SIDEBAR

Bibliometric Data and Terminology

Scopus Database. The counts, coauthorships, and citations discussed in the S&E output section are derived from information about research articles, conference papers, reviews, and short surveys (hereafter referred to collectively as "publications") that are published in peer-reviewed scientific and technical journals, books, and conference proceedings. This information characterizes the publication, the journal, proceedings, or book in which it appears; lists its author(s) and their institutional affiliations; and is collected in Elsevier's Scopus database. The publications exclude editorials, errata, letters, and other material whose purpose is not to present or discuss scientific data, theories, methods, apparatuses, or experiments. The publications also exclude working papers, which are not generally peer reviewed.*

The publications output database, also known as bibliometric data, undergoes a thorough review and processing to create the data for *Science and Engineering Indicators* (Science-Metrix 2017a). Before National Center for Science and Engineering Statistics (NCSES) analysis, a set of filters was applied to the Scopus database (see sidebar Bibliometric Data Filters). Additionally, for more information about the difference between Scopus and the data used in *Science and Engineering Indicators* before 2016, see sidebar New Data Source for Indicators Expands Global Coverage (NSB 2016). Although the full text of publications in Scopus does not need to be written in English, the publication's name and the abstract must be available in English for indexing.

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

Book Selection. The books included in the Scopus database are fully referenced and represent original research. They are selected based on publisher characteristics. These include the reputation and impact of the publisher, the size and subject area of the booklist, the publication and editorial policies, the quality of content, and the robustness of peer review.

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 on the selection of journals, books, and conferences is found at https://www.elsevier.com/online-tools/scopus/content-overview and https://www.elsevier.com/solutions/scopus/content/content-policy-and-selection.

Using the Scopus database, NCSES adds additional classifications and creates metrics listed as follows:

Field Classification. NCSES's WebCASPAR (Integrated Science and Engineering Resources Data System) classifies bibliometric data into the 13 broad fields of S&E. The WebCASPAR taxonomy classifies journals, and each publication is tagged with the field and subfield of the title under which it appears. However, many titles in Scopus are not classified in WebCASPAR; therefore, NCSES extends Scopus by associating these additional publication titles with the WebCASPAR fields with which they have the greatest affinity (methodological details available in http://www.science-metrix.com/sites/default/files/science-metrix/publications/science-metrix_sei_2016_technical_documentation_0.pdf). Appendix Table 5-25 shows data for these fields and their subfields, and Appendix Table 5-26 shows data grouped by regions, countries, or economies.

Publication Counts. Counts are the number of peer-reviewed publications produced by a given region, country, economy, or institutional sector. Publications coauthored by multiple countries or institutional sectors are counted in two ways. *Fractional* counting divides the publication count by the proportion of each of the countries or institutional coauthors named on the publication. Fractional counting enables the counts to sum up to the number of total publications (Appendix Table 5-27 through Appendix Table 5-41). *Whole* counting (also called *full* or *integer* counting) assigns one count to each country or institutional sector involved in coauthoring the publication, irrespective of their proportionate involvement in authorship (Appendix Table 5-42 through Appendix Table 5-51). Whereas fractional counting aims to assess the proportionate contributions of countries or sectors, whole counting aims instead to assess the participation of countries or sectors. One result of this difference is that with whole counting, the sum of publications from countries or institutional sectors will exceed the total number of publications. For the United States in 2016, there were 408,985 publications in the Scopus database as measured on a fractional-count basis (Appendix Table 5-41) and 519,289 as measured on a whole-count basis (Appendix Table 5-42).

Average Annual Change. The average annual change (also known as the *compound annual growth rate*) provides a measure of growth over time that accounts for compounding effects year over year. A stable rate of increase year-over-year will lead to exponential growth, and the average annual change reflects what level of annual increase would account for a measured exponential increase. Note that the underlying year-to-year growth may show considerable fluctuation; the average annual change smooths rates that fluctuate over time.

Coauthorship. Coauthorship measures collaboration across countries, regions, economies, and institutional sectors. Publication counts of coauthorship use whole counting, resulting in a full count being assigned to each country or institutional sector contributing to the publication. A publication is considered an international coauthorship when there are institutional addresses for authors from two or more different countries. Appendix Table 5-42 through Appendix Table 5-46 show international coauthorship by field of science.

Index of International Collaboration. Coauthorship or collaboration between countries is more likely between countries with large shares of publication output; thus, international collaboration measures are normalized using each country's total publication output to better reflect the propensity to collaborate. The index of international collaboration assesses each collaboration relationship, accounting for the size of each country's contribution to internationally coauthored publications. The result is a scaled index for country-country pairs, reflecting their tendency to collaborate with each other, relative to their tendency toward international collaboration overall. The measure is indexed to 1.00, meaning that if the pair collaborates more than their international average, the resulting score will be greater than 1.00; if they collaborate less than their international average, it will be less than 1.00.

