Using Conventional Bibliographic Databases for Social Science Research: Web of Science and Scopus are not the Only Options

ESTHER ISABELLE WILDER (iD)

WILLIAM H. WALTERS (iD)

*Author affiliations can be found in the back matter of this article

Levy
Library
Press

ABSTRACT

Although large citation databases such as Web of Science and Scopus are widely used in bibliometric research, they have several disadvantages, including limited availability, poor coverage of books and conference proceedings, and inadequate mechanisms for distinguishing among authors. We discuss these issues, then examine the comparative advantages and disadvantages of other bibliographic databases, with emphasis on (a) discipline-centered article databases such as EconLit, MEDLINE, PsycINFO, and SocINDEX, and (b) book databases such as Amazon.com, Books in Print, Google Books, and OCLC WorldCat. Finally, we document the methods used to compile a freely available data set that includes five-year publication counts from SocINDEX and Amazon along with a range of individual and institutional characteristics for 2,132 faculty in 426 U.S. departments of sociology. Although our methods are time-consuming, they can be readily adopted in other subject areas by investigators without access to Web of Science or Scopus (i.e., by faculty at institutions other than the top research universities). Data sets that combine bibliographic, individual, and institutional information may be especially useful for bibliometric studies grounded in disciplines such as labor economics and the sociology of professions.

Policy highlights

- While nearly all research universities provide access to Web of Science or Scopus, these databases are available at only a small minority of undergraduate colleges. Systematic restrictions on access may result in systematic biases in the literature of scholarly communication and assessment.
- The limitations of the largest citation databases influence the kinds of research that can be most readily pursued. In particular, research problems that use exclusively bibliometric data may be preferred over those that draw on a wider range of information sources.

CORRESPONDING AUTHOR:
William H. Walters

Mary Alice & Tom O'Malley Library, Manhattan College, US

william.walters@manhattan. edu

- Because books, conference papers, and other research outputs remain important in many fields of study, journal databases cover just one component of scholarly accomplishment. Likewise, data on publications and citation impact cannot fully account for the influence of scholarly work on teaching, practice, and public knowledge.
- The automation of data compilation processes removes opportunities for investigators to gain first-hand, in-depth understanding of the patterns and relationships among variables. In contrast, manual processes may stimulate the kind of associative thinking that can lead to new insights and perspectives.

1. INTRODUCTION

Although large citation databases are used extensively in research on scholarly communication and assessment, they have limitations that make them less than ideal for certain kinds of projects.[1] Neither Scopus nor Web of Science is available to faculty at most undergraduate colleges, for instance. Neither provides good coverage of books and conference proceedings, and neither has adequate mechanisms for distinguishing among authors. Likewise, Google Scholar has its own unique disadvantages. This paper describes how bibliographic databases other than the large citation databases can be used to create new data files for use in bibliometric research.

Our primary goal is to demonstrate that widely available databases such as SocINDEX and Amazon.com can be useful to scholars who do not have access to Web of Science or Scopus, and that these information sources offer distinct advantages that make them especially appropriate for research centered on particular disciplines or particular author groups. Data sets that combine bibliographic information with information on the characteristics of authors and their institutions can be uniquely valuable for research on the determinants of scholarly productivity.

Our secondary goal is to present a data set that illustrates these principles and to describe the methods used in its construction. Our data file (Wilder & Walters 2020a), freely available through Zenodo, includes five-year publication counts (2013–2017) for 2,132 professors and associate professors in 426 departments of sociology in the United States, along with institutional and individual covariates such as institution type, department size, academic rank, gender, Ph.D. year, and Ph.D. institution. It has already been used to evaluate the impact of institution type, gender, and other characteristics on the publishing productivity of American sociologists (Wilder & Walters 2020b, in press). The details of the data compilation procedure, presented in the Appendix, may be helpful to other researchers, especially if they promote the consistency of methods that is important for comparisons over time. Similar procedures can be used with other disciplines and other time periods.

2. LIMITATIONS OF THE LARGE CITATION DATABASES
2.1. LIMITED AVAILABILITY

For many scholars, the biggest disadvantage of Web of Science and Scopus is simply that neither resource is available to them. Although faculty at the major research universities often have access to at least one of these databases, the situation is very different elsewhere. Apart from those institutions in the Carnegie R1 and R2 categories, just 25% of American four-year colleges and universities provide access to either Scopus or Web of Science.[2] In contrast,

---

1   For overviews of the use of Scopus and Web of Science in bibliometric research, see Baas et al. (2020) and Birkle et al. (2020).

2   As used here, *Web of Science* refers to the four component databases most often held by research universities: Science Citation Index Expanded (SCI), Social Sciences Citation Index (SSCI), Arts & Humanities Citation Index, and Journal Citation Reports (JCR). Related resources such as Book Citation Index, Conference Proceedings Citation Index, Current Contents, Data Citation Index, Derwent Innovations Index, Emerging Sources Citation Index, and SciELO are usually acquired separately, so they are not included in our definition. Likewise, we exclude the conventional disciplinary databases that are sometimes hosted on the Web of Science (formerly Web of Knowledge) platform, such as BIOSIS, Inspec, and MEDLINE.

65% have current subscriptions to SocINDEX or Sociological Abstracts,[3] and the most popular disciplinary databases such as ABI/INFORM, EconLit, ERIC, MEDLINE, and PsycINFO are likely to be even more widely available.

Our experience at several universities suggests that for institutions with 2,000 to 10,000 students, Scopus costs three to four times as much as SocINDEX or Sociological Abstracts. Moreover, Web of Science generally costs more than Scopus. Cost is not the only factor that influences library holdings, of course. The disciplinary databases each have a clear constituency—an academic department or school with a strong interest in maintaining access to specific databases and journals. Although each multidisciplinary database may be of some interest to a large number of faculty, no single group is likely to feel a compelling need to choose Scopus (for instance) over Biological Abstracts or MathSciNet. When a small group with well-defined interests and a larger group with more diffuse interests compete, the small group is likely to prevail (Elhauge 1991; Olson 1971).

Finally, there is a perception among some librarians and faculty that Scopus and Web of Science are not especially attractive to undergraduates—that the advantages of these databases (such as size, multidisciplinary scope, and citation-tracing capabilities) are offset by disadvantages such as complicated interfaces and marketing strategies that target expert users rather than undergraduates. Faculty may appreciate the breadth of Web of Science and Scopus, but students often adopt a more constrained approach to database selection. As a student at Manhattan College stated during web site usability testing, 'If my paper is for a psychology class, I look for a database with "psych" in the name.'

Because Scopus and Web of Science are held by relatively few U.S. colleges and universities, many researchers who use bibliographic or bibliometric data must look elsewhere.[4] One might argue that this is not a major problem—that bibliometric research is more likely to be conducted by research-university faculty than by those at other institutions. However, this raises a question: Can the dominance of the R1 universities within fields such as information science be attributed to the long-term centrality of expensive databases such as Web of Science? Faculty at bachelor's colleges may avoid bibliometric research simply because they lack access to the tools that are closely associated with it. The situation has at least two potentially negative consequences. First, the perspectives of scholars at undergraduate colleges and universities may not be fully represented within the literature. Second, undergraduate science faculty may miss an opportunity to undertake empirical research that is less expensive than most lab research, that does not normally require external funding, and that lends itself to multidisciplinary faculty-student collaboration.

