The MIT Press

RESEARCH ARTICLE

Field-level differences in paper and author characteristics across all fields of science in Web of Science, 2000–2020

Jens Peter Andersen [iD]

Danish Centre for Studies in Research and Research Policy, Aarhus University, Aarhus, Denmark

ABSTRACT

With increasing availability of near-complete, structured bibliographical data, the past decade has seen a rise in large-scale bibliometric studies attempting to find universal truths about the scientific communication system. However, in the search for universality, fundamental differences in knowledge production modes and the consequences for bibliometric assessment are sometimes overlooked. This article provides an overview of article and author characteristics at the level of the OECD minor and major fields of science classifications. The analysis relies on data from the full Web of Science in the period 2000–2020. The characteristics include document type, median reference age, reference list length, database coverage, article length, coauthorship, author sequence ordering, author gender, seniority, and productivity. The article reports a descriptive overview of these characteristics combined with a principal component analysis of the variance across fields. The results show that some clusters of fields allow interfield comparisons, and assumptions about the importance of author sequence ordering, while other fields do not. The analysis shows that major OECD groups do not reflect bibliometrically relevant field differences, and that a reclustering offers a better grouping.

1. INTRODUCTION

The vastness of available bibliographic metadata has created a fertile field for bibliometric and scientometric research, and related areas. These areas have recently grown, not just in paper volume, diversity in methods, and measurements but also very much so in the amount and availability of data used in often global analyses of publication data. This has created a paradox of sorts, with the opportunities of the data on one hand, and the limitations and complexities of the same data on the other. Problems that are solvable in small-scale analyses, such as author name disambiguation or precise field delineation, become more difficult as the quantity of data grows. Assumptions about universal comparability across data also become harder as the magnitude and inclusivity of data increase.

Bibliometric research has often relied on an argument of *random errors*, which are assumed to even out when the lens is shifted from the individual object (where one may find missing and erroneous citations and references, or deliberate differences in the intentions of a reference (e.g., criticism)) to the statistical aggregate (Van Raan, 1998). This is a meaningful approach, but it requires that the objects that are aggregated are more or less the same type of objects. Naïvely, we could assume that scientific publications are all the same; however, it is

well established that there are rather large differences in, for example, referencing practices between fields (Glänzel & Schubert, 2003; Leydesdorff & Bornmann, 2011). This makes direct comparisons of citation counts and averages between fields somewhat meaningless. One might even argue that unstable changes even make comparisons within fields somewhat meaningless over longer time spans, whether those changes are due to changes in database inclusion or an absolute change in publication intensity (Nielsen & Andersen, 2021; Petersen, Pan et al., 2019). This is not new, and sophisticated indicators have been developed to account for some of these field differences in citation impact (Waltman & van Eck, 2019).

However, variation in citation intensity is not the only difference between fields of research. Fundamentally, there is a difference between the objects studied, which is linked to the ontological and epistemological assumptions of a field. This influences the norms and modes of knowledge production and authorship, along with varying norms in referencing and credit distribution, as well as expectations for outputs and impacts. We know a lot about bibliometrically relevant field differences already, such as about longer citing half-lives in certain fields (Glänzel & Schoepflin, 1999; Zhang & Glänzel, 2017), large collaborations in others (Piro, Aksnes, & Rørstad, 2013), latent rules about authorship positions (e.g., Burrows & Moore, 2011), and differences in preferred publishing venues and publication types (Bourke & Butler, 1996; Kulczycki, Engels, & Nowotniak, 2017; Sigogneau, 2000). We also know a great deal about the people behind the research, such as their career paths and how factors such as gender play a large role in the careers of scientists (e.g., Jagsi, Guancial et al., 2006; Lerchenmüller, Lerchenmueller, & Sorenson, 2018; Xie & Shauman, 1998). However, most of our knowledge about these field characteristics focuses on one or few particular aspects, and/or on specific fields. This previous research has established our knowledge about these field differences, and the narrowness of focus has been necessary, but we are lacking a broad overview allowing field comparisons on the global scale, with a wide range of characteristics across the board.

This article provides a comparative overview of field differences, as they can be measured from publications, as well as an analysis of field groupings based on these characteristics. It provides both paper and author characteristics at the OECD fields of science minor level (OECD, 2007), which corresponds rather well with traditional university departments. Researcher-level (i.e., career or personal) characteristics are not included in this article. Complete publication data from Web of Science (WoS), covering the period 2000 to 2020, are used to provide a primarily descriptive analysis. We also address the question of whether common field delineations, and especially groupings on the major level, such as the OECD fields of science are meaningful for bibliometric research.

As a final caveat regarding the analytical units of this article: While many field differences are found on the individual level (i.e., that of the individual researcher), this is truly the core of much new research on the sociology and science of science, and beyond the scope of this article. However, by documenting differences on the paper and author levels (i.e., the *abstract metadata* "authors" of articles, not the *physical person* behind this), I hope to provide information for additional research on all three of these levels of research. It should also be noted that this article will focus solely on journal articles, which is the predominant publication form in most fields of science. Additional information about other publication types will be included where relevant. This is detailed further in Section 3. It should also be noted that several of the variables mentioned in the following are commonly employed in bibliometric research, although they may not have all been studied as a property of a scientific field.

2. LITERATURE

Bibliometric comparisons between fields have always been problematic. It is long established that the "expected" average citations to an article vary greatly from field to field, and absolute numerical comparisons are not possible (e.g., Leydesdorff & Bornmann, 2011). This has been addressed in different ways, either by calculating normalized scores relative to the mean (e.g., Lundberg, 2007; Waltman, Calero-Medina et al., 2012), or as percentile-based indicators (Pudovkin & Garfield, 2009). While these approaches also have limitations (e.g., they depend somewhat on the size and activity as well as proper delineation of fields), they certainly increase comparability. But this is only the case for direct comparisons between large aggregates of publications (e.g., on the university level).

When it comes to sociological analyses of science, or research on the many other facets of science than citations, other field differences also matter. As an example, many analyses of academic careers rely on the position of authors in the byline to infer seniority (Jian & Xiaoli, 2013; Milojević, Radicchi, & Walsh, 2018; Perneger, Poncet et al., 2017), however, in some fields the order of the byline is commonly alphabetical (Henriksen, 2019; Mongeon, Smith et al., 2017; Waltman, 2012). Also, in terms of productivity, the expected output of a researcher varies a lot from field to field, depending on which types of publications are produced, how many coauthors work together on creating a publication, and how long such a publication is. For such types of analyses to make sense, we need to select meaningful fields to limit analyses on and resist the temptation of using universal publication data sets, simply because they are available. In the following, I will list those characteristics at the article and author level that have been found to influence such field differences, or which have obvious consequences for field selection. These will also be the characteristics that are included in the current analysis. In addition to obvious characteristics, I also include the (inferred) gender of authors as a variable. While there are many well-documented gender differences in academia (e.g., Allison & Stewart, 1974; Lerchenmüller et al., 2018; Xie & Shauman, 1998), it is perhaps less obvious why this should be a field characteristic, as the gender distribution is probably not causally related to, for example, knowledge production modes, publication types, and reference list lengths. However, it matters greatly for field selection and expectations around gender differences for quantitative studies in the sociology of science.

In the final part of this article, field characteristics on both the article and author level are used to explore a regrouping of fields through a principal component analysis of the included variables. In this analysis, gender will not be included, based on the same reasoning as above.

2.1. Article Characteristics

2.1.1. Document types

Citation analysis, especially focused on the natural and health sciences, is often restricted to particular publication types, namely original research articles in journals, reviews, and sometimes letters, while omitting editorial material, comments, books, book chapters, reports, and many other publication types. This is often a meaningful choice, as these are the document types that are most commonly—in those sciences—part of the scientific communication system. However, in parts of the social sciences and the humanities, monographs, edited volumes, art reviews, and a plethora of other document types are important. With different document types comes variation in the length and work put into them, and the interpretation in terms of productivity, the use and meaning of citations, collaboration forms, and in general just the role they play in producing and communicating knowledge. Bourke and Butler (1996) documented such field norms, showing the differences in which publication types were

common in the sciences, social sciences, and humanities (broadly categorized). This is not a static image though, as changes in norms have occurred over time, such as in the natural sciences, where changing trends in the use of WoS publication type registrations were seen in physics in the 1990s (Sigogneau, 2000). Even larger changes have recently been observed in the social sciences and humanities, with substantial national differences in the distribution of types as well as over time (Kulczycki et al., 2017; Kulczycki, Engels et al., 2018). Similar broad differences were also observed by Piro et al. (2013).

2.1.2. Cited half-life

The cited half-life is the median age of an article's references (i.e., the number of years between when the article was published and the cited reference was published). This gives us an indication about the speed of research in a field (i.e., the degree to which knowledge is generated in a cumulative mode). This indicator has been included in the *Journal Citation Reports* for many years. Many studies include this indicator in some shape, and Leydesdorff (2009) argues that it adds an additional dimension to traditional citation-based indicators. The majority of articles using the indicator operationalize it rather than studying field differences. Important exceptions to this are the studies by Glänzel and Schoepflin (1999) and more recently Zhang and Glänzel (2017). In Glänzel and Schoepflin (1999), selected fields from the natural and social sciences are compared, using mean reference age as one measure, highly similar to cited half-life. The same measure is used in Zhang and Glänzel (2017), and supplemented by the median reference age, comparing changes in this measure across a number of fields from 1992 to 2014. They show differences across fields of more than a factor of two, as well as some increase in mean reference age over time, mostly due to an increased referencing rate of very old papers and decreased referencing rate of very new papers.

2.1.3. Number of references and reference coverage

In addition to the cited half-life, the quantity of references in a field is one of the determining factors in how many citations a given publication can be expected to receive. Differences between fields in the reference list length is thus a central norm to consider when selecting and especially comparing fields. This was a central part of the arguments of both Glänzel, Schubert et al. (2011) and Leydesdorff and Bornmann (2011) for how to perform field delineations when field-normalizing citation-based indicators. But not only are there differences in the length of reference lists, the proportion of references covered by WoS also varies between fields (e.g., Kulczycki et al., 2018), which is both a question of document type distributions (because journals are covered better than books and conference proceedings in WoS) and field-specific journal coverage (Kulczycki et al., 2018).

2.1.4. Article length

Article length is one of the variables that have been studied as a driver for citations, for example in management science (Stremersch, Verniers, & Verhoef, 2007) and accounting research (Meyer, Waldkirch et al., 2018), with small, yet positive correlations. I would argue that while there may be some theoretical sense behind this correlation (i.e., longer articles potentially have additional content or are more thorough, and thus of higher quality), there is some speculation about referencing behavior in such an argumentation that does not align well with known reasons for referencing scientific work (Garfield, 1996). Nonetheless, there seems to be a correlation, and article length is also important for other reasons; both Stremersch et al. (2007) and Meyer et al. (2018) looked at within-field article length variation, which is a sensible delineation in both cases, but we should expect to see much greater between-field

variation, as this is something that is expectedly tied to the disciplinary norms of a field (Fanelli & Glänzel, 2013). While potential effects on expected citation outcomes between fields can be solved through field normalization, it is less common to make such normalizations for the productivity of researchers (i.e., the number of articles published by a scientist), and in this case an assumption about how much work goes into the making of an article. Page count is far from a perfect proxy for this, as a given page contains highly varying amounts of information (e.g., tables, figures, and formulas in contrast to text quotations), and the assumption of work amount is also highly field specific, as, for example, experimentation or field work does not directly influence article length. However, it is the only straightforward and available proxy allowing us to analyze field differences in article length, and is included as a part of the puzzle of figuring out which variables inform us about the bibliometrically relevant differences in knowledge production between fields.

2.1.5. Coauthorship

Following naturally from the above, the work that goes into making an article also depends on the number of collaborators, and it is well known that this is a norm with extreme variation between fields. As an example, in Nielsen and Andersen (2021), physics and astronomy was excluded from the analysis, as the number of coauthors (>1,000) on some articles skewed the overall trends of all included fields. More systematically, Piro et al. (2013) studied productivity and collaboration in terms of both whole counts and fractional counts across 37 fields in the humanities, social sciences, natural sciences, medicine, and technology. Excluding the more extreme cases (here >100 authors), there were still clear field differences, but with medicine rising to be more collaborative than the natural sciences (which fell from 19.6 to 5.6 authors per paper on average). Both Nielsen and Andersen (2021) and Fanelli and Larivière (2016) have found that productivity per author has increased greatly when whole-counting publications, but remained stable when taking into account the number of coauthors.