For example, the United States participated in 38.6% of the world's internationally coauthored publications in 2016. Looking at collaborations specifically between the United States and China, 46.1% of China's internationally coauthored publications in 2016 had a U.S. coauthor. Dividing the actual U.S. share of China's internationally coauthored publications by the U.S. international average overall yields an index value of 1.19. Thus, China coauthors with the United States 19% more often than the pair's international average. The country pair index is always symmetrical, so the United States coauthors with China 19% more often than the pair's international average, just as China collaborates with the United States 19% more often than the pair's international average. Appendix Table 5-44 contains the data for calculating the 2006 and 2016 indices, shown in Appendix Table 5-43. Appendix Table 5-45 shows the U.S. sector publications coauthored with other U.S. sectors and foreign institutions for 2006 and 2016. Appendix Table 5-46 shows the U.S. coauthorship by number of authoring domestic and foreign institutions, by field, for 2003–16.

Sector Coding. NCSES codes each U.S. author's institutional address by economic sector into one of the following six categories: academic, federal government, state or local government, private nonprofit, federally funded research and development centers (FFRDCs), and industry. Additionally, the academic sector was divided into private and public subcomponents. Elsevier itself provides a correspondence table to match some organizations to their sector; other official lists (such as the NCSES Higher Education Research and Development Survey file, the Integrated Postsecondary Education Data System, the Carnegie Classification of Institutions of Higher Education, and the Medicare Hospital Compare data set) have also been cross-referenced. Considerable further coding has been undertaken using a combination of manual and automated approaches. For instance, organizations with "university" or "polytech" in their names have been coded as academic, whereas those with "Inc." or "Corp." in their name have been coded as industry. Manual validation and quality-control measures are also applied to ensure an acceptable level of quality. Given the diversity of U.S. organizations found in the database, sector coding is not exhaustive, and some U.S. organizations remain tagged as unknown (roughly 7%).

Citations and Relative Citation Scores. Citations of S&E publications by other S&E publications provide an indication of the impact of publications and of the flow of knowledge or linkage between sectors or geographic locations. The relative citation (RC) score for each publication adjusts its citation count to its respective year and research subfield, facilitating meaningful comparisons across contexts (Narin and Hamilton 1996; Wang 2012). For instance, average citation counts for 2014 ranged from less than 1 citation per publication in some subfields to more than 15 citations per publication in others. Citation scores presented in *Science and Engineering Indicators 2018* refer to the year in which a publication appears, not the citation year. At least 3 years must elapse after publication before one can reliably assess citation levels (and more years are preferable; see [Wang 2012]), so *Science and Engineering Indicators 2018* presents RC scores for 1996–2014. From 2 to 3 years of citation data are used to compute international citations (Appendix Table 5-47) and the RC index between country pairs (Table 5-28). By contrast, more years of data are used (where available) to compute RC scores elsewhere in this report, noting that normalization of scores by publication year ensures comparability across years, even when publications have not had the same amount of time to collect citations. Because of the need for timely results, citation data for the most recent year (i.e., 2014) are based on individual citation windows ranging from 24 to 36 months, depending on the month in which each publication was released.

Average of Relative Citations. With the relative citation (RC) scores computed, the average of relative citations (ARC) can be computed for each sector or geographic area, offering a view of the average impact within the scientific community, accounting for citation differences over time and across fields of research. The ARC is indexed to 1.00, which represents the world level, meaning that a score greater than 1.00 shows that the entity's publications are cited more than the global average, whereas a score less than 1.00 shows that the entity's publications are cited less than the global average.

The ARC reflects the average impact of a country's total publications. ARC does not account for how many publications each country produces, nor does it measure total influence within the scientific literature. For example, Guinea's ARC of 4.67 compared to the U.S. ARC of 1.42 reflects that the *average* citation level for individual publications is higher for Guinea than it is for the United States. The United States accounts for a larger share of the global citation total than Guinea does, which is unsurprising, given that the U.S. output volume of publications is much larger than that of Guinea.

Science and Engineering Indicators 2018 uses the ARC to assess average impact by region, country, or economy. ARCs are calculated using all data available—noting, once again, that a minimum window of 3 years is imposed—so results are not calculated for publications appearing later than 2014. Figure 5-29 shows changes in the U.S. ARC index by field from 2004 to 2014. Appendix Table 5-49 shows ARCs for U.S. fields of S&E, and Appendix Table 5-50 shows ARCs for regions, countries, or economies.

Highly Cited Publications. Citations to S&E publications are concentrated on a small portion of the total number of cited items. These measures follow a power law, where a relatively small share of the publications gathers a relatively large share of the impact. In these highly skewed distributions, the average is substantially different from the median or "typical" behavior. Thus, average counts alone offer only a partial reflection of the impact of S&E activities, and highly cited publications are shown to round out the assessment. Scores on this indicator are computed as a share of the top percentile of publications based on RC scores, relative to total publication output volume; results are available in Appendix Table 5-48. As noted previously, RC scores are not computed in *Science and Engineering Indicators 2018* for publications appearing after 2014; all available citation data are used to compute scores for the highly cited publications indicator, and results are normalized to the year of publication and the subfield of research.

Measurement Limitations of Bibliometric Data. The Scopus database indexes peer-reviewed S&E publications collected and curated by Elsevier to conform to a set of quality standards, including the stipulation that the abstracts have been written in English. 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. This linguistic bias has been found to be more substantial in social sciences than in physical sciences, engineering, and mathematics (Archambault et al. 2009). In addition, contribution levels among authors typically varies, but these differences are not captured in the database.

* For more information, see https://www.elsevier.com/solutions/scopus/content.