2.2. POOR COVERAGE OF BOOKS AND CONFERENCE PROCEEDINGS

A second limitation of the large citation databases is their relatively poor coverage of books and conference proceedings. Scopus, Web of Science, and Google Scholar are all devoted primarily to journal articles. For instance, Scopus covers more than 39,000 journals but just 1,628 book series, 514 conference proceedings, and no books issued independently (i.e., not as part of a series). Likewise, the three most readily available Web of Science databases—SCI, SSCI, and Arts & Humanities Citation Index—include about 1/20 as many books and chapters as journal articles. The Web of Science book and proceedings databases are much smaller, offered as separate products, and not widely held, even by research universities. Because Google Scholar makes no distinction between articles, conference papers, books, and other online resources that 'look scholarly' to its web crawling mechanisms, the methods used in its construction do improve its coverage of conference papers and other non-journal documents. Google Scholar's coverage of books and chapters is still limited, however, perhaps because books are far less likely than journals to be indexed and made available online (Harzing 2019; Martín-Martín et al. 2018).

---

3   Institutions in the Carnegie R1 and R2 categories (*doctoral universities: very high research activity* and *doctoral universities: high research activity*) comprise just 17% of four-year colleges and universities in the United States. The estimated values reported here for all other four-year colleges and universities are based on a random sample of 80 public and nonprofit institutions in the other six Carnegie *baccalaureate, master's,* and *doctoral* categories. We searched the library web site of each institution in the sample.

4   Admittedly, the situation may be different in Europe and Asia, where undergraduate colleges are less common and universities are more likely to maintain doctoral programs.

The importance of conference proceedings is well established, especially in rapidly changing fields such as computer science (Bar-Ilan 2010; Larsen & von Ins 2010; Lisée et al. 2008). Likewise, books remain central to many areas of inquiry in the social sciences and humanities (Engels et al. 2018; Giménez Toledo 2020; Giménez Toledo et al. 2013; Moksony et al. 2014; Nederhof et al. 1989). It is also important to realize that where particular social science disciplines are concerned, the exclusion of books can lead not just to the underestimation of scholarly productivity and impact, but to systematic bias on the basis of research area, institution, and institution type. Recent data for U.S. sociology faculty reveal that nearly all research-active authors can be readily categorized as article authors or book authors; the correlation between article and book counts during the 2013–2017 period is just 0.13. Likewise, nearly all academic departments can be categorized as article departments or book departments; the department-level correlation between articles per faculty member and books per faculty member is just 0.23 (Wilder & Walters 2020b, in press). These distinctions within sociology, based at least partly on subfield (e.g., demography vs. critical theory), parallel the broader distinction between the article-based sciences and the book-based humanities.

Although faculty at the major research universities have far higher average article counts than those at other institutions, the highest average *book* counts can be found among faculty at the top liberal arts colleges (Wilder & Walters 2020b, in press). This suggests that liberal arts faculty may have a distinctive role in consolidating, synthesizing, and popularizing sociological research. It also demonstrates the advantages of evaluating books and articles separately— something that is not possible if Web of Science, Scopus, or Google Scholar are used as the sole information sources.

2.3. INADEQUATE MECHANISMS FOR DISTINGUISHING AMONG AUTHORS

A third limitation of the large citation databases is the difficulty of distinguishing among authors with similar names. To some extent, the problem can be traced to bibliographic errors in the databases themselves. For instance, Web of Science is not always consistent or reliable in its reporting of author names and institutional affiliations (Walters & Wilder 2016). The *address* field, which includes the author's institution and department, is sometimes incomplete or difficult to interpret, and the information appears to be compiled just from the first page of each article, so relevant information presented elsewhere (in a biographical statement at the end of the article, for instance) is omitted. Our informal investigations suggest that authors' first names, rather than initials, are not provided consistently for papers indexed prior to 2008, and Harzing (2013) has reported that the *document type* field does not always accurately distinguish between peer-reviewed articles and other contributions. Similar problems can be seen with Scopus, and the bibliographic errors associated with Google Scholar have been well documented (Delgado López-Cózar et al. 2019; Jacsó 2005, 2008, 2010; Orduña-Malea et al. 2017).[5]

Although the difficulty of resolving authors' names can be addressed through author identifier systems such as ORCID and ResearcherID, many authors are not included in either system. In compiling author information for our data set, we looked for databases that (a) provide complete and accurate author information, (b) maintain their own author identifier systems (although there are no such databases for sociology), (c) cover a limited range of subject areas and therefore minimize the number of instances in which 'unwanted' authors appear in the search results, and (d) provide, for each record, the full text of the title page and any other pages on which bibliographic or author information is likely to appear. We also conducted each search manually rather than relying on automated procedures. This last point is discussed more fully in section 4.3.

Accurate name disambiguation is especially important for *productivity studies*—those in which the investigator starts with a well-defined list of authors and counts all their publications or citations, wherever they may have appeared. When conducted manually, the compilation of data for a productivity study usually involves a large number of author searches, and the correct identification of individuals is central to the process. In contrast, *contribution studies* are

5   Microsoft Academic is likely to exhibit the same kinds of errors as Google Scholar, since it is compiled using similar methods. Two other potential data sources, Crossref and Dimensions, rely on publisher-supplied bibliographic information but are also subject to omissions and inaccuracies (Harzing 2019; Hendricks et al. 2020; Herzog et al. 2020; Visser et al. 2021).

those in which the investigator starts with a well-defined list of publications and records the contributions to that literature by all authors, whomever they might be. As might be expected, the distinction between productivity studies and contribution studies has implications for the kinds of research questions that can be addressed (Wilder & Walters 2019).

3. OTHER SOURCES OF BIBLIOGRAPHIC DATA
3.1. JOURNAL ARTICLE DATA

When evaluating the publishing productivity of American sociologists, we found no database that provided adequate coverage of both journal articles and books. (Other scholarly works, such as conference papers, were not included in our analysis.) We chose SocINDEX as our primary source of journal article data after evaluating Google Scholar, Scopus, Web of Science, SocINDEX, and Sociological Abstracts on the basis of five criteria:

1. Covers a large number of sociology journals, including the more prominent ones
2. Includes the journals of related fields in which sociologists routinely publish (e.g., criminology, demography, social statistics, and the social aspects of public health)
3. Provides reliable, easily compiled bibliographic information
4. Includes the author's full first name—not just the initial—as part of the searchable name field(s)
5. Excludes subjects areas unrelated to sociology, to minimize the need for investigation and clarification of matching and near-matching names.

Despite their very broad coverage, none of the three large citation databases (Google Scholar, Scopus, or Web of Science) are as comprehensive as SocINDEX and Sociological Abstracts with regard to sociology (criteria 1 and 2). For instance, the Scopus *sociology and political science* category includes 1,269 journals, about the same number that are indexed cover-to-cover in SocINDEX (1,257). SocINDEX also provides partial coverage of more than 1,500 other journals, however, teasing out individual articles of sociological interest from journals such as *Crime and Delinquency*, *Ethnicity*, and the *Journal of Biosocial Science*. Finally, SocINDEX provides better coverage of the sociology journals that aim to influence teaching and practice rather than academic research—the same journals that may be especially receptive to the work of authors at bachelor's and master's institutions. It is also likely to include more journals of local or regional interest—those with regional influence disproportionate to their overall citation impact (Etxebarria & Gomez-Uranga 2010).

Google Scholar, with its idiosyncratic presentation of results, did not satisfy criterion 3. None of the three large citation databases satisfied criterion 5.

