Davis and Price, Wilson and Li, and Wiersma[22] that the extent to which download statistics double-counts downloads varies widely among publishers.

Table 9: Estimated Publisher Effects and Ratios of PDF to Total Downloads

	Estimated Relative Publisher Effect	Ratio of Total to PDF Downloads
NPG: Nature	1.93	2.80
NPG: Other	1.05	2.93
Elsevier	1.00	2.67
ACS	0.89	1.19
Springer	0.61	1.47
IEEE	0.58	1.04
Wiley	0.51	1.42
Taylor & Francis	0.45	1.27

The ways in which publishers manage to induce users to download multiple copies of the same

article are not entirely obvious. The links provided by most of the journals published by the the seven publishers in our study appear remarkably similar. For each publisher, if one seeks access to an article from the table of contents of the volume in which it appears, one will see the article title and 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 one has the option of also opening it as a pdf. If one clicks the option pdf initially, one does not see a copy of the html file. While this setup is likely to lead to some inadvertent double counting, it is hard to see why the extent of this double counting would differ substantially between publishers.[23]

The platform one encounters when accessing an article through a search engine, or through Crossref appears to be much more variable. For some journals, the first link that the search engine points to will open an html copy immediately. For others it will take you to a page offering an option to download a pdf before it opens an html. Sometimes the first link will take you directly to a pdf file. Perhaps this variation explains a significant portion of the variation among publishers. There may be other factors that result in differences in the way publishers report downloads. Since the publishers do not give libraries direct access to the log files from which the COUNTER statistics are compiled, these differences remain mysterious.

6 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 .75 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 estimates, however, 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 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 provides circumstantial evidence that reported 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.

If the amount of double-counting were 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 can not be used to compare the value of similar journals published by different publishers, at least without an adjustment factor to account for publisher effect.

For example, if we assume that the publisher effects found in our Table 8 are due to the way publishers record downloads and not to actual usage, then an appropriate measure of usage would weight reported downloads with weights inversely proportional to the coefficients found in Table 8. This finding suggests that COUNTER has not achieved the objective stated on their web-site:[24]

> "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."

Currently, download data are collected by publishers and reported to subscribing libraries in summary form, often subject to a confidentiality clause that prevents them from sharing the data. A small step that would improve the credibility and reliability of data would be for subscribing libraries to demand access to the original web server logs of downloads from their own IP addresses. If download records are to become a credible and reliable tool for estimating usage, it may also be advisable for libraries to develop a uniform interface for downloaded articles from all publishers, and to maintain their own records of journal downloads, which they would share as public information.

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): 169-179.

[2]John Gibson, David Anderson, and John Tressler, "Which journal rankings best explain academic salaries? Evidence from the University of Calfiornia," *Economic Inquiry* 52, no. 4 (2014): 1322-1340; Glenn Ellison, "How does hte 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, U.K.: 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): 1088-1097; Joanna Duy and Liwen Vaughan, "Can electronic journal usage data replace citation data as a measure of journal use? An emprical examination." *Journal of Academic Librarianship* 32, no. 5 (2006): 512-517; 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): 555-566; Juan Gorraiz, Christian Gumpenberger, and Christian Schloegl, "Usage versus citation behaviours in four subject areas," *Scientometrics* 101, no. 2 (2014): 1077-1095; 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): 412-431; Liwen Vaughan, Juan Tang, and Rongbin Yang, "Investigating disciplinary differences in the relationships between citations and downloads," *Scientometrics* 111, no. 3 (2017): 533-1545.

[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): 1060-1072.

[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): 1243-1248; Philip Davis, "The article game," *The Scholarly Kitchen Blog* October 27, 2008: available online at https://scholarlykitchensspnet.org/2008/10/27/article-download-gaming.

[11]The number of citations reported by SCImago, and also by Web of Science, includes citations to all documents, not only "citable documents." The "number of articles" used by Web of Science in calculating impact factor is essentially the same as SCImago's citable documents. Elsevier's *CiteScore* calculates an

impact factor that uses the equivalent of SCImago's total documents.

[12]The CDL chose not to classify journals that are so rarely downloaded or cited that classification is unreliable. Because these journals are rarely cited and rarely downloade, their omission will have little impact.

