NATIONAL SCIENCE BOARD
SCIENCE & ENGINEERING INDICATORS 2020

R&D

Publications Output: U.S. Trends and International Comparisons

Technical Appendix

NSB-2020-6

December 17, 2019

This publication is part of the *Science and Engineering Indicators* suite of reports. *Indicators* is a congressionally mandated report on the state of the U.S. science and engineering enterprise. It is policy relevant and policy neutral. *Indicators* is prepared under the guidance of the National Science Board by the National Center for Science and Engineering Statistics, a federal statistical agency within the National Science Foundation. With the 2020 edition, *Indicators* is changing from a single report to a set of disaggregated and streamlined reports published on a rolling basis. Detailed data tables will continue to be available online.

Table of Contents

List of Figures

Technical Appendix

Publication Output Data and Methodology

The *Science and Engineering Indicators 2020* report "Publication Output: U.S. Trends and International Comparisons" utilized a large database of publication records (i.e., bibliometric data). The database allows researchers to search the records of journal articles and conference papers. The present analysis treats the bibliometric data as a source of administrative data that serves as an indicator of research output. Administrative data come from the operation of administrative systems, often by public sector agencies collecting death/birth records, tax records, and others. Bibliometric data are collected by private companies to create searchable catalogs of research articles containing each article's title, author(s), authors' institution(s), citation, and journal title as they become available. The National Center for Science and Engineering Statistics (NCSES) uses Elsevier's Scopus bibliometric database to examine national and global scientific publication-related activity.[1]

The administrative nature of bibliometric data provides a benefit because not all countries release comparable data on research metrics. But the administrative nature also poses some limitations to the conclusions drawn from the bibliometric data. For example, counting publications and citations masks unmeasured variables including the density of the knowledge in each article, data sets that may accompany a publication, and any country-specific incentives for academic publication.

This appendix discusses the bibliometric data used in the report and the classification of journals and articles into fields of science.

Data

The counts, coauthorships, and citations presented in the publication output report are derived from information about research articles and conference papers (hereafter referred to collectively as "articles") published in peer-reviewed scientific and technical journals and conference proceedings. The articles exclude editorials, errata, letters, and other material that do not present or discuss scientific data, theories, methods, apparatuses, or experiments. The articles also exclude working papers, which are not generally peer reviewed.

The bibliometric data undergo review and processing to create the data presented in *Science and Engineering Indicators* (Science-Metrix 2019). In 2016, the National Science Board (NSB) addressed the differences between Scopus and the pre-2016 data used in *Science and Engineering Indicators* (NSB ***Indicators 2016***: **New Data Source for Indicators Expands Global Coverage**).

The next three sections present potential biases in the data: inclusion of non-peer-reviewed articles; English-language bias; and bias to the citation index caused by conference papers. The bias associated with non-peer-reviewed journals is ameliorated by filtering, but the other two biases persist in the data presented in the report.

Database Composition

Journal Selection. Elsevier selects journals for the Scopus database based on evaluation by an international group of subject matter experts who examine a candidate journal's editorial policy, content quality, peer-review policies, peer-review process and capacity, citations by other publications, editor standing, regularity of publication, and content availability.

Conference Selection. Elsevier selects conference materials for the Scopus database by subject field based on quality and relevancy, including the reputations of the sponsoring organization and the publisher of the proceedings.

More information about the selection of journals and conference papers is available at **https://www.elsevier.com/online-tools/scopus/content-overview** and **https://www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection**.