In many academic disciplines, a single database such as EconLit or PsycINFO is widely accepted as the foremost source of bibliographic information. In contrast, sociology has two contenders: SocINDEX and Sociological Abstracts. Although the two are comparable in many ways, SocINDEX has a broader subject scope and provides more thorough coverage of peer-reviewed journals (Tyler et al. 2017). Each indexes a similar number of articles each year (52,000 for SocINDEX vs. 51,000 for Sociological Abstracts), but SocINDEX covers more journals in their entirety (1,257 vs. 922). Our impression, based on extensive experience with both databases, is that SocINDEX provides especially good coverage of the social science fields to which both sociologists and other scholars contribute (e.g., criminology, demography, gender studies, gerontology, organizational behavior, social psychology, and social work). However, Sociological Abstracts seems to offer more comprehensive coverage of sociological topics that border on the natural sciences or the humanities, in fields such as area studies, environmental studies, geography, history, law, and philosophy.

Each discipline has its own subject databases with unique characteristics, of course. We recommend that investigators seeking the most appropriate data sources consider the literature of the relevant disciplines, the scope/coverage information available at publishers' web sites, and the database reviews and comparative reports that have appeared in the library and information science (LIS) literature. LISTA, perhaps the most prominent LIS journal database, is freely available online (EBSCO 2021).

3.2. BOOK DATA

We evaluated nine source of bibliographic data for books. Six were withdrawn from consideration early in the process due to obvious gaps in coverage. Specifically,

1. We considered counting the books cited in key disciplinary journals, but those books are not necessarily representative of the literature as a whole.
2. Likewise, the books reviewed (or received for review) by key disciplinary journals are not representative. That list is also likely to be biased by publishers' actions and by the policies and preferences of the journals' editorial boards.
3. Book Citation Index includes just 60,000 books. It provides good coverage of highly cited books but poor coverage otherwise (Clarivate Analytics 2021; Torres-Salinas et al. 2014).
4. Books in Print is limited to titles currently in print, thereby excluding a substantial number of recently published books.
5. CHOICE Reviews covers only those books that are appropriate for liberal arts colleges, and, with few exceptions, only those that receive favorable reviews (Kousha & Thelwall 2015).
6. Google Books includes only those books that are available in digital format or cited online—a serious limitation, since fewer than half of all current print books are available in any digital format (Czechowski 2011; Pomerantz 2010; Walters 2013).

Three data sources are more comprehensive, however.

7. Amazon.com includes nearly all the books currently or recently available for purchase in the United States, new or used, in print or digital format (Kousha & Thelwall 2016).
8. OCLC WorldCat includes the books held by nearly 17,000 libraries worldwide as well as those available for purchase through major library vendors such as GOBI Library Solutions (White et al. 2009).
9. The GOBI database includes the books available through the largest U.S. academic library book vendor or through a selection of prominent used book dealers.

Of these nine information sources, we chose Amazon.com for our study of sociologists' publishing productivity. Although Amazon, WorldCat, and GOBI each provide good coverage of scholarly books, Amazon has a more user-friendly interface and includes title page images, which provide for reliable verification of bibliographic information. A more serious problem with WorldCat and GOBI is that they often present multiple records for a single title, making it difficult to distinguish between new (original) works and revised editions. (As noted in the Appendix, section A.7, our goal was to include new books but to exclude new editions, translations, and reprints.) With WorldCat, in particular, a single title may be represented by dozens or even hundreds of records that represent related works or that reflect the application of different libraries' cataloging practices to works that are identical in every respect. A difficulty we recently encountered, while not typical, reveals the extent of the problem. To estimate the number of libraries with access to the Web of Science database, we initially attempted to count the number of holding libraries listed in WorldCat.[6] A WorldCat title phrase search for *Web of Science* returned 192 records, including 109 that correspond to the entire database or to one of its primary components (SCI, SSCI, Arts & Humanities Citation Index, or JCR). Some of those records are virtually identical (e.g., one might use an ampersand rather than 'and' in a key field), some have been modified to comply with the cataloging standards of particular institutions or regions, and others were clearly created in error but never removed from the database. A significant number of the near-duplicate records appear to reflect disagreements among catalogers about best practices, and quite a few are unique due to local holdings information that should not have been added to the bibliographic record itself.

Although Amazon.com was the single best source of book data, we relied on additional sources—university web sites, publishers' web sites, personal web sites, Google Scholar, and OCLC WorldCat—to verify and clarify the information for about 20% of the faculty on our list. Many sources included helpful information, but none were comprehensive and many omitted the authors' most recent publications. In particular, authors' online CVs were often incomplete, out of date, or potentially misleading. For instance, many listed edited volumes as if they were single-authored books.

---

6   The attempt was unsuccessful, since WorldCat lists not just those libraries with access to the current data, but those with any of the previous editions—the SSCI volumes once issued in print or on microfiche, for instance.

4. NEW DATA ON THE PUBLISHING PRODUCTIVITY OF AMERICAN SOCIOLOGISTS

4.1. THE DATA FILE

As noted in section 1, our data file includes five-year publication counts (2013–2017) for 2,132 professors and associate professors in 426 U.S. departments of sociology. Publication counts and related data are presented separately for individuals and for academic departments. The data file, in .xlsx format, is freely available through both Zenodo and openICPSR (Wilder & Walters 2020a).[7]

The data compilation procedures, described fully in the Appendix, result in a data set with several unique advantages:

1. Data for six distinct institution types allow for the investigation of publishing productivity across the full range of U.S. colleges and universities: top research universities, other R1 universities, other doctoral universities, master's institutions, top liberal arts colleges, and other bachelor's institutions.
2. Four productivity measures—articles, articles in high-impact journals, books, and books from high-impact publishers—allow researchers to identify book- and article-centered departments and individuals, and to explore the relationships between book and article counts.[8]
3. The inclusion of key institutional and individual variables (e.g., department size, academic rank, gender, Ph.D. year, and Ph.D. institution) facilitates investigation of the correlates/determinants of scholarly productivity. Likewise, the identification of individuals and institutions allows for the linking of these data to the variables found in other data sets.
4. Multiple data sources and careful data cleaning/standardization procedures provide for a data set that is reliable and consistent in format. The data were compiled manually from authoritative sources, without relying on surveys or other instruments that might be subject to response bias.

Our data do have five significant limitations, however. First, assistant professors are not included. Section A.5 of the Appendix presents the rationale for this decision. Second, the data for institutions in the R1 and *master's* categories are sample data rather than population data. Complete population data are provided for the institutions in the other four categories, however; see the Appendix, sections A.1 and A.3. Third, our reliance on a list of authors working in departments of sociology, and on SocINDEX, limits the extent to which our data can be used to evaluate interdisciplinary topics. Databases such as Scopus and Web of Science continue to have a clear advantage in that respect. Fourth, our methods cannot capture the most recent publications in a reliable way. Some books appear in Amazon months before their publication date while others appear only afterward, and this uneven coverage creates the potential for bias unless the publication cut-off date is set at least a few months before the start of the data compilation process.[9] Finally, while our data include the journal name, journal CiteScore, and year of every article, article titles and DOIs are not included. This makes it difficult to link each individual article to the article data available elsewhere—to article-specific citation counts, for instance. For books, the title, publisher, and year are provided.

Overall, six types of variables are included in the data file: general data on academic institutions (sociology departments), general data on individuals (sociology faculty), data on each journal article, data on each book, productivity data for individuals, and productivity data for academic institutions. For a list of the variables, see the Appendix, section A.9.

---

7   The Zenodo site includes the associated user notes while the openICPSR site does not. Moreover, Zenodo can be accessed anonymously while openICPSR requires registration.