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

[14]A brief history of the science citation index and the impact factor appears in Garfield. Research on the use of citations is surveyed by Lutz Bornmann and Hans-Dieter Daniel, "What do citation counts measure? A review of studies on citing behavior," *Journal of Documentation* 64, no. 1 (2008): 45-80.

[15]To calculate recent downloads, we sum downloads over items published in three years, including year of downloading and the two previous years, while the impact factor sums citations in the citing year to items published in the three years prior to the year in which citing occurred.

[16]The magnitude of these ratios depend on the fact that we use downloads from University of California campuses only, while our citation measure counts citations from researchers from all institutions, worldwide.

[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 the years from 2013 onward we have data for all seven publishers.

[19]This number is the coefficient F_j on an indicator for discipline j when fitting (2) using fine categories to denote fields. These coefficients are normalized so that the mean coefficient social science journals is set to 1.

[20]Chan Li and Jacqueline Wilson, "Inflated journal value rankings: Pitfalls you should know about HTML and PDF usage," (paper presented at the American Library Associaton Conference, 2015).

[21]The JR5 data that we have used does not separately report html and pdf downloads, but the JR1 data for recent years does report separate numbers of html and pdf downloads. The JR1 data, however, simply reports the total number of times during a specified time interval that any volume of a journal is downloaded. It does not specify the year in which the downloaded material was published. 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.

[22]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): 183-192.

[23]According to Counter Project Release 4, (2017): available online at https://www.projectcounter.org/about, 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.

[24]"About Counter," (2017): available online at https://www.projectcounter.org/about.

References

About Counter. 2017*b*. *https://www.projectcounter.org/about/*, accessed 2-23-2018.

Althouse, Benjamin M., Jevin D. West, Carl T. Bergstrom, and Theodore Bergstrom. 2009. "Differences in impact factor across fields and over time." *Journal of the American Society for Information Science and Technology*, 60(1): 27–34.

Anauati, Victoria, Sebastian Galiani, and Ramiro H. Gálvez. 2016. "Quantifying The Life Cycle Of Scholarly Articles Across Fields Of Economic Research." *Economic Inquiry*, 54(2): 1339–1355.

Bergstrom, Carl T., Jevin D. West, and Marc A. Wiseman. 2008. "The Eigenfactor™ Metrics." *Journal of Neuroscience*, 28(45): 11433–11434.

Bollen, Johan, Herbert Van de Sompel, Joan A. Smith, and Rick Luce. 2005. "Toward Alternative Metrics of Journal Impact: A Comparison of Download and Citation Data." *Inf. Process. Manage.*, 41(6): 1419–1440.

Bornmann, Lutz, and HansDieter Daniel. 2008. "What do citation counts measure? A review of studies on citing behavior." *Journal of Documentation*, 64(1): 45–80.

Bouabid, Hamid. 2011. "Revisiting Citation aging: a model for citation distribution and life-cycle prediction." *Scientometrics*, 88: 199–211.

Brody, Tim, Stevan Harnad, and Leslie Carr. 2006. "Earlier Web usage statistics as predictors of later citation impact." *Journal of the American Society for Information Science and Technology*, 57(8): 1060–1072.

Card, David, and Stefano DellaVigna. 2013. "Nine Facts about Top Journals in Economics." *Journal of Economic Literature*, 51(1): 144–161.

Coughlin, Daniel M., and Bernard J. Jansen. 2015. "Modeling journal bibliometrics to predict downloads and inform purchase decisions at university research libraries." *Journal of the Association for Information Science and Technology.*

Coughlin, Daniel M., Mark C. Campbell, and Bernard J. Jansen. 2013. "Measuring the value of library content collections." *Proceedings of the American Society for Information Science and Technology*, 50(1): 1–13.

Counter Project Release 4. 2017. "Code of Practice, Release 4." *https://www.projectcounter.org/code-of-practice-sections/data-processing/*, Discussion in Data Processing section. Accessed 2-17-2018.

Davis, Phil. 2008. "The Article Game." *https://scholarlykitchen.sspnet.org/2008/10/27/article-download-gaming*, Published in The Scholarly Kitchen blog, 10/27/2008.

Davis, Philip M., and Jason S. Price. 2006. "eJournal Interface Can Influence Usage Statistics: Implications for Libraries, Publishers, and ProjectCounter." 57(9): 1243–1248.