Database Filtering

NCSES undertakes additional filtering of the Scopus data to ensure that the statistics presented in *Indicators* measure original and high-quality research publications (Science-Metrix Technical Documentation 2019). Around 2011, librarians and bibliometric experts noted an increase in articles in the database from electronic journals and conference proceedings lacking substantive peer review.[2]

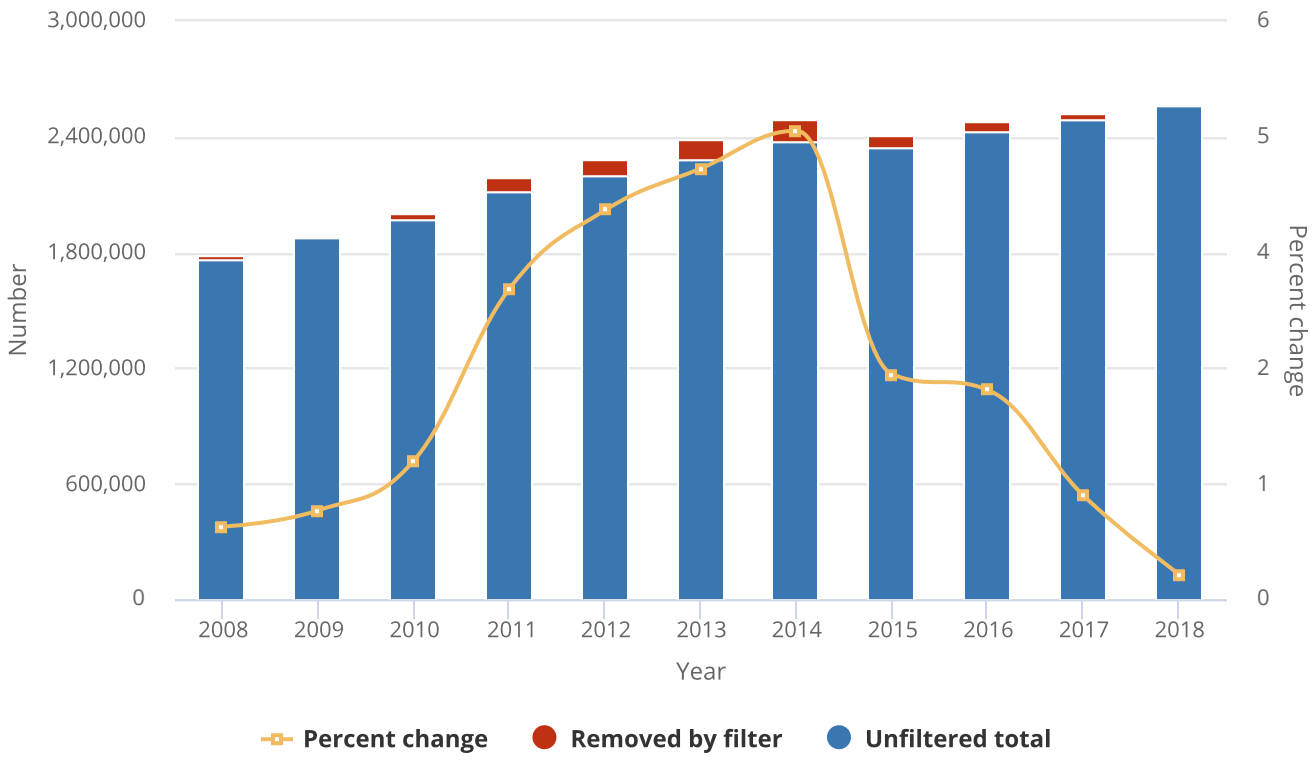
To exclude these publications from the bibliometric data used in this report NCSES removed two sets of data from the Scopus database:

1. Journals and conference papers flagged by the Directory of Open Access Journals (DOAJ) for failing to adhere to its list of best practices or being suspected of editorial misconduct.[3]
2. Titles that Elsevier removed from the Scopus database from 2014 onward are removed retroactively from the *Indicators* database for all publication years.[4]

As a result, NCSES removed 1% or fewer articles from the Scopus database for most years, then over 3% (more than 65,000 articles) in 2011 and 4%–5% (89,000–116,000 articles) each year from 2012 to 2014 (**Figure SA5a-1**). The number of articles filtered for the *Indicators* database dropped back down to the 2% range in 2015–16 as Elsevier began instituting filters on the Scopus database (**Figure SA5a-1**).