8   Our methods account for scholarly productivity and for the relative standing of particular journals and book publishers, but not for the citation impact of each individual article or book. Notably, neither the large citation databases nor our data sources capture other important dimensions of scholarly impact, such as influence on teaching, practice, and public knowledge.

9   Likewise, our procedures cannot fully capture the most recent changes in authors' affiliations or characteristics.

4.2. USING THE DATA IN SOCIAL SCIENCE RESEARCH

Although at least 25 studies have rated or ranked the scholarly output of sociologists and sociology departments since 1970, just three post-2000 analyses include rankings of at least 40 U.S. sociology departments based on articles published in a wide range of journals (Wilder & Walters 2019, Table 1). Of the three, two deal exclusively with research universities (Ostriker et al. 2011; ShanghaiRanking Consultancy 2021) and one deals exclusively with liberal arts colleges (Hartley & Robinson 2001). None account for publications other than journal articles. Our data set therefore provides for a broader approach to the assessment of particular authors and departments. However, its real value lies in the inclusion of key covariates and the provision of identifying information that can be used to link these data to information obtained elsewhere. That is, our data set can be used to study the relationships between individual characteristics, institutional characteristics, and publishing productivity.[10] For instance, the first study to use these data (Wilder & Walters 2020b) presented several new findings:

1. The productivity differential between faculty at the major research universities and those at other institutions appears to have declined over time. More generally, the link between institution type and publishing productivity is weaker now than in the past.
2. Although men are more productive than women at the R1 universities, women are more productive than men at the top liberal arts colleges, other bachelor's institutions, and universities in the *other doctoral* category. While there are several possible explanations for this, previous studies show that men are especially likely to gain entry-level positions at the major research universities. It is therefore likely that women who would otherwise be working at those institutions can instead be found among the most productive faculty at the other types of colleges and universities.
3. Although the major research universities have the highest average article counts, the highest average *book* counts can be found at the top liberal arts colleges. Article authors and book authors can be readily distinguished from one another, as can article departments and book departments.
4. In general, especially high publication counts can be found among associate professors (rather than full professors), faculty with fewer than 17 years' experience, and authors with doctorates from the most prestigious universities.
5. There is high variation in productivity among institutions and individuals within each of the six institution types.

The second study based on these data focused on the most productive faculty at various types of colleges and universities, yielding further evidence in support of the second finding, above (Wilder & Walter, in press). It also revealed that while the most productive authors, as a group, tend to publish in the same journals as other faculty, they are especially likely to publish repeatedly in their own preferred journals, which vary with each individual.

We encourage others to use our data set, either to explore new areas of research or to investigate these same topics in greater detail or from different perspectives. As noted in section 4.1, these data may be of limited value for multidisciplinary research. At the same time, however, our data—and our methods, more generally—are well suited to research on academic or professional groups that can be clearly delineated on the basis of individuals' characteristics or the characteristics of their publication outlets. Although bibliometric research often deals with interdisciplinary or multidisciplinary groups, an approach centered on particular disciplines or occupations may be more appropriate for research in fields such as labor economics and the sociology of professions.

4.3. MANUAL AND AUTOMATED DATA COMPILATION METHODS

The methods used to compile and clean the data are described in the Appendix. To summarize, we began by identifying the sociology departments in the population of interest—four-year public and nonprofit colleges and universities in the United States—and compiling rosters of all the faculty with professor or associate professor rank. We then searched SocINDEX, Amazon, and other publicly accessible sources (e.g., course catalogs, Google Scholar, the IPEDS Data

---

10   This emphasis follows the tradition of bibliometric research grounded in the social sciences, an approach established by authors such as Cole & Cole (1967, 1973); see Franssen & Wouters (2019).

System, OCLC WorldCat, personal web sites, ProQuest Dissertation Express, publishers' web sites, and Scopus Sources) to compile information on institutional characteristics, individuals' characteristics, and publishing productivity.

These methods are time-intensive, of course. Although we did not systematically record the time spent on each particular task (i.e., the average time per individual for the faculty rosters or the average time per article for the journal article data), we can provide a general sense of the time commitment required. Working 15–20 person-hours per week on data compilation, we spent 11 weeks compiling departmental rosters for 426 departments (2,132 individuals), 14 weeks compiling data on 4,928 journal articles, and 10 weeks compiling data on 598 books. That's roughly 5 minutes per person record, 3 minutes per article record, and 18 minutes per book record.[11] In general, information about faculty at the major research universities could be obtained far more readily than information about faculty elsewhere. Moreover, a relatively small number of individuals—perhaps 15%—accounted for much of the total time spent investigating and verifying author and publication information.

Many information science researchers use automated methods to compile bibliographic and bibliometric data. For instance, application programming interfaces (APIs) can be used to harvest data from Google Scholar, Scopus, and WorldCat (Elsevier 2021; OCLC 2021; SerpAPI 2021), and some data sources are specifically designed for use with APIs (Hendricks et al. 2020; Visser et al. 2021). There are four problems with the use of APIs and other automated methods, however. The first is that automated searching relies heavily on the accuracy of the data, especially the subject codes. For instance, if the subject codes are defined poorly or applied inconsistently, they will not extract the records that the researcher desires. As discussed elsewhere (Walters 2017), the large citation databases sometimes group two or more distinct fields of study under a single code. Journals from a single discipline are sometimes split across multiple subject categories, some subject categories are much narrower than others, and some do not seem to correspond to coherent research areas or disciplines. Moreover, the large citation databases are simply prone to error. Scopus once classified *Developmental Psychology* as a demography journal, for instance, and Web of Science once listed *Financial Research Letters* in the *infectious diseases* category (Jacsó 2011, 2012).

A second problem is the existence of multiple records for a single research contribution. In some cases, the papers are variants such as a published article and the corresponding manuscript or preprint, but the database provides no mechanism by which the researcher can choose to count these variants as the same work or as multiple works. In other cases, the exact same paper (e.g., the same PDF file) is represented by multiple records. As discussed in section 3.2, this is a particular problem with WorldCat, which has 109 records for Web of Science and its key components. Disentangling the relationships among these records requires considerable effort, and it cannot be done through an automated search mechanism. The same problem is readily apparent with Google Scholar.

A third difficulty is that many automated methods fail to retrieve all relevant records, often due to a lack of standardization in the underlying data. For an investigation of the citation impact of papers in predatory accounting journals (in progress), we evaluated the performance of the Publish or Perish search tool (Harzing 2016) in retrieving Google Scholar records for articles published in the *International Journal of Accounting and Financial Reporting* from 2015 through 2018. The *Journal*'s web site lists 209 papers within that date range, and manual searches of Google Scholar for each article title retrieved 193 of them. However, a July 2020 Publish or Perish journal title search retrieved just 108. Further investigation revealed at least five reasons for the discrepancy: (a) many journal titles are very similar, and some of them fully incorporate the titles of other journals; (b) many Google Scholar records use abbreviated journal titles, often with several different abbreviations for a single journal; (c) journal titles are not always consistent, even at the publishers' web sites;[12] (d) automated journal title searches exclude variants of the

---

11   Most of the time spent on book records went into verifying authorship, determining whether particular books should be counted (e.g., distinguishing between new books and revised editions), and verifying bibliographic information (e.g., resolving discrepancies between the publication dates provided by two different sellers). A particular difficulty was that many authors had identical or near-identical names.

12   For instance, the journal name that appears on the web site may be different from that on the article PDFs, even for well-established journals. Inconsistencies in the use of ampersands, commas, and British or American spelling are especially common.

article that are not labeled with the journal name, such as pre-acceptance manuscripts and working papers; and (e) some individual articles are simply not included in Google Scholar.

