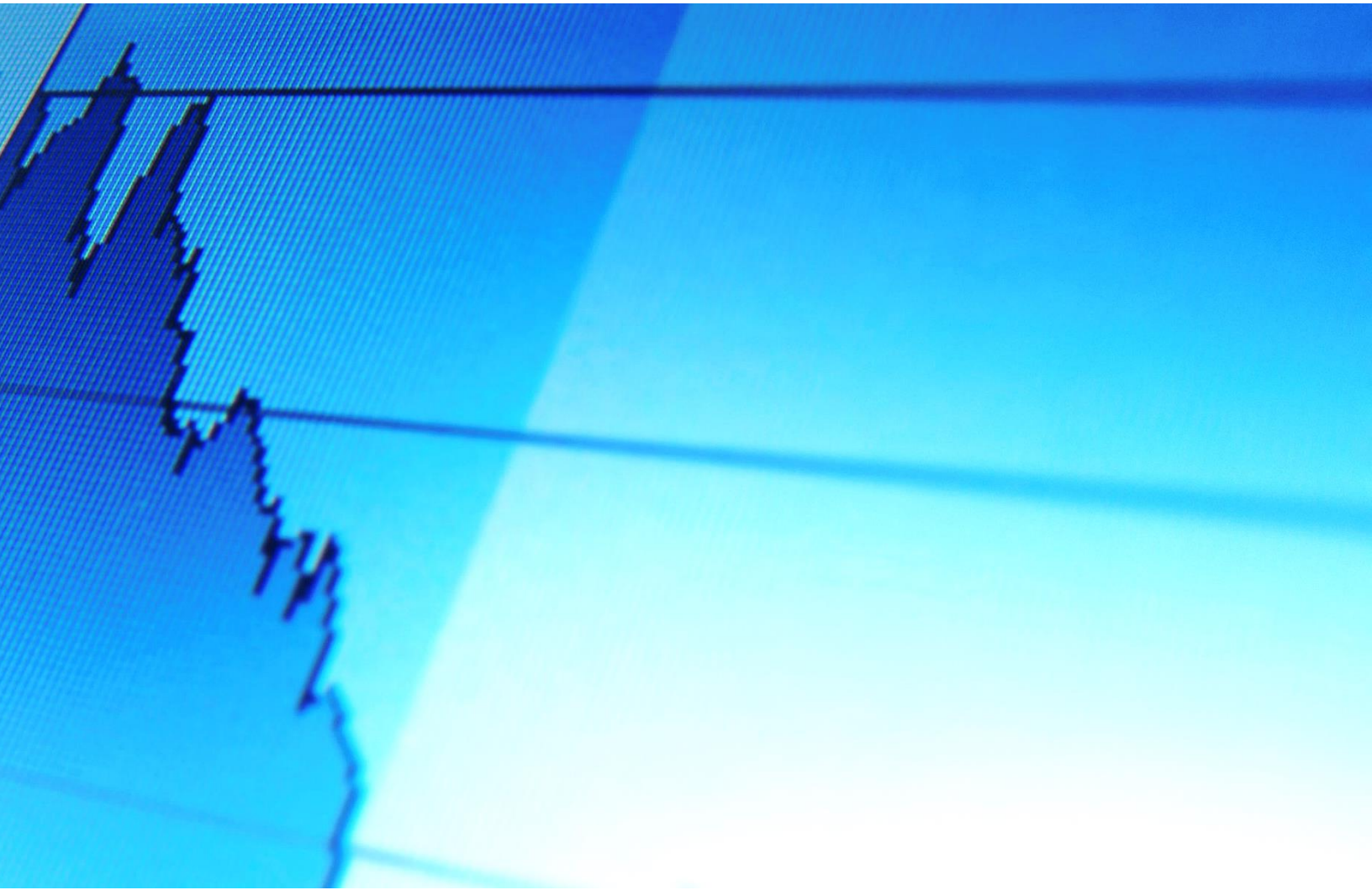
Science-Metrix

# Analytical Support for Bibliometrics Indicators

## Development of bibliometric indicators to measure women's contribution to scientific publications



Science-Metrix

# Analytical Support for Bibliometrics Indicators

## Development of bibliometric indicators to measure women's contribution to scientific publications*

**Final Report**

January 2018

***By:***


**Science-Metrix Inc.**
1335 Mont-Royal E. ▪ Montréal ▪ Québec ▪ Canada ▪ H2J 1Y6
1.514.495.6505 ▪ 1.800.994.4761
info@science-metrix.com ▪ www.science-metrix.com



*This work was funded by the National Science Foundation's (NSF) National Center for Science and Engineering Statistics (NCSES). Any opinions, findings, conclusions or recommendations expressed in this report do not necessarily reflect the views of NCSES or the NSF. The analysis for this research was conducted by SRI International on behalf of NSF's NCSES under contract number NSFDACS1063289.

# Contents

# Tables



# Figures

# 1 Introduction

Leading research nations worldwide are recognizing gender issues and acting to improve gender balance throughout the research ecosystem.[1] Several European countries have established research policies promoting gender equality in research through their national science organizations, including the Austrian Science Fund, the Academy of Finland, the German Research Foundation, the Netherlands Organisation for Scientific Research, the Research Council of Norway, the Science Foundation Ireland, the Swedish Research Council, the Swiss National Science Foundation, and the U.K. Research Councils. For example, some of these organizations are planning or have already performed studies and monitoring activities on gender equality in research funding.[2]

In North America, the Canadian federal government supports initiatives such as the Society for Canadian Women in Science and Technology,[3] among others, with other programs conducted at the provincial level. The U.S. National Science Foundation (NSF) has implemented several research policies dealing directly with the gender dimension in research, covering aspects such as career and funding opportunities,[4] the place of women in decision-making[5] and equal opportunity in research funding.[6] The U.S. also supports many initiatives in gender diversity, such as those administered by the USDA[7] and USDoE.[8] The White House Office of S&T Policy also collaborates with the White House Council on Women & Girls to support women pursuing education and professions in science, technology, engineering and mathematics (STEM).[9] National academies have also been following this issue,[10] and

[1] European Commission, & Directorate-General for Research and Innovation. (2016). *She Figures 2015*. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf#view=fit&pagemode=none

[2] European Commission, & Directorate-General for Research. (2009). *The gender challenge in research funding: Assessing the European national scenes*. EUR 23721 EN. Retrieved from http://ec.europa.eu/research/science-society/document_library/pdf_06/gender-challenge-in-research-funding_en.pdf

[3] http://www.scwist.ca/

[4] "All NSF directorates are participating in the Career-Life Balance (CLB) Initiative. In addition to direct financial support to reduce the career barriers related to dependent care, NSF activities have included harmonizing family-friendly policy language in collaboration with NIH and the development and use of an implicit bias informational briefing for NSF program directors, reviewers, and principal investigators (Ward, 2013)." Citation taken from Diversity Fueling Excellence in Research and Innovation Conference Report, Gender Summit—North America, Washington, DC, November 13–15, 2013. Retrieved from http://www.nsf.gov/od/iia/activities/gendersummit/GS3-ConfReport.pdf

[5] "Directorate for Biological Sciences at NSF has a policy statement that reinforces the inclusion of women and others from underrepresented groups in the planning activities and program agenda when seeking support for workshops and conferences (Ward, 2013)." Citation as per footnote 4 above.

[6] "NSF has integrated the legal requirement of IX with its core value of being broadly inclusive, to give a real focus on women commensurate with future excellence in NSF programs, panels and awards (Ward, 2013; Wise, 2013)." Citation as per footnote 4 above. Title IX aims to end sex discrimination in education.

[7] https://nifa.usda.gov/program/women-and-minorities-science-technology-engineering-and-mathematics-fields-grant-program

[8] http://energy.gov/diversity/services/minority-education-and-community-development/minority-educational-institution

[9] https://www.whitehouse.gov/administration/eop/ostp/women

[10] http://sites.nationalacademies.org/PGA/cwsem/PGA_049131; Council of Canadian Academies, & Expert Panel on Women in University Research. (2013). *Strengthening Canada's research capacity: the gender dimension*. Ottawa, ON: Council of Canadian Academies. Retrieved from

private organizations such as the L'Oréal Foundation are becoming involved,[11] not only in the U.S. but in many countries.[12]

This paper presents a promising approach to measure the proportion of women's authorship in scientific publications and to develop other indicators related to the participation of women in science. This approach uses the given name and surname of each author to determine the probability of the author being a man or a woman. The approach is highly accurate for U.S. authors, enabling the computation of robust indicators at highly disaggregated levels. The U.S. can also be compared with other leading countries at a more aggregated level. Results show that the proportion of scientific authorship by women is increasing globally and in the majority of countries. Relative to the top 50 most publishing countries, the U.S. is in the midrange in regard to the proportion of women scientific authors, after having steadily increased over the course of a decade, from 26% in 2006 to 32% in 2015. The U.S. trails countries with the highest proportion of women scientific authors, such as Thailand (46%) and Serbia (46%), but leads countries on the lower end of the spectrum such as Saudi Arabia (15%) and Japan (14%).

This report also shows that the proportion of women scientific authors is already high in some disciplines, mainly in the domains of health sciences and social sciences. Conversely, women scientific authors are less represented in disciplines within the domains of economics, applied sciences and natural sciences (although within most of these disciplines, the proportion of women scientific authors is increasing faster than the average across all disciplines combined).

## 1.1 Current approaches to assessing gender diversity in research

Existing quantitative methods for assessing gender diversity within the research community are primarily based on education and labor force statistics. These include assessments of the proportions of women among the following groups:

- undergraduate, master's and doctoral students, often assessed specifically for STEM fields, individually or collectively;
- graduating cohorts from each of these levels of study;
- cohorts of newly hired S&T employees, in academia or in specified occupations in the public and private sectors;
- total S&T employees, again in academia or in specified occupations in the public and private sectors;
- total S&T employees by level of seniority, often covering various steps along professional paths toward top-level research and top-level administrative positions.

Additionally, existing quantitative measures assess financial gaps between men and women, at various stages of career progression, considering both employment income and grant-based research funding.

http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20news%20releases/women_university_research/wur_fullreporten.pdf.pdf

[11] http://www.lorealusa.com/csr-commitments/the-l%E2%80%99or%C3%A9al-corporate-foundation/science/l%E2%80%99or%C3%A9al-usa-for-women-in-science-program

[12] https://www.womeninscience.co.uk/

Other measures focus predominantly on gender differences in work–life balance, often assessing these differences along the education, labor and financial dimensions described above. Furthermore, qualitative assessments are undertaken using a wide variety of approaches, often in an attempt to understand the underlying mechanisms that explain these quantitative findings.

## 1.2 Gender equality and research excellence

Important gains have been made in gender equality in research, such as increases in the number of women enrolling in and completing STEM education, and in the professional engagement of women in STEM occupations—these gains can be observed both in raw numbers and as a proportion of overall students, graduates and professionals. Nonetheless, gender disparities persist in the research ecosystem and are more and more acute the higher one looks in the professional hierarchy.[13] Additionally, there still appear to be important discrepancies in employment income and research funding, with female researchers lagging behind their male counterparts; these discrepancies are found even in the more proactive countries in this matter—namely, the Nordic countries.[14]

To some extent, discrepancies in grant-based funding might be attributable to the rise in the emphasis countries have placed on promoting research excellence through the implementation of numerous programs specifically designed to support the most outstanding researchers. One such example is the funding provided through the European Research Council, which was first instigated by the European Commission under the Seventh Framework Programme to support research excellence throughout Europe. Such programs might, in their current implementation, prove to work counter to initiatives fostering gender equality in research funding. For instance, if the definition of "excellence" for a granting program is outlined in such a way that it intensifies existing advantages, then the presently unequal distribution of advantages will serve to disproportionately favor male researchers. Ultimately, such initiatives can dampen the effect of diversity and equality initiatives.

The concern about excellence initiatives entrenching present advantages, and ultimately undermining diversity efforts, exists along many diversity lines within the research community. A further concern, more specific to gender diversity, is the effect of parenthood on career progression, a concern that disproportionately affects women researchers. Once again in this case, initiatives that entrench existing advantages will have a disproportionate effect based on gender.

The reliance on bibliometric statistics in research assessment exercises and in grant competitions is rising worldwide. Consequently, to increase their chances of securing funding, or to increase the amount of funding they manage to gather, researchers must be increasingly competitive in relation to the number of scientific papers they publish, as well as the scientific impact/quality of those papers; these pressures are especially acute in the context of grant competitions targeted at excellence. Hence, if women are at a

[13] http://www.catalyst.org/knowledge/women-sciences; Council of Canadian Academies, & Expert Panel on Women in University Research, *Strengthening Canada's research capacity*. Retrieved from http://www.scienceadvice.ca/uploads/eng/assessments%20and%20publications%20and%20news%20releases/women_university_research/wur_fullreporten.pdf.pdf

[14] Louët, S. (2014). *Research funding gap: Her excellence dwarfed by his excellence*. Retrieved from http://euroscientist.com/2014/06/research-funding-gap-excellence-dwarfed-excellence/

disadvantage relative to their male counterparts in terms of the number of publication outputs, then women might very well get stuck in a vicious circle—a smaller scientific and technological production reduces the chance of being funded and/or reduces the actual amount of funding secured, which would in turn reduce capacity to increase research output.

In addition, there are exponential relationships between the size of a researcher's publication portfolio and the citations to that portfolio by other researchers within the scientific community, where citations are the widespread bibliometric proxy for research excellence. That is to say, a larger research portfolio is likely to derive a double benefit, being both larger in size and more often cited. These two parameters are both important bibliometric proxies for excellence in many granting processes. In this way, bibliometric assessments of excellence tend strongly toward entrenching existing advantages; grants awarded on such bases thus tend to focus primarily on past performance rather than future potential. This pressure inherently tends to perpetuate the status quo, sustaining existing disparities along gender lines (and other lines of diversity).

The operation of these mechanisms is not purely speculative; there is ample evidence in the scientific literature to confirm the presence of a Matthew effect in scientific publication—that is, "papers by already-prestigious scientists usually receive far more attention than articles by scientists still on the way up, regardless of the intrinsic merit of such contributions."[15] In fact, in a 1999 paper, Katz[16] revealed the presence of a power-law relationship between publishing size (i.e., the number of papers) and recognition (i.e., number of citations), whereby a 10% increase in publishing size leads to a 12.7% increase in recognition. In other words, the gain in a researcher's citation impacts gets larger as his or her pool of papers gets larger, in a similar manner to the phenomenon of the rich getting richer and the poor getting poorer. Thus, if women currently lag behind men in terms of production size, then the size-dependent dynamics described here make it more likely that women also trail behind in terms of scientific impact, as this dimension is typically measured through citation counts. (In this case, the relevant bibliometric indicators are both those that count citations to individual articles and those that count citations to the journal in which the paper is published.)

Similarly, there is considerable evidence in the scientific literature of a link between international co-authorships and the scientific impact of papers as measured through citation counts. For instance, Science-Metrix has shown how the citation impact of papers rises as the number of authors and countries involved on scientific papers increases.[17] Consequently, it is also of interest to investigate whether there is a gap between women and men in terms of the extent to which their research is performed in international partnerships. If women currently trail behind men in their propensity to collaborate with

---

[15] Goldstone, J. A. (1979). A deductive explanation of the Matthew effect in science. *Social Studies of Science*, *9*(3), 385–391.

[16] Katz, J. S. (1999). The self-similar science system. *Research Policy*, *28*(5), 501–517.

[17] Campbell, D., Côté, G., Haustein, S., Lefebvre, C., & Roberge, G. (2014). *Bibliometric study in support of Norway's strategy for international research collaboration*. Report prepared for the Research Council of Norway. Retrieved from http://www.forskningsradet.no/servlet/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheadername1=Content-Disposition%3A&blobheadervalue1=+attachment%3B+filename%3D%22SMBibliometricsRCNInterimAnalythicalReport.pdf%22&blobkey=id&blobtable=MungoBlobs&blobwhere=1274503843081&ssbinary=true

research partners abroad, this gap can also lead to a higher assessment of excellence for research published by men. Once again, such a mechanism can serve to consolidate the current situation.

Larivière and colleagues[18] released a study in 2013 in which they showed that women still lag behind men in terms of the size and impact[19] of scientific production; furthermore, their findings show that women are less frequently involved in international co-authorships than their male colleagues. The authors suggest that the observed gaps between women and men might very well relate to differences of professional advancement:

> As is well known, the academic pipeline from junior to senior faculty leaks female scientists, and the senior ranks of science bear the imprint of previous generations' barriers to the progression of women. Thus, it is likely that many of the trends we observed can be explained by the under-representation of women among the elders of science. After all, seniority, authorship position, collaboration and citation are all highly interlinked variables.

They then go on to conclude that policies aimed at fostering international collaboration for female researchers could help reduce observed gaps, as co-publishing with international partners can help raise production size and impact—thus giving women opportunities to demonstrate potential for research excellence, along the lines discussed.

In short, women seem to have fewer publication outputs, receive fewer citations, and collaborate less frequently in international partnerships; all three of these dimensions are considered important features of research excellence, and each of these three parameters is correlated with the others. It is noted that the result reported above is drawn from a single study, and that findings across studies are not unanimous on the connection between gender and international collaboration. Nonetheless, the various dimensions of the Matthew effect outlined here highlight the importance of monitoring this situation, and of taking appropriate policy action to promote gender diversity within the research community.

## 1.3 Emerging bibliometric measures of gender

One difficulty at present is that whereas education, labor and financial measures have developed to characterize and monitor the evolving participation of women in the research ecosystem, bibliometric measures of gender have not seen widespread adoption. Accordingly, the understanding of gender dimensions in research output and impact is comparatively undeveloped. The *She Figures* report series, produced by the European Commission, can be seen as an important step to develop and implement such measures, bringing them into more mainstream conversations about women in research.

In the development of new indicators, one should consider the following:

- An identification of policy issues surrounding gender in science, research and innovation (e.g., horizontal segregation, vertical segregation, and the funding gap)
- An analysis of the availability of timely, representative and validated time-series data

[18] Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*, 211–213.

[19] A recent unpublished study suggests that self-citation may play a role in explaining the greater citation rate of men's research as opposed to women's; see King, M. M., et al. (unpublished). *Men set their own cites high: Gender and self-citation across fields and over time.* Retrieved from http://www.eigenfactor.org/projects/gender/self-citation/SelfCitation.pdf

- An analysis of international availability and completeness, as well as cross-country comparability
- An assessment of the scalability of these indicators, to determine whether they can be applied to the large data sets commonly used for bibliometric analysis—highlighted dimensions here include accuracy, as well as cost of data preparation and enrichment

The present project builds on the recent gender indicator development work carried out by Science-Metrix for the *She Figures 2015*[20] publication, which was developed in collaboration with European Commission officials, the Helsinki Group on Gender in Research and Innovation and the Helsinki Group's Statistical Correspondents in 41 countries. Detailed descriptions of how these quality dimensions were met and how the indicators are computed are provided in the *She Figures 2015 Handbook*[21] (which provides a succinct description) and in a companion methodology document[22] specific to the new bibliometric indicators (which provides a comprehensive description).