2.2. Author Characteristics

2.2.1. Sequence ordering

It is obvious that the amount of work from any given author on an article with very long coauthor lists is different from that of articles from small groups or individual researchers. One of the ways in which bibliometric research has approached this problem is by attributing greater weight to the main author(s) of a publication. However, this relies on an assumption that it is possible to infer who the main author is. In some cases, this is the corresponding author; however, recent research has shown that the number of corresponding authors per paper has increased over time and the position in the byline of the corresponding author(s) varies between fields (Chinchilla-Rodríguez, Costas et al., 2022). Other research has shown that intentional alphabetical ordering is highly field dependent, and generally decreasing over time (Henriksen, 2019; Mongeon et al., 2017; Waltman, 2012), while in other fields there is a tendency to assign special meaning to the first and last author position, namely that the first author produced the majority of the work, wrote the article, and performed the experiments, and the last author supervised, secured funding, or had similar leadership roles (Larivière, Desrochers et al., 2016; Larivière, Pontille, & Sugimoto, 2021). In these studies, high agreement about the role of first authors was found but less so on the last authors, which varied more in their role. In this article, I will report both the probability of intentional alphabetical ordering and the difference in academic age (seniority) and productivity of first and last authors to quantify the tendency of a field to use equal distribution of credit (alphabetical ordering) and to assign special value to the first and last author position.

2.2.2. Seniority

Seniority and academic age have been used in numerous studies as explanatory variables for differences in, for example, productivity and citedness. Milojević (2012) has shown connections between the time from first publication (academic age) and productivity, collaboration, and the referencing behavior of scientists. Also the position in the network changes over time, where more senior scientists tend to become more central in their collaboration network (Wang, Yu et al., 2017).

2.2.3. Gender

Quantitative studies of gender differences in academia have existed at least since the 1970s (Allison & Stewart, 1974; Cole, 1979; Cole & Zuckerman, 1984). Landmark studies since then have shown, for example, evidence of systemic differences in the productivity of men and women, the differences in expectations for promotions, and the consequences for academic careers (Xie & Shauman, 1998), and the very long lag of women in senior author positions even after parity has been reached in the entry to scientific research in fields as large as clinical medicine (Jagsi et al., 2006). The literature leaves no doubt about differences in both the productivity and citedness between men and women in academia, although there are quite different results around the direction, magnitude, and underlying mechanisms of the difference (e.g., Andersen, Schneider et al., 2019; Caplar, Tacchella, & Birrer, 2017; Larivière, Vignola-Gagné et al., 2011; Lerchenmüller et al., 2018; Nielsen, 2016, 2017; Pagel & Hudetz, 2011; Thelwall, 2020). Some of these studies are also good examples of cases where author position is used to assume additional importance of the first and sometimes last author (Andersen et al., 2019; Jagsi et al., 2006; Lerchenmüller et al., 2018; Thelwall, 2020).

3. DATA AND METHODS

The analyses presented in this article are based on bibliographic data from the WoS, specifically through the in-house implementation at CWTS, Leiden University, which is a structured, relational database implementation of WoS. WoS is one of the largest bibliographical databases of scientific research, including citation indices. The major parts of the analyses are limited to publications classified as journal articles published between 2000 and 2020 ($n = $ 24,715,351), except for the analysis of document types, which uses all documents in WoS, for the same period ($n = $ 37,806,737). The analyses are based on data from the WoS citation indices: Arts & Humanities Citation Index, Book Citation Index—Social Sciences & Humanities, Book Citation Index—Science, Current Chemical Reactions, Emerging Sources Citation Index, Index Chemiculus, Conference Proceedings Citation Index—Social Sciences & Humanities, Conference Proceedings Citation Index—Science, Science Citation Index Expanded, and the Social Sciences Citation Index. While some of these indices include very few journal articles, they all provide reference data.

In addition to article metadata about the publication type, reference list, journal, publication year, page count, and author list, we also require an appropriate field classification and disambiguated author sets, in order to know their academic age and productivity as well as their inferred gender. These variables are considerably more difficult to collect than the metadata, which is why I will describe the process thereof below.

3.1. Author Disambiguation

The process of correctly assigning all the publications a scientist has written to a set symbolizing that scientist's oeuvre is difficult. Sometimes scientists will change their name, or use, for

example, middle names interchangeably, and quite often more than one person has the same name. Basing such an assignment purely on the author's name will create multiple sets for each name variant and group publications by different authors using the same name. More advanced algorithms have been developed, using additional evidence, such as author affiliations, coauthor network, keywords of their articles, bibliographic coupling, and cocitations to improve both the recall (the number of an author's publications assigned to one set) and the precision (the share of publications assigned to one set that are actually authored by this author) of disambiguated author sets. Through the CWTS implementation of WoS, there is access to the results of the author disambiguation algorithm by D'Angelo and van Eck (2020), which reports 93.7% precision and 94.1% recall. This is higher than the earlier version of the algorithm (Caron & van Eck, 2014), which was found to be the best performing disambiguation algorithm by Tekles and Bornmann (2020).

The errors that do arise in this disambiguation are almost entirely through creation of singletons (i.e., author sets with just one publication) due to a change in an author's record (e.g., a single article published with a different affiliation and with a different group of coauthors). In addition to providing substantially more precise author sets than the raw WoS data would, this approach also gives us data about the affiliation and first name of an author prior to 2008, which is when WoS started systematically registering affiliations of authors.

3.2. Alphabetization

I use the alphabetization approach of Waltman (2012), where all papers with two or more coauthors are first assigned a value, $a_i = 1$ if their byline is alphabetically ordered and $a_i = 0$ if not. As shorter bylines especially may be ordered alphabetically by chance, or as a result of contribution-based ordering, Waltman (2012) calculates the probability of intentional alphabetical ordering. For this purpose, Waltman defines a set of N publications, where publication N_i has n_i authors. Furthermore, p_i is the probability that a potential alphabetical ordering is intentional, which is an estimate based on the number of authors. Waltman defines this estimator, $\hat{p}_i$ as

$$\hat{p}_i = \frac{a_i - \left(1/n_i!\right)}{1 - \left(1/n_i!\right)}$$

and the overall probability of N using intentional alphabetical ordering, $\hat{p}$, thus is

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N}\hat{p}_i$$

The ratio of papers with alphabetical ordering, $\bar{a}$, is simply

$$\bar{a} = \frac{1}{N}\sum_{i=1}^{N}a_i$$

and can be observed directly from the bibliographic metadata. We expect that fields with high degrees of intentional ordering will have distributions of $\hat{p}$ over time close to the distribution of $\bar{a}$; however, for fields with norms for limited coauthorship, there may still be a difference between the values, as intentionality is difficult to estimate in publications with few authors.

I use a strict ordering, based on surnames as they are reported, which could potentially introduce some errors in cases where authors use, for example, dual or compound surnames such as "Van der Waal." Obviously, this may diverge, in either direction, from the authors'

intention; however, as we are not interested in minor differences but broad field trends, this is an acceptable error.

3.3. Author Productivity

This article reports author productivity at the point of publication (i.e., how many other articles each author in the byline has written when a given article is published). This is measured on a per-publication basis as the number of publications by the author up until the year in which it was published plus a share of the publications published in the same year, as exact publication dates are not available for all publications. This fraction avoids counting the same article multiple times while also allowing for the counting of publications from the same year. We can express this productivity ρ of a given author in year y as

$$\rho_y = \sum_{i=y_0}^{y-1} N_i + \frac{1}{N_y}$$

where N_i is the number of articles coauthored by this author in year i. As we are looking at within-field differences between first and last author, publication counts are not fractionalized. There may be a potential bias here, as more senior scientists (with higher productivity) are potentially more likely to join larger collaborations, regardless of field.

3.4. Gender Inference

When inferring the gender of an author, it is important to note that designations of "man" and "woman" are probability-based estimates based on population statistics, and not necessarily correct assumptions at the individual level. This is also why name-based gender inference does not consider the biological sex of the person, or other genders than the two most common. The inferred genders used in this study are from the same data source as the Leiden Ranking[1], which uses a combination of genderize.io[2] and gender-API[3] to give the best estimate with a combination of first name and country (Boekhout, van der Weijden, & Waltman, 2021). Some names are not possible to infer gender from, because they are very rare (insufficient statistical material), gender ambiguous (between-country ambiguity can be resolved; however, within-country ambiguity is problematic) or from countries with naming traditions that do not typically use gendered personal names (e.g., China and South Korea). In these cases, the gender of an author is considered "unknown." For all countries, the gender can be inferred reliably for 70.5% of all authors. The majority of unidentified authors are from China, meaning that analyses of gender do not apply to this country. Previous studies have established very high representativity and reliability for other countries (e.g., Madsen, Nielsen et al., 2022; Nielsen, Andersen et al., 2017). The same research also established that a manually estimated distribution of men and women in the "unknown" category was comparable to that of the inferred genders (Nielsen et al., 2017; Santamaría & Mihaljević, 2018).

3.5. Field Classifications

This article uses the OECD fields of science classification, as defined in the OECD Revised Fields of Science and Technology (OECD, 2007). The assignment of papers to fields is achieved through the WoS journal subject category to OECD field of science translation table

---

[1] https://www.leidenranking.com/information/indicators\#gender-indicators.
[2] https://genderize.io/.
[3] https://gender-api.com/.

provided by Clarivate (2012). This classification operates on two levels, where the top level covers the major fields natural sciences, technical sciences and engineering, medical sciences, agricultural sciences, social sciences, and humanities and art. On the second level, 39 minor fields are arranged below the six major fields. For a full overview of the included fields, and the number of publications registered with each, see Table 1.

Table 1. OECD major and minor fields, and number of publications in Web of Science, 2000–2020. The following fields are intentionally omitted in the Clarivate InCites conversion schema for WoS subject categories to OECD field: 4.04 Agricultural Biotechnology, 3.04 Health Biotechnology, and 3.04 Other Medical Sciences

1 Natural sciences	n	%	4 Agricultural sciences	n	%
1.1 Mathematics	885,407	3.6	4.1 Agriculture. forestry. fisheries	357,027	1.5
1.2 Computer and information sciences	519,595	2.1	4.2 Animal and dairy science	104,679	0.4
1.3 Physical sciences and astronomy	24,165,34	9.9	4.3 Veterinary science	185,062	0.8
1.4 Chemical sciences	2,775,960	11.4	4.5 Other agricultural science	114,125	0.5
1.5 Earth and related Environmental sciences	1,298,582	5.3			
1.6 Biological sciences	2,881,592	11.8	**5 Social sciences**	**n**	**%**
1.7 Other natural sciences	631,035	2.6	5.1 Psychology	414,749	1.7
			5.2 Economics and business	517,731	2.1
2 Engineering and technology	**n**	**%**	5.3 Educational sciences	139,983	0.6
2.1 Civil engineering	170,404	0.7	5.4 Sociology	165,219	0.7
2.2 Electrical engineering. Electronic engineering. Information engineering	767,350	3.2	5.5 Law	73,688	0.3
2.3 Mechanical engineering	125,121	0.5	5.6 Political science	102,061	0.4
2.4 Chemical engineering	213,997	0.9	5.7 Social and economic geography	135,656	0.6
2.5 Materials engineering	920,869	3.8	5.8 Media and communication	64,381	0.3
2.6 Medical engineering	125,121	0.5	5.9 Other social sciences	60,074	0.2
2.7 Environmental engineering	284,952	1.2			
2.8 Environmental biotechnology	132,160	0.5	**6 Humanities**	**n**	**%**
2.9 Industrial biotechnology	41,579	0.2	6.1 History and archaeology	112,954	0.5
2.10 Nano-technology	14,733	0.1	6.2 Languages and literature	147,673	0.6
2.11 Other engineering and technologies	338,819	1.4	6.3 Philosophy. ethics and religion	92,115	0.4
			6.4 Art	44,106	0.2
3 Medical and Health sciences	**n**	**%**	6.5 Other humanities	25,727	0.1
3.1 Basic medical research	1,590,483	6.5			
3.2 Clinical medicine	4,306,049	17.7			
3.3 Health sciences	1,031,921	4.2			

There are many problems associated with journal-based classification systems, such as field delineation and correct assignment of inter-, trans-, and multidisciplinary research published in monodisciplinary journals. In this case, I argue that the broadness and wide acceptance of the field definitions in the OECD system provides a helpful overview fitting of the aim of the article. However, we must be aware that the broadness of the fields also hides some interfield differences in characteristics. This is the case in psychology, for example, which is a field that historically as well as currently encompasses great methodological and epistemological diversity, spanning from purely clinical psychology (closer to clinical medicine) to theoretical psychology (closer to the humanities) and social and behavioral psychology (closer to sociology). With that in mind, the classification schema is still found valid, as the high level of the OECD fields serves to give the kind of overview required for the present analysis, and the high aggregation level and number of publications reduce problems of misassignment to mostly random errors. It will, however, be important to keep especially interfield differences in mind for future research.