Finally, automated searches may lead researchers to bypass the mechanisms that would otherwise make them more familiar with the patterns and idiosyncrasies that exist within the data. We believe manual searching can give the investigator a deeper understanding of the relationships among variables, including facets of those relationships that might not be detected by the more common statistical methods. (For example, is there a threshold level at which institutional prestige begins to influence book productivity? Is the relationship between gender and article productivity conditional on a third characteristic? Do the a priori delineations of the variables capture the most important distinctions between institutions and groups, or would alternative specifications be more appropriate?) Likewise, the experience of compiling the data may suggest hypotheses, explanations for findings that emerge later in the research process, or caveats related to data interpretation that might not have come to light through automated searching. For example, when compiling the data we noticed that many mid-ranked bachelor's and master's universities have just one 'superstar' faculty member with far higher productivity than the others—and that a disproportionate number of those superstars are women. We interpreted our initial statistical results with this idea in mind, then later developed more careful, formal methods of evaluating the situation (Wilder & Walters 2020b, in press). We cannot claim that manual data compilation is always a cost-effective approach, but it does appear useful as a mechanism for stimulating the kind of associative thinking that can lead to new insights and perspectives (Benedek et al. 2012; DeHaan 2011; Mednick 1962; Verhaeghen et al. 2017).

APPENDIX. DATA COMPILATION PROCEDURES
A.1. THE POPULATION OF INTEREST

Our institutional population is based on the set of all four-year public and nonprofit colleges and universities in the United States (National Center for Education Statistics 2017). However, it is restricted to institutions that award degrees in sociology. (See section A.2.) Broad interdisciplinary degrees (e.g., social sciences) were not counted for this purpose, nor were degrees in fields such as criminology and gerontology, even if based in departments of sociology. However, degrees that combine sociology with one other discipline (sociology and anthropology, for instance) were counted as sociology degrees. The data for Cornell University and the University of Wisconsin (Madison) include only faculty in the Departments of Sociology—not those in Development Sociology, Community and Environmental Sociology, or related departments.

The individuals in the population of interest include full-time faculty with the rank of (full) professor or associate professor. Faculty with endowed chairs or distinguished professor rank were included, as were those on sabbatical or temporary leave. Adjunct (part-time) faculty were excluded, as were instructors, lecturers, and assistant professors. (Section A.5 explains the exclusion of assistant professors.) Likewise, we excluded emeritus faculty as well as faculty with current, non-interim administrative appointments at the dean level or higher (e.g., provosts, vice provosts, and deans). Associate deans and department chairs were included, however. For departments with faculty from two or more academic disciplines, we included only the sociologists—those who hold doctorates in sociology *or* who teach more courses in sociology than in any other field.

A.2. SIX INSTITUTION TYPES

For interpretive and sampling purposes, we identified six types of colleges and universities:

1. TopR—Top research universities: The top 26 doctoral programs in sociology, based on the subjective ratings assigned by department chairs and graduate program directors at doctorate-granting institutions (U.S. News & World Report 2017b).
2. R1—Other R1 universities: Other institutions with a Carnegie classification of *doctoral universities: highest research activity* that award the doctorate in sociology (Carnegie Foundation 2017).
3. OD—Other doctoral universities: Institutions with a Carnegie classification of *doctoral universities: higher research activity* or *doctoral universities: moderate research activity* that award the doctorate in sociology.

4. M—Master's institutions: Institutions in the three Carnegie *master's* categories that award bachelor's or master's degrees in sociology.
5. TopLA—Top liberal arts colleges: The top 50 national liberal arts colleges that award the bachelor's degree in sociology. 'Top 50' status is based on alumni giving rate, class size, faculty salaries, financial resources, graduation rate, percentage of faculty with terminal degrees, retention rate, student selectivity, and undergraduate academic reputation (U.S. News & World Report 2017a).
6. B—Other bachelor's institutions: All other institutions with a Carnegie classification of *baccalaureate colleges: arts & sciences focus* or *baccalaureate colleges: diverse fields* that award the bachelor's degree in sociology.

Despite the labels used by the Carnegie Foundation, none of the six types are defined on the basis of publishing productivity.

A.3. SAMPLING

The data for four of the six institution types (TopR, OD, TopLA, and B) include the entire populations of interest. For those four institution types, Table A1 shows the base population (the number of institutions/departments in the relevant Carnegie classifications), the population of departments (the number that met the other criteria presented in section A.2), and the population of full and associate professors in those departments. As the table reveals, the population of departments is sometimes much smaller than the base population. For example, there are 201 universities with Carnegie classifications that place them in the OD category, but just 21 of them award the doctorate in sociology.

	TOPR	R1	OD	M	TOPLA	B
Population or sample	P	S	P	S	P	P
Base population of institutions[a]	26	89	201	695	50	469
Number of departments checked[b]	26	42	201	185	50	469
Population of departments[c]	26	64	21	406	41	200
Sample of departments[d]	—	30	—	108	—	—
Population of faculty[e]	546	867	205	1,518	165	403
Sample of faculty	—	409	—	404	—	—

Table A1 Population and sample sizes for six institution types.
a. Includes institutions without sociology programs and those without full or associate professors of sociology.
b. Number of departments checked to arrive at the necessary sample size (for R1 and M) or to cover the entire population (for TopR, OD, TopLA, and B).
c. For R1 and M, this is an estimate equal to Base population of institutions × (Sample of departments/Number of departments checked).
d. Number of checked departments that have sociology programs and full or associate professors of sociology.
e. For R1 and M, this is an estimate equal to Population of departments × (Sample of faculty/Sample of departments).

Our data for the R1 and M institutions are sample data that include roughly 47% and 27% of the corresponding populations. For the R1 group, we began by identifying the 89 universities in the relevant Carnegie classification, excluding those already placed in the TopR group. To obtain the R1 sample, we listed those institutions in random order, then went down the list and compiled data for the departments that met our criteria (offers doctorate in sociology and has one or more full or associate professors) until the sample included at least 400 faculty. We had to check 42 departments before reaching the desired sample size, and about 71% of those departments—30 departments with 409 faculty—met our criteria. Based on that proportion, we estimated a population size of 64 departments with 867 professors and associate professors. (See Table A1.) These same procedures were used with the institutions in the M category.

Because the data file includes sample rather than population data for two of the six institution types, it is necessary to apply case weights when estimating population values that account for all six institution types combined. Weights of 2.1198 for the R1 group, 3.7574 for the M group, and 1.0 for the other four groups will result in unbiased estimates for the population. To arrive at a sample that is representative of the entire population without inflating the sample size—when undertaking significance tests, for instance—use case weights of 1.2201 for R1, 2.1628 for M, and 0.5756 for all other cases.

A.4. GENERAL CHARACTERISTICS OF INDIVIDUALS AND DEPARTMENTS

Basic institutional data—institution name, location, and control (public, private nonreligious, Roman Catholic, Protestant, or other religious)—were obtained from the IPEDS Data System (National Center for Education Statistics 2017).

Department rosters were compiled in the first three months of 2018 from university web sites, course catalogs, OCLC WorldCat, personal web sites, ProQuest Dissertation Express, and other

publicly available sources. For each institution, we noted the highest sociology degree offered. For each individual, we recorded name, academic rank (professor or associate professor), gender (female or male), Ph.D. year, and Ph.D. institution. There are no missing values, although Ph.D. year was estimated for 8 of the 2,132 individuals.