Duy, Joanna, and Liwen Vaughan. 2006. "Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination." *The Journal of Academic Librarianship*, 32(5): 512–517.

Ellison, Glenn. 2013. "How Does the Market Use Citation Data? The Hirsch Index in Economics." *American Economic Journal: Applied Economics*, 5(3): 63–90.

Galiani, Sebastian, and Ramiro H. Gálvez. 2017. "The life cycle of scholarly articles across fields of research." National Bureau of Economic Research working paper.

Gallagher, John, Kathleen Bauer, and Daniel M. Dollar. 2005. "Evidence-based librarianship: Utilizing data from all available sources to make judicious print cancellation decisions." *Library Collections, Acquisitions, and Technical Services*, 29(2): 169–179.

Garfield, Eugene. 2007. "The evolution of the science citation index." *International Microbiology*, 10: 65–69.

Gibson, John, David L. Anderson, and John Tressler. 2014. "Which Journal Rankings Best Explain Academic Salaries? Evidence From The University Of California." *Economic Inquiry*, 52(4): 1322–1340.

Gorraiz, Juan, Christian Gumpenberger, and Christian Schlögl. 2014. "Usage versus citation behaviours in four subject areas." *Scientometrics*, 101(2): 1077–1095.

Gould, William. 2011. "Use Poisson Rather than Regress, Tell a Friend." *The STATA Blog.*

Guidelines for News and Views articles. 2017*a*. *http://ridl.cfd.rit.edu/products/press/nature/Nature%202014/N&V%%20Guidelines%20ANA%20LOPES.pdf*, Accessed: 2017-12-1.

Hazelkorn, Ellen. 2015. *Rankings and the Reshaping of Higher Education.* Palgrave Macmillan.

Kurtz, Michael J., and Edwin A. Henneken. 2017. "Measuring metrics–a 40-year longitudinal cross-validation of downloads and peer review in asstrophysics." *Journalof the Association for INformation Science and technology*, 68(3): 695–708.

Kurtz, Michael J., and Johan Bollen. 2010. "Usage Bibliometrics." *Annual review of information science and technology*, 44(1): 1–64.

Kurtz, Michael J., Gunther Eichhorn, Alberto Accomazzi, and Stephen S. Murray. 2005. "The effect of use and access on citation." *Information processing & management*, 41(6): 1395–1402.

Li, Chan, and Jacqueline Wilson. 2015. "Inflated Journal Value Rankings: Pitfalls you should know about HTML and PDF Usage." Slides for talk delivered at American Library Association Annual Conference.

McDonald, John D. 2007. "Understanding Journal Usage, A statistical analysis of citation and use." *Journal of the American society for information science and technology*, 58(1): 39–50.

Moed, Henk F. 2005. *Journal of the American society for information science and technology*, 56(10): 1088–1097.

Moed, Henk F., and Gali Halevi. 2015. "Multidimensional assessment of scholarly research impact: The Multidimensional Assessment of Scholarly Research Impact." *Journal of the Association for Information Science and Technology*, 66(10): 1988–2002.

Moed, Henk F., and Gali Halevi. 2016. "On full text download and citation distributions in scientific-scholarly journals." *Journal of the Association for Information Science and Technology*, 67(2): 412–431.

Perneger, Thomas V. 2004. "Relation between online "hit counts" and subsequent citations: prospective study of research papers in the BMJ." *BMJ*, 329(7465): 546–547.

Vaughan, Liwen, Juan Tang, and Rongbin Yang. 2017. "Investigating disciplinary differences in the relationships between citations and downloads." *Scientometrics*, 111(3).

Wan, Jin-kun, Ping-huan Hua, Ronald Rousseau, and Xiu-kun Sun. 2010. "The journal download immediacy index (DII): experiences using a Chinese full-text database." *Scientometrics*, 82(3): 555–566.

West, Jevin, Theodore Bergstrom, and Carl T. Bergstrom. 2010. "Big Macs and Eigenfactor scores: Don't let correlation coefficients fool you." *Journal of the American Society for Information Science and Technology*, 61(9): 1800–1807.

Wiersma, Gabriella. 2016*a*. "Report of the ALCTS CMS collection evaluation and assessment interest group meeting. American Library Association Conference, San Francisco, June 2015." *Technical Services Quarterly*, 33(2): 183–192.