4. RESULTS

In the following we report the descriptive results on a per-variable basis, first for article, then for author characteristics. This section concludes with a principal component analysis of the intervariable variance, and a reclustering of the OECD fields of science based on their bibliographic characteristics.

4.1. Article Characteristics

4.1.1. Document types

WoS utilizes a large number of document types. However, many of them have very low frequency. In our data, 35 document types were identified, with the first six document types

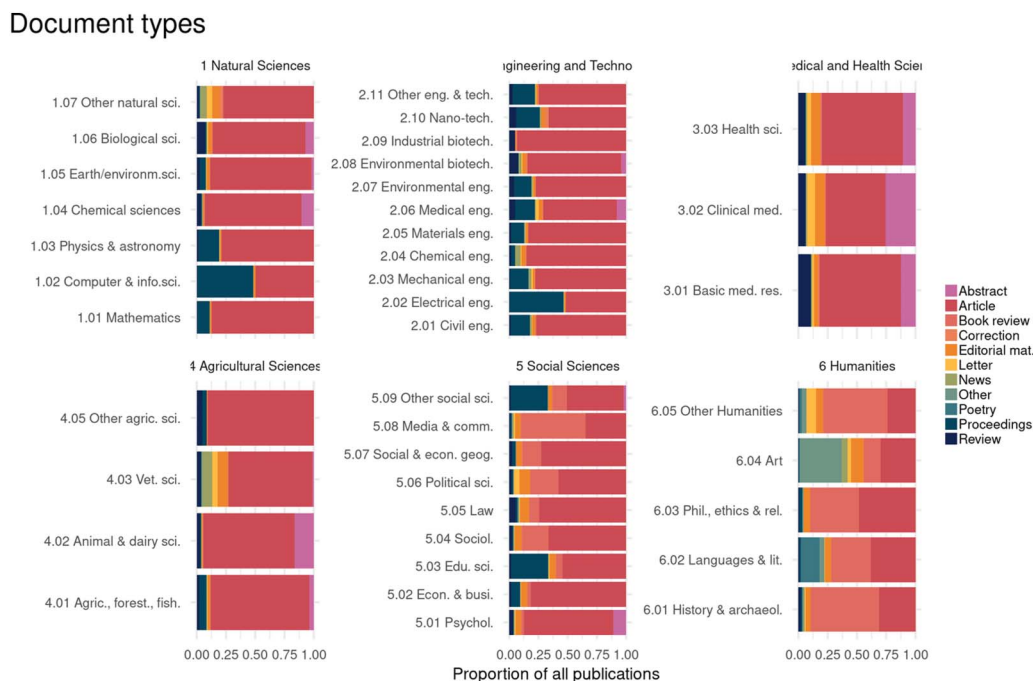Document types

Figure 1. Proportion of document types, as classified by Web of Science, per OECD minor field of science, shown in panels per OECD major field of science. The "Other" category covers less than 99% of all publications in total.

accounting for more than 95% of all publications, and the first 10 accounting for more than 99%. As can be seen in Figure 1, almost all documents in the remaining 1% of publications (with 25 categories) are found in art, where this category ("Other") covers 36.1% of all publications. In the natural sciences, agricultural sciences, and medical and health sciences, as well as engineering and technology, journal articles are the most frequent document types, but also abstracts and proceedings are common, especially in computer and information science, some engineering fields (especially electrical engineering), and education science. In the humanities, but also media and communication and to some extent sociology, book reviews play an important role (Zuccala & van Leeuwen, 2011). The exception to this is art, as mentioned above. In the medical and health sciences, and especially in clinical research, reviews are generally more present than in the other disciplines, although single fields (chemical, biological, and animal & dairy sciences, as well as psychology and medical engineering and environmental biotechnology) also have some degree of secondary literature.

4.1.2. Reference age

In Figure 2, we see the proportion of references in all papers of a given age up to 40 years, where the age is the difference in publication years between the citing and the cited
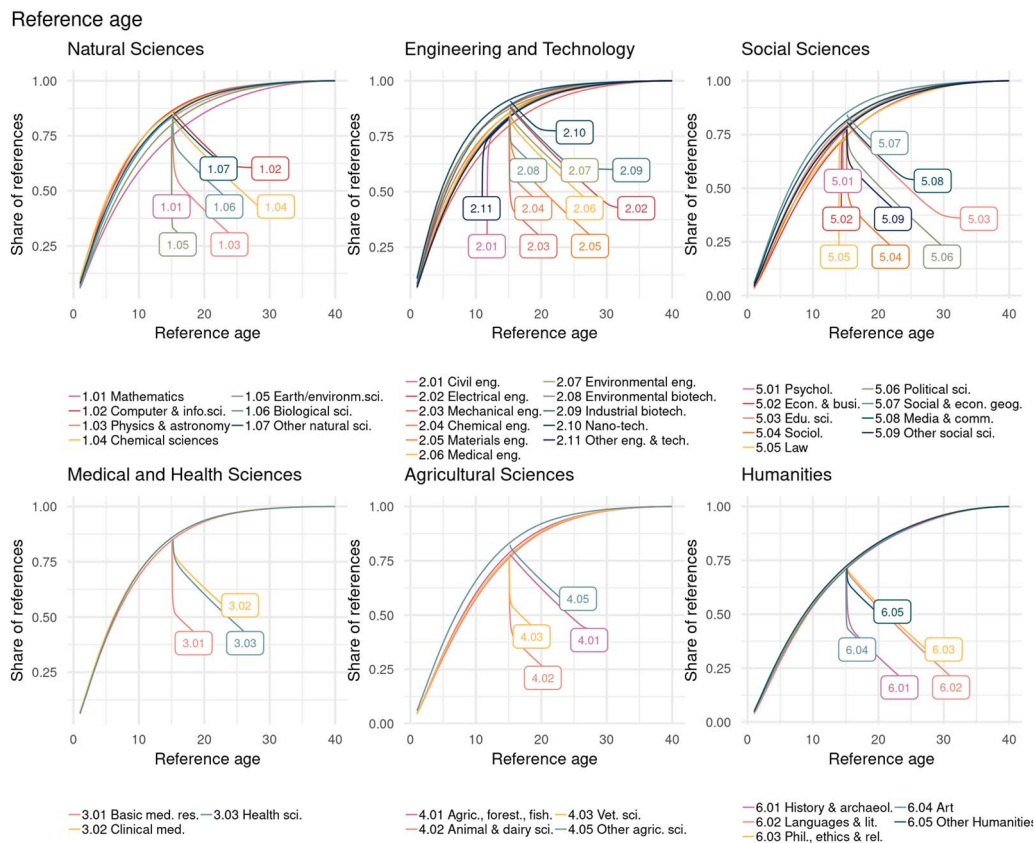

Figure 2. Share of references by reference age (difference between publication years of citing and cited document) for all OECD minor fields of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in legend.

publication. In the figure, marker lines from the field codes (colored labels) to the curves all meet the curves at 15 years reference age (arbitrarily chosen value for visual aid), highlighting the field differences in these curves. In all disciplines except for health science and the humanities, there are considerable within-discipline differences. The humanities also have considerably longer reference age curves than the other disciplines, and fields such as mathematics and economics are closer to the fields in the humanities in this regard. The median reference age, or cited half-life, can be read as the reference age corresponding to the 0.5 proportion of references, and is shown over time in Figure 3.

The variation between median reference age is large within all disciplines, except medical and health sciences and the humanities. Variations between fields are even greater and growing for almost all fields. Fields in engineering appear most stable over time, some even with declines in median reference age in the latter part of the publication period (e.g., mechanical engineering and materials engineering). In the natural sciences, computer science is also stable throughout most of the period with a decline by the end, meaning computer scientists tend to reference newer material today than earlier, which is likely explained by rapid growth in this field. In particular, the humanities have long and growing median reference ages. The high growth in this discipline is potentially an artificial effect of WoS coverage of this area in particular.
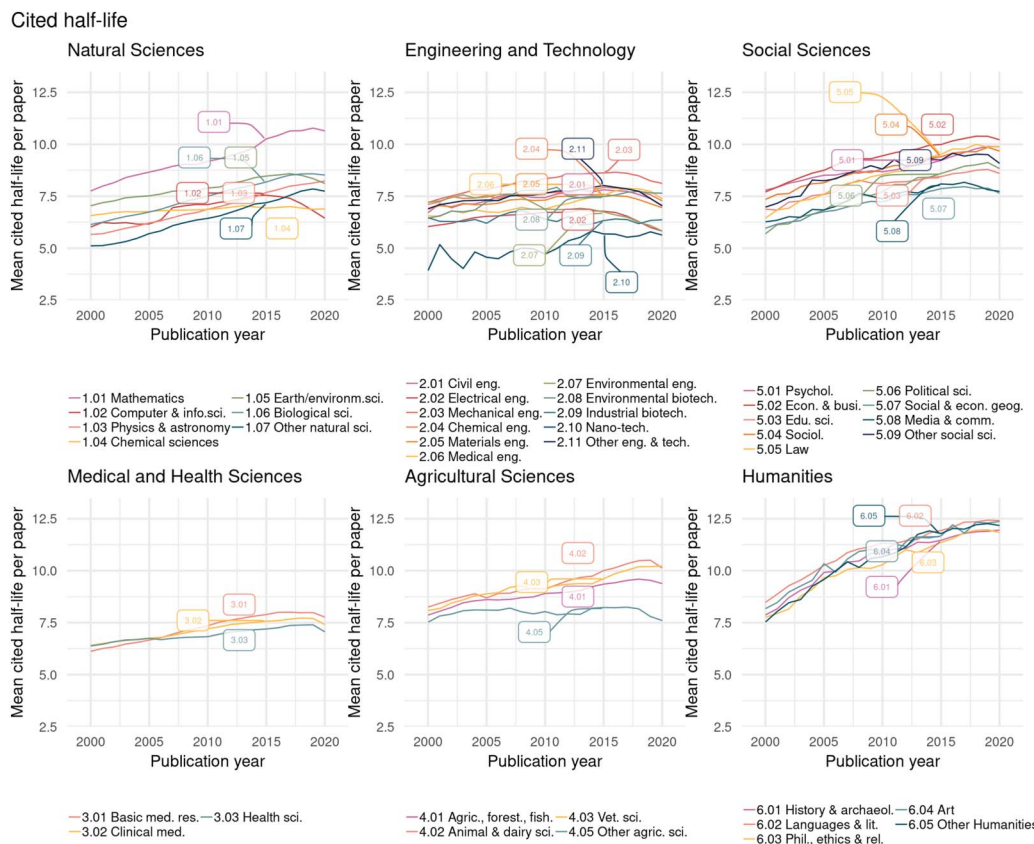
Figure 3. Annual mean value of the cited half-life or median reference age (difference between publication years of citing and cited document), per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in legend.

4.1.3. Reference list length and coverage

The length of reference lists per field is shown in Figure 4, illustrating the differences both between and within fields, as well as the general growth in reference list length. Fields in the medical and health sciences generally grow more slowly than other fields, and fields in the social sciences (as well as history and archaeology) have the longest reference lists. The longest reference lists (mean over the entire period, σ_r) are found in history and archaeology ($\sigma_r = 50.8$) which is 2.5 times longer than the shortest in electrical engineering and mathematics ($\sigma_r = 20.4$). Journal guidelines may play a role in limiting reference list lengths in some cases (e.g., some medical journals have traditionally limited the number of references that could be included).