FIGURE SA5A-1

Filtered and unfiltered publications in Scopus, by year: 2008–18



Note(s)

Percent change is computed as the difference in number of publications between the filtered and the unfiltered approaches divided by the number of publications in the unfiltered approach.

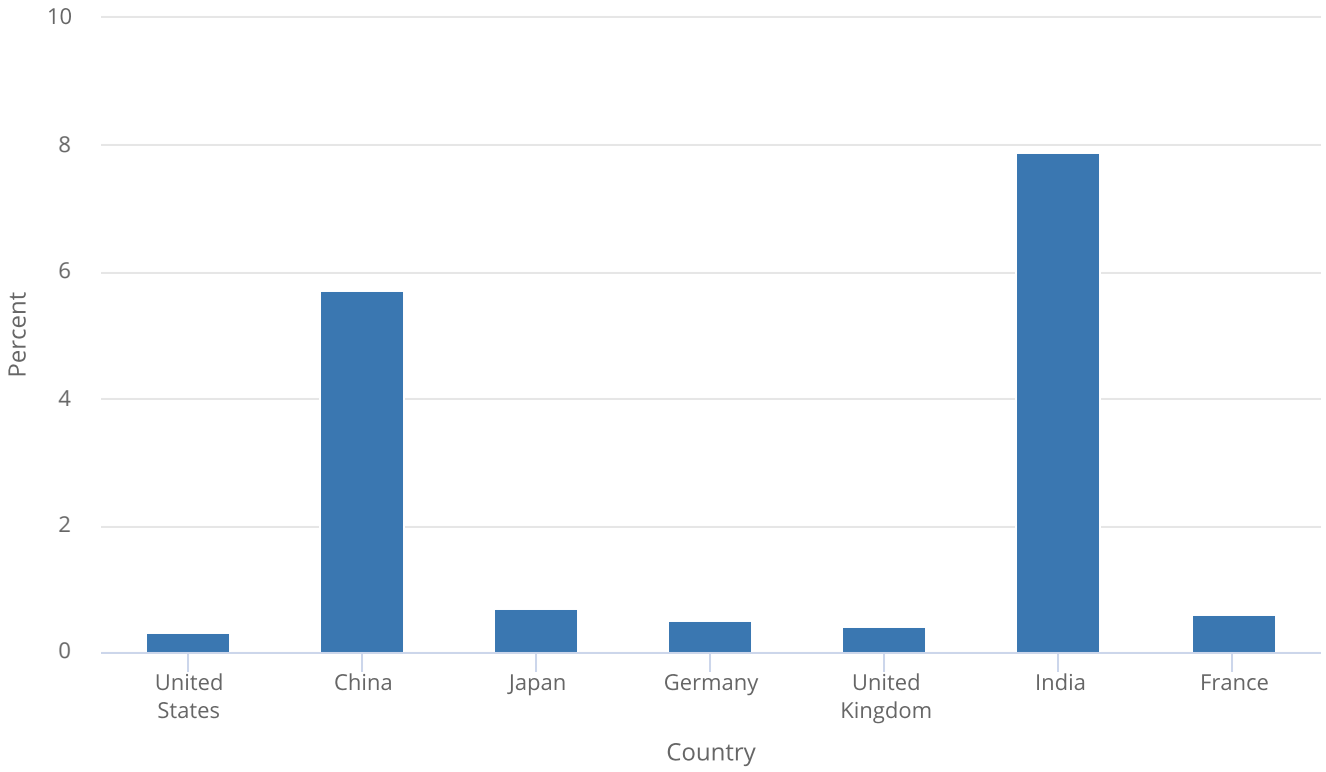
Source(s)

National Center for Science and Engineering Statistics, National Science Foundation; Science-Metrix; Elsevier, Scopus abstract and citation database, accessed June 2019.