Wiersma, Gabrielle. 2016*b*. "Report of the ALCTS CMS Collection Evaluation & Assessment Interest Group Meeting. American Library Association Annual Conference, San Francisco, June 2015." *Technical Services Quarterly*, 33(2): 183–192.

A Statistical Methods

The number of downloads is a count variable taking non-negative 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 non-integer 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 non-integer 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 $\mathbb{E}[D_{j,y}|x_{j,y}] = \mu_{j,y}$, hence $\mu_{j,y} > 0$ because downloads are restricted to be non-negative. 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 observations 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.[25] 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).[26] 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

$$\hat{V}(\hat{\beta}|x) = (\sum_{i=1}^{n} \hat{\mu_i} x_i x_i^{\mathrm{T}})^{-1} \ (\sum_{i=1}^{n} (D_i - \hat{\mu_i})^2 x_i x_i^{\mathrm{T}}) \ (\sum_{i=1}^{n} \hat{\mu_i} x_i x_i^{\mathrm{T}})^{-1}, \tag{8}$$

where $\hat{\mu_i} = exp(x_i^{\mathrm{T}} \hat{\beta})$.[27]

B Coefficients for narrowly-defined disciplines

Tables 10-13 record discipline effects on downloads for each of the narrowly-defined disciplines within each of the four broadly-defined subject areas. The second column of each table shows the ratio of downloads to citations. The third column shows the coefficient F_j of an indicator for discipline j when fitting equation 2. Both of these columns have been normalized so the coefficient on social sciences journals is one.

Table 10: Discipline Effects for Arts and Humanities

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Arts and Humanities	2.69	2.12
Architecture	1.41	1.05
Dance	13.98	21.05
Drama	3.25	8.57
Film	7.55	6.16
Fine Arts	2.55	3.61
Languages	3.50	2.39
Literature	4.98	4.49
Music	9.27	18.67
Philology & Linguistics	1.34	1.77
Philosophy	2.00	3.30
Religion	4.66	3.27
Visual Arts	5.64	7.26

Table 11: Discipline Effects for Life and Health Sciences

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Life and Health Sciences	0.83	0.97
Agriculture	0.44	0.55
Alternative Medicine	1.23	1.05
Anatomy	0.72	1.12
Animal Behavior	0.78	1.54
Animal Sciences	0.44	0.77
Bioethics	0.96	1.81
Biology	1.12	1.72
Biophysics	0.89	1.74
Botany	0.59	0.95
Cardiovascular Diseases	6.58	0.88
Clinical Endocrinology	0.77	0.81
Clinical Immunology	0.78	0.88
Cytology	1.08	2.33
Dentistry	0.91	1.08
Dermatology	0.66	1.40
Diet & Clinical Nutrition	0.69	0.88
Ecology	0.51	0.93
Emergency Medicine	1.51	1.42
Food science	0.29	0.41
Forestry	0.39	0.64
Gastroenterology	0.50	0.79
Genetics	1.70	1.13
Geriatrics	0.70	0.97
Gynecology & Obstetrics	0.98	1.31
Hematologic Diseases	0.74	0.97
Infectious Diseases	0.89	0.93
Internal Medicine	0.69	1.03
Invertebrates & Protozoa	0.56	0.78
Marine Science	0.62	0.94
Medical Research	0.70	0.95
Medicine	1.04	1.20
Microbiology & Immunology	0.77	1.20
Musculoskeletal System Diseases	2.02	0.94
Nephrology	1.73	0.70
Neurology	0.86	1.63
Neuroscience	1.27	1.58
Nursing	1.49	1.48
Occupational Therapy & Rehabilitation	1.30	0.94
Oncology	1.13	0.98
Ophthalmology & Optometry	0.97	1.38
Otorhinolaryngology	1.45	1.28
Pathology	2.17	0.94
Pediatrics	1.15	1.28
Pharmacy, Therapeutics, & Pharmacology	0.67	0.80
Physical Therapy	1.62	1.20
Physiology	0.69	0.88
Plant Physiology	0.46	1.02
Plant Sciences	0.47	0.70
Psychiatric Disorders, Individual	0.70	0.95
Psychiatry	0.82	1.17
Psychotherapy	0.71	1.19
Public Health	1.08	1.08
Radiology, MRI, Ultrasonography & Medical Physics	0.86	1.03
Sciences	0.58	0.75
Surgery & Anesthesiology	1.28	1.24
Surgery and By Type	1.01	1.05
Urology	0.99	0.95
Vertebrates	0.87	1.45
Veterinary Medicine	1.39	1.21
Zoology	0.69	0.89