The present project also draws on previous work by several scholars to provide a robust picture of research outputs (i.e., number of papers, international co-publishing rate and scientific quality) by gender. In particular, the methods published by Larivière and colleagues (2013) have been improved in *She Figures 2015* through the computation of confidence intervals accounting for widely recognized biases resulting from the use of some of the most comprehensive databases of peer-reviewed scientific publications.

These confidence intervals are further examined in the present study, in view of developing robust indicators for potential consideration for *SEI 2018*. These new indicators will shed light on some of the gender dimensions within current research systems, in the hope of contributing to the identification of facilitators and barriers facing women in research and, ultimately, to the successful development of science policy to overcome these issues. One example of this might be funding policies that are designed to account for gender differences among various dimensions of performance.

## 1.4 Outline of the report

Section 2 of this report presents the general approach pursued in the development of the bibliometric indicators on gender, and points to the approach's limitations when using either the Web of Science (WoS) or Scopus databases. This section also provides some recommendations on the implementation of the indicators to provide robust statistics, including the calculation of confidence intervals.

Section 3 presents the results that have been compiled in Scopus for the 2006–2015 period. These results present an interesting picture of the participation of women in scientific publications in various fields of research, in different countries, and how this picture is changing over time.

---

[20] European Commission, & Directorate-General for Research and Innovation, *She Figures 2015*. Retrieved from https://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf

[21] Campbell, D. et al. (2015). *She Figures Handbook 2015*, produced for the European Commission. Retrieved from http://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015_Handbook_final.pdf#view=fit&pagemode=none

[22] Campbell, D. et al. (2015). *She Figures 2015: Comprehensive methodology – New research & innovation output indicators*. Retrieved from http://www.science-metrix.com/files/science-metrix/publications/she_figures_2015_comprehensive_methods_on_ri_outputs_final.pdf

# 2 Description of the approach

The gender of authors is not disclosed on peer-reviewed publications. As a result, using articles' author name information is usually the only tangible means available to large-scale studies to evaluate the contribution of women to scientific production. The determination of gender based on author name is not something new, nor specific to bibliometric studies. It is a part of the larger field of research of *onomastics*, which is the study of the origin, history and use of proper names. For instance, this technique has been used for marketing purposes and, more generally, for all kinds of studies related to gender equity or that are gender-concerned. The majority of these studies are based on statistical approaches that use large existing data sets (census data, governmental data, etc.) or web mining tools to generate lists of target publics with their appropriate gender. As one would anticipate, the given or first name of a person typically enables discerning their gender. However, in some cultures it may be the surname or last name that changes according to gender, so at times this may also be taken into account.

The proposed approach for measuring the gender of scientific authors is thus based mainly on the analysis of their given names, while also using their surnames when applicable. To infer a gender from a name, Science-Metrix employs an approach that combines the use of existing and established name/gender lists and the use of a powerful commercial tool designed to determine the gender of names, taking into account different elements such as given name, surname, ethnicity and country.

There exist several challenges and/or limitations related to determining the gender of scientific researchers using their names. Section 2.1 presents these challenges and the proposed approaches to overcome them. Section 2.2 presents a comparative analysis of the WoS and Scopus in terms of given name coverage, which further supports the methodology and the choice of indicators proposed by Science-Metrix. Section 2.3 describes the commercial tool for gender assignation based on names selected by Science-Metrix and provides a thorough assessment of its performance by testing it on real data sets of names for which the gender is already known. Section 2.4 presents the steps for using the tool for the attribution of authors in scientific publications databases. Section 2.5 and Section 2.6 present the approach for the calculation of the indicator and the confidence intervals.

## 2.1 Challenges and limitations

### 2.1.1 Given names are not always available in databases of scientific literature, and this availability varies over time

The two most complete databases of peer-reviewed publications are Scopus (Elsevier) and the Web of Science, or the WoS (initially by Thomson Reuters, but now operating as a separate company called Clarivate Analytics). These databases also enable computing statistics on authors, output, impact and collaboration, among others. However, the full given name of authors is not always accessible in these two sources. In many cases, only the initials of given names are provided, making it impossible to determine the gender of authors. Although more and more publishers seem to be aware of the importance of providing the full name of authors when publishing papers, this is still not a generalized practice. Section 2.2.1 presents a timely analysis and comparison of given name coverage in Scopus and the WoS.

Two approaches can be used to overcome this problem. The first way is to compute statistics on gender using all authors for which the given name is available—that is, to produce statistics based on a subset of authors for each level of aggregation desired (subfield, field or domain, as well as country), while ensuring that these samples are representative of the populations of interest and enable robust estimation of the proportion of men's and women's contribution to the scientific literature. The statistics can also be computed only on the corresponding (or reprint) author of a publication. The full name is more frequently indexed for this author, and as Science-Metrix demonstrated in the *She Figures 2015* study, measures based on this author provide a good proxy for the contribution of the lead author on scientific publications. More detail on this approach is provided in Section 2.2.

### 2.1.2 The proportion of papers for which the given name of authors is available varies significantly across subfields and between countries

Different editorial policies or different cultures across scientific disciplines may explain the significant variations observed in the coverage of the full given name of authors (as opposed to only initials) across disciplines. Among other things, the availability of given names is higher in subfields where women are more present (e.g., subfields of the social sciences and humanities [SSH]). If this bias is not properly dealt with when producing aggregated statistics for all subfields combined, the share of women authors in all publications will be overestimated, indicating that the gender gap in production is less pronounced than it actually is. For the *She Figures 2015* study, this problem was resolved by first estimating the indicators at the subfield level along with their margins of error (because only a sample of the papers had information on the given names of the corresponding authors) prior to aggregating those estimates (and their margins of error) at a higher aggregation level. This way, reliable estimates could be obtained at all levels of aggregation, along with information on their actual level of accuracy by country (through the margins of error). Variations in first name availability is also observed across countries.

### 2.1.3 Not all given names are gender specific

It is well known that some given names apply for both women and men (e.g., Ashley, Kim, Riley, Lee, Claude). Although this particularity is rather uncommon globally, the discriminating power of given names drops significantly for Asian names, especially those from China and the Republic of Korea. For instance, many given names are as common for women as they are for men in China, a large producer of research papers.[23] Science-Metrix is attempting to devise an approach for producing robust statistics for such countries by relying on the gender-specific given names available.

In cases of non-gender-specific names outside Asia, the last name can provide direct information on the author's gender for some ethnicities (e.g., the termination of the last name can change by gender for some Slavic languages). The approach used in this study is capable of handling most of these cases with high accuracy.

[23] The names are less ambiguous when written in Chinese characters, but considerable information that would be useful in determining the gender is lost when Chinese names are romanized.

Another specific situation concerns given names that could refer to a man for one nationality and to a woman for another. One such name is Andrea, which is a female name in most English-speaking countries, but a male name in Italy. For these cases, the surname is used in order to assess the ethnicity of the person, which gives insight into the likelihood of the given name being that of a woman or a man in this ethnic context. Here again, the approach used is designed to handle these cases with high accuracy.

### 2.1.4 There is no link between author names and their address in the WoS prior to 2006

Both Scopus and the WoS provide the list of authors as well as the list of addresses for each author, including the department, the name and address of the institution, and the country. However, in the WoS, in most cases, the link between an author and his or her institution is not provided for the years preceding 2006. Therefore, it is impossible to produce statistics by institution and country related to a specific author in the WoS prior to this time. However, the WoS almost always provides the address of the corresponding or reprint author, as this information is registered in a separate field from the normal list of authors. By applying the gender analysis solely to the corresponding author, the production of gender statistics can be calculated for a longer period when using the WoS. Section 2.2.2 presents further analysis of this situation.

## 2.2 Comparative analysis of Scopus and the WoS in terms of their amenability to support indicators based on authors' names

The potential for developing a robust indicator of gender contribution can be tested in both Scopus and the WoS; however, the methods used in each database may differ because of their specific features and limitations. Across both databases, the given name of authors is not always available and can vary over time. Although the practice of journals including the full given names of authors on publications is growing, many journals still only record the initial of an author's given name. Also, as previously mentioned, the link between authors and addresses is not provided by the WoS for the years preceding 2006, which prevents the calculation of statistics based on the geographic location of authors for this period.

The implication of this is that the gender indicator is calculated based only on a part of the total population of articles—that is, the portion for which the given names of all authors are provided. The reprint author is registered separately in the database, and this field is more likely to provide the author's full name (first name and surname) than the fields for the typical list of authors on the publication. Also, in the WoS, the reprint author is associated with an address in most cases. This information is of interest as the reprint author is often in a leading position—that is, as the principal investigator and usually the researcher to whom the project grant was awarded, or as a researcher who was highly involved in experimental research. The idea of computing gender based only on the reprint author presents the advantage of better discriminating the position of women versus men in research, as opposed to computing gender using all authors. For example, the latter case makes it impossible to determine whether women were leading the research or instead held the role of assistant.

Note also that another way of computing data on the lead author is to consider the first author on the list of authors on a publication.[24] However, notwithstanding the fact that the given name is not always provided, this approach is prone to some biases. For example, in some fields of research, authors are typically listed alphabetically. In such cases, the first author does not relate to a leading position at all, either as the team leader or as the main contributor. Additionally, the team leader often appears as the last author when the main contributor, placed as first author, is a graduate student. From a methodological standpoint, the use of the first author is also less desirable because the share of publications for which it is possible to assign a gender and a country to the first author is smaller than when using the reprint author.[25]

The approach using the reprint author is still imperfect though, as graduate students can sometimes appear as reprint authors. It is also possible that within teams involving multiple researchers (excluding graduate students), women might face stronger barriers than men in taking the place of the reprint (or lead or corresponding) author. If this is the case, the ratio of women-to-men authorships based on the reprint author might, to some degree, underestimate the contribution of women researchers.

In brief, when comparing the methods of computing a gender indicator based on all authors and based on the reprint author only, both have their pros and cons, and each can lead to different interpretations of the results obtained. With this in mind, the following subsections present a comparative analysis of Scopus and the WoS in terms of their coverage of the given names of authors and the reprint author, as well as their capacity to link authors to their addresses. This supports the assessment of the best method (considering all authors or only the reprint author) for computing a gender-based indicator for each database.

### 2.2.1 Completeness/quality of authors' names

Because the method used to identify the gender of authors is mostly based on their first names, it is important to verify the availability of this information in the databases. Figure 1 presents statistics on the availability of the first name for all authors in Scopus and the WoS for the 1996–2015 period. Since 2007, both databases have been converging, and this may indicate that they are indexing the information when it is available and that the trends are now mostly influenced by the behavior of the publishers. Given that no database offers the first name (or given name) for all authors, the proportion of men and women must be inferred from the portion of records for which this information is known. At this point, Science-Metrix believes it would be very hazardous to conduct the analysis prior to 2006 using the WoS because less than 1% of its records include the author first name in this period.

[24] Larivière, V., Ni, C., Gingras, Y., Cronin, B., & Sugimoto, C. R. (2013). Global gender disparities in science. *Nature*, *504*, 211–213.

[25] This holds true for the last author.

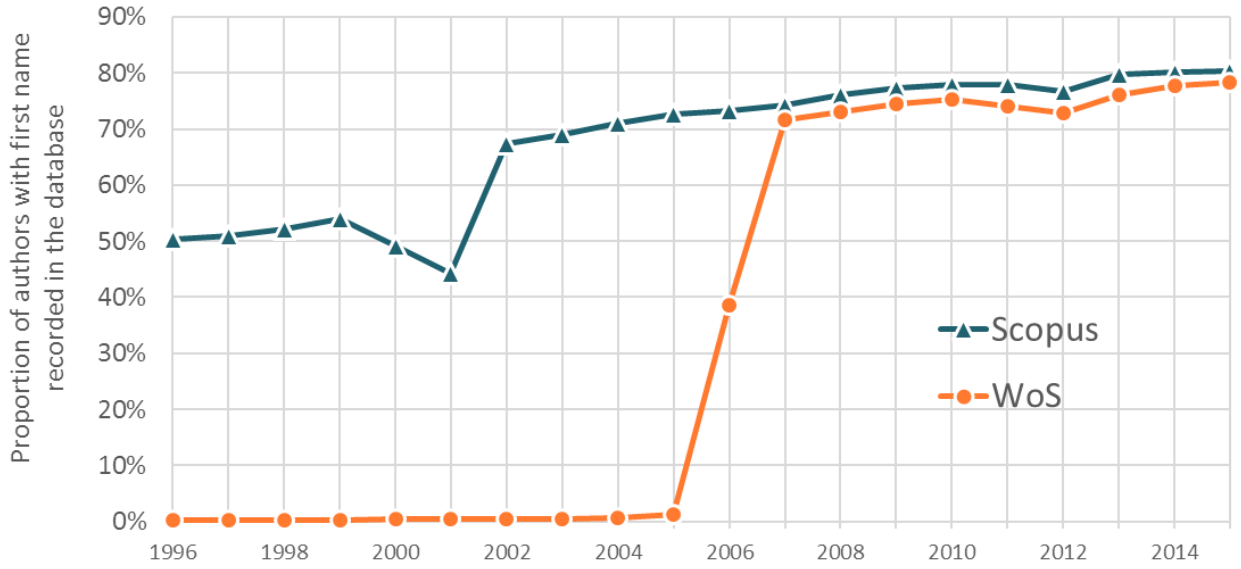



Figure 1 Availability of authors' first names in Scopus and the WoS
Note: Underlying data available in Table XIII.
Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics) and Scopus (Elsevier)

### 2.2.2 Link between author and address

The link between the author and the address is always available in Scopus, no matter the year of publication, and this information can be used to determine the country of the author. The WoS only links authors and addresses for the majority of articles published in 2008 and later (Figure 2); if a study is designed to encompass all authors on each publication, the WoS cannot be used to produce indicators on gender at country level before 2008.

However, if the study is performed using only the corresponding or reprint author in the WoS, then the link between this author and his or her address is almost always available, see Figure 3. The given name of the corresponding author in the WoS is not systematically available though, as shown in Figure 4, which combines linked author and address fields and the availability of given names for both reprint and all authors in the WoS.

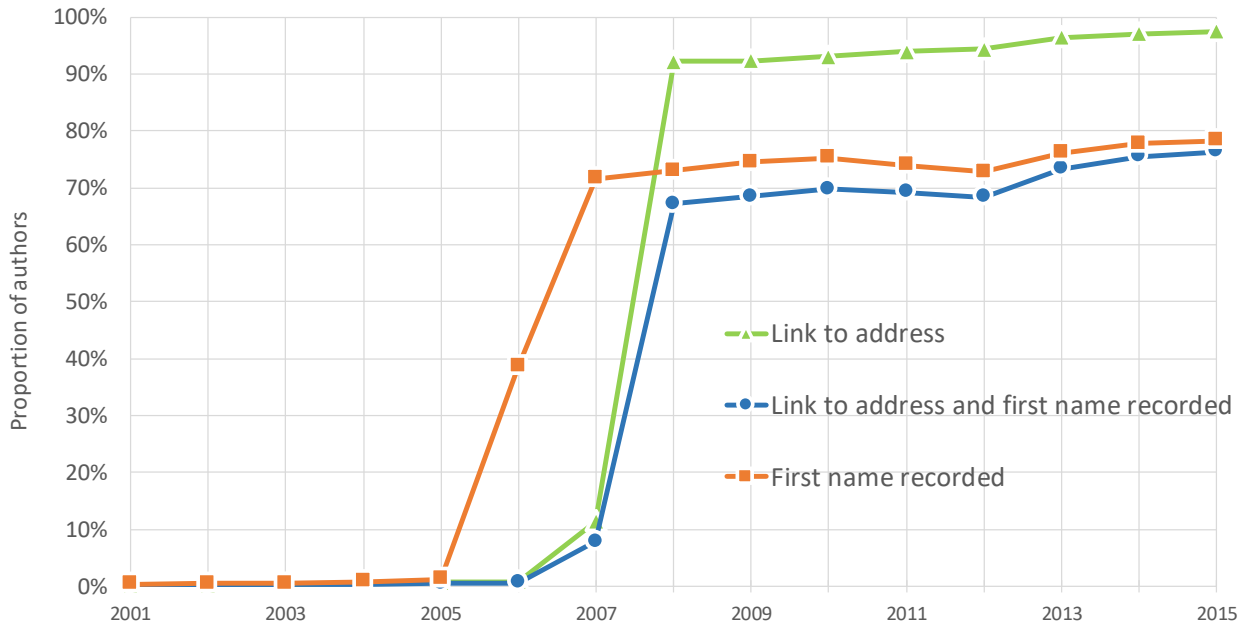



Figure 2 Availability of the link between the author and author address in the WoS

Note: Underlying data available in Table XIV.
Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics)

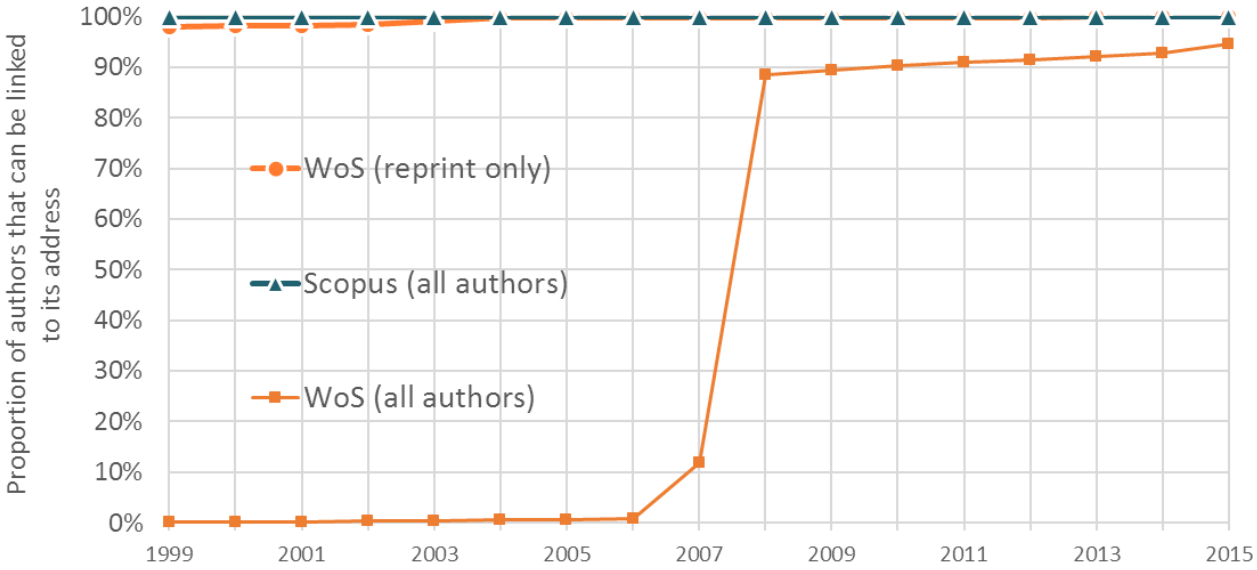



Figure 3 Availability of the link between the author and author address in Scopus and in the WoS for all authors, and in the WoS for corresponding authors only