The filtering has different impacts by country and field of science. NCSES has examined this filtering to better understand any potential bias. **Figure SA5a-2** shows the numerical impact of the filters by country or economy. During the last 11 years, 2008–18, China had the most articles removed (more than 249,000 articles removed, approximately 6% of China's total article count and accounting for 43% of all removed articles), followed by India (over 83,000 articles removed, 8% of India's article total and accounting for nearly 14% of all removed articles) (**Figure SA5a-1** and **Figure SA5a-2**). Other countries notably affected by this filtering (but not shown in **Figure SA5a-2**) include Iran and Malaysia; each had approximately 20,000 articles removed. In the case of Malaysia, this accounted for more than 10% of its total article output. Beyond these, only Russia and South Korea had more than 17,000 articles removed (about 3.5% of all articles removed from each) (NSB ***Indicators 2018***: **Sidebar Bibliometric Data Filters**).

FIGURE SA5A-2

Percent reduction in article count from removing low-quality publications from Scopus, by selected countries: 2008–18



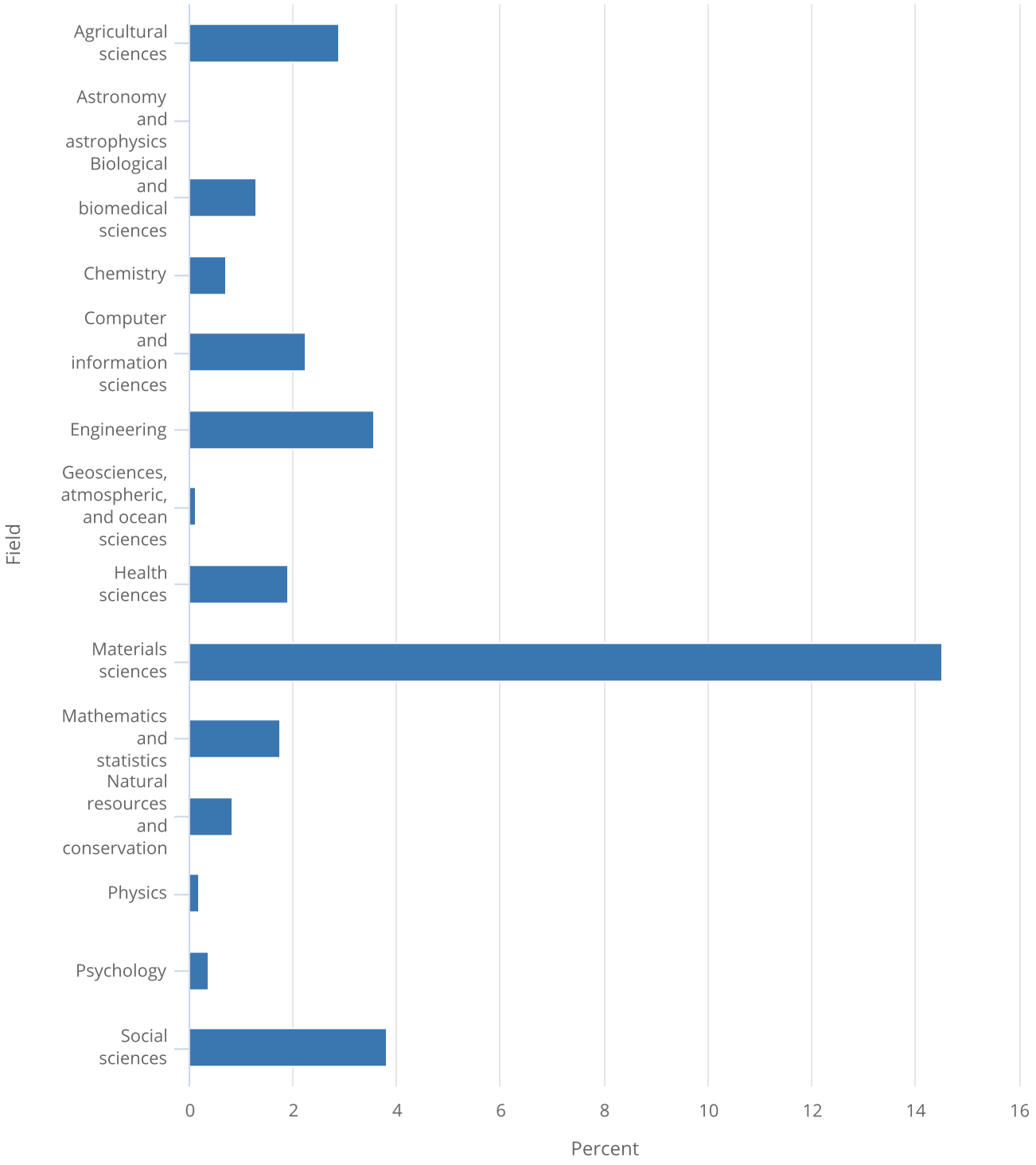
Source(s)
National Center for Science and Engineering Statistics, National Science Foundation; Science-Metrix; Elsevier, Scopus abstract and citation database, accessed June 2019.