When taking into account which references refer to works included in WoS, the distributions change fundamentally, as seen in Figure 5. Where the medical and health sciences were on the low range of the number of references, almost all of these are covered in both basic and clinical medical research. All fields in the humanities have very low coverage. All fields here are below mean coverage, $\sigma_c = 25\%$ and as low as 13.4% (art). In the social sciences and economics & business, as well as psychology are considerably better covered than the other fields, which are all covered between 25% and 50%. In both the natural sciences and the technical and engineering sciences, there is great within-discipline variation, with well-covered fields such as industrial biotech ($\sigma_c = 84.4\%$), nanotechnology ($\sigma_c = 80.2\%$),
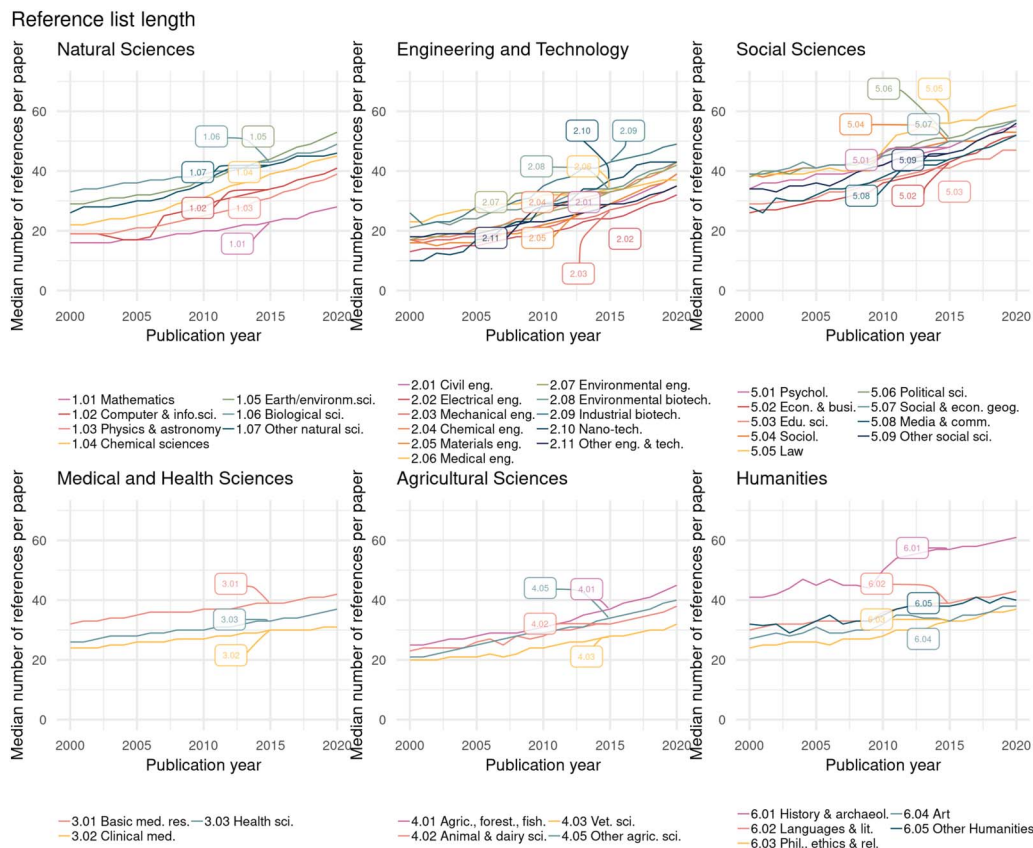

Figure 4. Median number of references per article over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in the legend.
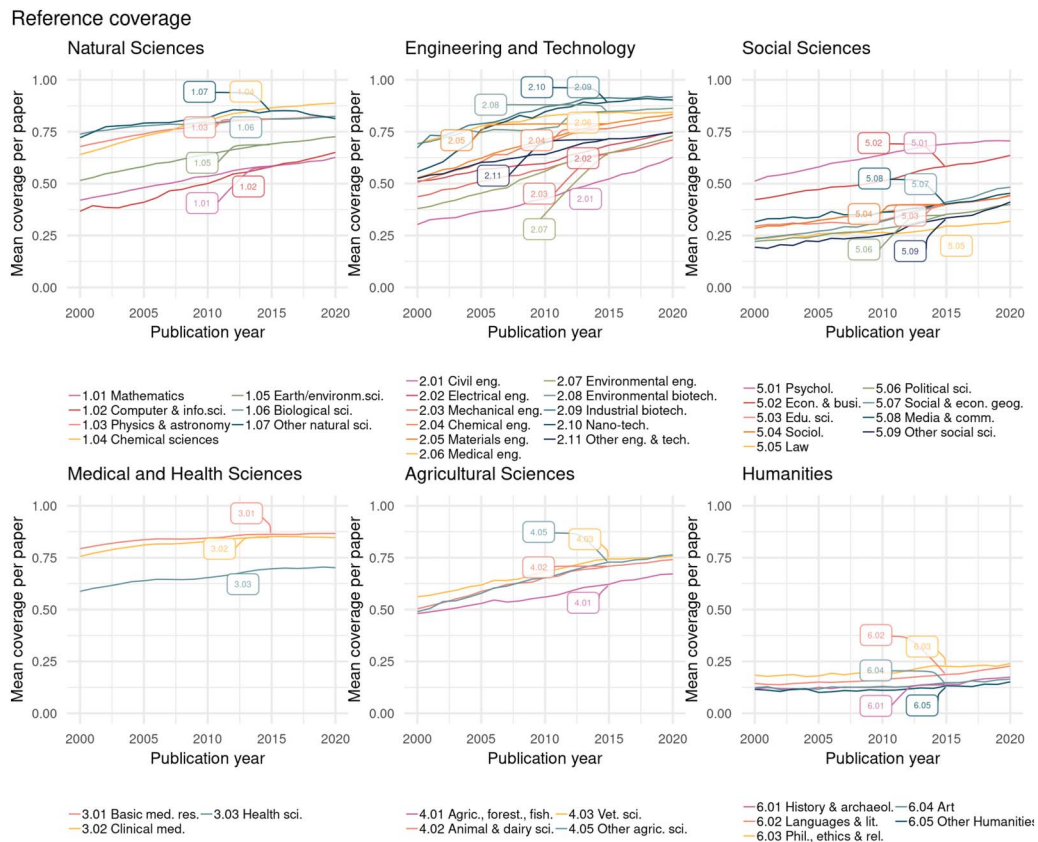
Figure 5. Mean percentage of references in reference lists identifiable in Web of Science over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in legend.

biological sciences (σ_c = 79.1%), and chemical sciences (σ_c = 79.0%), and much less covered fields such as civil engineering (σ_c = 44.5%) and computer & information science (σ_c = 50.4%). Most fields have growing coverage over the time period, except for the humanities, where fields either remain low in coverage or grow very slowly (e.g., languages & literature, and philosophy, ethics, & religion).

4.1.4. Article length

One fundamental difference between publications from different fields, even though they are here all considered original research articles in journals, is the typical length of an article. As illustrated in Figure 6, both medical sciences and agricultural sciences, and somewhat also engineering and technology, are on the low end of the scale, with 5–10 pages per publication (although increasing to a little more than 10 for some fields in recent years), while the natural sciences spread over the middle range with 5–20 pages per publication, and mathematics and computer science have by far the longest articles in the discipline. The humanities generally produce the longest articles, quite closely concentrated around 20 pages, while the social sciences fall between 10 and 20 pages, although law stands out as having by far the longest articles, at 30 pages on average. Both the mean (solid lines) and median (dotted lines) are shown in the figure, illustrating that for most fields, these two statistics are very close, while for some (e.g., mathematics and law), a tail of very long articles creates a mean somewhat higher than the median.

Figure 6. Mean (solid lines) and median (dashed lines) number of pages per article over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in the legend.
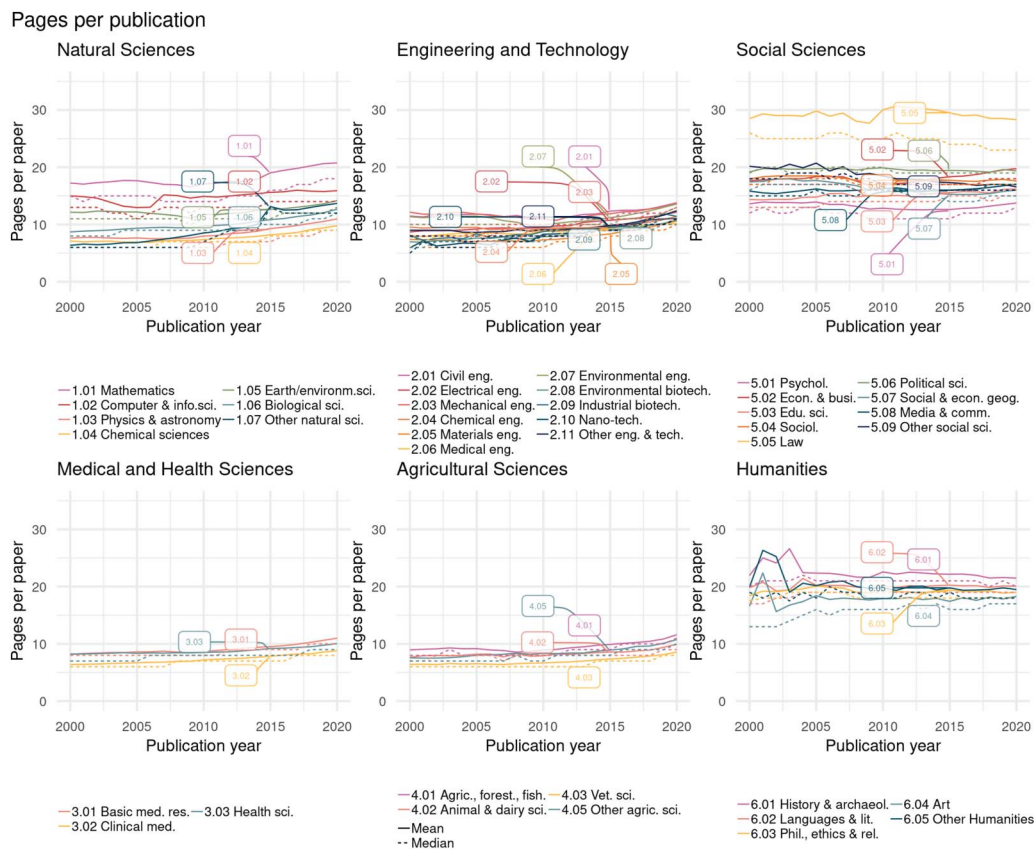
4.1.5. Reference density

Both the length of an article and the number of references can be seen as potential symbols of the amount of intellectual work that has gone into an article. While there is certainly no direct causality between these two indicators and the quality or intellectual value of an article, the combination of references per article page (reference density) can indicate whether articles in a given field are more information dense than others, assuming that increased information per article requires further references to underline arguments and give credit to prior claims, etc. Figure 7 shows the changes over time in average number of references per page across OECD fields of science, and the trends must of course be seen in relation to trends in article length (Figure 6) and reference list length (Figure 4), of which this is a ratio. We see that fields in the agricultural, medical and health sciences, and humanities, as well as mathematics and biological sciences, have stable reference density or only very small changes over time, but also very different baseline densities (from around one reference per page in mathematics to two in the humanities and four in the agricultural, biological, and medical and health sciences). It is unclear why the category "other natural sciences" decreases over time, except that a corresponding growth in article length occurs concurrently. All other fields have rising trends, with a growth of more than one reference per page over the period. A considerable part of this change is likely explained by expansive growth in these fields, providing more relevant references to include.

Figure 7. Mean number of references per page per article over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in legend.

The interpretation of these values is not obvious. However, with differences of this magnitude, there is no doubt that there are quite different norms for how to write articles and how to use references across fields, with direct consequences for bibliometric evaluation. This raises the question of whether comparisons between the most different fields, despite normalizing citations, for example, are meaningful.