Note: Underlying data available in Table XV.
Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics) and Scopus (Elsevier)

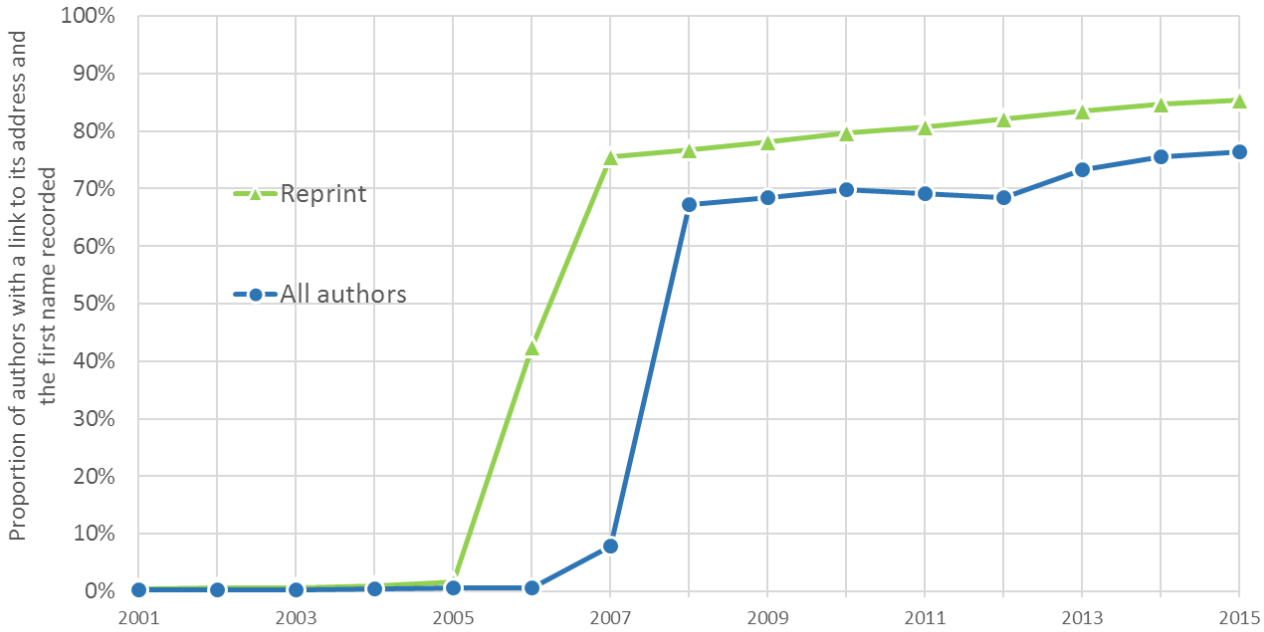



Figure 4 Availability of the link between the author and author address, combined with availability of the given name in the WoS, for reprint author versus for all authors

Note: Underlying data available in Table XVI.
Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics)

These analyses show that a gender study using the WoS would be a little more robust when using the reprint author only, and should be restricted to 2007 or later if statistics are to be produced at the country level. In Scopus, the availability of the first name for the corresponding authors only is not markedly higher than for all the authors on the publication. Therefore, the indicators on gender can be computed on all authors as well as for the corresponding author only. This would support the assessment of both women's authorship and women as lead author.

In the context of this pilot study, the gender indicators are computed using Scopus and are based on the list of all available authors, not only the corresponding author.

## 2.3 Gender name inference

Many solutions are available on the market to determine gender based on an author's first name and other available information (e.g., last name, ethnicity, location); this study used a solution developed by NamSor™. NamSor is a European designer of name recognition software committed to promoting diversity and equal opportunity.[26] NamSor was selected for this study because it offers a very high degree of accuracy and recall, and a global coverage.

[26] http://www.namsor.com/

NamSor claims to cover all languages, alphabets, countries and regions. In addition to using the data mining approaches that are behind most of the solutions available (e.g., using national lists of baby names), NamSor works with linguists, anthropologists and historians to increase their products' accuracy in various cultural contexts. They also develop solutions to infer the origin or ethnicity of individuals based on their names, and these developments reinforce the quality of gender estimation. Because a surname may change depending on gender in some cultures, the API automatically recognizes if gender can be inferred from the first name (e.g., Carl) or the last name (e.g., Sololova). Finally, the API is quite tolerant of typographic errors and multiple names, a feature that is very handy given the significant number of input errors in the publication databases. As of October 2016, NamSor API had processed over 824 million names since it was launched in February 2014.

NamSor was also selected for its capacity to handle very large volumes of data. To determine the gender of authors in the WoS and Scopus, nearly 8 million combinations of first name and last name have been processed quite rapidly and at a reasonable cost.

NamSor have implemented a rigorous protocol to assess the quality of their tool, demonstrating that it can achieve a high recall (i.e., there are very few unknowns) and accuracy (i.e., there are very few false positives) in the United States, Canada, Mexico, Russia, Japan and most European countries. Their validation procedure relies on the use of directories listing names along with their geographic location (i.e., country) and specified gender of titles (Mr. for men and Ms. for women). Using the known gender of individuals, they validate whether their algorithm attributes the correct gender.

As noted in an article published on their website, NamSor used data from the Official Directory of the European Union to validate their Gender API, as well as to study the gender gap in the European Union. In application, they reported data on vertical segregation as it relates to the positions being occupied by women relative to men.[27] Science-Metrix has tested the tool on a newer edition of the directory, reproducing the validation procedure.[28] The names of all employees listed in the directory (10,419) are preceded by Ms. for women and Mr. for men, which provides a strong benchmark to evaluate the accuracy and recall of the API. The API was able to provide a gender for the clear majority of names; only about 1% of names were unknown.

For each combination of first name and last name, the API returns a score between -1 and 1. A score of -1 indicates a man with a certainty of 100%, whereas a score of 1 is returned for a woman, again with 100% certainty. A score of 0 denotes that the gender can't be determined at all. In fact, the API does not provide a gender for all values between -0.1 and 0.1. This score has been designed to improve the estimation of the proportion of women or men by providing less weight in the estimation for lower scores.

Table I presents the results of this assessment. Because the goal of this assessment is to evaluate the robustness of the API to provide the right gender for a given pair of first name and last name, only the employees for which the API was able to provide a gender are kept for the analysis (both for the

[27] https://blog.namsor.com/2014/09/09/whats-the-gender-gap-in-the-european-union-whoiswho/

[28] This to validate the validation performed by NamSor and to gain a better understanding of the tool.

estimations and the calculation of the "real" values). The API caused very few errors in attributing gender. Although some women were erroneously coded as men (79 errors, 1.5% of women) and some men were coded as women (102 errors, 2.5% of men), these errors almost balanced themselves out. This resulted in a very good estimation of the distribution of women (35.3% compared to 35.1%) and men (64.7% compared to 64.9%).

These first estimates are computed using a dichotomous coding of gender: when the score is between -1 and -0.1, the individual is coded as man, and when it is between 0.1 and 1, the individual is coded as woman. If instead we use the score as intended to weight the attribution of each individual, we obtain what is supposed to be an improved estimate (34.9% for women and 65.1% for men). In this case, the improved estimate is as good as the first estimate: the improved estimate of the proportion of women is 0.59% lower than the real proportion, whereas the first estimate was 0.63% higher. In any case, the accuracy of the API is very high for this data set.

Table I Validation of NamSor API using data from the Official Directory of the European Union

| | Gender (estimate) | | |
|---|---|---|---|
| **Gender (real)** | **Women** | **Men** | **TOTAL** |
| **Women** | 3,552 | 79 | 3,631 |
| **Men** | 102 | 6,618 | 6,720 |
| **Total** | 3,654 | 6,697 | 10,351 |

| | Women | Men |
|---|---|---|
| **Real** | 35.1% | 64.9% |
| **Estimate** | 35.3% | 64.7% |
| **Recall** | 97.8% | 98.5% |
| **Precision** | 97.2% | 98.8% |
| **Improved estimate** | 34.9% | 65.1% |

Source: Compiled by Science-Metrix using NamSor and the Directory of the European Union 2015, PDF version, http://europa.eu/whoiswho/public/

These results provide a high level of confidence for the API, at least in a European context. Although this directory included employees from various cultural backgrounds, they were mostly European. The list included very few names of Asian origin, and as previously discussed, these names are the most ambiguous. It was therefore necessary to test the robustness of the API in providing gender on a data set with a more balanced geographical distribution, or at least a data set more closely aligned with what is found in the databases of scientific publications. The most extensive and balanced data set found was the list of Olympic medalists from 1896 to 2008.[29] It offers a good international representation, and because the country of the athlete is recorded, it can support an assessment of the accuracy and recall for various countries. Table II presents the results of the estimation for all medalists, and for the United States and China. Overall, the estimation is very close to the real values. The estimation based on NamSor data is that 32.3% of medalists on the list are women, whereas the figure is in fact 31.4%. NamSor was unable to identify a gender for only 5% of the medalists (786). When the proportions were computed using the NamSor score, the estimation slightly improved (31.9% for the proportion of women).

[29] Retrieved from www.theguardian.com/sport/datablog/2012/jun/25/olympic-medal-winner-list-data

The estimation for U.S. medalists is also good, with a first estimation of the proportion of women at 37.4% when it is in fact 37.9%. The precision and recall are very high for the U.S. When trying to improve the estimations by using the weighted score, the estimations of the proportions are a little less accurate.

When examining the estimation for Chinese athletes, however, the difficulty in inferring the gender of Chinese names is confirmed. In fact, the tool could provide a gender for less than 17% of the names. When examining the few cases for which the tool could assign a gender, the precision and recall are low, which results in a significant underestimation of the proportion of women (real = 62.2%, estimate = 51.2%). However, when using the score provided by the API to weight the attributions, the estimated proportion of women (and men) is fairly close to the real value. The improved estimated proportion of women is 60.9%, which is quite close the 62.2% observed.

Table II Validation of NamSor API using data from Olympic medalists (1960–2008)

**All countries**

| Gender (real) | Gender (estimate) W | Gender (estimate) M | TOTAL |
|---|---|---|---|
| W | 4,045 | 135 | 4,180 |
| M | 251 | 8,871 | 9,122 |
| Total | 4,296 | 9,006 | 13,302 |

| | W | M |
|---|---|---|
| Real | 31.4% | 68.6% |
| Estimate | 32.3% | 67.7% |
| Recall | 96.8% | 97.2% |
| Precision | 94.2% | 98.5% |
| Improved estimate | 31.9% | 68.1% |

**United States**

| Gender (real) | Gender (estimate) W | Gender (estimate) M | TOTAL |
|---|---|---|---|
| W | 632 | 27 | 659 |
| M | 18 | 1,062 | 1,080 |
| Total | 650 | 1,089 | 1,739 |

| | W | M |
|---|---|---|
| Real | 37.9% | 62.1% |
| Estimate | 37.4% | 62.6% |
| Recall | 95.9% | 98.3% |
| Precision | 97.2% | 97.5% |
| Improved estimate | 36.4% | 63.6% |

**China**

| Gender (real) | Gender (estimate) W | Gender (estimate) M | TOTAL |
|---|---|---|---|
| W | 31 | 20 | 51 |
| M | 11 | 20 | 31 |
| Total | 42 | 40 | 82 |

| | W | M |
|---|---|---|
| Real | 62.2% | 37.8% |
| Estimate | 51.2% | 48.8% |
| Recall | 60.8% | 64.5% |
| Precision | 73.8% | 50.0% |
| Improved estimate | 60.9% | 39.1% |

Source: Compiled by Science-Metrix using NamSor and a list of Olympic medalists from www.theguardian.com/sport/datablog/2012/jun/25/olympic-medal-winner-list-data

It appears that NamSor is generally accurate, especially when using the weighted score to improve estimations. Table III presents a comparison of the estimated value versus the real value for the proportion of women among the medalists for a selection of countries. Again, because the objective here is to assess the accuracy of the attribution, only the medalists for which the API is able to provide a gender are kept for the comparison.

The quality of the estimation stands at country level. Although the difference is a little more pronounced for some countries (Table III). The worst case presented in the table is New Zealand, for which the estimation is 2.2 percentage points below the real value—a point estimation that falls within a 95% confidence interval for the API (see Section 2.5).

Table III Accuracy of the estimation of proportion of women among Olympic medalists using NamSor for a selection of countries

| Country | Proportion of women | | |
|---|---|---|---|
| | Real | Estimate | Δ |
| **Overall** | **31.4%** | **31.9%** | **0.5%** |
| France | 17.9% | 19.5% | 1.6% |
| Bulgaria | 35.7% | 38.0% | 2.3% |
| Rep. of Korea | 41.2% | 43.4% | 2.2% |
| Russia | 47.5% | 50.0% | 2.5% |
| Ukraine | 48.4% | 50.4% | 2.0% |
| Poland | 16.7% | 17.3% | 0.6% |
| Argentina | 26.4% | 27.3% | 0.9% |
| Great Britain | 27.2% | 28.0% | 0.8% |
| Spain | 17.4% | 17.9% | 0.5% |
| Romania | 50.3% | 51.6% | 1.4% |
| Brazil | 28.6% | 29.4% | 0.8% |
| Sweden | 20.5% | 21.0% | 0.5% |
| Italy | 13.4% | 13.7% | 0.3% |
| Denmark | 32.7% | 33.4% | 0.7% |
| Germany | 42.1% | 42.7% | 0.5% |
| Netherlands | 51.6% | 51.6% | 0.0% |
| Japan | 36.2% | 36.0% | -0.2% |
| Norway | 66.9% | 66.2% | -0.7% |
| Canada | 42.9% | 42.2% | -0.7% |
| Australia | 41.3% | 40.6% | -0.8% |
| China | 62.2% | 60.9% | -1.3% |
| Hungary | 23.8% | 23.0% | -0.7% |
| United States | 37.9% | 36.4% | -1.5% |
| Cuba | 21.3% | 19.5% | -1.9% |
| New Zealand | 13.4% | 11.2% | -2.2% |

Source: Compiled by Science-Metrix using NamSor and a list of Olympic medalists from www.theguardian.com/sport/datablog/2012/jun/25/olympic-medal-winner-list-data

## 2.4 Name disambiguation in publication databases

In order to conduct a large-scale assessment of the potential to develop new bibliometric indicators on the gender of scientific authors, Science-Metrix aimed to determine the gender of all authors on all documents indexed in the WoS (1980–) and Scopus (1996–). In total, 161,898,256 complete names (first name + surname) were extracted from the two databases. After cleaning and de-duplication, 14,353,648 combinations of first name and last name remained. To save time and money (NamSor charges for each combination), very safe first names were tested without the API to determine their gender in a first round. For example, it was not necessary to use NamSor to determine that Michael, David, Peter and John are men and Maria, Barbara, Christine and Jennifer are women.

Science-Metrix used statistics on the first names of all people who had asked for a social security number in the U.S. since 1950, in an effort to identify the likelihood of a given name being associated with a woman or a man. To limit the bias inherent in using a U.S. database, similar statistics were derived from the data on nearly 2 million names that were previously genderized for the She Figures study. This latter

data set is international in scope. All the names that had more than 98% of their instances associated with one of the genders in the two databases were identified.

The resulting list of first names was used to determine the gender of nearly 7 million of the 14.35 million complete name combinations. A sample of 20,000 combinations (10,000 women and 10,000 men) was genderized in NamSor to compare the attribution of the API with the attribution based on the known first names. The results were highly similar: in the infrequent cases where the two approaches were not providing the same result, the number of errors with the approach based on the known first name was equivalent to the number of errors resulting from NamSor.

The gender of the remaining 7.5 million combinations was determined using the NamSor API, making it possible to provide a gender for 88% of the 14.35 million combinations; the remaining 12% were coded as unknown. This information was then used to code all authors on all papers in Scopus and the WoS. Authors are coded as woman, man, unknown or non-genderizable.

Presented in Table IV is a statistical overview of the capacity of each database to support the genderization of authors. Data were only presented from 2006 to 2015 in the WoS because, as noted in the previous section, the information on the first name of authors and the addresses of authors is not available prior to 2006. Data are available in Scopus dating back to the inception of the database in 1996. In any case, both databases have encouraging numbers for the 2007–2015 period. Although the share of genderizable author names is a little lower in the WoS, the proportion of authorships for which a gender is derived is highly similar in both databases and stable at roughly 60%. Indeed, even when the full name is available, the gender cannot necessarily be inferred. Most of the time, this is because the name is ambiguous, usually because the name is used as frequently by men as women. It can also be that the name is very uncommon, and the algorithm is not capable of guessing a gender with a sufficient level of certainty.

It seems that although the availability of author first names is increasing, the increasing proportion of Chinese authors is thwarting this improvement in gender identification because of a high proportion of ambiguous Chinese names. And although perhaps eventually all first names will be available, it is now time to start the development of novel methods to identify the gender of Chinese names and those of other ethnicities for which the results are unreliable.

Please note that Table IV also presents the share of women and men in authorship by year. However, these data have not been corrected for the wide variation in the number of genderized names between the various fields, countries and years. The results presented in Section 3 take this into account and are thus more reliable than the raw results presented here.