Because the names of institutions and individuals were standardized, the personal names listed in our data file are not necessarily those used professionally by each individual. We may have used a full middle name, for instance, in order to provide for more reliable identification or to differentiate between individuals with similar names. Gender was determined through names, pronouns, and photographs, as presented on personal web sites, university web sites, and in sources such as RateMyProfessors. We found no cases in which our information sources suggested a gender category other than female or male.

A.5. PUBLISHING PRODUCTIVITY

Four measures were used to represent publishing productivity over the 2013–2017 period:

1. Articles: Number of articles in journals indexed by SocINDEX.
2. HI articles: Number of articles in high-impact journals indexed by SocINDEX.
3. Books: Number of books listed in Amazon.
4. HI books: Number of books from high-impact publishers listed in Amazon.

The consideration of both article and book counts is essential, since both forms of publication remain important within sociology. See sections A.6 and A.7 for notes on the delineation of high-impact journals and publishers.

Our four measures of publishing productivity all represent five-year productivity (January 2013 through December 2017) rather than lifetime productivity. While this constraint limits investigators' ability to directly examine long-term trends, it also helps avoid two significant problems. First, by focusing on five-year productivity and excluding assistant professors, we ensure a five-year period of potential productivity for everyone in the population of interest. That is, we avoid the need to pro-rate scholarly productivity based on the number of research-active years. (Active engagement in research may or may not pre-date the Ph.D. year, so the inclusion of assistant professors would have required us to determine a 'first research year' for each faculty member with less than five years' experience.) Second, the use of a five-year period minimizes the potential impact of name changes and avoids the difficulty of using older bibliographic records that sometimes list just initials rather than first names. Our use of multiple information sources gave us confidence in matching authors to scholarly works over a five-year period. That task would have been more difficult and less reliable if we had tried to match authors and works over a period of several decades.

A.6. ARTICLE SEARCHES AND THE HIGH-IMPACT ARTICLE DESIGNATION

SocINDEX searches for individual articles were conducted over a four-month period beginning in March 2018. Each SocINDEX search was limited to peer-reviewed journals and to contributions with a document type of *article* rather than *book review, editorial, letter*, and so on. Because SocINDEX uses the *article* designation for some items that are not actually articles, every item of six or fewer pages was evaluated individually to determine whether it fit that description. Items longer than six pages were also excluded if the *article* designation was obviously incorrect. Finally, two magazines intended for general audiences—*Focus* (Institute for Research on Poverty, University of Wisconsin-Madison) and *Contexts: Understanding People in Their Social Worlds* (American Sociological Association)—were excluded from the article counts even though SocINDEX lists them as peer reviewed.

To ensure comprehensiveness, we searched for multiple variants of each author's name: the full name and the short form (e.g., 'Christopher' and 'Chris'), with the middle initial and without, with the full middle name (if known) and without. We also searched without the first name if there was any reason to believe the author did not use it consistently. Hyphenated last names were searched as written, with a space instead of a hyphen, with the first component alone, and with the second component alone.

As described in section A.5, the data file includes separate counts for *all articles* and *articles in high-impact journals*. The 44 high-impact journals are those with a Scopus CiteScore of 2.35 or

greater *and* a CiteScore rank of 95th percentile or better within the category in which the journal is ranked highest. (A particular journal may be listed in multiple Scopus subject categories.) This two-part standard ensures that the high-impact journals have high citation impact in both absolute and relative terms. The high-impact journals represent 9% of the journals in the sample but 25% of the articles.

A.7. BOOK SEARCHES AND THE HIGH-IMPACT PUBLISHER DESIGNATION

Our Amazon searches were conducted in June, July, and August 2018. We used the *Books— Advanced* search and checked multiple variants of each author's name, as with SocINDEX.

Our book counts include only new books with initial publication dates from January 2013 through December 2017. We excluded chapters in edited volumes, editorships of edited volumes, new editions, translations, and re-publications such as paperback editions of titles originally issued in hardcover. We also excluded self-published books and books of fewer than 60 pages.

As noted in section A.5, the data file includes separate counts for *all books* and *books from high-impact publishers*. The high-impact publishers—the top 25 in terms of average citations per book—include 18 university presses (Belknap, California, Cambridge, Chicago, Clarendon, Columbia, Cornell, Duke, Harvard, Johns Hopkins, Manchester, MIT, North Carolina, Oxford, Pennsylvania, Princeton, Stanford, and Yale) and 7 commercial publishers (Basic Books, Berg, Knopf, Penguin, Polity Press, Verso, and W.W. Norton) (Zuccala et al. 2015). Together, they account for 45% of the books in the sample.

A.8. WEIGHTING PRODUCTIVITY MEASURES TO ACCOUNT FOR CO-AUTHORSHIP

For all four productivity measures, we used harmonic weighting to assign credit for works with two or more authors. This method accounts for the number of authors as well as each individual's place in the author list. Specifically, the credit assigned to each author of a paper is $1/i$ divided by $(1/1 + 1/2 + 1/3 + ... + 1/N)$, where N is the number of authors and i is the author's place (1 for first author, 2 for second author, etc.). For instance, the first author of a paper with three authors receives 0.545 credits; the second, 0.273 credits; and the third, 0.182 credits. Authorship credits calculated in this way correspond well to the subjective weights assigned by scholars in the natural and social sciences (Hagen 2010, 2013, 2014a, 2014b). In particular, they match scholars' subjective assessments more closely than either whole counting or fractional counting. Although harmonic weighting does not account for the practice of listing a senior author last, that approach is more common in the natural sciences than in the social sciences (Abramo et al. 2013). For information on alternative weighting methods, see Wilder and Walters (2020b, in press).

A.9. VARIABLES INCLUDED IN THE DATA FILE

Apart from certain identifier variables and note fields, six types of variables are included in the data set:

1. General data on academic institutions (sociology departments): institution name, institution type, location, public or private status, highest sociology degree offered, number of faculty in the department, percentage who are full rather than associate professors, percentage who are female, average years since Ph.D. (in 2018), percentage who earned the Ph.D. within each of three date ranges, and percentage with doctorates from top-25 universities.
2. General data on individuals (sociology faculty): name, faculty rank, gender, Ph.D. year, Ph.D. institution, doctorate from a top-25 university (yes or no), and non-U.S. doctorate (yes or no).
3. Data on each journal article, matched to the general institutional and individual data: year of publication, journal name, number of authors, author's place in the author order, article credit (adjusted for coauthorship), CiteScore of journal, percentile rank corresponding to the CiteScore, high-impact journal (yes or no), and high-impact article credit (adjusted for coauthorship; see section A.8).

4. Data on each book, matched to the general institutional and individual data: year of publication, title, publisher, number of authors, author's place in the author order, fewer than 60 pages (yes or no), book credit (adjusted for coauthorship), high-impact publisher (yes or no), and high-impact book credit (adjusted for coauthorship).

5. Productivity data for individuals: article credit, high-impact article credit, book credit, and high-impact book credit (all adjusted for coauthorship).