4.1.6. Coauthorship

As noted in Section 2, coauthorship is one of those variables that have seen considerable change over time. This is confirmed in Figure 8, illustrating general growth in both the mean (solid lines) and median (dotted lines) number of coauthors per article in all fields, except for mathematics and the humanities (which may appear to see a small growth in the most recent years only). While the bump observed in 2010 for "other engineering and technology" is likely an outlier, the extreme change in mean number of coauthors occurring in physics and astronomy from 2011 and onward is not by chance, but quite stable up until 2020. These may still to some degree be considered outliers, as the publications underlying the change are relatively few articles with several hundred, sometimes more than a thousand, coauthors. These articles mainly stem from the CERN high-energy particle experiments, but there are also some astrophysics papers. Computer science and mathematics are the only natural science fields with less than five coauthors per paper on average in the most recent observations, and the figure

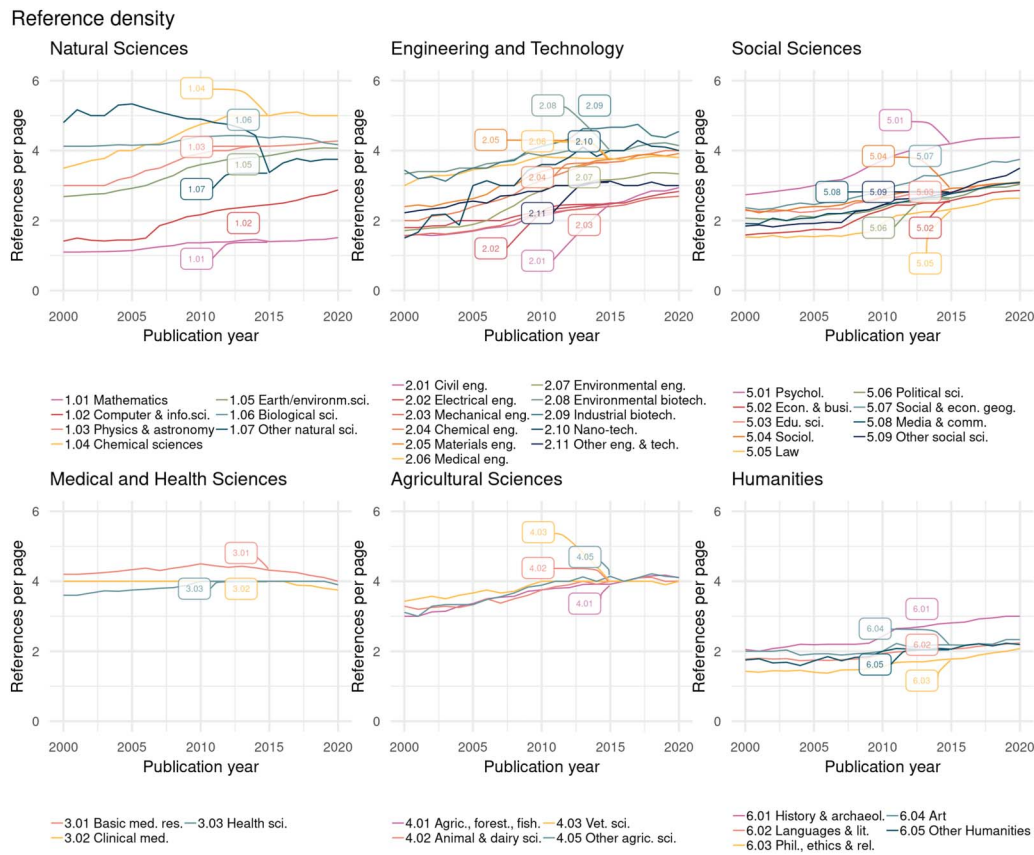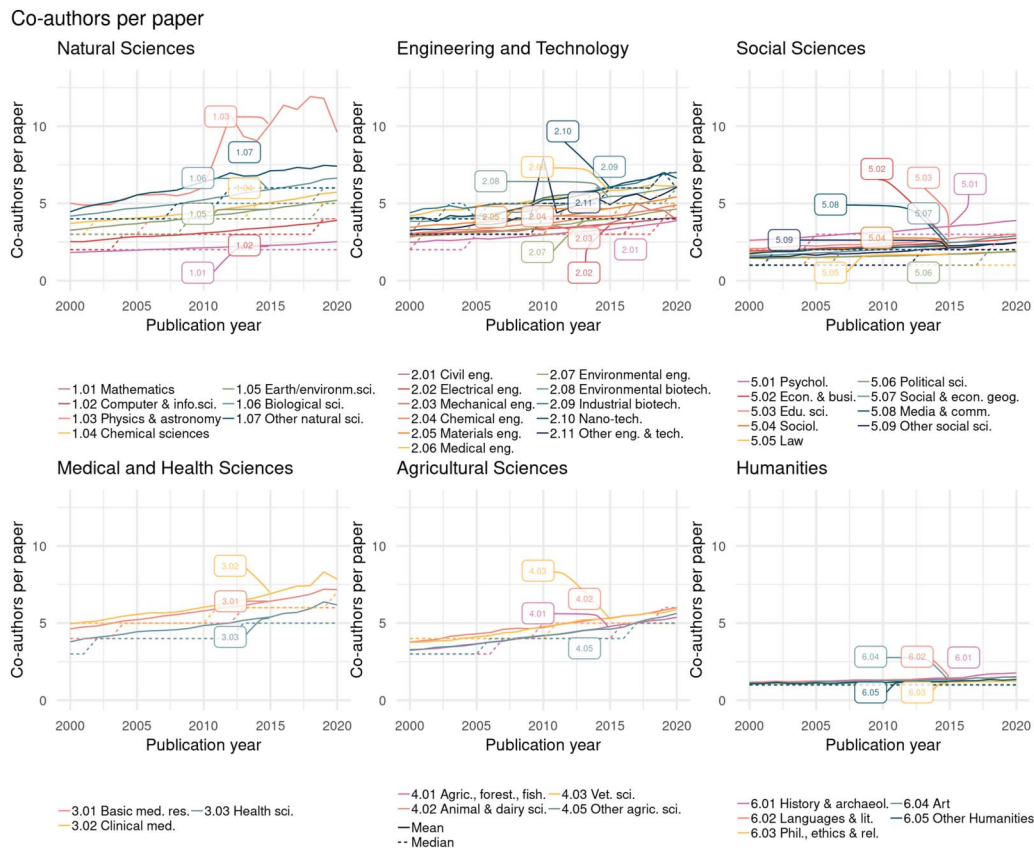**Figure 8.** Mean (solid lines) and median (dashed lines) number of coauthors per article over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in the legend.

is particularly low for mathematics. Fields in the social sciences and humanities are also much less collaborative than those in other disciplines. While tradition may play a role in these differences, it seems more likely that the growth across the hard sciences is a result of new requirements for data collection and generation that require larger teams (big science), clearly illustrated by the bump in coauthors for physics. While parts of the social sciences are taking steps in the same direction, the data show there are still many studies conducted by individuals or small teams.

4.2. Author Characteristics

In this section, results on the author characteristics of fields are reported. All analyses shown here are calculated on information based on the disambiguated author data described in the methods, and thus rely on a larger publication data set than used in the article characteristics section. The analyses are still done on the same sample, but they use information on, for example, total number of papers per author at the time of publishing a given paper, or an author's first year of publication.

4.2.1. Alphabetization

In Figure 9 we see the rate of alphabetical ordering (solid lines) and the probability of intentional alphabetical ordering (dashed lines) according to the methodology proposed by
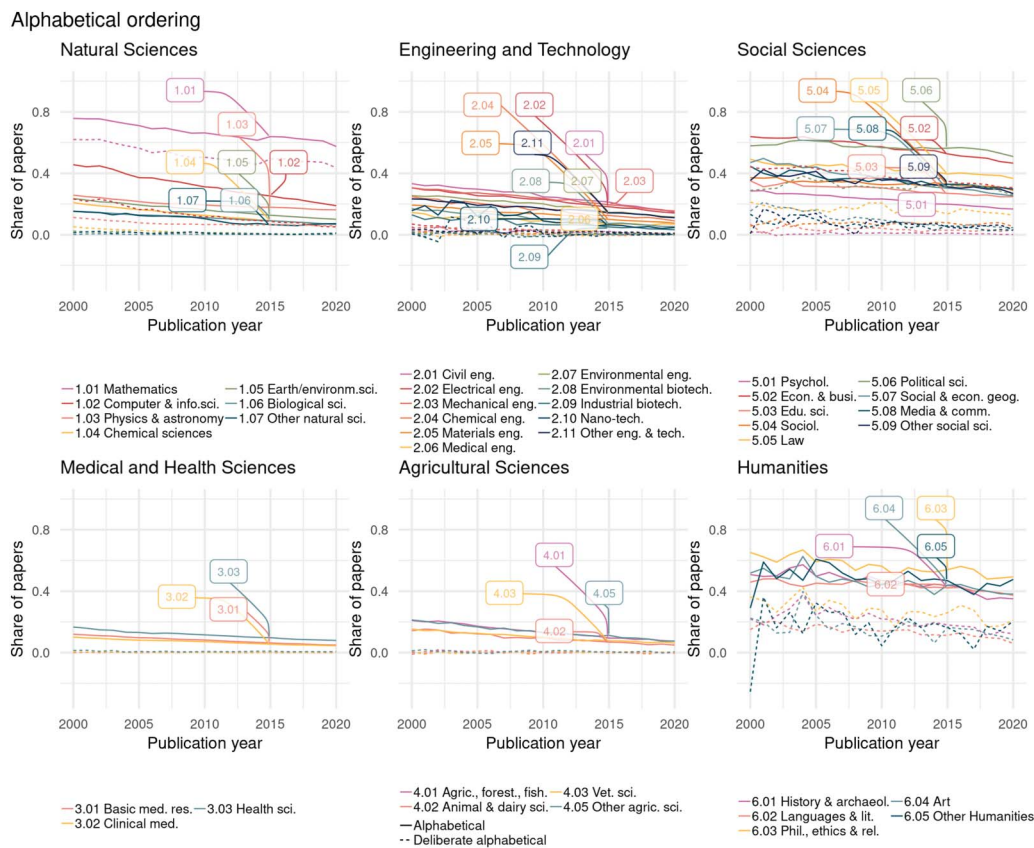
Figure 9. Share of papers with alphabetical (solid lines) and probability of intentionally alphabetical (dashed lines) author sequence order over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in the legend.

Waltman (2012), across OECD fields of science. The analysis only includes articles with two or more authors. As noted in the methods, fields with low collaboration may have somewhat high differences between the two curves, even though the rate of alphabetical ordering is high. This is clearly the example in mathematics, which goes from 80% to 60% alphabetical ordering, the highest of all fields, but is also one of the fields with the lowest number of coauthors, making it difficult to estimate statistically whether the ordering is intentional, however, with values as high as here, it is safe to assume that alphabetical ordering is a common but not exclusive form of attributing authorship in mathematics. Similar statements can be made for computer and information science, although the tendency to use alphabetical ordering is lower and decreasing more over time. This decrease is found across most fields, but less so in the social sciences and humanities. General trends and field differences for fields are similar to those in Waltman (2012) and Mongeon et al. (2017), confirming previous research on the topic.

4.2.2. Gender distribution

Figure 10 shows the proportion of authors with reliably inferred gender who are estimated to be women, across all fields. The proportion reflects not headcount but proportion of authorships in a given year attributable to women. Very few fields have more than half their
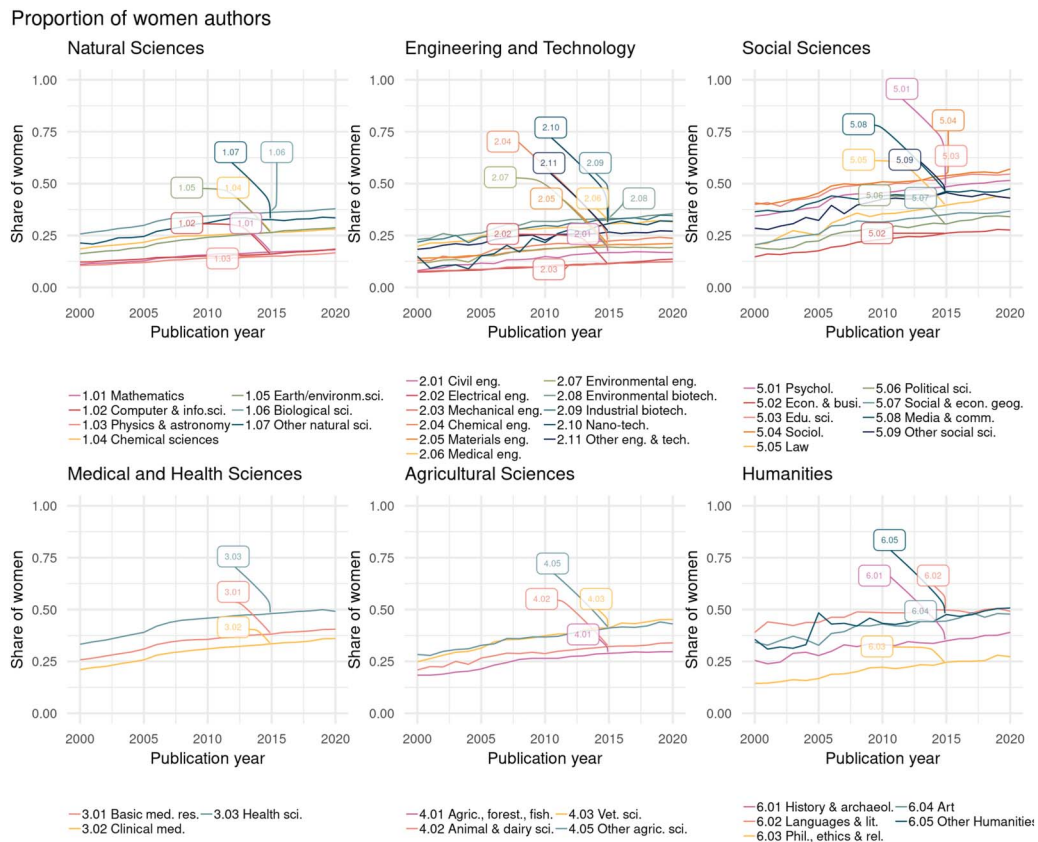
Figure 10. Proportion of women as authors per field over time, per OECD minor field of science, shown in panels per OECD major field of science. Numbers in boxes correspond to field codes in the legend.

authorships from women, but most of the humanities and some social sciences, as well as health science, reach parity at the end of the period, and fields such as sociology and education science even have more female authorships than male. In engineering and technology and the natural and agricultural sciences, as well as in basic and clinical medical research, most fields have low representation of women at the beginning of the period, in some cases down to 10%, but with increasing shares over time. Fields such as mathematics, computer and information science, nanotechnology, and civil, mechanical, and electrical engineering have the lowest representation of women and also have very low growth. In the social sciences and humanities, the lowest representation is found in economics and business, social and economic geography, political science, and philosophy, ethics, and religion.