Table 12: Discipline Effects for Physical Sciences and Engineering

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Physical Sciences and Engineering	0.42	0.52
Aeronautics Engineering & Astronautics	1.80	0.56
Algebra	0.39	0.66
Analytical Chemistry	0.32	0.54
Applied Mathematics	0.45	0.54
Applied Physics	0.33	0.50
Astronomy & Astrophysics	0.38	0.63
Atomic Physics	0.36	0.66
Biochemistry	0.73	1.04
Bioengineering	0.62	1.04
Biomedical engineering	0.62	0.85
Calculus	0.34	0.42
Chemical Engineering	1.41	0.37
Chemistry	0.45	0.86
Civil Engineering	0.37	0.53
Computer Science	0.61	0.39
Crystallography	0.33	0.45
Electrical Engineering	0.59	0.60
Electricity & Magnetism	0.46	0.57
Electrochemistry	0.33	0.49
Energy & Fuels	1.12	0.42
Engineering	0.35	0.46
Environmental Engineering	0.36	0.59
Environmental Sciences	0.51	0.66
Geology	0.41	0.62
Geometry	0.57	0.67
Geophysics	0.48	0.88
Industrial & Management Engineering	0.28	0.30
Information Technology	0.46	0.66
Inorganic Chemistry	0.39	0.62
Light & Optics	0.58	0.53
Materials Science	0.33	0.52
Mathematical Statistics	0.43	0.84
Mathematical Theory	0.51	0.45
Mathematics	0.82	0.81
Mechanical Engineering	0.80	0.41
Metallurgy & Mineralogy	0.59	0.36
Meteorology & Climatology	0.48	0.81
Nanotechnology	0.48	0.71
Nuclear Physics	0.35	0.30
Operations Research	0.43	0.58
Organic Chemistry	0.56	1.00
Paleontology	0.62	0.87
Physical & Theoretical Chemistry	0.35	0.52
Physical Geography	0.38	0.65
Physics	0.49	0.59
Polymers	0.32	0.40
Spectroscopy	0.47	0.67
Technology	0.60	0.35
Telecommunications	0.28	0.36
Transportation Engineering	0.51	0.65

Table 13: Discipline Effects for Social Sciences

	Download/Citation Ratio Relative to Social Science	Discipline Coefficient Relative to Social Science
Social Sciences	1.00	1.00
Agricultural Economics	0.73	1.37
Anthropology	1.40	2.47
Archaeology	1.59	1.64
Atlases & Maps	0.57	0.93
Child & Youth Development	0.93	1.97
Commerce	0.56	0.66
Communities	1.28	1.37
Criminology, Penology & Juvenile Delinquency	0.95	1.00
Demography	1.08	2.18
Economic History	0.90	0.96
Economic Theory	0.83	1.24
Economics	0.68	0.91
Education	0.85	1.11
Education, Special Topics	1.69	2.06
Ethnic & Race Studies	3.95	6.08
Finance	0.49	0.76
Gender Studies & Sexuality	2.59	2.68
Geography	0.57	0.80
Government	0.52	1.02
History	4.42	3.86
Industries	0.52	0.55
International Relations	1.66	2.29
Journalism & Communications	2.01	1.59
Labor & Workers' Economics	0.69	0.60
Law	1.75	1.57
Library & Information Science	0.87	0.44
Management	0.48	0.45
Marketing & Sales	0.36	0.38
Military & Naval Science	3.51	3.17
Office & Personnel Management	0.33	0.30
Political Science	2.44	2.11
Psychology	0.90	1.39
Real Estate, Housing & Land Use	0.47	0.76
Recreation & Sports	0.34	0.62
Regions & Countries	3.45	2.70
Social & Cultural Anthropology	2.54	2.41
Social Change & Social Conditions	1.28	2.22
Social Sciences	1.37	1.70
Social Welfare & Social Work	0.75	1.42
Statistics	0.70	1.13
Transportation Economics	0.87	1.09