Science and Engineering Indicators

Conference papers accounted for about 40% of articles removed. For example, cases where publishers posted new conference proceedings every day (with each post containing many papers) sent a clear red flag concerning robustness, originality, and peer review (Van Noorden 2014). In addition, filtering had the largest impact on the field of materials science, where the filtering process removed almost 15% of the articles (**Figure SA5a-3**). This is because conference proceedings comprised both a large share of the removed articles (40%) and a large share of the materials science articles (33%).

FIGURE SA5A-3

Percent reduction in article count from removing low-quality publications from Scopus, by TOD field: 2008–18



TOD = Taxonomy of Disciplines.

Source(s)
National Center for Science and Engineering Statistics, National Science Foundation; Science-Metrix; Elsevier, Scopus abstract and citation database, accessed June 2019.

Science and Engineering Indicators

English-Language Bias

Scopus contains an unmeasurable bias because the database only registers articles with an English-language title and abstract. Scopus uses English because it is the assumed global language of science (Amano, González-Varo, and Sutherland 2016). Bibliometric researchers have found an own-language preference in citations (Liang, Rousseau, and Zhong 2012). Thus, the indexing of publications with English-language abstracts can undercount citations associated with non-English publications. Social sciences exhibit more substantial linguistic bias than physical sciences, engineering, and mathematics (Archambault et al. 2009).

Conference Paper Bias in Highly Cited Article Index

Conference papers are included in the database analyzed in the report both for output and highly cited article (HCA) computation. Conference papers may bias HCAs because of uneven inclusion in the database[5] and widely different citation patterns compared with journal articles.

The impact on performance comes from the imbalance between percentage of output in conference proceedings across countries, and the fact that depending on the normalization approach, the score of countries can be heavily impacted compared to others simply because conference papers represent a larger share of their output. The issue is demonstrated in a simplified two-country example, both copublishing 1,000 journals articles, but with one also publishing 10 conference papers, and the other one publishing 200 conference papers. Assume that based on the 1,000 journal articles, both countries have the same impact. However, if conference papers are added into the computation, the entity with 200 conference papers will present a smaller combined HCA score as the citation scores associated with its conference papers will be lower. Therefore, in this case, two entities with similar impact in research published in journals may present much different impact because of the propensity of one to also send people to conferences. This potentially reduces the HCA for the country who participates more in conferences.

The impact across different fields is not uniform. Some fields of science publish and cite conference proceedings at different rates. In these cases, conference papers with low numbers of citations may yield high normalized HCA because the average is low for citing conference proceedings. For example, if the average number of citations stands at 1, and a conference paper receives 2 citations, its normalized impact will be 2.0, which is quite high. Adding these high impact conference papers may boost the score of a country specializing in that field and submitting conference papers.

The *Indicators 2020* report "Publication Output: U.S. Trends and International Comparisons" keeps conference proceedings in the analysis because for some fields and countries, conference proceedings are an important component of their output. NCSES will further explore the impact of retaining conference proceedings in the HCA.