4.2.3. Author sequence

In the following three figures, data on the productivity (Figure 11), gender (Figure 12), and academic age (Figure 13) for first, last, and middle authors are reported. These three figures together inform us about further demographic differences and expectations around the meaning of authorship in academic fields. It also helps us understand which assumptions are reasonable to make about different author positions, and which fields it is meaningful to apply author weighting by sequence position in.
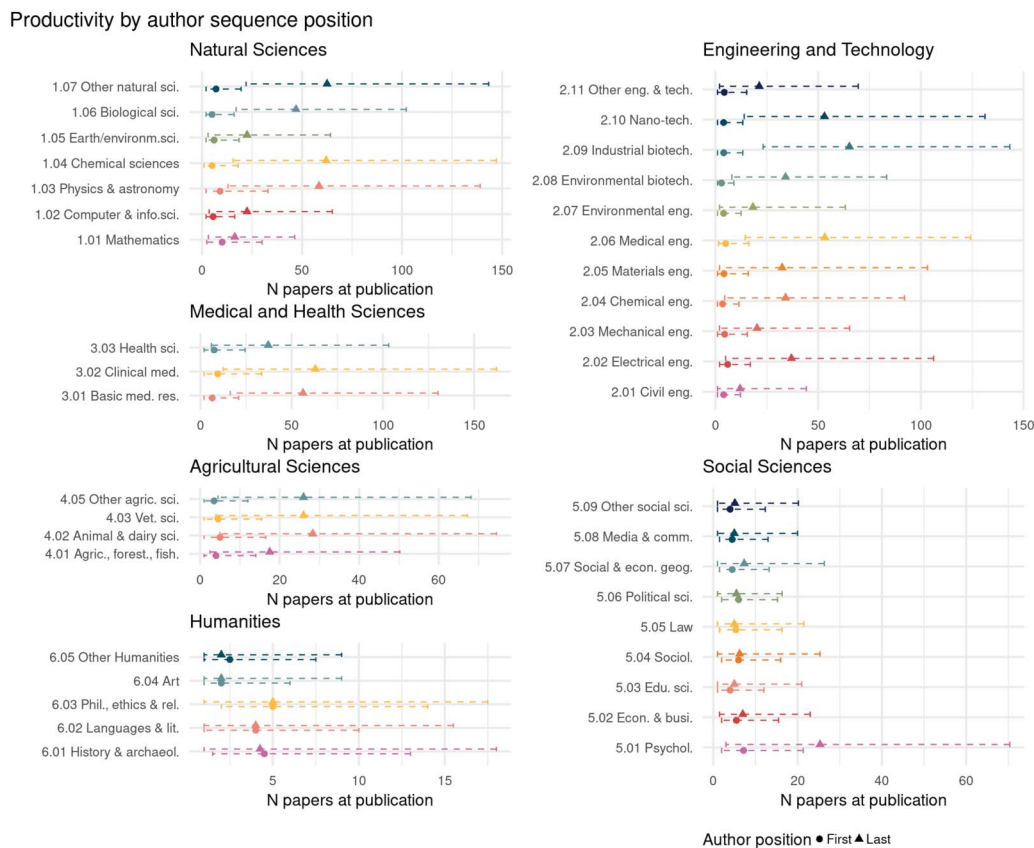
Figure 11. Median productivity at publication year (ρ_y) for first (circle) and last (triangle) authors. Dashed lines show first (low endpoint) and third (high endpoint) quartile productivity. Distributions represent the entire period 2000–2020, per OECD minor field of science, shown in panels per OECD major field of science.

With regard to productivity by author position, Figure 11 shows productivity variation between first and last authors, using the productivity indicator, ρ. We see from the figure that most fields outside of the social sciences and humanities have a wide variation between the two positions. Evaluating this figure should take into account that data points show median productivity and the low and high end points of the dashed bars show first and third quartile variance. Considering this, mathematics is the only field outside the social sciences and humanities that has no real difference between first and last author, although some other fields only have small differences in the median productivity (e.g., computer and earth sciences, as well as civil, mechanical and environmental engineering). In the social sciences and humanities, the only field with median productivity difference between author positions is psychology, although several fields have some difference in the third quartile. This is an interesting observation taken together with the trends in alphabetization, indicating that many of these fields have a dual modus for authorship interpretation, where at times the sequence is alphabetical, and when it is not, the last author is more likely to be more senior (higher productivity).

Figure 12 shows both the absolute and the proportional share of female authors by author sequence position. The absolute share is calculated in the same way as the distributions in Figure 10 and are of course dependent on how many women are active researchers in the

Proportion women by author sequence position
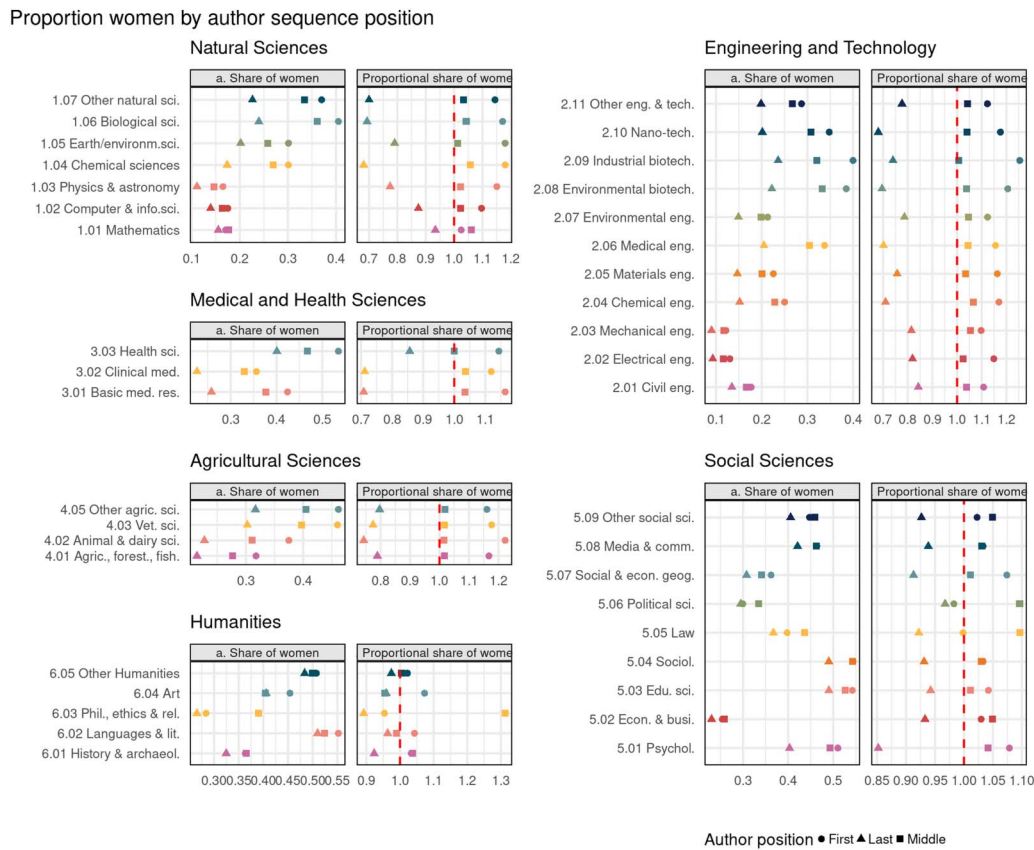
Figure 12. Percentage share and proportional share, with respect to overall distribution, of women authorships as first (circle), last (triangle), or middle (square) author. Proportional share is the ratio between women at a given position and the ratio of women overall. Distributions represent the entire period 2000–2020, per OECD minor field of science, shown in panels per OECD major field of science.

different fields. The proportional share, however, is the share at each author position compared to the overall share in that field. This proportional share indicates that there are differences in the roles and seniority of men and women across *all* fields. Even in fields that tend to use alphabetical ordering, there are more women in the first author position and fewer women in the last author position than expected, given the overall share of women. The only three exceptions to this are law, political science, and philosophy, ethics, and religion, where only the middle author position is overrepresented by women, while the first author position is at 1 (law) or below (political science and philosophy, ethics, and religion). Fields with higher degrees of alphabetical ordering tend to have more narrow proportional spread (values closer to 1).

In Figure 13, we see the academic age of authors, at the time of publishing a publication, as either first, last, or middle author. The academic age is calculated as the number of years between the first publication recorded for a specific author set and the current one. Thus, one individual author may be present several times in the data set, with varying ages over time. For each field, the circles represent the median age at publication for first authors, triangles represent last authors, and squares represent middle authors. These distributions correlate highly with those in Figure 11, as all fields outside of the social sciences and humanities have large differences between first and last author age, except for mathematics, although there is

Figure 13. Median academic age (difference between first publication year of an author and current publication year) for first (circle), last (triangle), and middle (square) authors per article. Lower and upper bounds of dashed lines show first and third quartile academic age. Distributions represent the entire period 2000–2020, per OECD minor field of science, shown in panels per OECD major field of science.

a small difference here too. The other fields with smaller productivity variation are also the fields with smaller age variation. Middle author age varies a lot between fields, indicating that there is less certainty about their roles compared to the first and last author roles; they may be both senior and early career scientists, although in most fields they tend to be closer in age to the first author. There are no differences between author position in the humanities, and in the social sciences the only large difference is found in psychology. Economics and business and social and economic geography also have small differences, but only of a few years.

4.3. Field Clustering

In the previous two sections, we have seen several bibliographical and demographic differences between fields, with variation both within and between disciplines. For some variables, individual fields have stood out, and some fields (e.g., mathematics) have consistently had different distributions than other fields in their discipline, while other disciplines have had very similar distributions between their fields. To explore if these differences can be systematically grouped, principal component analysis (PCA) is used to cluster variables.

For this purpose, we use the mean values of all variables except for those around gender. The gender variables were included above to inform researchers about expected populations prior to selecting fields to study, and not as a bibliometric variable or a scientific communication variable, as also described in the methods section. This leaves 11 variables, which are on very different scales. To compensate for this, PCA is performed in R, using the built-in prcomp function, with scale = TRUE to rescale variables. In the resulting list of components, the first two explain 78.9% of the variance between variables, as shown in Figure 14a. To recluster fields based on their composition, the Euclidean distance between vectors composed of the component loadings is calculated and used for hierarchical clustering. A dendrogram showing these clusters of fields can be seen in Figure 14b. The number of clusters identified from this approach is a subjective but informed choice, and tests were performed with other cluster numbers and compared to the underlying data. The first of these comparisons is through a biplot of the variables' loadings on the first two components, shown in Figure 15, using the original OECD major fields to group minor fields, and in Figure 16 showing the same fields but grouped by the clusters identified in the dendrogram.