6. Productivity data for academic institutions (sociology departments): the same four measures, each presented as the total, average, median, and standard deviation for the department.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Esther Isabelle Wilder *orcid.org/0000-0002-2232-9371*
Lehman College, The City University of New York, US

William H. Walters *orcid.org/0000-0001-9490-4032*
Mary Alice & Tom O'Malley Library, Manhattan College, US

REFERENCES

Abramo, G., D'Angelo, C. A., & **Rosati, F.** (2013). Measuring institutional research productivity for the life sciences: The importance of accounting for the order of authors in the byline. *Scientometrics, 97*(3), 779–795. DOI: *https://doi.org/10.1007/s11192-013-1013-9*

Baas, J., Schotten, M., Plume, A., Côté, G., & **Karimi, R.** (2020). Scopus as a curated, high-quality bibliometric data source for academic research in quantitative science studies. *Quantitative Science Studies, 1*(1), 377–386. DOI: *https://doi.org/10.1162/qss_a_00019*

Bar-Ilan, J. (2010). Web of Science with the Conference Proceedings Citation Indexes: The case of computer science. *Scientometrics, 83*(3), 809–824. DOI: *https://doi.org/10.1007/s11192-009-0145-4*

Benedek, M., Könen, T., & **Neubauer, A. C.** (2012). Associative abilities underlying creativity. *Psychology of Aesthetics, Creativity, and the Arts, 6*(3), 273–281. DOI: *https://doi.org/10.1037/a0027059*

Birkle, C., Pendlebury, D. A., Schnell, J., & **Adams, J.** (2020). Web of Science as a data source for research on scientific and scholarly activity. *Quantitative Science Studies, 1*(1), 363–376. DOI: *https://doi.org/10.1162/qss_a_00018*

Carnegie Foundation for the Advancement of Teaching. (2017). *Standard listings: Basic classification.* Retrieved from *https://carnegieclassifications.iu.edu/lookup/standard.php* (11 July 2021).

Clarivate Analytics. (2021). *Editorial selection process: Web of Science Core Collection.* Retrieved from *https://clarivate.com/webofsciencegroup/solutions/editorial/* (11 July 2021).

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

Cole, S., & **Cole, J. R.** (1967). Scientific output and recognition: A study in the operation of the reward system in science. *American Sociological Review, 32*(3), 377–390. DOI: *https://doi.org/10.2307/2091085*

Czechowski, L. (2011). Problems with e-books: Suggestions for publishers. *Journal of the Medical Library Association, 99*(3), 181–182. DOI: *https://doi.org/10.3163/1536-5050.99.3.001*

DeHaan, R. L. (2011, December 16). Teaching creative science thinking. *Science, 334*(6062), 1499–1500. DOI: *https://doi.org/10.1126/science.1207918*

Delgado López-Cózar, E., Orduña-Malea, E., & **Martín-Martín, A.** (2019). Google Scholar as a data source for research assessment. In W. Glänzel, H. F. Moed, U. Schmoch & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 95–127). DOI: *https://doi.org/10.1007/978-3-030-02511-3_4*

EBSCO. (2021). *Library, Information Science and Technology Abstracts.* Retrieved from *https://www.ebsco.com/products/research-databases/library-information-science-and-technology-abstracts* (11 July 2021).

Elhauge, E. R. (1991). Does interest group theory justify more intrusive judicial review? *Yale Law Journal, 101*(1), 31–110. Retrieved from *https://digitalcommons.law.yale.edu/cgi/viewcontent.cgi?article=7390&context=ylj* (11 July 2021). DOI: *https://doi.org/10.2307/796935*

Elsevier. (2021). *Elsevier developer portal.* Retrieved from *https://dev.elsevier.com/* (11 July 2021).

Engels, T. C. E., Starčič, A. I., Kulczycki, E., Pölönen, J., & **Sivertsen, G.** (2018). Are book publications disappearing from scholarly communication in the social sciences and humanities? *Aslib Journal of Information Management, 70*(6), 592–607. DOI: *https://doi.org/10.1108/AJIM-05-2018-0127*

Etxebarria, G., & Gomez-Uranga, M. (2010). Use of Scopus and Google Scholar to measure social sciences production in four major Spanish universities. *Scientometrics*, *82*(2), 333–349. DOI: *https://doi.org/10.1007/s11192-009-0043-9*

Franssen, T., & Wouters, P. (2019). Science and its significant other: Representing the humanities in bibliometric scholarship. *Journal of the Association for Information Science and Technology*, *70*(10), 1124–1137. DOI: *https://doi.org/10.1002/asi.24206*

Giménez Toledo, E. (2020). Why books are important in the scholarly communication system in social sciences and humanities. *Scholarly Assessment Reports*, *2*, article 6. *https://www.scholarlyassessmentreports.org/articles/10.29024/sar.14/*

Giménez Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, *22*(1), 64–77. DOI: *https://doi.org/10.1093/reseval/rvs036*

Hagen, N. T. (2010). Harmonic publication and citation counting: Sharing authorship credit equitably—not equally, geometrically or arithmetically. *Scientometrics*, *84*(3), 785–793. DOI: *https://doi.org/10.1007/s11192-009-0129-4*

Hagen, N. T. (2013). Harmonic coauthor credit: A parsimonious quantification of the byline hierarchy. *Journal of Informetrics*, *7*(4), 784–791. DOI: *https://doi.org/10.1016/j.joi.2013.06.005*

Hagen, N. T. (2014a). Counting and comparing publication output with and without equalizing and inflationary bias. *Journal of Informetrics*, *8*(2), 310–317. DOI: *https://doi.org/10.1016/j.joi.2014.01.003*

Hagen, N. T. (2014b). Reversing the byline hierarchy: The effect of equalizing bias on the accreditation of primary, secondary and senior authors. *Journal of Informetrics*, *8*(3), 618–627. DOI: *https://doi.org/10.1016/j.joi.2014.05.003*

Hartley, J. E., & Robinson, M. D. (2001). Sociology research at liberal arts colleges. *The American Sociologist*, *32*(3), 60–72. DOI: *https://doi.org/10.1007/s12108-001-1028-1*

Harzing, A.-W. (2013). Document categories in the ISI Web of Knowledge: Misunderstanding the social sciences? *Scientometrics*, *94*(1), 23–34. DOI: *https://doi.org/10.1007/s11192-012-0738-1*

Harzing, A.-W. (2016). *Publish or perish*. Retrieved from *https://harzing.com/resources/publish-or-perish* (11 July 2021).

Harzing, A.-W. (2019). Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science? *Scientometrics*, *120*(1), 341–349. DOI: *https://doi.org/10.1007/s11192-019-03114-y*

Hendricks, G., Tkaczyk, D., Lin, J., & Feeney, P. (2020). Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, *1*(1), 414–427. DOI: *https://doi.org/10.1162/qss_a_00022*

Herzog, C., Hook, D., & Konkiel, S. (2020). Dimensions: Bringing down barriers between scientometricians and data. *Quantitative Science Studies*, *1*(1), 387–395. DOI: *https://doi.org/10.1162/qss_a_00020*

Jacsó, P. (2005). Google Scholar: The pros and the cons. *Online Information Review*, *29*(2), 208–214. DOI: *https://doi.org/10.1108/14684520510598066*

Jacsó, P. (2008). Google Scholar revisited. *Online Information Review*, *32*(1), 102–114. DOI: *https://doi.org/10.1108/14684520810866010*

Jacsó, P. (2010). Metadata mega mess in Google Scholar. *Online Information Review*, *34*(1), 175–191. DOI: *https://doi.org/10.1108/14684521011024191*

Jacsó, P. (2011). The h-index, h-core citation rate and the bibliometric profile of the Scopus database. *Online Information Review*, *35*(3), 492–501. DOI: *https://doi.org/10.1108/14684521111151487*

Jacsó, P. (2012). The problems with the subject categories schema in the Eigenfactor database from the perspective of ranking journals by their prestige and impact. *Online Information Review*, *36*(5), 758–766. DOI: *https://doi.org/10.1108/14684521211276064*

Kousha, K., & Thelwall, M. (2015). Alternative metrics for book impact assessment: Can *Choice* reviews be a useful source? In A. A. Salah, Y. Tonta, A. A. Akdag Salah, C. Sugimoto, & U. Al (Eds.), *Proceedings of ISSI [International Society for Scientometrics and Informetrics] 2015* (pp. 59–70). Retrieved from *https://pdfs.semanticscholar.org/c69a/38d1ac5bafd750a3f54411452e8ac6d5f79d.pdf* (11 July 2021).