Table IV Availability of given name (genderizable), proportion of authorship genderized (known gender) and preliminary statistics on proportions of women and men in the Web of Science and Scopus

**Web of Science**

| Year | Papers | Authorships (AU) | Authors/ Paper | Genderizable (GDZ) | | Known Gender (KG) | | | Women | | | | Men | | | | Ambiguous | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | n | % of AU | n | % of AU | % of GDZ | n | % of AU | % of GDZ | % of KG | n | % of AU | % of GDZ | % of KG | n | % of AU | % of GDZ |
| 2006 | 980,477 | 4,295,038 | 4.38 | 1,654,664 | 38.5% | 1,427,377 | 33.2% | 86.3% | 412,286 | 9.6% | 24.9% | 28.9% | 1,015,091 | 23.6% | 61.3% | 71.1% | 227,287 | 5.3% | 13.7% |
| 2007 | 1,050,083 | 4,586,885 | 4.37 | 3,282,938 | 71.6% | 2,823,458 | 61.6% | 86.0% | 840,753 | 18.3% | 25.6% | 29.8% | 1,982,705 | 43.2% | 60.4% | 70.2% | 459,480 | 10.0% | 14.0% |
| 2008 | 1,129,441 | 4,952,147 | 4.38 | 3,614,497 | 73.0% | 3,077,655 | 62.1% | 85.1% | 939,977 | 19.0% | 26.0% | 30.5% | 2,137,678 | 43.2% | 59.1% | 69.5% | 536,842 | 10.8% | 14.9% |
| 2009 | 1,183,706 | 5,283,497 | 4.46 | 3,938,012 | 74.5% | 3,313,561 | 62.7% | 84.1% | 1,031,078 | 19.5% | 26.2% | 31.1% | 2,282,483 | 43.2% | 58.0% | 68.9% | 624,451 | 11.8% | 15.9% |
| 2010 | 1,226,929 | 5,691,863 | 4.64 | 4,281,031 | 75.2% | 3,572,897 | 62.8% | 83.5% | 1,131,541 | 19.9% | 26.4% | 31.7% | 2,441,356 | 42.9% | 57.0% | 68.3% | 708,134 | 12.4% | 16.5% |
| 2011 | 1,308,110 | 6,408,631 | 4.90 | 4,742,711 | 74.0% | 3,903,912 | 60.9% | 82.3% | 1,257,598 | 19.6% | 26.5% | 32.2% | 2,646,314 | 41.3% | 55.8% | 67.8% | 838,799 | 13.1% | 17.7% |
| 2012 | 1,375,335 | 7,177,501 | 5.22 | 5,225,136 | 72.8% | 4,240,632 | 59.1% | 81.2% | 1,388,553 | 19.3% | 26.6% | 32.7% | 2,852,079 | 39.7% | 54.6% | 67.3% | 984,504 | 13.7% | 18.8% |
| 2013 | 1,451,327 | 7,531,312 | 5.19 | 5,734,421 | 76.1% | 4,567,630 | 60.6% | 79.7% | 1,515,679 | 20.1% | 26.4% | 33.2% | 3,051,951 | 40.5% | 53.2% | 66.8% | 1,166,791 | 15.5% | 20.3% |
| 2014 | 1,490,237 | 7,839,751 | 5.26 | 6,097,782 | 77.8% | 4,754,163 | 60.6% | 78.0% | 1,593,179 | 20.3% | 26.1% | 33.5% | 3,160,984 | 40.3% | 51.8% | 66.5% | 1,343,619 | 17.1% | 22.0% |
| 2015 | 1,455,361 | 7,903,934 | 5.43 | 6,186,711 | 78.3% | 4,736,841 | 59.9% | 76.6% | 1,608,721 | 20.4% | 26.0% | 34.0% | 3,128,120 | 39.6% | 50.6% | 66.0% | 1,449,870 | 18.3% | 23.4% |

**Scopus**

| Year | Papers | Authorships (AU) | Authors/ Paper | Genderizable (GDZ) | | Known Gender (KG) | | | Women | | | | Men | | | | Ambiguous | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | n | % of AU | n | % of AU | % of GDZ | n | % of AU | % of GDZ | % of KG | n | % of AU | % of GDZ | % of KG | n | % of AU | % of GDZ |
| 1996 | 923,659 | 3,201,125 | 3.47 | 1,610,328 | 50.3% | 1,486,878 | 46.4% | 92.3% | 349,249 | 10.9% | 21.7% | 23.5% | 1,137,629 | 35.5% | 70.6% | 76.5% | 123,450 | 3.9% | 7.7% |
| 1997 | 949,662 | 3,362,015 | 3.54 | 1,710,665 | 50.9% | 1,572,540 | 46.8% | 91.9% | 375,925 | 11.2% | 22.0% | 23.9% | 1,196,615 | 35.6% | 70.0% | 76.1% | 138,125 | 4.1% | 8.1% |
| 1998 | 954,944 | 3,429,186 | 3.59 | 1,791,302 | 52.2% | 1,636,482 | 47.7% | 91.4% | 398,166 | 11.6% | 22.2% | 24.3% | 1,238,316 | 36.1% | 69.1% | 75.7% | 154,820 | 4.5% | 8.6% |
| 1999 | 965,222 | 3,509,623 | 3.64 | 1,895,528 | 54.0% | 1,722,582 | 49.1% | 90.9% | 432,697 | 12.3% | 22.8% | 25.1% | 1,289,885 | 36.8% | 68.0% | 74.9% | 172,946 | 4.9% | 9.1% |
| 2000 | 1,013,189 | 3,732,268 | 3.68 | 1,829,671 | 49.0% | 1,648,457 | 44.2% | 90.1% | 411,032 | 11.0% | 22.5% | 24.9% | 1,237,425 | 33.2% | 67.6% | 75.1% | 181,214 | 4.9% | 9.9% |
| 2001 | 1,034,607 | 3,848,014 | 3.72 | 1,705,746 | 44.3% | 1,510,315 | 39.2% | 88.5% | 383,811 | 10.0% | 22.5% | 25.4% | 1,126,504 | 29.3% | 66.0% | 74.6% | 195,431 | 5.1% | 11.5% |
| 2002 | 1,083,671 | 4,076,342 | 3.76 | 2,749,549 | 67.5% | 2,414,483 | 59.2% | 87.8% | 625,271 | 15.3% | 22.7% | 25.9% | 1,789,212 | 43.9% | 65.1% | 74.1% | 335,066 | 8.2% | 12.2% |
| 2003 | 1,163,320 | 4,428,525 | 3.81 | 3,059,138 | 69.1% | 2,653,304 | 59.9% | 86.7% | 701,487 | 15.8% | 22.9% | 26.4% | 1,951,817 | 44.1% | 63.8% | 73.6% | 405,834 | 9.2% | 13.3% |
| 2004 | 1,296,211 | 5,012,165 | 3.87 | 3,561,718 | 71.1% | 3,007,139 | 60.0% | 84.4% | 796,879 | 15.9% | 22.4% | 26.5% | 2,210,260 | 44.1% | 62.1% | 73.5% | 554,579 | 11.1% | 15.6% |
| 2005 | 1,484,047 | 5,777,781 | 3.89 | 4,196,847 | 72.6% | 3,437,891 | 59.5% | 81.9% | 923,605 | 16.0% | 22.0% | 26.9% | 2,514,286 | 43.5% | 59.9% | 73.1% | 758,956 | 13.1% | 18.1% |
| 2006 | 1,586,737 | 6,262,054 | 3.95 | 4,584,025 | 73.2% | 3,708,763 | 59.2% | 80.9% | 1,023,693 | 16.3% | 22.3% | 27.6% | 2,685,070 | 42.9% | 58.6% | 72.4% | 875,262 | 14.0% | 19.1% |
| 2007 | 1,687,677 | 6,758,090 | 4.00 | 5,023,032 | 74.3% | 4,019,671 | 59.5% | 80.0% | 1,130,156 | 16.7% | 22.5% | 28.1% | 2,889,515 | 42.8% | 57.5% | 71.9% | 1,003,361 | 14.8% | 20.0% |
| 2008 | 1,788,987 | 7,180,347 | 4.01 | 5,458,109 | 76.0% | 4,304,169 | 59.9% | 78.9% | 1,241,408 | 17.3% | 22.7% | 28.8% | 3,062,761 | 42.7% | 56.1% | 71.2% | 1,153,940 | 16.1% | 21.1% |
| 2009 | 1,902,144 | 7,716,151 | 4.06 | 5,968,948 | 77.4% | 4,642,440 | 60.2% | 77.8% | 1,364,375 | 17.7% | 22.9% | 29.4% | 3,278,065 | 42.5% | 54.9% | 70.6% | 1,326,508 | 17.2% | 22.2% |
| 2010 | 2,011,013 | 8,328,408 | 4.14 | 6,492,141 | 78.0% | 4,999,158 | 60.0% | 77.0% | 1,493,645 | 17.9% | 23.0% | 29.9% | 3,505,513 | 42.1% | 54.0% | 70.1% | 1,492,983 | 17.9% | 23.0% |
| 2011 | 2,156,203 | 9,217,743 | 4.27 | 7,178,292 | 77.9% | 5,472,296 | 59.4% | 76.2% | 1,674,635 | 18.2% | 23.3% | 30.6% | 3,797,661 | 41.2% | 52.9% | 69.4% | 1,705,996 | 18.5% | 23.8% |
| 2012 | 2,232,912 | 10,090,486 | 4.52 | 7,732,517 | 76.6% | 5,885,109 | 58.3% | 76.1% | 1,831,376 | 18.1% | 23.7% | 31.1% | 4,053,733 | 40.2% | 52.4% | 68.9% | 1,847,408 | 18.3% | 23.9% |
| 2013 | 2,314,550 | 10,462,374 | 4.52 | 8,332,036 | 79.6% | 6,262,408 | 59.9% | 75.2% | 1,985,530 | 19.0% | 23.8% | 31.7% | 4,276,878 | 40.9% | 51.3% | 68.3% | 2,069,628 | 19.8% | 24.8% |
| 2014 | 2,318,257 | 10,787,377 | 4.65 | 8,647,846 | 80.2% | 6,429,953 | 59.6% | 74.4% | 2,074,898 | 19.2% | 24.0% | 32.3% | 4,355,055 | 40.4% | 50.4% | 67.7% | 2,217,893 | 20.6% | 25.6% |
| 2015 | 1,643,360 | 8,132,251 | 4.95 | 6,540,139 | 80.4% | 4,899,252 | 60.2% | 74.9% | 1,636,349 | 20.1% | 25.0% | 33.4% | 3,262,903 | 40.1% | 49.9% | 66.6% | 1,640,887 | 20.2% | 25.1% |

Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics) and Scopus (Elsevier)

When examining the proportion of authors for which a gender is inferred for the top 20 leading countries (the countries producing the most papers in the 2006–2015 period), important variations can be observed (Table V). China has the lowest proportion of defined gender (17%), even though it is the country with the lowest share of authors for which a full name is not recorded (7%). This is because the API is not able to infer a gender from the majority of Chinese names. Similarly, Korean names are difficult for the API to genderize.

For the other countries presented, the potential of inferring a gender is mostly determined by the availability of full names. Russia has the largest share of authors for which the given name is not recorded (77%), followed by Switzerland, India, France and Italy. These variations must be taken into account when examining gender equity at world level. For example, with only 17% of names for which a gender can be inferred, China's score would not be weighted appropriately in the world total if only the genderized authors are used in the calculation.

Table V Gender identification in Scopus for a selection of countries (2006–2015) (top 20 countries with most publications in the period)

| Country | Papers | Papers.Authors | Known Gender (Woman or Man) | Unisex | Non-Genderizable |
|---|---|---|---|---|---|
| WORLD | 19,641,840 | 84,935,281 | 60% | 18% | 22% |
| United States | 4,934,146 | 17,563,237 | 74% | 9% | 17% |
| China | 3,724,801 | 14,966,013 | 17% | 76% | 7% |
| United Kingdom | 1,357,912 | 3,959,097 | 67% | 3% | 30% |
| Germany | 1,295,761 | 4,565,462 | 65% | 1% | 33% |
| Japan | 1,167,430 | 5,252,734 | 83% | 2% | 14% |
| France | 947,169 | 3,369,799 | 62% | 1% | 37% |
| India | 784,201 | 2,523,138 | 59% | 1% | 41% |
| Canada | 771,521 | 2,199,932 | 75% | 6% | 19% |
| Italy | 767,525 | 3,428,248 | 63% | 0% | 37% |
| Spain | 649,784 | 2,378,312 | 66% | 0% | 33% |
| Australia | 601,341 | 1,690,765 | 75% | 6% | 19% |
| Rep. of Korea | 584,414 | 2,422,565 | 28% | 63% | 10% |
| Brazil | 475,886 | 1,920,824 | 81% | 1% | 18% |
| Netherlands | 428,185 | 1,405,872 | 71% | 1% | 28% |
| Russia | 402,333 | 1,375,696 | 21% | 2% | 77% |
| Switzerland | 316,755 | 1,147,753 | 56% | 1% | 43% |
| Poland | 299,478 | 909,526 | 74% | 0% | 26% |
| Turkey | 298,438 | 1,053,454 | 83% | 0% | 17% |
| Iran | 288,270 | 913,863 | 70% | 1% | 29% |
| Sweden | 279,863 | 787,985 | 76% | 3% | 21% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

Table VI also shows important variations between fields of research in the proportion of authorships for which a gender can be inferred.[30] Although the gender is defined for more than 90% of the authors in psychology, social sciences and arts & humanities, the capacity to infer a gender is very low in physics & astronomy. For this field, the full name is not provided for 60% of authors. And within this field, some subfields exhibit an even lower availability of full names. For example, the lowest availability of given names is in the subfield of nuclear & particle physics, for which only about 10% of the authorships include a full given name. Variation in the availability of given names appears to be linked to editorial decisions at the journal level, which also appears to be influenced by field traditions.

Important differences can also be observed between fields in the proportion of ambiguous names. Prevalence is particularly high in applied physics. Because countries specialize in some fields, the proportion of given names available and the proportion of ambiguous names at field level is at least partially determined by these proportions at country level and vice versa. For example, the high prevalence of ambiguous names in applied sciences can be explained, at least partially, by the strong specialization of China and the Republic of Korea in this domain. And, somewhat similarly, the high proportion of unavailable first names for Russian authors can likely be linked to Russia's significant specialization in physics, a field where the given name is often not recorded.

[30] The fields of research used and presented in this report stem from a journal-based classification of research developed by Science-Metrix. All papers in Scopus and the Web of Science are classified into 6 domains, 22 fields and 176 subfields. For more details about this classification, please visit: http://www.science-metrix.com/en/classification

Table VI Gender identification in Scopus, by field (2006–2015)
(top 20 countries with most publications in the period)

| Domain / Field | papers | papers.authors | Known Gender (Woman or Man) | Ambiguous | Non-Genderizable |
|---|---|---|---|---|---|
| TOTAL | 19,641,840 | 84,935,281 | 60% | 18% | 22% |
| Natural Sciences | 4,970,785 | 24,101,475 | 44% | 17% | 39% |
| Biology | 685,839 | 2,726,496 | 67% | 12% | 20% |
| Chemistry | 1,272,000 | 5,821,004 | 62% | 27% | 12% |
| Earth & Environmental Sciences | 591,381 | 2,427,210 | 54% | 18% | 28% |
| Mathematics & Statistics | 454,960 | 992,230 | 58% | 21% | 21% |
| Physics & Astronomy | 1,966,605 | 12,134,535 | 28% | 12% | 60% |
| Applied Sciences | 6,671,842 | 24,189,663 | 52% | 28% | 20% |
| Agriculture, Fisheries & Forestry | 587,263 | 2,530,081 | 56% | 13% | 31% |
| Built Environment & Design | 158,927 | 445,856 | 63% | 19% | 18% |
| Enabling & Strategic Technologies | 2,016,339 | 8,404,515 | 45% | 32% | 23% |
| Engineering | 2,045,865 | 6,976,457 | 49% | 31% | 20% |
| ICT | 1,863,448 | 5,832,754 | 63% | 28% | 9% |
| Health Sciences | 6,195,163 | 31,838,827 | 74% | 12% | 15% |
| Biomedical Research | 1,180,508 | 6,631,056 | 74% | 14% | 12% |
| Clinical Medicine | 4,209,235 | 22,251,847 | 72% | 12% | 16% |
| Psychology & Cognitive Sciences | 333,226 | 1,073,147 | 91% | 4% | 5% |
| Public Health & Health Services | 472,194 | 1,882,777 | 86% | 5% | 9% |
| Economic & Social Sciences | 1,094,646 | 2,342,627 | 86% | 9% | 4% |
| Economics & Business | 471,882 | 1,073,548 | 82% | 14% | 4% |
| Social Sciences | 622,764 | 1,269,079 | 90% | 6% | 4% |
| Arts & Humanities | 356,739 | 522,530 | 92% | 4% | 4% |
| General | 352,665 | 1,940,159 | 70% | 25% | 5% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

## 2.5 Estimation of the proportion of women

Because the full name is not available for all authors, and because of the imprecision of the genderization tools, it is not possible to measure the real proportion of women's authorship in publications, but only an estimation of this proportion. Therefore, the statistics have more analytical value if accompanied by confidence intervals. These provide a sense of the level of certainty associated with the different statistics prepared across countries, disciplines and years. Two major sources of error have been identified in the estimation of the proportion of women: measurement error and sampling error.

### Measurement error

The first source of error is related to the accuracy of the approach to determine the gender of each author based on the combination of first name and last name: inaccuracy causes measurement error. One of the benefits of using NamSor is that it already includes a built-in measure of accuracy. As mentioned in Section 2.3, results from the NamSor API not only provide the most likely gender for a given combination of first and last name, they also provide a statistic on the reliability of this attribution. This statistic ranges from -1 to 1, with 1 being the highest certainty that the name belongs to a woman and -1 being the highest

level of certainty that the name belongs to a man. For any value between -0.1 and 0.1, no gender is provided.[31] The authors for which a gender is provided constitute the sample. A linear transformation of this statistic is used as an approximation of the probability for each author to be a woman. These probabilities will be used to calculate the point estimates and a margin of error of the proportion of women in the sample.

### Sampling error

The second source of errors is sampling error. Because the first name is not available for all authors, and because NamSor is not able to provide a gender for all combinations, the proportion of women in the population is estimated on the sample of authors for which a gender can be inferred. Importantly, this sample of authors is not picked randomly. Inclusion in the sample is determined by the availability of the full name and the capacity of NamSor to identify the gender from this name. Hence, this is a non-probability sample, and extrapolation to the population cannot be made by simply using standard probability-sampling approaches.

Even if the proportion of authors for which the gender can be derived in Scopus offers a very large sample to estimate the global population, there is a risk that the sample is not representative of this population. There are two categories of authors for which the gender is not identified:

- those for which the full name is available but the gender cannot be identified by the tool; and
- those for which the full name is not available.

For the cases for which the full name is recorded but the API is not able to determine a gender, it is possible that the prevalence of ambiguous names is not distributed equally between men and women. For example, it could be that more women have ambiguous names than men do, or that women more often have uncommon names for which the API cannot determine a gender.

Science-Metrix has not thoroughly tested these hypotheses yet, but a first inspection was conducted to assess potential representativity of the sample. The gender of 1,056 authors, selected randomly from all the authors in Scopus, was validated manually. In this test sample, 26.3% of authors coded as unisex were in fact women, a proportion that is highly similar to what is observed for the whole sample (27.2%). This finding does not prove that the gender of authors with ambiguous names is distributed similarly to the overall population, but it supports the idea that if there were a bias, it would be small.

For most countries, the percentage of authors with names of ambiguous gender is low, and this assumption, if proved wrong, would only slightly affect the accuracy of the estimation. Thus, it seems that the impact of a potential sampling error here is low. This conclusion applies even to China, for which ambiguous authors represent 76% of the population, because in the test sample representativity was even higher for Chinese authors: 21.9% of Chinese authors with an ambiguous name were women, whereas 22.0% of Chinese authors in the overall sample were women. In this paper, Science-Metrix assumes that this bias is negligible, and that the sample is representative of the authors with ambiguous names. Still, this validation has only been performed using a fairly small test sample. Manual validation using a stratified

[31] When the gender has been inferred solely on the first name (see Section 2.4) the probability associated with the first name is used. The robustness of these attributions is high and possible values range between 0.98 and 1.

random sampling, with appropriate sample size for various strata, would be useful to validate that the representativity holds for all subfields, countries and years. Such extensive validation would require a huge amount of time and is beyond the scope of the present work.