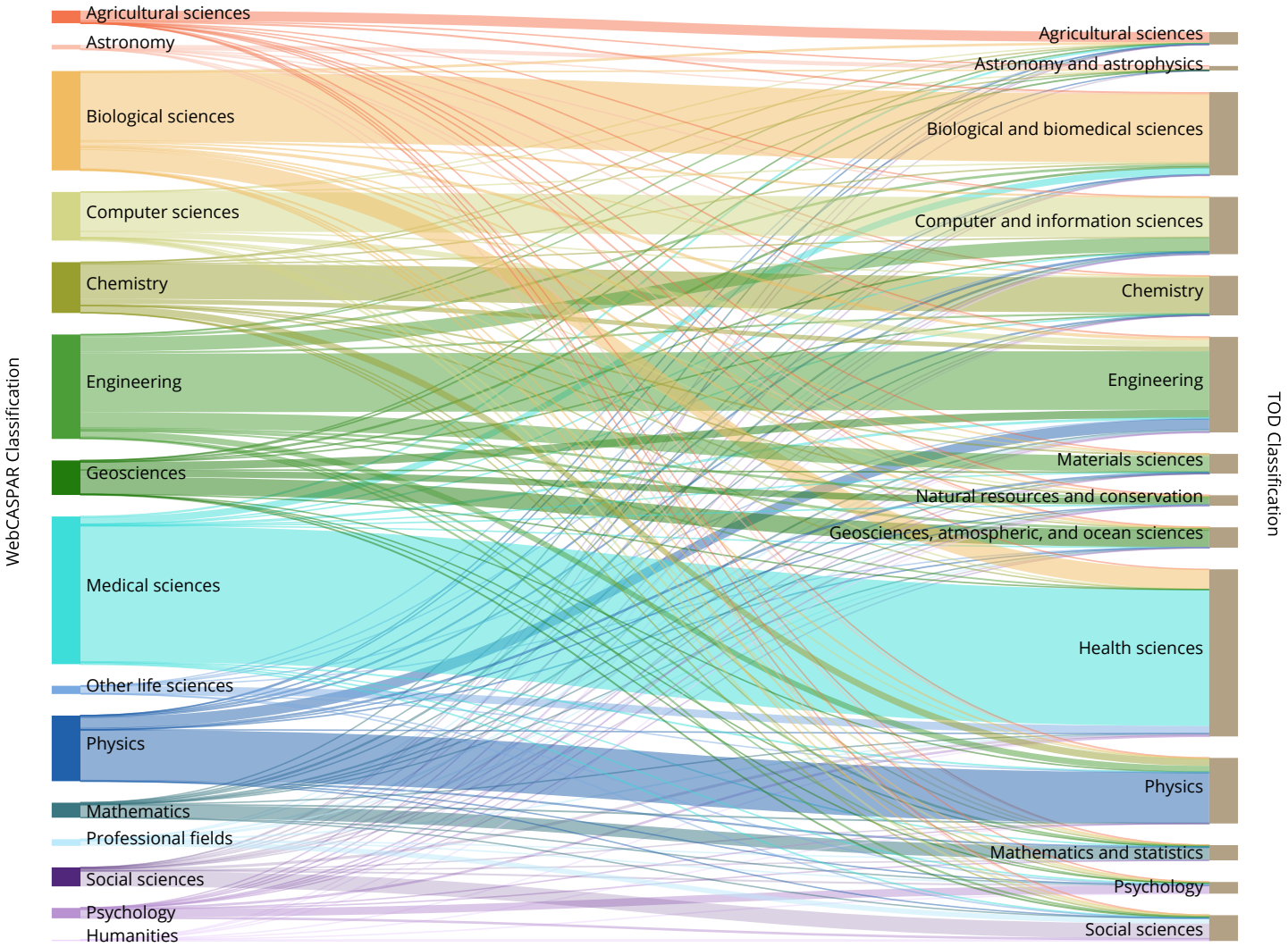
Fields of Science Classification

Beginning with the present report, NCSES updated the fields of science used to classify articles. The prior fields of science classification was designed in the 1970s. Since then, some areas of science have grown into distinct focus areas (e.g., materials science has grown apart from engineering) while others have tended to cluster (e.g., health sciences has combined with medical and other life sciences). The new taxonomy allows direct matching to NCSES's surveys such as the Higher Education Research and Development (HERD) Survey, Survey of Doctorate Recipients (SDR), and Survey of Earned Doctorates (SED).