Figure 15 shows the loadings on the first component on the *x*-axis and on the second component on the *y*-axis. The two factors together do not explain all the variance in the data, which is why the clusters shown in Figure 16 appear to diverge slightly from the clusters in the dendrogram. This is most clear for the pink-colored cluster containing political science, law, and social and economic geography (we will refer to this cluster as HumSoc, as it is the part of sociology closest to the humanities in this plot), which in the biplot appears to overlap with the teal SocSoc cluster (named as a social science middle ground between the humanities and hard sciences). This apparent overlap is due to the limitations in plotting just two components, even when they explain almost 80% of the variance.

Figure 14. Proportion of cumulative variance explained by principal components (panel a, left), and dendrogram of field clusters based on Euclidean distance between fields represented by component loadings.
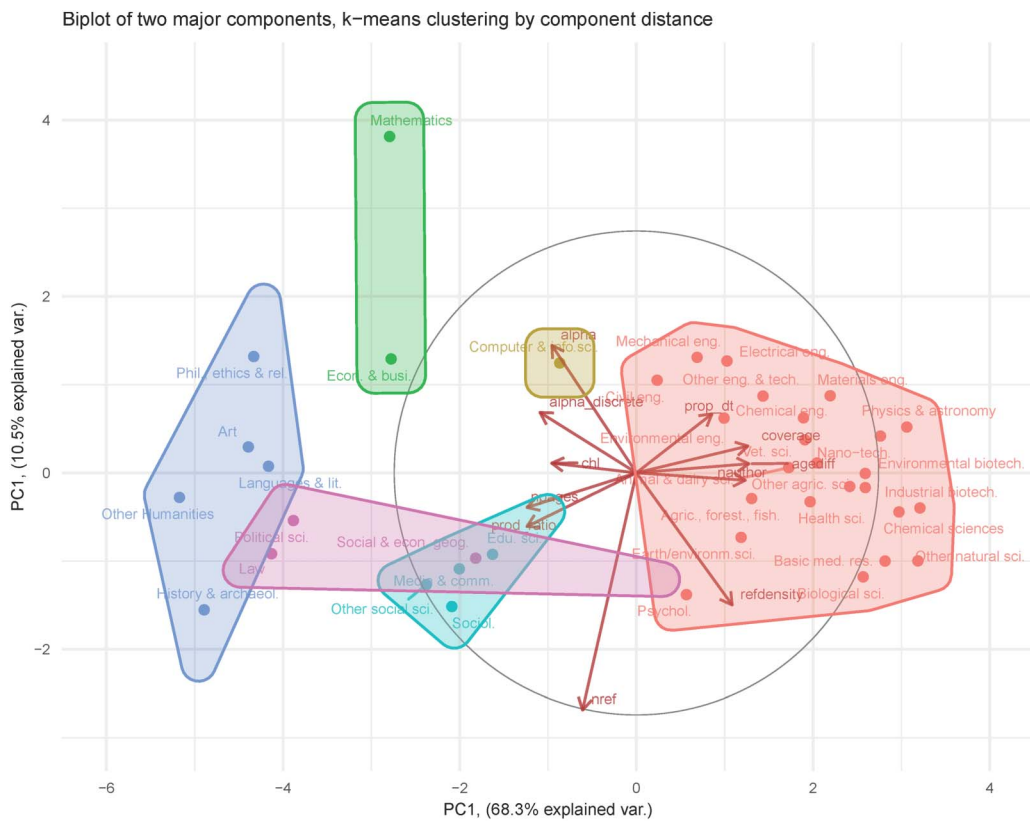
Figure 15. Distribution of fields based on loadings on the two components explaining the most variance (PC1 and PC2). Arrows show the relation between field characteristics and components. Hulls around fields represent the OECD major field groupings.

The other clusters quite neatly group the fields they consist of on the first two components, as shown by the colored hulls around the field point estimates. The humanities are unaffected by the regrouping, while most of the natural, medical, agricultural sciences, and engineering and technology are grouped into one cluster. Psychology is added to this cluster, due to aforementioned characteristics closer to those of clinical medicine. Computer and information science is sufficiently different from the other hard sciences to form a singleton cluster, and mathematics and economics form a small cluster on their own.

Figure 17 confirms the split of the social sciences, with the HumSoc and SocSoc clusters having quite different loadings on some of the components. While the hard sciences have a considerable amount of variance in the PC1 loadings, they are also the only cluster loading positively on this component, explaining why these fields are grouped despite the high variance. All in all, this shows us that a wide range of the hard sciences can be used for quantitative studies of bibliometric and sociological patterns of science, and that it is reasonable to use assumptions about author position, productivity, and reference norms in sample selection procedures. It also shows us that much more care must be used when working with fields outside of the hard sciences, and that these assumptions do not hold here. Rather, one should assume lower coverage, less meaning in author sequence position, fewer coauthors, longer median reference age, and a lower share of journal articles.

Figure 16. Distribution of fields based on loadings on the two components explaining the most variance (PC1 and PC2). Arrows show the relation between field characteristics and components. Hulls around fields represent the clusters generated from component loadings.
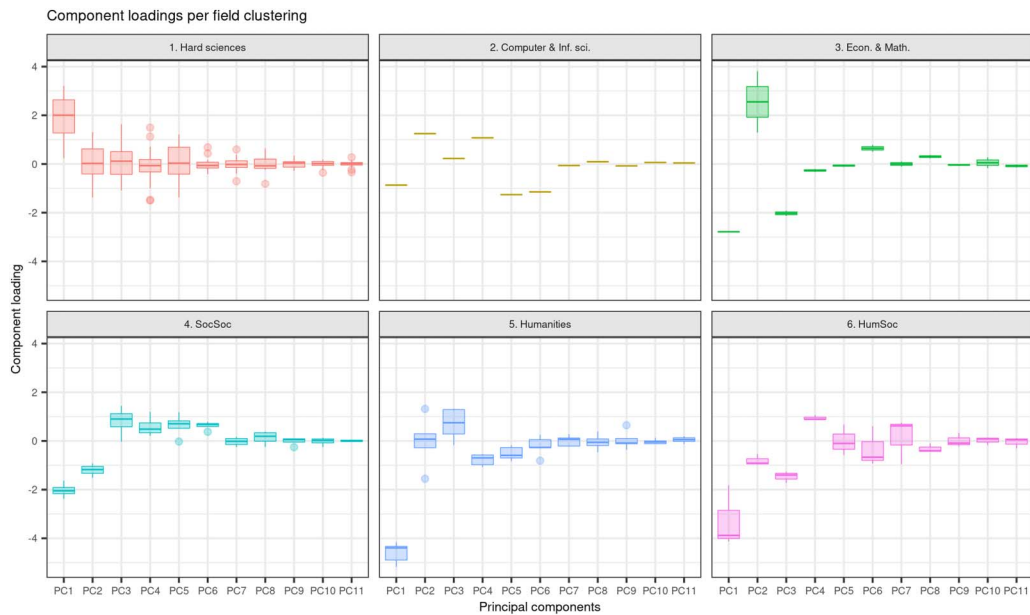

Figure 17. Distribution of component loadings per cluster as Tukey-boxplot.

5. DISCUSSION

The descriptive analyses of article and author characteristics in this article confirm previous research on individual fields and single measurements but bring them together in a comprehensive overview. The results show that from a bibliometric and sociology of science point of view, some fields in the natural and social sciences cannot be compared to other fields in those disciplines. This is particularly the case for mathematics, economics, computer science, and psychology. The first three of these are very different from all other fields in general, perhaps having most in common with each other, whereas psychology tends to look more like clinical medicine than the other social sciences in terms of how they produce knowledge.

The data shown in this paper provide an overview of the limitations and possibilities for bibliometric research in specific fields, and the reclustering of fields offers insights into which fields can reasonably be grouped together in bibliometric and quantitative sociological analyses of the scientific communication system. At the same time, it raises the question of whether it is meaningful to aggregate, for example, citation scores (also field-normalized and percentile-based ones) across fields that are as dissimilar as many of the fields are.

Summarizing the findings per characteristic, I find that

- Articles are the main publication type in all disciplines, except for the Humanities. Some fields use conference articles a lot, but most focus on journal articles.
- Cited half-life has large between-field variation, but also within-field variation in most disciplines. This is likely linked to how cumulative the knowledge production modes of the fields are.
- Average reference list length varies by more than a factor of two between fields and grows considerably over time in some fields.
- WoS coverage of references in the reference list is very low in the humanities (10–20%) and low for many of the social sciences, as well as some technical and natural sciences.
- Article length varies much across all fields and stays quite stable over time.
- The number of references per page grows in some fields but is mostly stable. Humanities and most social sciences, as well as mathematics and some engineering fields, are least dense.
- The humanities have very few coauthors per paper. The social sciences are becomingly increasingly collaborative but do not have the same team size as other disciplines. Physics and astronomy have by far the highest mean number of coauthors per paper.
- Alphabetical author lists are common in mathematics and computer science, as well as most of the social sciences and all of the humanities; however, due to less collaboration in these fields, estimates are less precise.
- Many of the natural sciences and engineering fields have very few women scientists, while health science, some of the social sciences and parts of the humanities have reached parity.
- There is a strong connection between author position and the author's number of papers written in fields without alphabetical ordering, except in the humanities and social sciences (although psychology is more like medicine in this regard).
- Even in fields with parity, and in fields with alphabetical ordering (with few exceptions only), it is more common for women to be first authors (more likely to be early career) and men to be last authors (more likely to be senior).

Aggregating research to fields, based on journals, and on the levels used in this study, is meaningful for interpreting results, but also has limitations. One problem in this respect is

the incorrect classification of articles originating from one field but published in a journal belonging to another field. This is especially a problem on the microlevel, and less so on the macrolevel, as we can reasonably expect that only a low number of articles out of the entire population are erroneously classified. More substantially, some of the included fields cover a very broad range of methodologies and topics, and therefore also tend to communicate findings differently, depending on the subfield. One of the examples that can be seen is psychology, which includes social, behavioral, theoretical, and clinical psychology. Also, the trends in coauthorship in physics mask the theoretical physics subfield(s), with much different coauthorship patterns.

The analysis of author characteristics is, like other current large-scale analyses, limited by the quality of the author disambiguation and gender inference algorithms and data availability. For bibliometric researchers wanting to work with (near-)complete sets of publication and author data, the challenge of addressing these limitations in the data is of the same significance as addressing field differences in publishing norms. However, the solutions to these problems are not trivial, and should be a priority for future research. This becomes even more the case as China becomes an increasingly important science producer, as Chinese names are also among the most difficult to disambiguate and infer gender from, at least when using the pinyin transliteration seen in the majority of bibliographic databases (Sebo, 2021).

In summary, the results presented in this article reiterate concerns about both normalization techniques and their application to aggregates composed of very diverse disciplines, and the use of very large data sets for bibliometric analysis without controlling for field. Across all variables included here, at least some fields were fundamentally different from other fields, even within disciplines. This shows us that great care must be taken when trying to make universal claims about science and that perhaps we should strive less for universality in these types of studies and more for well-defined, delimited samples.

REFERENCES

Allison, P. D., & Stewart, J. A. (1974). Productivity differences among scientists: Evidence for accumulative advantage. *American Sociological Review*, *39*(4), 596–606. https://doi.org/10.2307/2094424

Andersen, J. P., Schneider, J. W., Jagsi, R., & Nielsen, M. W. (2019). Gender variations in citation distributions in medicine are very small and due to self-citation and journal prestige. *eLife*, *8*, e45374. https://doi.org/10.7554/eLife.45374, PubMed: 31305239

Andersen, J. P. (2023). De-identified article and author characteristics for a large data set of Web of Science (Version 1) [Data set]. *Zenodo.* https://doi.org/10.5281/zenodo.7573523

Boekhout, H., van der Weijden, I., & Waltman, L. (2021). Gender differences in scientific careers: A large-scale bibliometric analysis. *arXiv:2106.12624.* https://doi.org/10.48550/arXiv.2106.12624

Bourke, P., & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics,* 37(3), 473–494. https://doi.org/10.1007/BF02019259

Burrows, S., & Moore, M. (2011). Trends in authorship order in biomedical research publications. *Journal of Electronic Resources in Medical Libraries,* 8(2), 155–168. https://doi.org/10.1080/15424065.2011.576613

Caplar, N., Tacchella, S., & Birrer, S. (2017). Quantitative evaluation of gender bias in astronomical publications from citation counts. *Nature Astronomy,* 1(6), 0141. https://doi.org/10.1038/s41550-017-0141

Caron, E., & van Eck, N. J. (2014). Large scale author name disambiguation using rule-based scoring and clustering. In *Proceedings of the Science and Technology Indicators Conference 2014* (pp. 79–86).