For the cases for which the full name is not available, it is possible that these cases are not distributed similarly between men and women. If this is the case, it is not possible to make a robust inference from these measurements to the whole population (in this case, all authorships in Scopus). In many cases, the author provides a full name, and the first name is dropped further along in the journal publishing process or in the collection of data for the bibliographic databases. This deletion of the given name does not have any obvious link with a preference of the authors themselves, so it should be gender-neutral. However, there are also likely cases when the author does not provide their first name. There is no easy means to systematically assess the occurrence of voluntary omissions by the author as opposed to the deletion or truncation of first-name data, which would be taking place later in the processing of metadata by the journal editor or the database publisher.

For Scopus as a whole, we have counted the number of papers where the given names of co-authors are either all recorded, all not recorded, or a mix of the two. In more than 95% of the papers with at least two authors, the given names were either all recorded or not recorded. So, it seems that the presence of the full name in Scopus is more determined by the way metadata are handled from the submission of a paper to its indexation in the database than by a personal decision of the author. Still, there are about 5% of papers for which the full name is recorded only for some of the co-authors but is not recorded for others. This is likely a telltale sign of voluntary omission.

In a statistical context, these "anonymous" authors are highly similar to the non-respondents in a survey setting. Their decision not to provide their first name may or may not be related to their gender. Hypotheses can be put forward, though, that would connect such a decision to gender. For example, it is possible that some women would not reveal their first name when submitting a paper out of a fear of gender discrimination by the reviewers. A study based on the frequency distribution of the first letter of names for men and women in authorship supports this hypothesis.[32] The author points to an important caveat to her analysis, though. The protocol assumes that the first-initial frequency distribution is homogeneous across countries and fields, which is unlikely to be the case.

Although historically inference to the population necessitated random sampling, because it is seldom possible to obtain a truly random sample, new methods are emerging to infer statistics from non-probability samples.[33,34] In particular, various approaches have been developed to assess and reduce

[32] https://fivethirtyeight.com/datalab/are-female-scientists-hiding/

[33] Baker, R. et al. (2013) Summary report of the AAPOR Task Force on Non-probability Sampling, *Journal of Survey Statistics and Methodology, 1*(2), 90–143. Retrieved from https://doi.org/10.1093/jssam/smt008

[34] Brick, J. M. (2011). The future of survey sampling. *Public Opinion Quarterly, 75*(5), 72–888. Retrieved from https://doi.org/10.1093/poq/nfr045

nonresponse bias in surveys.[35] It some cases, the nonprobability sample is compared with a smaller random sample to assess the extent of the nonresponse bias.[36]

In order to verify whether the relative occurrence of women (and men) is similar when the full name is recorded and when it is not, the same test sample of 1,056 authors was used. Within the whole test sample, 27.2% of authors are women, whereas 23.1% of authors with no full name recorded are women. The difference, although not huge, is still non-negligible. However, because of the relatively small sample size, part of the difference observed can be attributed to a sampling error in the testing protocol.[37] Therefore, it is impossible on this basis to reject the hypothesis that women may be systematically more often omitting their full first name. If that hypothesis holds, the percentage of women would be underestimated. However, because voluntary omission by authors of their first name appears to be infrequent, and because the proportion of women in anonymous authorships is similar to the proportion of women in the overall sample, any such bias would be expected to have only have a small effect on the precision of the estimations calculated here.

Importantly, a substantial share of the correlation between the gender and the availability of the given name is not explained by a direct relationship between these parameters, but rather by collinearity of these two parameters with other variables. Those other variables explain the apparent connection. As presented above, the availability of full names varies across countries, years and subfields. The proportion of women also varies across these same variables. If, for example, there are more women involved in authorship in countries and/or in subfields for which the number of genderizable authors is lower, then the proportion of women in the study population would be underestimated.

For this reason, Science-Metrix uses a post-sampling stratification. Each stratum is the combination of a specific country in a specific subfield for a specific publication year. The data set used for this study covers 199 countries, 159 subfields and 10 years, for a total of 163,492 strata with at least one publication. The proportion of women is calculated for each stratum and then aggregated to higher levels with a weighting based on the relative size of the strata within the population. Figure 5 presents the proportion of women in Scopus authorships with and without the re-weighting by strata. It seems that the post-sampling stratification corrects for an under-sampling of women authors, and thus the proportion of women in Scopus overall is higher than in the sample for which the gender can be determined.

[35] National Research Council. (2013). Chapter 2: Nonresponse Bias. *Nonresponse in Social Science Surveys: A Research Agenda*, pp. 40–50. Washington, D.C.: The National Academies Press. Retrieved from https://www.nap.edu/read/18293/chapter/4#49

[36] Billiet, J., Philippens, M., Fitzgerald, R., & Stoop, I. (2007) Estimation of nonresponse bias in the European Social Survey: Using information from reluctant respondents. *Journal of Official Statistics, 23*(2), 135–162.

[37] Only 1,056 authorships are included in the test sample for a population of 85 million authorships, which leads to large margins of error.

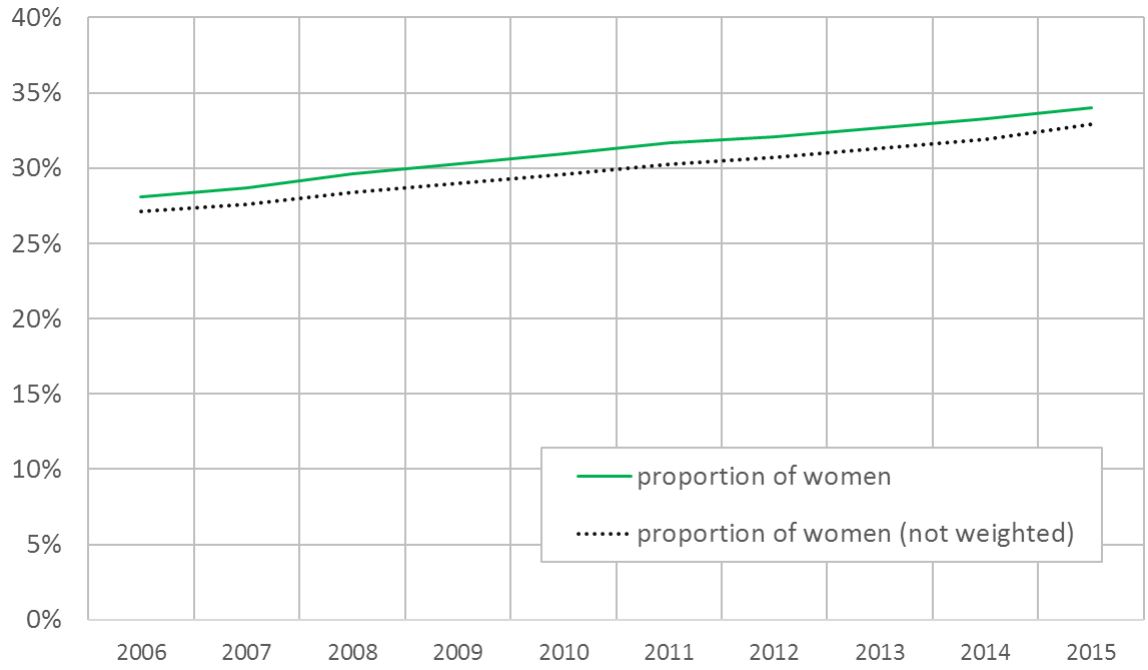



Figure 5 Effect of weighting when calculating the proportion of women at aggregated level

Note: Underlying data available in Table XVII.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

It would be possible to investigate more thoroughly the potential biases linked to this nonprobability sampling approach; however, such investigations would require substantial investment in terms of time and resources. Especially given that the full name is now recorded on more than 95% of authorships when using fractional counting, and that this proportion is increasing rapidly, such investments do not seem warranted at this point. In addition, some promising avenues are presented in Section 4.1 to improve the gender identification in the databases for the cases where the full name is not recorded. In sum, we will soon be able to work on the full population (or nearly so), with little need for calculation of sampling errors.

For this report, we suggest assuming that the samples are unbiased, and a simple calculation of the margins of error can be used based on sample size, proportion of each gender in the sample, and population size. Statistics, and margins of error, are then aggregated to prepare indicators at higher levels. This approach leads to relatively higher margins of error at disaggregated levels. Usually, a sample size of about 1,000 items leads to a margin of error of about 3 percentage points when the population is large—above 100,000, for instance. However, for smaller populations, sampling a large share of the population will lead to margins higher than 3 percentage points. For instance, for a population of 100, sampling half the population (50 items) will lead to a margin of error of roughly 10 percentage points.

An interval estimate is computed by combining the margins of error from both the measurement and the sampling errors. A point estimate and interval estimate are calculated for each stratum and then aggregated at a higher level with a weighting based on the relative size of the strata within the population. Because the margin of error calculated for the measurement is derived from an approximation of the

probability of the gender attribution, and not from a real probability, and because it can't be confirmed that the samples are not biased, this interval estimate will be named "reliability interval" instead of "confidence interval" to avoid confusion.

## 2.6 Calculation of proportions and reliability intervals

The calculation of proportion of men and women is based on fractional counting. With fractional counting, each author on a paper is attributed the same fraction of the paper. For example, on a paper with four authors, each author is attributed 0.25 of a paper. This way, on a paper authored by 99 men and 1 woman, the paper would be considered as 99% men and 1% women instead of counting the paper once for men and once for women. In addition to providing a more precise measure of the contribution of men and women on a paper, using fractional counting also limits the influence of publications with several authors. For example, if a paper lists 1,000 authors and a full count is attributed to each author and their gender, the paper will have 1,000 times more weight in calculating an aggregated score on gender compared to a paper with a single author.

For each stratum (combination of year, country and subfield), the sum of fractions for each author is calculated. Following a binomial distribution, the point estimate for the proportion of women in the sample is given by

$$\hat{P}_w = \sum_{i=1}^{n} f_i p_i \Big/ \sum_{i=1}^{n} f_i$$

Where: $\hat{P}_w$ = proportion of women in the sample
$f_i$ = fraction of author i in the sample
$p_i$ = probability of being a woman for author i
n = number of genderized authors in the sample

Using the variance:

$$Var(\hat{p}_w) = \sum_{i=1}^{n} f_i\, p_i(1-p_i) \Big/ \sum_{i=1}^{n} f_i^{\,2}$$

The margin of error associated with the measurement error ($ME_m$) for a 95% reliability interval (z = 1.96) of this proportion is calculated as follows:

$$ME_m(\hat{p}_w) \;=\; 1.96\sqrt{Var(\hat{p}_w)}$$

To estimate the proportion of women for the population (all papers in a given year, subfield and country), a 95% margin of error ($ME_s$) is calculated for the sampling errors based on the point estimate of the proportion of women in the sample, the sample size (n) and the population size (N):

$$ME_s(\hat{p}_w) \;= 1.96\sqrt{\frac{\hat{p}_w(1-\hat{p}_w)(N-n)}{(N-1)n}}$$

A z-score of 1.96 is used for a 95% reliability interval. Note that a finite population correction factor is included in the computation to account for added accuracy gained by sampling a large percentage of the

population, which is of high importance in this context as the samples often account for a large share of the population.

Finally, the proportion of women in the population (Pw) is given by

$$Pw = \hat{P}_w \pm (ME_m + ME_s)$$

It is based on the point estimate for the sample, and the combination of the measurement error and sampling error.

# 3 Results

For this pilot study, Science-Metrix has computed the indicators using science and engineering (S&E) publications indexed in Scopus. This is firstly because Scopus offers a more balanced representation of the various fields of research, particularly because conference papers are included in the analysis and because Scopus offers a better representation of emerging/non-English-speaking countries. The second consideration is that Scopus makes it possible to use all authors on the publications for a longer period. Nevertheless, using the WoS starting in 2007 would provide comparable data, as shown in Table IV (and taking into account that the authors and their addresses can be linked more consistently from 2007 onward). In this section, apart from some exceptions, the results are presented with 95% reliability intervals for both the measurement error and the sampling error. A discussion on the source of errors and on the calculation of confidence intervals is presented in sections 2.5 and 2.6.

Figure 6 presents the evolution of the proportion of women in S&E publications' authorship. The proportion of women has increased constantly in the 2006–2015 period, starting at 28% in 2006 and reaching almost 34% in 2015. Although the margins of error are somewhat large, it is still possible to conclude with a high level of confidence that the proportion of women in S&E is increasing at the world level, with women accounting for about 31%–37% of authorship in 2015.

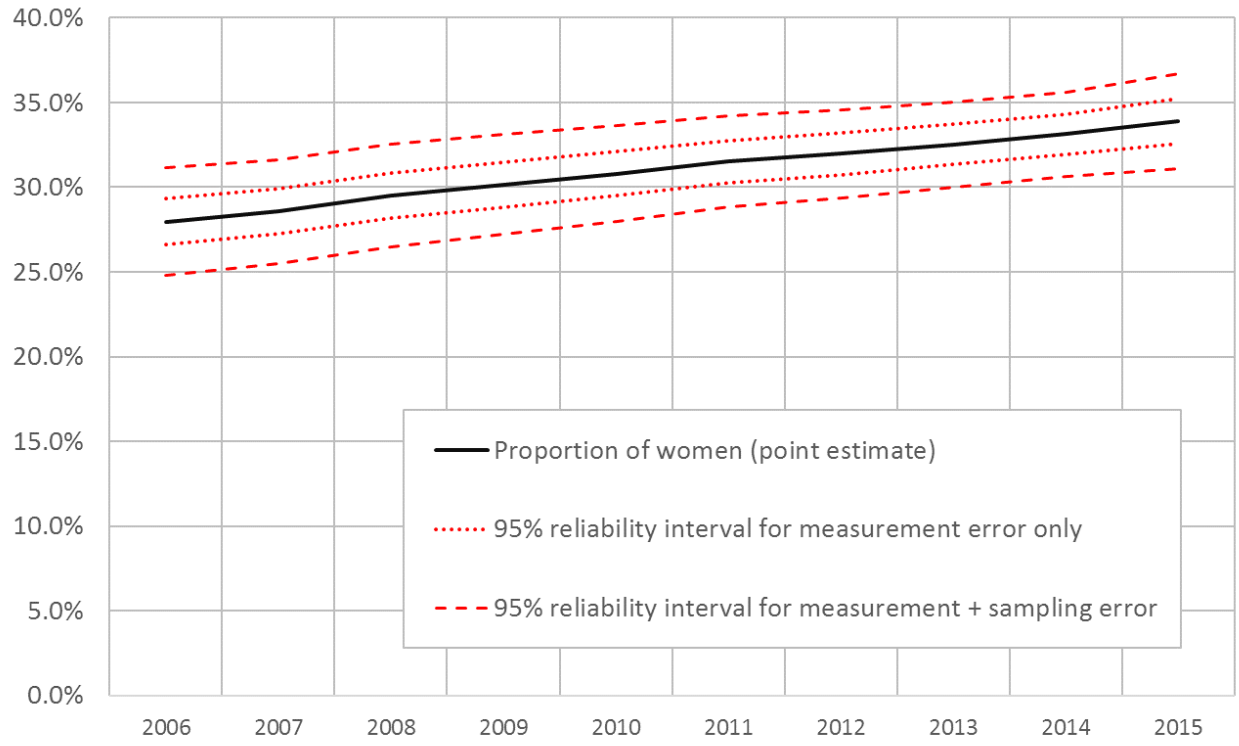



Figure 6 Trends in the proportion of women in authorship in S&E publications in Scopus

Note: Underlying data available in Table XVIII.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

For the U.S., the accuracy of the measurement is high, the proportion of genderizable names is also high, and so is the number of authorships. Therefore, the confidence intervals are smaller, as exemplified in Figure 7, which presents the trends in the proportion of women in U.S. authorships.

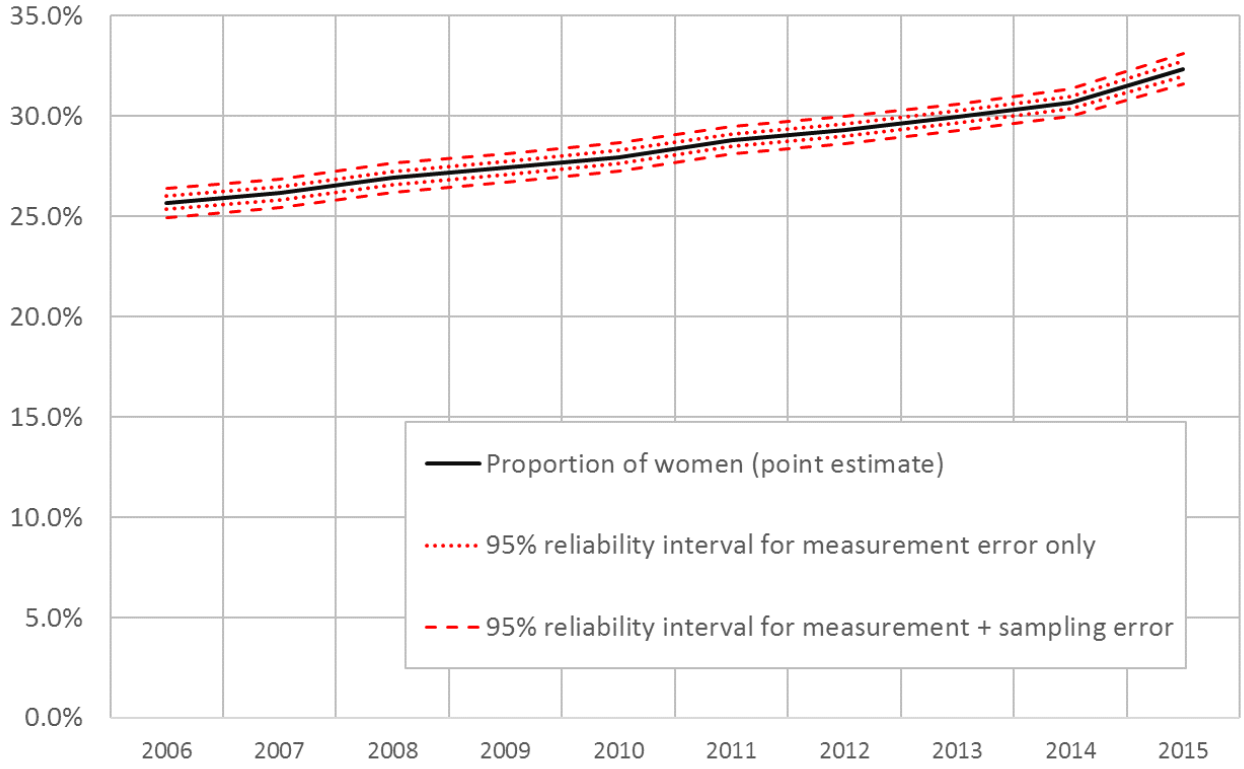



Figure 7 Trends in the proportion of women in authorship in the U.S.
Note: Underlying data available in Table XIX.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

Figure 8 presents the proportion of women in the various fields of research. The 95% reliability interval for the measurement is represented by the white rectangle, and the overall 95% reliability interval is represented by the blue band. A table presenting the data for the 158 S&E subfields is presented in Appendix A. The proportion of women is high in the fields related to health sciences, with the highest proportion in the field of public health & health services, where approximately 55% of authors were women in the 2006–2015 period. Proportions of women are also high in social sciences and in agriculture, fisheries & forestry. Within this last field, women are highly active mainly in the subfields of food science (42%) and veterinary sciences (42%). In turn, they are less well represented in most of the natural and applied sciences. They are in particularly low proportions in the subfields of economic theory (15%), computer hardware & architecture (16%), fluids & plasmas (17%), econometrics (17%), nuclear & particle physics (18%), mathematical physics (18%) and distributed computing (19%).