Previous *Indicators* reports used 13 major fields from NCSES's WebCASPAR system of databases (Narin, Stevens, and Whitlow 1991).[6] Beginning with *Indicators 2020*, NCSES is using 14 fields of science developed by linking among the 261 fields in the NCSES Taxonomy of Disciplines (ToD)—specifically the fourth level of the six-level ToD—and the 176 fields defined in the Science-Metrix Ontology (Archambault, Beauchesne, and Caruso 2011).[7] **Figure SA5a-4** shows how the fields of science for the 2018 articles map from WebCASPAR to the ToD.

FIGURE SA5A-4

Comparison in Scopus 2018: WebCASPAR to TOD



WebCASPAR = Integrated Science and Engineering Resources Data System; TOD = Taxonomy of Disciplines.

Note(s)

Article counts from a selection of journals in S&E from Scopus. The Sankey diagram shows how the WebCASPAR fields are redistributed across the TOD fields. The width of the lines is proportional to the number of articles in each field.

Source(s)

National Center for Science and Engineering Statistics, National Science Foundation; Science-Metrix; Elsevier, Scopus abstract and citation database, accessed June 2019.

Science and Engineering Indicators

The ToD provides a transparent and up-to-date fields of science categorization with the ability to leverage publication output data in additional NCSES analyses. NCSES adopted the ToD across multiple surveys, expanding cross-survey analytical capabilities. For example, the ability to match data from surveys such as the SDR to publication output opens new avenues for research and enhances further understanding of linkages between academic degrees (input to research and development [R&D]) and publications (output of R&D). In addition, the ability to match to HERD data enables linkages between R&D expenditures within higher education institutions in the United States and publications, a key output of academic R&D.

Key to Acronyms and Abbreviations

DOAJ: Directory of Open Access Journals

HERD: Higher Education Research and Development Survey

R&D: research and development

SDR: Survey of Doctorate Recipients

SED: Survey of Earned Doctorates

ToD: Taxonomy of Disciplines

WebCASPAR: Integrated Science and Engineering Resources Data System

References

Amano T, González-Varo JP, Sutherland WJ. 2016. Languages Are Still a Major Barrier to Global Science. *PLOS Biology* 14(12):e2000933. Available at **http://journals.plos.org/plosbiology/article?id=10.1371%2Fjournal.pbio.2000933**. Accessed 26 May 2016.

Archambault É, Campbell, D, Gringras, Y, Larivière V. 2009. Comparing Bibliometric Statistics Obtained from the Web of Science and Scopus. *Journal of the American Society for Information Science and Technology* 60(7), 1320-1326. Available at **https://onlinelibrary.wiley.com/doi/full/10.1002/asi.21062**. Accessed 11 October 2019.

Archambault É, Beauchesne O, Caruso J. 2011. Towards a Multilingual, Comprehensive and Open Scientific Journal Ontology. *In* Noyons B, Ngulube P, Leta J, editors, *Proceedings of the 13th International Conference of the International Society for Scientometrics and Infometrics (ISSI)*, pp. 66-77.

Elsevier. Scopus. Available at **https://www.elsevier.com/solutions/scopus.**

Liang L, Rousseau R, Zhong Z. 2012. Non-English Journals and Papers in Physics: Bias in Citations? *Scientometrics* 95(1): 333–50.

Narin F, Stevens K, Whitlow E. 1991. Scientific Co-Operation in Europe and the Citation of Multinationally Authored Papers. *Scientometrics* 21(3):313–23.

National Science Board (NSB), National Science Foundation. 2016. *Science and Engineering Indicators 2016*: New Data Source for Indicators Expands Global Coverage. Alexandria, VA. Available at **https://www.nsf.gov/statistics/2016/nsb20161/#/sidebar/chapter-5/new-data-source-for-indicators-expands-global-coverage**.