Kousha, K., & Thelwall, M. (2016). Can *Amazon.com* reviews help to assess the wider impacts of books? *Journal of the Association for Information Science and Technology*, *67*(3), 566–581. DOI: *https://doi.org/10.1002/asi.23404*

Larsen, P. O., & von Ins, M. (2010). The rate of growth in scientific publication and the decline in coverage provided by Science Citation Index. *Scientometrics*, *84*(3), 575–603. DOI: *https://doi.org/10.1007/s11192-010-0202-z*

Lisée, C., Larivière, V., & Archambault, E. (2008). Conference proceedings as a source of scientific information: A bibliometric analysis. *Journal of the American Society for Information Science and Technology*, *59*(11), 1776–1784. DOI: *https://doi.org/10.1002/asi.20888*

Martín-Martín, A., Orduna-Malea, E., Thelwall, M., & Delgado López-Cózar, E. (2018). Google Scholar, Web of Science, and Scopus: A systematic comparison of citations in 252 subject categories. *Journal of Informetrics*, *12*(4), 1160–1177. DOI: *https://doi.org/10.1016/j.joi.2018.09.002*

Mednick, S. (1962). The associative basis of the creative process. *Psychological Review, 69*(3), 220–232. DOI: *https://doi.org/10.1037/h0048850*

Moksony, F., Hegedus, R., & **Császár, M.** (2014). Rankings, research styles, and publication cultures: A study of American sociology departments. *Scientometrics, 101*(3), 1715–1729. DOI: *https://doi.org/10.1007/s11192-013-1218-y*

National Center for Education Statistics. (2017). *IPEDS integrated postsecondary education data system: Compare institutions.* 2016–17 data. Retrieved from *https://nces.ed.gov/ipeds/use-the-data* (11 July 2021).

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & **Dekker, P. J.** (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social and behavioural sciences: A comparative study. *Scientometrics, 15*(5–6), 423–435. DOI: *https://doi.org/10.1007/BF02017063*

OCLC. (2021). *WorldCat Search API.* Retrieved from *https://www.oclc.org/developer/develop/web-services/worldcat-search-api.en.html* (11 July 2021).

Olson, M. (1971). *The logic of collective action: Public goods and the theory of groups.* Revised edition.

Orduña-Malea, E., Ayllón, J. M., Martín-Martín, A., & **Delgado López-Cózar, E.** (2017). The lost academic home: Institutional affiliation links in Google Scholar citations. *Online Information Review, 41*(6), 762–781. DOI: *https://doi.org/10.1108/OIR-10-2016-0302*

Ostriker, J. P., Kuh, C. V., & **Voytuk, J. A.** (Eds.). (2011). *A data-based assessment of research-doctorate programs in the United States.* Retrieved from *https://www.nap.edu/download/12994* (11 July 2021).

Pomerantz, S. (2010). The availability of e-books: Examples of nursing and business. *Collection Building, 29*(1), 11–14. DOI: *https://doi.org/10.1108/01604951011015240*

SerpAPI. (2021). *Google Scholar API.* Retrieved from *https://serpapi.com/google-scholar-api* (11 July 2021).

ShanghaiRanking Consultancy. (2021). *ShanghaiRanking's global ranking of academic subjects— Sociology.* Retrieved from *http://www.shanghairanking.com/rankings/gras/2021/RS0505* (11 July 2021).

Torres-Salinas, D., Robinson-García, N., Cabezas-Clavijo, Á., & **Jiménez-Contreras, E.** (2014). Analyzing the citation characteristics of books: Edited books, book series and publisher types in the book citation index. *Scientometrics, 98*(3), 2113–2127. DOI: *https://doi.org/10.1007/s11192-013-1168-4*

Tyler, D. C., Cross, J., & **DeFrain, E.** (2017). Sociological Abstracts vs. SocINDEX for graduate students in sociology: Comprehensive enough to satisfy? *Library Philosophy and Practice,* 1520. Retrieved from *http://digitalcommons.unl.edu/libphilprac/1520* (11 July 2021).

U.S. News & World Report. (2017a). *National liberal arts colleges 2017.* Retrieved 1 December 2017; no longer available online.

U.S. News & World Report. (2017b). *Sociology 2017.* Retrieved 1 December 2017; no longer available online.

Verhaeghen, P., Trani, A. N., & **Aikman, S. N.** (2017). On being found: How habitual patterns of thought influence creative interest, behavior, and ability. *Creativity Research Journal, 29*(1), 1–9. DOI: *https://doi.org/10.1080/10400419.2017.1263504*

Visser, M., van Eck, N. J., & **Waltman, L.** (2021). Large-scale comparison of bibliographic data sources: Scopus, Web of Science, Dimensions, Crossref, and Microsoft Academic. *Quantitative Science Studies, 2*(1), 20–41. DOI: *https://doi.org/10.1162/qss_a_00112*

Walters, W. H. (2013). E-books in academic libraries: Challenges for acquisition and collection management. *Portal: Libraries and the Academy, 13*(2), 187–211. DOI: *https://doi.org/10.1353/pla.2013.0012*

Walters, W. H. (2017). Citation-based journal rankings: Key questions, metrics, and data sources. *IEEE Access, 5*, 22036–22053. DOI: *https://doi.org/10.1109/ACCESS.2017.2761400*

Walters, W. H., & **Wilder, E. I.** (2016). Disciplinary, national, and departmental contributions to the literature of library and information science, 2007–2012. *Journal of the Association for Information Science and Technology, 67*(6), 1487–1506. DOI: *https://doi.org/10.1002/asi.23448*

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & **Cole, F. T.** (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology, 60*(6), 1083–1096. DOI: *https://doi.org/10.1002/asi.21045*

Wilder, E. I., & **Walters, W. H.** (2019). Quantifying scholarly output: Contribution studies and productivity studies in sociology since 1970. *The American Sociologist, 50*(3), 430–436. DOI: *https://doi.org/10.1007/s12108-018-9400-6*

Wilder, E. I., & **Walters, W. H.** (2020a). *New data on the publishing productivity of American sociologists.* Data file available at *Zenodo* (DOI: *https://doi.org/10.5281/zenodo.3892308*) and at *openICPSR* (*https://doi.org/10.3886/E119867V1*).

Wilder, E. I., & **Walters, W. H.** (2020b). Publishing productivity of sociologists at American colleges and universities: Institution type, gender, and other correlates of book and article counts. *Sociological Perspectives, 63*(2), 249–275. DOI: *https://doi.org/10.1177/0731121419874079*

Wilder, E. I., & **Walters, W. H.** (In press). Characteristics of the most productive U.S. sociology faculty and departments: Institution type, gender, and journal concentration. *Sociological Quarterly,* in press. DOI: *https://doi.org/10.1080/00380253.2020.1775530*

Zuccala, A., Guns, R., Cornacchia, R., & **Bod, R.** (2015). Can we rank scholarly book publishers? A bibliometric experiment with the field of history. *Journal of the Association for Information Science and Technology, 66*(7), 1333–1347. DOI: *https://doi.org/10.1002/asi.23267*