Still at the level of subfields, women are more active in gender studies (76%), nursing (74%), social work (60%), family studies (60%) and developmental & child psychology (58%). Outside of health sciences and social sciences, women are well represented in food science (42%), veterinary sciences (42%), sport, leisure & tourism (39%), industrial relations (37%), medical informatics (37%) and biotechnology (37%).

In the natural sciences, they have the highest proportion in plant biology & botany (36%), medicinal & biomolecular chemistry (36%), environmental sciences (35%), analytical chemistry (35%) and marine biology & hydrobiology (35%).

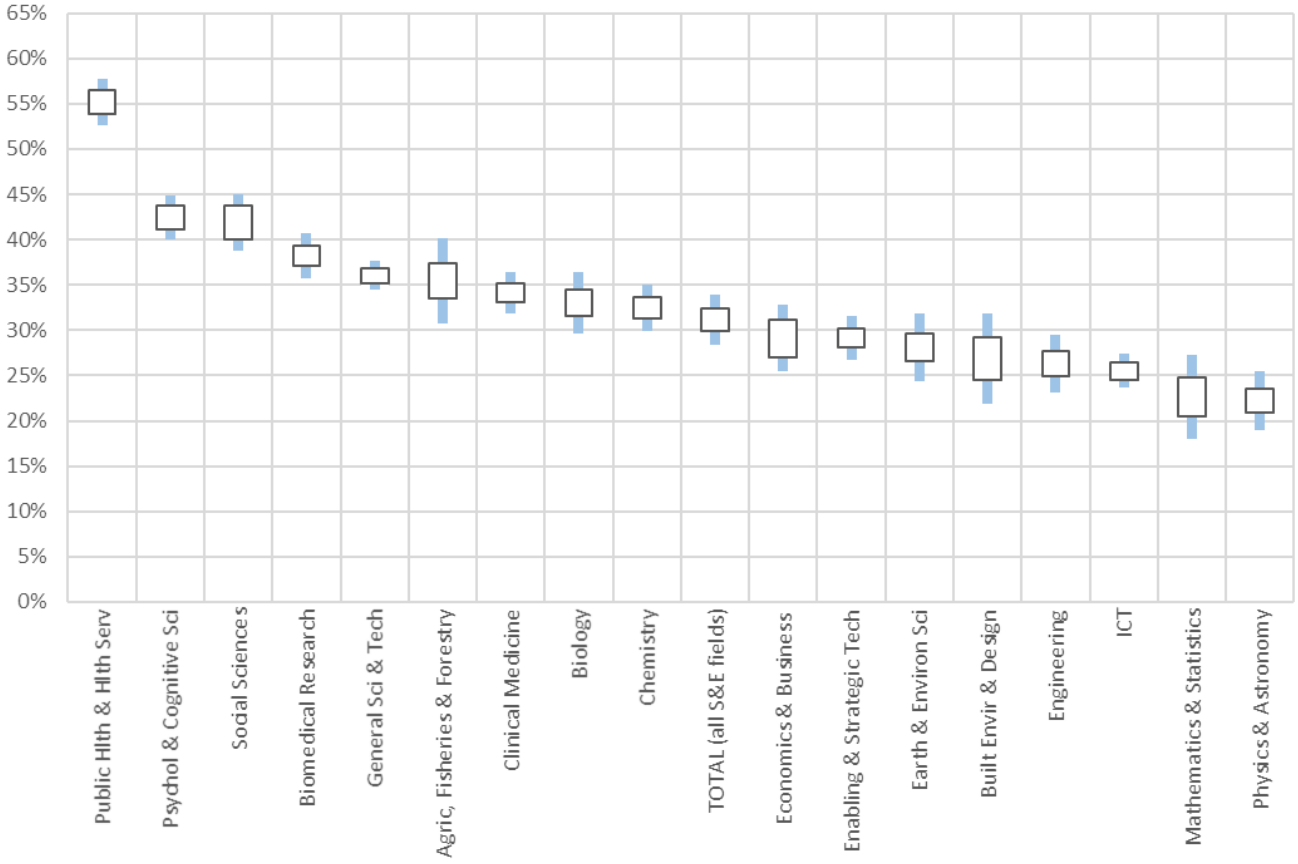



Figure 8 Proportion of women in authorship in Scopus, by S&E field (2006–2015)

Note: White rectangle = 95% reliability interval for measurement only; blue band = overall 95% reliability interval (measurement + sampling error). Underlying data available in Table XX.

Source: Compiled by Science-Metrix using Scopus (Elsevier)

As shown in Table VII, the fields with low proportions of women are also the fields where the proportions of women increase faster. Women are gaining ground in all fields, but we can observe a catch-up growth whereby they usually gain ground faster in fields where they started from a lower proportion. The field of ICT is a clear exception, having the third-lowest proportion of women's authorship but a growth that is lower than the overall growth in all S&E fields.

Table VII Proportion of women by field of research, and growth, 2006–2015

| Field | GR | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **All S&E fields** | **1.19** | 28.0% | 28.6% | 29.5% | 30.2% | 30.8% | 31.5% | 32.0% | 32.5% | 33.1% | 33.9% |
| Public Hlth & Hlth Serv | 1.06 | 53.1% | 53.8% | 54.1% | 54.6% | 55.1% | 55.0% | 55.9% | 56.3% | 56.5% | 56.5% |
| Psychol & Cognitive Sci | 1.12 | 40.2% | 40.0% | 41.3% | 41.6% | 41.4% | 42.5% | 42.5% | 43.3% | 44.0% | 46.2% |
| Social Sciences | 1.13 | 38.8% | 39.5% | 40.2% | 40.9% | 41.7% | 41.9% | 42.2% | 42.8% | 43.4% | 45.1% |
| Biomedical Research | 1.12 | 35.4% | 36.3% | 36.8% | 37.3% | 38.2% | 38.6% | 39.2% | 39.6% | 39.8% | 40.3% |
| Agric, Fisheries & Forestry | 1.17 | 31.9% | 33.1% | 33.9% | 34.3% | 34.9% | 35.5% | 36.2% | 37.2% | 37.5% | 38.4% |
| Clinical Medicine | 1.17 | 30.9% | 31.5% | 32.4% | 33.0% | 33.7% | 34.4% | 35.1% | 35.8% | 36.2% | 36.9% |
| Biology | 1.12 | 30.5% | 31.3% | 32.0% | 32.7% | 32.3% | 32.9% | 33.8% | 34.2% | 34.6% | 34.8% |
| Chemistry | 1.14 | 29.2% | 30.8% | 31.5% | 31.9% | 32.1% | 32.9% | 33.2% | 33.7% | 33.8% | 34.3% |
| Economics & Business | 1.24 | 24.8% | 25.7% | 26.6% | 28.7% | 29.3% | 30.1% | 29.7% | 30.6% | 31.0% | 31.8% |
| Enabling & Strategic Tech | 1.23 | 25.0% | 25.8% | 26.7% | 27.8% | 28.7% | 30.3% | 30.5% | 30.6% | 31.1% | 31.2% |
| Earth & Environ Sci | 1.24 | 24.4% | 25.2% | 26.4% | 26.6% | 27.7% | 29.1% | 28.8% | 29.3% | 30.6% | 31.0% |
| Built Envir & Design | 1.23 | 24.8% | 23.8% | 25.5% | 25.5% | 25.4% | 27.2% | 27.2% | 28.0% | 29.3% | 30.4% |
| Engineering | 1.23 | 22.9% | 23.4% | 24.4% | 24.7% | 25.6% | 27.1% | 27.2% | 28.1% | 28.7% | 28.2% |
| ICT | 1.17 | 22.5% | 22.9% | 24.5% | 26.0% | 26.6% | 26.3% | 26.2% | 25.6% | 26.1% | 26.9% |
| Physics & Astronomy | 1.23 | 19.7% | 20.1% | 21.0% | 21.4% | 22.4% | 22.2% | 23.0% | 23.7% | 24.0% | 24.8% |
| Mathematics & Statistics | 1.25 | 19.4% | 20.2% | 21.0% | 22.1% | 22.0% | 22.6% | 23.2% | 24.1% | 24.8% | 24.5% |

Note: Growth ratio = average of 2014 and 2015 divided by average of 2006 and 2007
Source: Compiled by Science-Metrix using Scopus (Elsevier)

The share of women's authorship in the top 50 most publishing countries in the 2006–2015 period is presented in Figure 9. It is very interesting here to examine the two types of margin of error for each country. The countries for which the measurement is less precise are Nigeria (±7.3%), Thailand (±5.9%), Singapore (±4.6%), the Republic of Korea (±4.1%) and Tunisia (±3.9%). The countries with the largest sampling errors are Nigeria (±7.7%), Ukraine (±7.7%), Tunisia (±4.8%), South Africa (±4.5%), Slovakia (±4.4%), Singapore (±4.4%) and Russia (±4.0%). This sampling error is itself determined by three things: the availability of first names in the database, the proportion of these first names that can be genderized (non-ambiguous names), and the number of authorships by country. So, for example, even if Chinese names are highly ambiguous, because the availability of the first name in Scopus is very high for China and the number of papers from Chinese authors is also very high, then the margins of error are relatively small. For some countries with a large number of papers and infrequent ambiguous or missing names, the margins of error are very small. When combined with a high accuracy of genderization in NamSor, the overall reliability interval of the indicator is small, as exemplified by the U.S. and Japan.

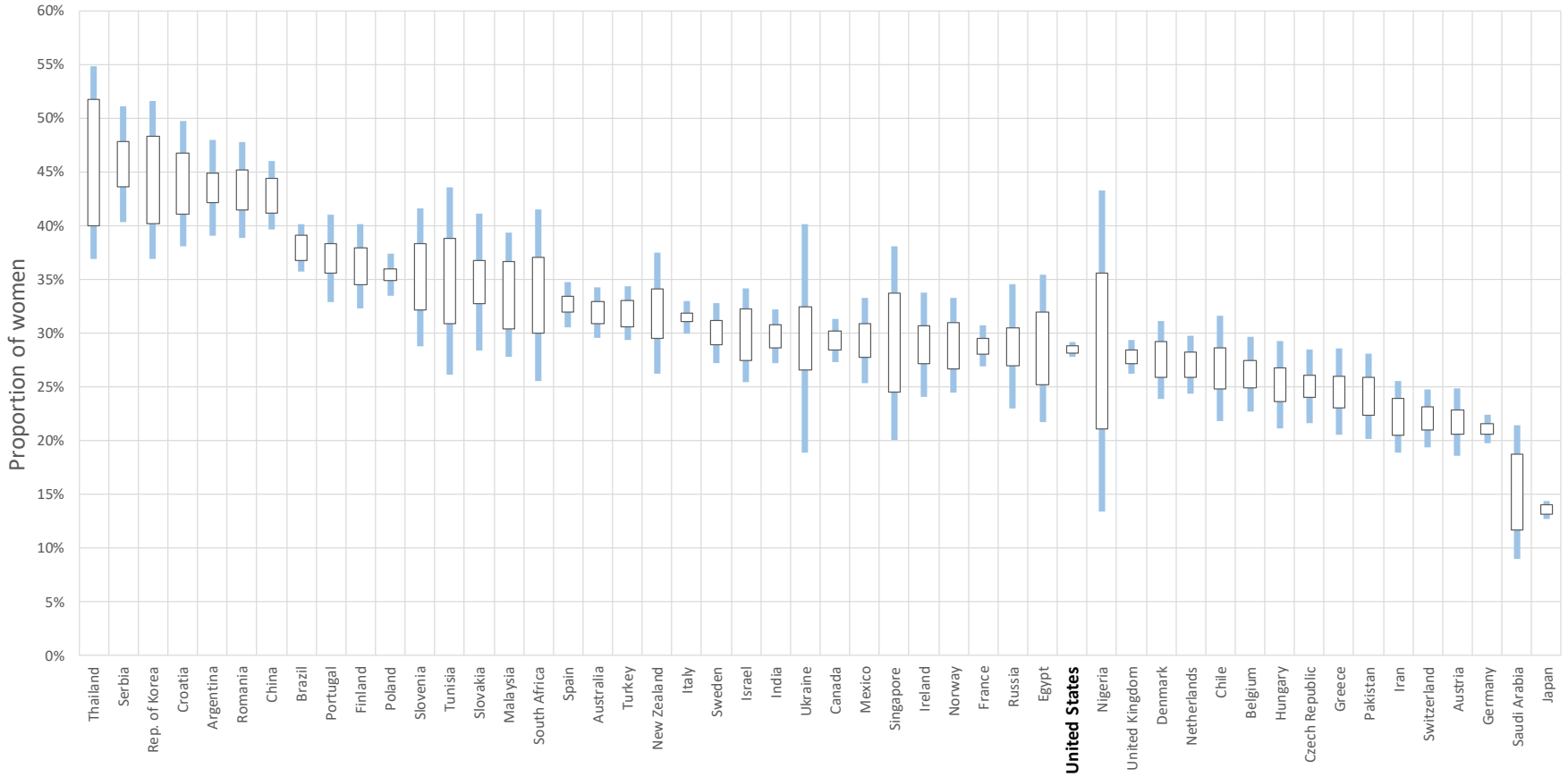



Figure 9 Proportion of women in authorship in S&E publications in the 50 most publishing countries (2006–2015)

Note: White rectangle = 95% CI for measurement only; blue band = overall 95% confidence interval (measurement + sampling error). Underlying data available in Table XXI.

Source: Compiled by Science-Metrix using Scopus (Elsevier)

Because there is a high level of imprecision in the measurement of the proportion of women at country level, it is not possible to precisely rank countries with this indicator, nor it is simple to compare the proportion of women for the U.S. with similarly ranking countries, such as Nigeria, Egypt, Russia, France and Norway. It can be concluded with high confidence, though, that the U.S. has a higher proportion of women in scientific publications than Japan, and a lower proportion than Serbia.

Seven countries clearly demonstrate a higher participation of women: Thailand, Serbia, the Republic of Korea, Croatia, Argentina, Romania and China. Remarkably, three of these seven countries are from Southeastern Europe, and many countries from Eastern Europe and the Balkans are ranking highly in terms of the participation of women.[38] At the other end of the scale, two countries demonstrate the lowest proportion of women in scientific publications: Japan and Saudi Arabia, both with about 15% of authorships by women. Women are also not highly present in Iran and in three German-speaking countries: Germany, Austria and Switzerland. In fact, almost all Germany's neighbors exhibit a relatively low proportion of women's authorship. This is in high contrast to other studies showing these countries among the top ranking in regard to women's opportunities. For example, in a report prepared by Save the Children International, these countries top the list when ranked with the Girls Opportunity Index, a composite indicator based on various indicators related to girls' opportunities.[39]

The U.S. is undoubtedly not among the leaders for the proportion of women in scientific publications, but also not among the clear laggards. With a proportion of roughly 28% women, the U.S. is ranking somewhere between the bottom third and the bottom half of the 50 countries presented.

Analogous to observations at the field level, many countries with a high proportion of women in scientific publications are also the countries for which the indicator exhibits the slowest growth over the 2006–2015 period (Table VIII). The proportion of women was highest in 2006 in the Republic of Korea (45%), China (42%), Brazil (36%) and Poland (34%), all countries for which the proportion increased feebly over the period. Similarly, countries with a lower proportion of women at the beginning of the period are also among the countries where women are gaining ground rapidly. The exception to this statement is Japan, for which growth is not very high, although it was starting with the lowest proportion of women within the top 50 most publishing countries.

Many other countries with a high presence of women are slowing down their progression. In the case of the Republic of Korea, the proportion of women has decreased in the period. It would be interesting to compare these trends with the trends based on other indicators on women in science (e.g., number of graduate students) to see if a similar slowing down can be observed there.

[38] Although not within the top 50 most publishing countries, Albania, Latvia, Bulgaria, Macedonia, Bosnia and Herzegovina, and Kosovo have point estimates above 40% for the proportion of women in the period.