Chinchilla-Rodríguez, Z., Costas, R., Larivière, V., Robinson-García, N., & Sugimoto, C. R. (2022). The relationship between corresponding authorship and author position. *Proceedings of the 26th International Conference on Science, Technology and Innovation Indicators (STI 2022).* https://doi.org/10.5281/zenodo.6957638

Clarivate. (2012). *OECD Category Scheme.* https://help.prod-incites.com/inCites2Live/filterValuesGroup/researchAreaSchema/oecdCategoryScheme.html

Cole, J. R. (1979). *Fair science: Women in the scientific community.* The Free Press.

Cole, J. R., & Zuckerman, H. (1984). The productivity puzzle. In *Advances in Motivation and Achievement* (Vol. 2, pp. 217–258). JAI Press Inc.

D'Angelo, C. A., & van Eck, N. J. (2020). Collecting large-scale publication data at the level of individual researchers: A practical proposal for author name disambiguation. *Scientometrics,* 123(2), 883–907. https://doi.org/10.1007/s11192-020-03410-y

Fanelli, D., & Glänzel, W. (2013). Bibliometric evidence for a hierarchy of the sciences. *PLOS ONE,* 8(6), e66938. https://doi.org/10.1371/journal.pone.0066938, PubMed: 23840557

Fanelli, D., & Larivière, V. (2016). Researchers' individual publication rate has not increased in a century. *PLOS ONE,* 11(3), e0149504. https://doi.org/10.1371/journal.pone.0149504, PubMed: 26960191

Garfield, E. (1996). When to cite. *The Library Quarterly: Information, Community, Policy,* 66(4), 449–458. https://doi.org/10.1086/602912

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management,* 35(1), 31–44. https://doi.org/10.1016/S0306-4573(98)00028-4

Glänzel, W., & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics,* 56(3), 357–367. https://doi.org/10.1023/A:1022378804087

Glänzel, W., Schubert, A., Thijs, B., & Debackere, K. (2011). A priori vs. a posteriori normalisation of citation indicators. The case of journal ranking. *Scientometrics,* 87(2), 415–424. https://doi.org/10.1007/s11192-011-0345-6

Henriksen, D. (2019). Alphabetic or contributor author order. What is the norm in Danish economics and political science and why? *Journal of the Association for Information Science and Technology,* 70(6), 607–618. https://doi.org/10.1002/asi.24151

Jagsi, R., Guancial, E. A., Worobey, C. C., Henault, L. E., Chang, Y., ... Hylek, E. M. (2006). The "gender gap" in authorship of academic medical literature—A 35-year perspective. *New England Journal of Medicine,* 355(3), 281–287. https://doi.org/10.1056/NEJMsa053910, PubMed: 16855268

Jian, D., & Xiaoli, T. (2013). Perceptions of author order versus contribution among researchers with different professional ranks and the potential of harmonic counts for encouraging ethical co-authorship practices. *Scientometrics,* 96(1), 277–295. https://doi.org/10.1007/s11192-012-0905-4

Kulczycki, E., Engels, T. C. E., & Nowotniak, R. (2017). Publication patterns in the social sciences and humanities in Flanders and Poland. *Proceedings of the 16th International Conference on Scientometrics and Informetrics* (pp. 95–104).

Kulczycki, E., Engels, T. C. E., Pölönen, J., Bruun, K., Dušková, M., ... Zuccala, A. (2018). Publication patterns in the social sciences and humanities: Evidence from eight European countries. *Scientometrics,* 116(1), 463–486. https://doi.org/10.1007/s11192-018-2711-0

Larivière, V., Desrochers, N., Macaluso, B., Mongeon, P., Paul-Hus, A., & Sugimoto, C. R. (2016). Contributorship and division of labor in knowledge production. *Social Studies of Science,* 46(3), 417–435. https://doi.org/10.1177/0306312716650046, PubMed: 28948891

Larivière, V., Pontille, D., & Sugimoto, C. R. (2021). Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). *Quantitative Science Studies,* 2(1), 111–128. https://doi.org/10.1162/qss_a_00097

Larivière, V., Vignola-Gagné, E., Villeneuve, C., Gélinas, P., & Gingras, Y. (2011). Sex differences in research funding, productivity and impact: An analysis of Québec university professors. *Scientometrics,* 87, 483–498. https://doi.org/10.1007/s11192-011-0369-y

Lerchenmüller, C., Lerchenmueller, M. J., & Sorenson, O. (2018). Long-term analysis of sex differences in prestigious authorships in cardiovascular research supported by the National Institutes of Health. *Circulation,* 137(8), 880–882. https://doi.org/10.1161/CIRCULATIONAHA.117.032325, PubMed: 29459476

Leydesdorff, L. (2009). How are new citation-based journal indicators adding to the bibliometric toolbox? *Journal of the American Society for Information Science and Technology,* 60(7), 1327–1336. https://doi.org/10.1002/asi.21024

Leydesdorff, L., & Bornmann, L. (2011). How fractional counting of citations affects the impact factor: Normalization in terms of differences in citation potentials among fields of science. *Journal of the American Society for Information Science and Technology,* 62(2), 217–229. https://doi.org/10.1002/asi.21450

Lundberg, J. (2007). Lifting the crown—Citation z-score. *Journal of Informetrics,* 1(2), 145–154. https://doi.org/10.1016/j.joi.2006.09.007

Madsen, E. B., Nielsen, M. W., Bjørnholm, J., Jagsi, R., & Andersen, J. P. (2022). Meta-research: Individual-level researcher data confirm the widening gender gap in publishing rates during COVID-19. *eLife,* 11, e76559. https://doi.org/10.7554/eLife.76559, PubMed: 35293860

Meyer, M., Waldkirch, R. W., Duscher, I., & Just, A. (2018). Drivers of citations: An analysis of publications in "top" accounting journals. *Critical Perspectives on Accounting,* 51, 24–46. https://doi.org/10.1016/j.cpa.2017.07.001

Milojević, S. (2012). How are academic age, productivity and collaboration related to citing behavior of researchers? *PLOS ONE,*

7(11), e49176. https://doi.org/10.1371/journal.pone.0049176, PubMed: 23145111

Milojević, S., Radicchi, F., & Walsh, J. P. (2018). Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences*, *115*(50), 12616–12623. https://doi.org/10.1073/pnas.1800478115, PubMed: 30530691

Mongeon, P., Smith, E., Joyal, B., & Larivière, V. (2017). The rise of the middle author: Investigating collaboration and division of labor in biomedical research using partial alphabetical authorship. *PLOS ONE*, *12*(9), e0184601. https://doi.org/10.1371/journal.pone.0184601, PubMed: 28910344

Nielsen, M. W. (2016). Gender inequality and research performance: Moving beyond individual-meritocratic explanations of academic advancement. *Studies in Higher Education*, *41*(11), 2044–2060. https://doi.org/10.1080/03075079.2015.1007945

Nielsen, M. W. (2017). Gender and citation impact in management research. *Journal of Informetrics*, *11*(4), 1213–1228. https://doi.org/10.1016/j.joi.2017.09.005

Nielsen, M. W., & Andersen, J. P. (2021). Global citation inequality is on the rise. *Proceedings of the National Academy of Sciences*, *118*(7), e2012208118. https://doi.org/10.1073/pnas.2012208118, PubMed: 33558230

Nielsen, M. W., Andersen, J. P., Schiebinger, L., & Schneider, J. W. (2017). One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis. *Nature Human Behaviour*, *1*(11), 791–796. https://doi.org/10.1038/s41562-017-0235-x, PubMed: 31024130

OECD. (2007). *Revised fields of science and technology (FOS) in the Frascati Manual* (p. 12). OECD.

Pagel, P. S., & Hudetz, J. A. (2011). An analysis of scholarly productivity in United States academic anaesthesiologists by citation bibliometrics. *Anaesthesia*, *66*(10), 873–878. https://doi.org/10.1111/j.1365-2044.2011.06860.x, PubMed: 21864299

Perneger, T. V., Poncet, A., Carpentier, M., Agoritsas, T., Combescure, C., & Gayet-Ageron, A. (2017). Thinker, soldier, scribe: Cross-sectional study of researchers' roles and author order in the *Annals of Internal Medicine*. *BMJ Open*, *7*(6), e013898. https://doi.org/10.1136/bmjopen-2016-013898, PubMed: 28647720

Petersen, A. M., Pan, R. K., Pammolli, F., & Fortunato, S. (2019). Methods to account for citation inflation in research evaluation. *Research Policy*, *48*(7), 1855–1865. https://doi.org/10.1016/j.respol.2019.04.009

Piro, F. N., Aksnes, D. W., & Rørstad, K. (2013). A macro analysis of productivity differences across fields: Challenges in the measurement of scientific publishing. *Journal of the American Society for Information Science and Technology*, *64*(2), 307–320. https://doi.org/10.1002/asi.22746

Pudovkin, A. I., & Garfield, E. (2009). Percentile rank and author superiority indexes for evaluating individual journal articles and the author's overall citation performance. *Collnet Journal of Scientometrics and Information Management*, *3*(2), 3–10. https://doi.org/10.1080/09737766.2009.10700871

Santamaría, L., & Mihaljević, H. (2018). Comparison and benchmark of name-to-gender inference services. *PeerJ Computer Science*, *4*, e156. https://doi.org/10.7717/peerj-cs.156, PubMed: 33816809

Sebo, P. (2021). How accurate are gender detection tools in predicting the gender for Chinese names? A study with 20,000 given names in Pinyin format. *Journal of the Medical Library Association*, *110*(2), 205–211. https://doi.org/10.5195/jmla.2022.1289, PubMed: 35440899

Sigogneau, A. (2000). An analysis of document types published in journals related to physics: Proceeding papers recorded in the Science Citation Index Database. *Scientometrics*, *47*(3), 589–604. https://doi.org/10.1023/A:1005628218890

Stremersch, S., Verniers, I., & Verhoef, P. C. (2007). The quest for citations: Drivers of article impact. *Journal of Marketing*, *71*(3), 171–193. https://doi.org/10.1509/jmkg.71.3.171

Tekles, A., & Bornmann, L. (2020). Author name disambiguation of bibliometric data: A comparison of several unsupervised approaches. *Quantitative Science Studies*, *1*(4), 1510–1528. https://doi.org/10.1162/qss_a_00081

Thelwall, M. (2020). Gender differences in citation impact for 27 fields and six English-speaking countries 1996–2014. *Quantitative Science Studies*, *1*(2), 599–617. https://doi.org/10.1162/qss_a_00038

Van Raan, A. F. J. (1998). In matters of quantitative studies of science the fault of theorists is offering too little and asking too much. *Scientometrics*, *43*(1), 129–139. https://doi.org/10.1007/BF02458401

Waltman, L. (2012). An empirical analysis of the use of alphabetical authorship in scientific publishing. *Journal of Informetrics*, *6*(4), 700–711. https://doi.org/10.1016/j.joi.2012.07.008

Waltman, L., Calero-Medina, C., Kosten, J., Noyons, E. C. M., Tijssen, R. J. W., … Wouters, P. F. (2012). The Leiden ranking 2011/2012: Data collection, indicators, and interpretation. *Journal of the American Society for Information Science and Technology*, *63*(12), 2419–2432. https://doi.org/10.1002/asi.22708

Waltman, L., & van Eck, N. J. (2019). Field normalization of scientometric indicators. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp. 281–300). Cham: Springer. https://doi.org/10.1007/978-3-030-02511-3_11

Wang, W., Yu, S., Bekele, T. M., Kong, X., & Xia, F. (2017). Scientific collaboration patterns vary with scholars' academic ages. *Scientometrics*, *112*(1), 329–343. https://doi.org/10.1007/s11192-017-2388-9

Xie, Y., & Shauman, K. A. (1998). Sex differences in research productivity: New evidence about an old puzzle. *American Sociological Review*, *63*(6), 847–870. https://doi.org/10.2307/2657505

Zhang, L., & Glänzel, W. (2017). A citation-based cross-disciplinary study on literature aging: Part I—The synchronous approach. *Scientometrics*, *111*(3), 1573–1589. https://doi.org/10.1007/s11192-017-2289-y

Zuccala, A., & van Leeuwen, T. (2011). Book reviews in humanities research evaluations. *Journal of the American Society for Information Science and Technology*, *62*(10), 1979–1991. https://doi.org/10.1002/asi.21588