National Science Board (NSB), National Science Foundation. 2018. *Science and Engineering Indicators 2018*: Bibliometric Data Filters. NSB-2018-1. Alexandria, VA. Available at **https://www.nsf.gov/statistics/2018/nsb20181/report/sections/academic-research-and-development/outputs-of-s-e-research-publications**.

Science-Metrix. 2017. Bibliometric and Patent Indicators for the *Science and Engineering Indicators 2018.* Technical documentation. Montreal, Canada: Science-Metrix. Available at **http://www.science-metrix.com/en/methodology-report**.

Science-Metrix. 2019. *Bibliometric and Patent Indicators for the Science and Engineering Indicators 2018. Technical Documentation.* Montreal, Canada: Science-Metrix. Available at **http://www.science-metrix.com/?q=en/publications/reports#/?q=en/publications/reports/bibliometric-indicators-for-the-sei-2020-technical-documentation**.

Van Noorden R. 2014. Publishers Withdraw More Than 120 Gibberish Papers. *Nature*. 24 February. Available at **http://www.nature.com/news/publishers-withdraw-more-than-120-gibberish-papers-1.14763**. Accessed 6 June 2017.

Notes

1 Because the bibliometric database is constantly updated, NCSES does not recommend comparing bibliometric data across different editions of *Indicators*. For each edition of the *Indicators*, NCSES uses a fixed snapshot of the database. This means that while trends are comparable, the exact number of articles, citations, and other data will vary across editions. Schneider J, van Leeuwen T, Visser M, Aagaard K. 2019. Examining National Citation Impact by Comparing Developments in a Fixed and Dynamic Journal Set. *Scientometrics* 119(2):973–85. Available at **https://doi.org/10.1007/s11192-019-03082-3**. Accessed 1 May 2019.

2 For an example of journals requiring robust and novel submissions, see **https://www.nature.com/authors/policies/peer_review.html**. For articles on low quality publications, see **https://www.nytimes.com/2016/12/29/upshot/fake-academe-looking-much-like-the-real-thing.html?_r=0**, **https://www.nytimes.com/2013/04/08/health/for-scientists-an-exploding-world-of-pseudo-academia.html**, **http://science.sciencemag.org/content/342/6154/60.full,** and **https://www.nature.com/news/predatory-publishers-are-corrupting-open-access-1.11385**.

3 For the DOAJ list of excluded journals see **https://docs.google.com/spreadsheets/d/183mRBRqs2jOyP0qZWXN8dUd02D4vL0Mov_kgYF8HORM/edit#gid=0**. Note that DOAJ also flags serials that are no longer available in open access (OA); although an important and evolving phenomenon in the research landscape, OA status is not associated here with any specific demarcation of quality, whether low or high. Thus, NCSES does not filter the titles flagged by DOAJ solely for OA-related reasons out of the *Indicators* database.

4 For Elsevier's principles of quality see **https://www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection**. During its periodic reevaluation of items flagged for follow-up, the Scopus Content Selection and Advisory Board elected to remove 42 titles as of 2014. NCSES retroactively removed the 42 titles from the *Indicators* database to create a valid time series for bibliometric analysis, even though Elsevier does not claim that these titles were necessarily of low quality before 2014.

5 Details about the inclusion of content in Scopus are available from **https://www.elsevier.com/solutions/scopus/how-scopus-works/content/content-policy-and-selection**.

6 NCSES established the 13 fields from 127 subfields recommended by Francis Narin and Mark Carpenter of Computer Horizons Inc. (CHI Research) in their 1976 research project for NCSES. The fields were computer sciences, engineering, geosciences, agricultural sciences, biological sciences, medical sciences, other life sciences, mathematics, astronomy, chemistry, physics, psychology, and social sciences.

7 In addition, Science-Metrix undertook article-level classification for general journals such as *Nature* and *PLOS ONE* using citation analysis.