[39] Save the Children. (2016). *Every last girl: free to live, free to learn, free from harm*. London, UK: Save the Children. Retrieved from https://www.savethechildren.org.uk/content/dam/global/reports/advocacy/every-last-girl.pdf

Table VIII Proportion of women in scientific authorship by country (top 20 countries with most publications in the period)

| Country | GR | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **World** | **1.19** | 28.0% | 28.6% | 29.5% | 30.2% | 30.8% | 31.5% | 32.0% | 32.5% | 33.1% | 33.9% |
| Rep. of Korea | 0.98 | 45.0% | 44.5% | 44.5% | 44.4% | 44.3% | 44.2% | 44.1% | 44.2% | 44.0% | 43.8% |
| China | 1.02 | 42.1% | 42.5% | 42.7% | 42.7% | 42.9% | 43.1% | 43.0% | 42.7% | 42.9% | 43.0% |
| Brazil | 1.08 | 35.5% | 36.6% | 36.8% | 37.6% | 37.6% | 38.3% | 38.5% | 38.7% | 39.1% | 39.1% |
| Poland | 1.10 | 34.1% | 33.3% | 33.7% | 35.0% | 34.7% | 36.0% | 35.8% | 36.6% | 36.1% | 38.2% |
| Australia | 1.14 | 30.5% | 30.6% | 31.5% | 31.4% | 32.3% | 32.4% | 33.3% | 33.9% | 34.4% | 35.4% |
| Spain | 1.09 | 30.5% | 30.1% | 30.8% | 31.4% | 31.9% | 32.4% | 32.1% | 32.2% | 32.7% | 33.5% |
| Italy | 1.14 | 29.4% | 29.7% | 30.4% | 30.7% | 30.9% | 31.4% | 32.1% | 32.2% | 32.8% | 34.4% |
| Turkey | 1.20 | 28.7% | 29.0% | 30.5% | 31.2% | 31.0% | 32.1% | 32.3% | 33.2% | 33.8% | 35.4% |
| Sweden | 1.15 | 27.8% | 28.7% | 29.1% | 29.1% | 29.1% | 29.8% | 30.5% | 31.0% | 31.7% | 33.2% |
| Russia | 1.10 | 27.8% | 28.3% | 28.5% | 28.7% | 29.3% | 29.6% | 30.0% | 30.2% | 30.7% | 31.2% |
| Canada | 1.11 | 27.3% | 27.7% | 28.2% | 28.3% | 28.3% | 28.6% | 29.2% | 29.6% | 30.2% | 31.0% |
| United States | 1.18 | 26.8% | 27.1% | 27.7% | 28.4% | 28.9% | 29.7% | 30.3% | 30.7% | 31.2% | 32.4% |
| Netherlands | 1.16 | 25.9% | 26.0% | 26.5% | 26.6% | 27.2% | 27.9% | 28.5% | 29.3% | 29.6% | 30.7% |
| India | 1.22 | 25.7% | 26.1% | 26.9% | 27.4% | 28.0% | 28.8% | 29.3% | 29.9% | 30.7% | 32.4% |
| France | 1.31 | 24.3% | 24.9% | 25.9% | 27.7% | 29.0% | 28.9% | 29.1% | 30.6% | 31.7% | 32.8% |
| United Kingdom | 1.37 | 22.6% | 23.2% | 24.4% | 25.2% | 26.5% | 27.2% | 28.5% | 29.5% | 30.6% | 32.1% |
| Switzerland | 1.24 | 19.9% | 19.8% | 20.7% | 21.0% | 21.8% | 22.1% | 22.5% | 23.4% | 23.9% | 25.2% |
| Iran | 1.28 | 18.7% | 19.7% | 20.0% | 19.8% | 20.9% | 21.7% | 22.4% | 23.7% | 24.2% | 25.0% |
| Germany | 1.27 | 18.3% | 19.1% | 19.4% | 20.0% | 20.8% | 21.3% | 21.8% | 22.4% | 23.1% | 24.4% |
| Japan | 1.19 | 12.5% | 12.7% | 12.9% | 13.3% | 13.3% | 13.4% | 13.8% | 14.3% | 14.6% | 15.2% |

Note: GI = Growth ratio = average of 2014 and 2015 divided by average of 2006 and 2007.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

# 4 Discussion and conclusion

This paper presents a promising approach to measure the proportion of women in scientific publications, and to develop other indicators related to the participation of women in science. It demonstrates that the proportions of women (and men) at various aggregation levels (year, country, subfield) can be estimated using the first name and last name of authors recorded in the publication databases. Although the first name is not recorded for all authors, and although some first names are too ambiguous to determine the gender, the samples for which a gender can be derived are fairly large; using a calculation of a 95% reliability interval, it is still possible to make robust comparisons and findings about the proportion of women in scientific publications.

The causes of imprecision in the indicator can be divided into measurement errors and sampling errors. The measurement errors stem from the incapacity to determine the gender from the combination of first and last names with 100% accuracy. Names that are too ambiguous are not genderized. But virtually no combination is totally unambiguous. For all the combinations of first and last names in its database, the tool used for genderization, NamSor, also computes a reliability score based on a name combination's likelihood of being that of a woman or a man. We use this score to compute the margins of error related to genderization.

Because the gender can only be determined for a subset of authors in the scientific publications, the proportions of women and men in the population are inferred from samples, and are therefore prone to sampling errors. The samples are convenience samples and thus cannot be considered probability samples. Because of this, the representativity of the samples is questioned. Based on preliminary tests, we assume that the sampling approach is not biased, or at least that the biases are negligible and can be ignored. Additional validation of this assumption would increase the robustness of the approach.

The sampling error is added to the measurement error to obtain the final 95% reliability interval for the indicator. Therefore, at any aggregation level, this confidence interval depends on (1) the number or share of authors for which the first name was recorded in the database, (2) the number of names for which a gender can't be derived (e.g., ambiguous or unknown name), (3) the specificity of the first name (i.e., the accuracy of the gender provided by NamSor), and (4) the number of authorships in the aggregate. At a highly aggregated level (e.g., country in all fields for the 2006–2015 period), the margins of error are fairly small in most cases, and interesting comparisons can be made. However, for smaller aggregates, the margins of error can rapidly become so large that it is impossible to conclude anything useful.

For the U.S., the accuracy of the tool is high, the proportion of genderizable names is also high, and so is the number of authorships (because the U.S. is still the largest producer of scientific publications). Therefore, the indicator at disaggregated levels still supports robust analyses. Thus, in future editions of the Science and Engineering Indicators, the overall proportion of women in science and engineering could probably be reported annually for leading countries. Also, indicators could probably be prepared at field level for the U.S. only, or for some leading countries where the numbers are more reliable.

Section 3 of this report presents some interesting findings. The results show that the proportion of women in scientific authorships is increasing at world level and in the majority of countries. However, although the proportion of women is increasing in most leading countries, it is slowly decreasing or stable

in many developing countries. And although in some fields the proportions of women are already high, mainly fields in the domains of health sciences and social sciences, women are in lower proportions in economics, applied sciences and natural sciences, but in most of these fields the proportion of women is increasing faster.

When compared with the top 50 most publishing countries, the U.S. is in the midrange among the leading countries when it comes to the proportion of women in scientific authorship. In fact, there are slightly more countries doing better than the U.S. than there are countries doing worse. It is difficult to forecast if this position will improve over time. On one hand, the proportion of women in the U.S. is increasing faster than in many countries that currently have higher proportions of women. On the other hand, many lagging countries are improving faster than the U.S. Trends can also be difficult to predict for proportions because the values are bonded between 0% and 100% and the proportion may change toward stabilization somewhere between these two extremes. Most likely, the evolution will be best described using an s-curve with a slightly exponential progression at the start, followed by an exponential decay toward a potential stable proportion. A model that would include other variables related to gender in research would probably be better at explaining the historical trends and at forecasting future trends. It would certainly be useful to include other variables to have a better understanding of the drivers that determine a greater participation of women in scientific activities.

At least two general factors contribute to the prevalence of women within scientific publications. The first is the number of women in research, and the second is the productivity of women compared to that of men, in terms of number of papers per researcher. Even if more women are trained in universities and more women turn to research, if the system is not supporting them as well as it is supporting men, then women may end up creating less new knowledge and they may have more difficulties publishing their results in peer-reviewed journals. Therefore, even with a 50/50 representation of gender in academic research, women may still be underrepresented in scientific publications. Because promotion and financial support often depend on an author's CV, it is easy to understand why things are not changing rapidly.

To measure the productivity of women compared to that of men, an accurate measure of the number of authors for each gender is necessary. At aggregate level, the number of PhDs employed by gender can be suggested as a reasonable estimation of the number of authors by gender. But in fact, many PhD graduates do not work in academic research, and this proportion of PhDs that do not work in research varies between fields, countries and over time. This is probably more so for women versus men.

A statistic on current employment in research could be a better alternative, if available. It would also be useful to measure the productivity of women at the field level, but this would be even more challenging as the number of employees by field of research would be necessary. This would also imply that a robust alignment between the field in employment statistics and the field in publication data could be established, which is further complicated by the reality that most researchers publish their papers in a variety of fields.

There is another factor to consider when discussing potential errors in the measurements prepared using the data indexed in the bibliometric databases, one that is not directly related to errors, but instead to coverage. Although Scopus and the WoS both encompass millions of documents, they do not cover all research. If some domains of research are not covered as exhaustively as others, and if gender biases are present in the share covered by the databases, this could lead to a miscalculation of gender proportions

aggregated at the field level and overall for all fields combined. For example, if there are more women in the social sciences compared to other areas and this domain is not adequately represented in the database, the proportion of women could be underestimated. At a disaggregated level, issues can also arise. For instance, it is well known that the domain of social sciences is not covered as extensively as other scientific disciplines, with many of the favored media of this domain not being covered by these databases (e.g., books). In a scenario where women or men tend to be more active on scientific papers as opposed to within these other media, statistics for the social sciences could be less reliable.

It is rather difficult to estimate what percentage of the total scientific output is covered in the databases, even more so when going down to the level of scientific domains. One way of estimating this coverage is to analyze the share of scientific references within each scientific domain that is also indexed in the databases. This makes it possible to compute a "visibility" indicator to show what share of references within a domain is covered in the databases. This acts as a proxy for the percentage of the total scientific literature that is covered in the databases.

This approach is quite appealing and interesting, but it is limited by the fact that not all references made in peer-reviewed publications are to scientific literature, though the extent to which this is the case would have to be investigated. Other non-peer-reviewed literature cited by the content of a bibliographic database could include conference presentations, scientific reports, and books. Thus, the misrepresentation of fields in the database using this information may be regarded as a way to account for the materials, which is of relevance to the development of a given field, rather than to only consider the core of peer-reviewed literature, which is not equally important across all areas. Cited material, be it part of the peer-reviewed literature or not, conveys important information to the development of an area as citations (either positive or negative ones; both contribute to advancing knowledge) are rarely given for free.

In a future exercise, corrections in the proportions of women by subfield/field and overall could be computed using the above information and compared with those presented in the current report. This could at least enable one to appreciate the potential effects of coverage biases in bibliographic databases. Additional work could then be done to de-duplicate references in the covered papers as well as to evaluate the relevance of cited materials. Preliminary findings show that Scopus offers a very good coverage in the natural sciences and engineering, and a fairly good coverage of social sciences as well. The coverage is lower in philosophy & theology, communication & textual studies, historical studies and visual & performing arts, all fields that will not be covered in the bibliometric indicators for SEI 2018.

Other avenues for related explorations presented themselves in the preparation of this report, but they had to be left for future research. The following section outlines items that should be carried forward, as candidates for building further on the foundations laid down here.

## 4.1 Future work for improving the gender identification approach

The present section covers potential avenues to improve the accuracy and recall of gender tagging procedures; it suggests ways to improve the tool being developed. Section 4.2 below highlights some further lines of analysis that would be interesting to explore based on the gender information collected through these approaches; it suggests potentially valuable applications for the tool.

### Characterization of gender for Chinese researchers with non-Asian first names

As detailed earlier in the report, NamSor was used to identify the gender for a large share of author names on publications indexed in both the Scopus and WoS databases. One of the limitations of NamSor is related to East Asian names, which are usually not assigned a gender by the algorithm. Although this behavior is somewhat understandable—considering that most Asian names are unisex in their Latin form—it also leads to the unnecessary neglect of a given range of cases. Specifically, the algorithm automatically passes over authors with East Asian last names but non-gender-neutral first names, when in fact sufficient material is available to identify their gender. For instance, although NamSor has no trouble identifying Sophia S. Smith as a woman, Sophia S. Wang would end up under the unassigned category. Although this limitation was partially bypassed by the first gender assignation step, which used highly common unambiguous names to pre-filter what was sent to NamSor, there remained a number of first names with lower frequencies that were not systematically assigned a gender, leading to cases such as the one presented above.

In future, improvements to the list of unambiguous first names could be made so that it would encompass an even larger number of names, which would limit the effects of this artifact in the NamSor treatment. Note also that this improvement is expected to have its strongest effect for authors of East Asian heritage but born in North America or Europe, where the combination of a more common occidental first name with an East Asian last name is frequent and these combinations constitute the range of cases for which this improvement would have the greatest effect. Thus, the effect of this improvement on accuracy and coverage of statistics computed for East Asian countries is likely to be much more modest.

### Gender identification using email information

Only author names, as indexed in the Scopus and WoS databases, were used to identify gender. However, as analysts were preparing material for this report, it was discovered that one additional parameter indexed in the databases could be used to gather information on gender: email addresses. Indeed, email addresses are often quite informative regarding gender, with many containing the full first names of the authors they are linked to. It is therefore conceivable to design a method to genderize authors using their email addresses.

Although this method would not reduce the number of cases with gender-ambiguous names, it would be highly effective for cases where only author initials are indexed in the database, as the email address might contain the full first name. Using a method similar to the first step of the genderization process, common and unambiguous first names could be searched for within the email field to assign a gender. Although there are some limitations to this approach, as email addresses are not provided for all authors and are more frequently provided on more recent publications (stable in the last decade though), a preliminary estimate based a random sample indicates that about 5%–10% of all cases not yet resolved could be successfully identified using email information, which would account for millions of additional publications (an estimated 2–3 million in Scopus; number to be determined in the WoS). Although this estimate was produced using a greatly expanded version of the list of unambiguous first names, it highlights the strong potential of this approach.

**Disambiguation of researchers' portfolios of publications to resolve unassigned cases**

An author's first initial is insufficient to identify their gender; however, the combination of a first initial, a last name and an institutional affiliation (along with perhaps some other information indexed in the database, such as areas or even specific topics of research, email addresses, and so forth) may be sufficient to uniquely identify a researcher as the author of a whole portfolio of articles. From there, if any single publication in the portfolio has sufficient information to determine the researcher's gender, this information can be expanded to cover the portfolio as a whole, being tagged to each paper therein. Thus, a combination of the gender identification tool and an author disambiguation tool can be used to draw on highly discriminant information about gender to tag publications, even where insufficient information is available to assign the publication using the gender tool alone.

Efforts to come up with researcher disambiguation tools have been made on various fronts. Scopus already includes an author ID that aggregates publications from researchers, but the accuracy and recall of this author ID are far from perfect. The Web of Science also has an author ID, but it is more recent, and Science-Metrix does not have access to it yet. Third-party approaches such as the ORCID project are also working toward providing solutions to this problem. Science-Metrix itself also developed an algorithm, in collaboration with the École Polytechnique de Montréal, to disambiguate authors in both the Scopus and Web of Science databases, with considerable success. Although identifying portfolios of researchers with uncommon names is quite easy overall when adding information on scientific topics and institutional addresses, it becomes much less reliable for authors with highly common names. This is especially a problem for East Asian researchers as a few East Asian names can account for millions of authors. Nevertheless, this approach appears to be quite promising and further work would enable estimating its full potential.

**Gender tagging using image-based gender recognition**

To independently validate the accuracy and recall of the different tools used to identify gender by author name (e.g., NamSor, list of unambiguous names), a manual verification of a random sample of authors was performed by analysts, using information in the database to retrieve author details online and searching for photos to visually identify genders. Although this process is quite labor intensive, as it relies on manual case-by-base searches by humans, it could be interesting to investigate whether image-recognition software could be calibrated to detect information on gender. For instance, using information from the databases, searches could be made automatically to retrieve photos associated with each query and image recognition could be used to determine the gender of a high volume of researchers for whom other methods have proven unsuccessful. Although this approach is by far the most complex of those described in the present section, it has enormous potential, yielding huge rewards if the technical intricacies of the approach can be unwound.

## 4.2 Future gender-specific analyses of interest

Whereas Section 4.1 covers potential avenues for improving the accuracy and recall of the gender-tagging tool, the present section outlines some potentially interesting analyses that one might carry out using the gender information collected with these tools. Gender analyses in bibliometric studies are currently quite limited, and the present work opens a lot of doors. Specifically, the full complement of bibliometric tools

already available can be combined with this new tool to determine whether gender dimensions are a relevant consideration in each of those other lines of analysis.

### Scientific impact and gender

An important part of scientific excellence is gaining recognition from colleagues for one's scientific accomplishments. Although this recognition can be expressed in many ways, references to scientific publications are often considered to be explicit acknowledgements of an intellectual contribution. As such, the more a scientific article is cited, the greater its impact on the scientific community, and the more likely it is to be an influential piece of work. This is the basic assumption that underlies the various indicators grouped here under "citation analysis" (e.g., citation counts, journal impact factors).

Because citation practices are different between subfields of science and over time, the preferred way to use citations to measure scientific impact is using normalized relative citation counts, which control for variations in citation practices. Citation indicators can be produced at various aggregation levels, such as for individual scientists, research groups, departments, institutions or countries. They can also be used to track the scientific impact of women and men, which is increasingly relevant given the growing reliance on bibliometric statistics for research evaluation purposes in research assessment exercises and grant competitions. If women have lower scientific impact than men due to factors other than their professional competencies (e.g., greater responsibilities in personal life, such as caring for children), this could reduce their chances of being funded or lower the value of the grants they do receive, which could in turn decrease their scientific impact, thereby creating a vicious circle.

Based on the approach described in this report to measure the proportion of women in scientific publications, there is a clear potential to develop a robust gender-disaggregated indicator of scientific impact. However, to date, it has not been possible. Firstly, it was already a challenge to infer the proportion of women in the whole of the Scopus database from the subset of papers for which the information on the gender of authors could be determined using the Gender API. For an indicator based on citations, gender cannot only be inferred from known gender proportions in the subset—it must be inferred in a way which accounts for the wide variations in the citation scores of papers in the database. While performing some tests, we discovered that even when measured at country/year/subfield levels, the citation patterns for the unknown authorships by gender are not similar to the patterns observed for the genderized authorships. It is therefore impossible to compute scores and confidence intervals with the same approach used for gender proportions.

We are currently working on a promising approach based on a semi-random attribution of gender to all authorships in the database. This attribution is based on

- the estimation of the proportion of women at subfield, country and year level, based on the subset of genderized names; and
- the probability of each name being a woman (or a man).

Once the gender has been determined for all authorships, the indicator of impact can be computed at any level by proportionally averaging the scores of women and men at the desired aggregation level. Because the attribution is somewhat random, the accuracy of the indicator increases with the number of authorships in a given aggregate group. In order to determine a confidence interval for the indicator, we

use the variations in the results of a bootstrapping approach in which the semi-random attribution of gender is performed several times.

The approach has been implemented in MS SQL Server, but because of the sheer size of the database, this implementation is too slow. Preliminary results on some small subsets clearly point to a viable solution. The next step will be to implement a faster version of the algorithm to enable faster execution time over thousands of bootstraps on the whole database and for various levels of aggregation. This will enable a better assessment of the validity of the results.

### Citation profiles and gender

An important consideration in normalizing citation counts is to determine the time to peak citation—the moment after which a clear signal can be read about the uptake of a given paper. Times to citation peak are known to vary from one area of research to another, usually ranging from two years to seven, depending on the area. At present, no assessment has been undertaken to determine whether gender differences are relevant here as well.

Such an assessment could determine whether research conducted by women (or by research groups predominantly constituted or led by women) is slower to be taken up in the research community. In addition to highlighting important questions about why these differences exist, such an analysis could also contribute to designing more equitable research evaluation practices, by helping to define more appropriate benchmarks for comparison across cases. For instance, if research conducted by women is taken up more slowly in the community, then assessing articles after two years may disproportionately affect researchers along gender lines. In such a case, longer citation windows might be a valuable avenue to level the field in research evaluation.

Another analysis under this heading would be to compare the impact of articles to the impact factor of the journals in which they are published, to determine whether gender influences the visibility of venues in which an author publishes, relative to the quality of their work. Comparing paper-level and journal-level indicators could determine, for instance, whether women are publishing in more prestigious or less prestigious journals than men for the same quality of work.

### Collaboration and gender

With the large proportion of authors being identified by gender in both the Scopus and WoS databases following work for this report, data on scientific collaboration and gender could easily be prepared to identify patterns in the interplay between gender and collaboration in research publication. Some potential lines for analysis could include a computation of how many papers are published by research groups consisting of men only, research groups consisting of women only, or mixed-gender research groups; in mixed teams, gender balance could also be assessed. Furthermore, these analyses could be dissected by scientific discipline, to identify any subject-specific patterns in terms of gender collaboration.

An additional layer would be to assess whether collaborations across national, disciplinary or sectoral boundaries are more likely or less likely to include a gender balance component. International research collaboration has been a facet of interest in the science policy community for some time, while interdisciplinary research is especially in focus right now, as are partnerships between the public and

private sectors. (These latter two are often cited as drivers of innovation.) With a robust indicator to integrate gender considerations into these analyses, the role of men and women in these types of collaborations can be studied.

### Research networks and gender

Another set of indicators that could be interesting to break down by gender would be social network analysis indicators. In this case, networks could be created by co-authorship links and by citation links; the former constitute networks of research collaboration, whereas the latter constitute networks of knowledge integration and additive creation. Social network indicators in the bibliometric context are used to identify clusters in the research community, researchers acting as cluster hubs, researchers acting as bridges between clusters, and so forth. Integrating a gender dimension into such analyses could help to determine whether gender differences exist in terms of the different roles that researchers occupy within such networks.

These analyses can be conducted retrospectively and longitudinally, to assess the evolution of network dynamics as women became and continue to become more and more integrated into the research community. For instance, do we see generational turnover in the network centrality of women, and are women as equally likely as men to occupy central roles in a network as they progress through their research careers? Do women and men in research tend to create mixed-gender clusters, or rather do clusters divide along gender lines?[40] As women move into more central network roles, have they reached these roles through rising through the ranks of mixed-gendered clusters, or have women moved up in clusters created by and for women?

### Emerging research topics and gender

Bibliometric tools exist to delineate individual research topics and can be used to identify topics that are emerging on national or global scales, in terms of high growth, interdisciplinarity, or other properties of interest. Within these topics, gender dimensions can be explored to determine whether women and men participate equally in defining the leading edge of research, or whether one gender or the other is more likely to explore established topics; whether women and men at the forefront of new explorations are exploring in the same area, or whether some emerging topics are particularly dominated by one gender; and so forth.

### Research funding and gender

Are men more often successful applicants on grants compared to women? Among successful applications, and taking into account the fields of science in which participants are involved, are men awarded larger amounts of money than women? Gender tagging methods used to prepare this report could be transposed easily to other data sources, such as grant information from the NIH and the NSF, to address these important issues.[41] Such analyses can also be crossed with interdisciplinarity indicators

[40] These social network indicators expand greatly on the collaboration analyses described above, which constitute a very simplified version of what is discussed here.

[41] The gender of applicants and grantees is not recorded in the public version of these databases.

and emerging topic indicators to determine whether one gender or the other has more funding success when proposing to break new ground in the research world.

### Patent uptake and gender

The current report, and the lines of potential follow-up analysis outlined here, focus on gender analysis of scientific production. However, similar analyses could also be prepared using patent information to elucidate topics pertaining to gender and innovation, which would prove quite useful and complementary in the context of research and development studies. Science-Metrix has already prepared similar analyses for the *She Figures 2015* report, which focused on European countries; with the tool now being calibrated to the U.S. context, similar analyses could be prepared in the context of the U.S. patenting market.

For instance, these analyses could determine whether research undertaken by men and women have equal chances of being cited in patent literature, and whether women are more likely or less likely than men to file and receive patents. These analyses could also be crossed with measures of interdisciplinarity and public–private partnership, and other factors contributing to innovation, to determine whether gender differences are relevant in these mechanisms.

### Evidence-based decision-making and gender

Other pathways from research to impact pass by different avenues. For instance, in healthcare, clinical guidelines make an important contribution to guiding practice, and these guidelines often cite a considerable list of peer-reviewed literature on which the guidelines are based. Similarly, policy documents are sometimes quite rigorous in citing the peer-reviewed and grey literature on which they are based (though clinical guidelines tend to be much more rigorous and more consistent in referencing than other types of policy documents). Interest is growing around the development of tools to automatically parse these reference lists and to match cited items to papers indexed in traditional bibliometric databases; in this way, bibliometric tools are evolving the ability to assess broader ranges of impacts, though these tools are not yet developed to the point of broad-based implementation.

When this work matures to an appropriate stage, lists of references in clinical guidelines and other policy documents could be assessed to determine whether there are important splits along gender lines. Is research by women any more likely or less likely to provide the evidential basis for policy positions? Do the genders participate equally in evidence-based decision-making? Are guidelines for women's health issues built primarily on the research conducted by men? Crossing these assessments with other administrative data could yield further lines of potential analysis, including an assessment of whether expert panels (or other decision-making bodies) that include women in their membership are more likely to draw on research produced by women.

### Social media uptake and gender

Looking beyond innovation and policymaking, other relevant avenues to impact might be highlighted. One such avenue might be public uptake, as measured through social media visibility and uptake. Interest in social media analyses is growing rapidly in the bibliometric and research policy communities, and these emerging tools could be crossed with gender analyses to determine, for instance, whether women

researchers are more active in promoting their research through social media, whether social media is more responsive to research stories posted on social media by one gender or another, and so on.

# Appendix A – Proportion of women by domain, field and subfield

Table IX Proportion of women in natural sciences, by field & subfield, 2006–2015

| Domain / Field / Subfield | 2006-2015 Papers | 2006-2015 P women | 2006-2015 95% RI | Proportion (P) of women 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **TOTAL (S&E in Scopus)** | **19,113,853** | **31.1%** | [28.4%, 33.9%] | **28.0%** | **28.6%** | **29.5%** | **30.2%** | **30.8%** | **31.5%** | **32.0%** | **32.5%** | **33.1%** | **33.9%** |
| **Natural Sciences** | **4,957,685** | **27.1%** | [23.8%, 30.3%] | **24.0%** | **24.8%** | **25.7%** | **26.2%** | **26.8%** | **27.3%** | **27.8%** | **28.5%** | **29.0%** | **29.7%** |
| **Biology** | **682,702** | **33.0%** | [29.7%, 36.4%] | **30.5%** | **31.3%** | **32.0%** | **32.7%** | **32.3%** | **32.9%** | **33.8%** | **34.2%** | **34.6%** | **34.8%** |
| Ecology | 146,173 | 31.4% | [28.9%, 34%] | 28.0% | 28.9% | 30.1% | 30.4% | 30.9% | 31.9% | 32.4% | 32.7% | 33.3% | 34.2% |
| Entomology | 60,786 | 30.1% | [25.2%, 35%] | 27.4% | 28.3% | 29.0% | 29.7% | 28.6% | 30.7% | 30.8% | 31.5% | 32.1% | 32.8% |
| Evolutionary Biology | 74,980 | 30.6% | [27.3%, 33.9%] | 28.1% | 29.2% | 28.8% | 30.4% | 30.2% | 31.2% | 30.6% | 32.2% | 31.9% | 32.5% |
| Marine Biology & Hydrobiology | 89,372 | 34.9% | [31%, 38.7%] | 32.0% | 33.3% | 32.5% | 33.9% | 34.6% | 35.9% | 36.3% | 36.2% | 36.9% | 37.1% |
| Ornithology | 15,099 | 23.9% | [20.4%, 27.4%] | 21.3% | 23.6% | 23.1% | 25.2% | 23.2% | 23.4% | 24.5% | 23.1% | 25.9% | 26.6% |
| Plant Biology & Botany | 248,443 | 36.4% | [33.3%, 39.5%] | 34.4% | 35.0% | 36.3% | 36.4% | 35.8% | 35.2% | 37.4% | 37.2% | 37.6% | 37.7% |
| Zoology | 47,849 | 27.4% | [22.9%, 32%] | 27.1% | 26.0% | 27.1% | 27.9% | 26.7% | 26.8% | 26.9% | 29.0% | 28.1% | 28.2% |
| **Chemistry** | **1,269,292** | **32.5%** | [29.9%, 35%] | **29.2%** | **30.8%** | **31.5%** | **31.9%** | **32.1%** | **32.9%** | **33.2%** | **33.7%** | **33.8%** | **34.3%** |
| Analytical Chemistry | 210,214 | 34.9% | [32.6%, 37.2%] | 32.9% | 33.8% | 34.2% | 34.3% | 34.3% | 34.6% | 35.0% | 36.2% | 36.2% | 36.5% |
| General Chemistry | 168,579 | 32.2% | [29.3%, 35.2%] | 29.1% | 30.8% | 31.0% | 31.5% | 32.1% | 33.1% | 32.9% | 33.5% | 33.8% | 33.4% |
| Inorganic & Nuclear Chemistry | 170,654 | 32.2% | [29.5%, 34.9%] | 30.4% | 31.4% | 31.9% | 32.3% | 32.1% | 33.1% | 32.9% | 32.4% | 32.1% | 33.0% |
| Medicinal & Biomolecular Chem. | 146,332 | 36.1% | [33.2%, 38.9%] | 32.8% | 33.7% | 35.0% | 35.2% | 35.6% | 36.4% | 36.7% | 37.5% | 38.0% | 37.9% |
| Organic Chemistry | 279,583 | 29.3% | [27.6%, 30.9%] | 25.6% | 26.2% | 27.0% | 27.8% | 27.9% | 29.2% | 29.8% | 30.6% | 31.4% | 32.6% |
| Physical Chemistry | 92,264 | 30.7% | [26.8%, 34.6%] | 29.4% | 29.5% | 30.6% | 30.9% | 30.1% | 31.2% | 31.5% | 31.0% | 31.0% | 30.9% |
| Polymers | 201,667 | 33.0% | [30%, 36%] | 27.5% | 31.3% | 32.2% | 32.6% | 33.3% | 33.8% | 34.1% | 34.8% | 34.8% | 35.7% |
| **Earth & Environmental Sciences** | **589,954** | **28.1%** | [24.4%, 31.8%] | **24.4%** | **25.2%** | **26.4%** | **26.6%** | **27.7%** | **29.1%** | **28.8%** | **29.3%** | **30.6%** | **31.0%** |
| Environmental Sciences | 166,618 | 35.3% | [32.1%, 38.5%] | 32.2% | 31.7% | 32.7% | 33.7% | 35.1% | 36.6% | 36.0% | 36.7% | 37.5% | 37.7% |
| Geochemistry & Geophysics | 154,609 | 26.8% | [23.3%, 30.2%] | 23.3% | 25.4% | 25.6% | 25.5% | 26.5% | 27.0% | 27.8% | 27.9% | 28.7% | 29.7% |
| Geology | 32,340 | 23.2% | [16%, 30.3%] | 20.8% | 21.5% | 22.1% | 22.3% | 22.4% | 22.9% | 23.8% | 24.1% | 25.5% | 24.6% |
| Meteorology & Atmospheric Sci. | 157,598 | 23.6% | [20.3%, 26.9%] | 20.1% | 20.8% | 22.3% | 22.3% | 23.3% | 24.2% | 24.4% | 24.9% | 26.3% | 25.8% |
| Oceanography | 33,727 | 24.8% | [19%, 30.5%] | 21.0% | 21.7% | 23.9% | 22.7% | 22.7% | 24.0% | 24.8% | 27.5% | 29.8% | 30.1% |
| Paleontology | 45,061 | 27.9% | [23.9%, 31.9%] | 26.3% | 25.3% | 26.6% | 26.6% | 27.5% | 27.7% | 27.8% | 28.7% | 30.6% | 30.8% |
| **Mathematics & Statistics** | **453,516** | **22.6%** | [18%, 27.2%] | **19.4%** | **20.2%** | **21.0%** | **22.1%** | **22.0%** | **22.6%** | **23.2%** | **24.1%** | **24.8%** | **24.5%** |
| Applied Mathematics | 94,061 | 23.0% | [17.2%, 28.9%] | 18.1% | 19.6% | 19.8% | 22.3% | 21.5% | 23.1% | 24.0% | 25.1% | 27.4% | 24.1% |
| General Mathematics | 216,198 | 21.1% | [17.4%, 24.9%] | 18.5% | 19.7% | 20.1% | 20.3% | 20.8% | 21.1% | 21.5% | 22.3% | 22.4% | 22.8% |
| Numerical & Computational Math. | 78,047 | 26.1% | [20.5%, 31.7%] | 21.9% | 21.7% | 24.6% | 26.2% | 24.8% | 26.0% | 26.4% | 28.4% | 28.7% | 29.1% |
| Statistics & Probability | 65,210 | 22.9% | [18.5%, 27.4%] | 20.8% | 20.7% | 21.4% | 22.4% | 23.4% | 22.8% | 23.7% | 23.9% | 24.1% | 25.1% |
| **Physics & Astronomy** | **1,962,220** | **22.2%** | [19%, 25.4%] | **19.7%** | **20.1%** | **21.0%** | **21.4%** | **22.4%** | **22.2%** | **23.0%** | **23.7%** | **24.0%** | **24.8%** |
| Acoustics | 96,907 | 19.4% | [15.8%, 23.1%] | 17.5% | 17.7% | 18.2% | 18.6% | 19.5% | 19.2% | 20.5% | 20.8% | 20.7% | 22.0% |
| Applied Physics | 526,947 | 24.8% | [22.3%, 27.3%] | 22.3% | 22.3% | 22.8% | 24.2% | 25.0% | 25.0% | 26.0% | 26.4% | 26.7% | 27.9% |
| Astronomy & Astrophysics | 125,150 | 20.9% | [16.5%, 25.3%] | 19.1% | 19.7% | 20.1% | 20.7% | 21.4% | 21.6% | 20.8% | 21.6% | 22.3% | 22.2% |
| Chemical Physics | 238,215 | 27.5% | [25%, 29.9%] | 23.9% | 25.3% | 26.1% | 26.3% | 27.5% | 27.7% | 28.8% | 28.6% | 29.3% | 29.6% |
| Fluids & Plasmas | 214,433 | 17.1% | [13.6%, 20.6%] | 15.9% | 15.9% | 16.4% | 16.5% | 17.2% | 16.9% | 17.0% | 17.7% | 19.0% | 18.6% |
| General Physics | 251,646 | 23.1% | [19.4%, 26.8%] | 19.4% | 21.4% | 22.5% | 22.1% | 24.2% | 23.7% | 23.7% | 24.8% | 25.1% | 24.5% |
| Mathematical Physics | 34,366 | 18.1% | [10.2%, 26%] | 15.9% | 17.6% | 17.1% | 19.4% | 18.1% | 17.3% | 18.1% | 18.9% | 18.8% | 20.2% |
| Nuclear & Particles Physics | 270,776 | 17.6% | [14%, 21.3%] | 15.6% | 15.3% | 16.6% | 16.6% | 17.2% | 17.8% | 19.2% | 19.6% | 18.9% | 18.9% |
| Optics | 203,782 | 22.6% | [19.9%, 25.3%] | 19.4% | 19.6% | 21.5% | 22.3% | 22.5% | 21.4% | 23.1% | 24.5% | 24.6% | 26.8% |

Note: 95% RI = lower and upper limit for a 95% reliability interval.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

# Table X Proportion of women in applied sciences, by field & subfield, 2006–2015

| Domain / Field / Subfield | 2006-2015 Papers | 2006-2015 P women | 2006-2015 95% RI | Proportion (P) of women 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **TOTAL (S&E in Scopus)** | **19,113,853** | **31.1%** | [28.4%, 33.9%] | **28.0%** | **28.6%** | **29.5%** | **30.2%** | **30.8%** | **31.5%** | **32.0%** | **32.5%** | **33.1%** | **33.9%** |
| **Applied Sciences** | **6,639,628** | **27.7%** | [25%, 30.5%] | **24.3%** | **24.8%** | **26.0%** | **26.9%** | **27.6%** | **28.6%** | **28.7%** | **29.0%** | **29.7%** | **30.0%** |
| **Agriculture, Fisheries & Forestry** | **584,840** | **35.4%** | [30.7%, 40.2%] | **31.9%** | **33.1%** | **33.9%** | **34.3%** | **34.9%** | **35.5%** | **36.2%** | **37.2%** | **37.5%** | **38.4%** |
| Agronomy & Agriculture | 149,645 | 31.1% | [27.2%, 34.9%] | 27.6% | 28.3% | 29.5% | 30.6% | 31.2% | 31.6% | 31.9% | 32.4% | 32.6% | 33.4% |
| Dairy & Animal Science | 101,360 | 35.8% | [28.8%, 42.9%] | 32.9% | 35.1% | 35.0% | 34.5% | 35.0% | 36.0% | 36.1% | 37.8% | 37.3% | 37.9% |
| Fisheries | 55,011 | 29.7% | [24.5%, 35%] | 25.9% | 27.0% | 27.6% | 28.1% | 29.9% | 29.9% | 30.2% | 31.1% | 32.9% | 33.2% |
| Food Science | 109,391 | 42.5% | [38.7%, 46.2%] | 38.2% | 39.6% | 40.7% | 41.3% | 41.4% | 42.2% | 43.4% | 44.1% | 44.7% | 44.9% |
| Forestry | 55,204 | 27.8% | [23.1%, 32.5%] | 24.5% | 25.2% | 25.8% | 27.0% | 27.6% | 28.2% | 28.5% | 30.1% | 30.0% | 31.1% |
| Horticulture | 16,583 | 34.6% | [27%, 42.2%] | 30.7% | 28.8% | 32.3% | 33.0% | 34.4% | 33.4% | 36.7% | 37.3% | 38.5% | 38.5% |
| Veterinary Sciences | 97,645 | 41.6% | [37.5%, 45.7%] | 38.9% | 40.2% | 41.0% | 40.7% | 40.8% | 41.4% | 42.1% | 43.3% | 43.5% | 44.2% |
| **Built Environment & Design** | **157,894** | **26.9%** | [21.8%, 31.9%] | **24.8%** | **23.8%** | **25.5%** | **25.5%** | **25.4%** | **27.2%** | **27.2%** | **28.0%** | **29.3%** | **30.4%** |
| Architecture | 5,237 | 31.9% | [22.1%, 41.7%] | 32.6% | 27.6% | 31.3% | 30.4% | 32.6% | 31.4% | 29.7% | 33.5% | 32.0% | 37.8% |
| Building & Construction | 80,364 | 25.4% | [20.4%, 30.5%] | 22.9% | 21.7% | 22.9% | 24.5% | 24.1% | 26.6% | 25.8% | 27.2% | 27.7% | 27.7% |
| Design Practice & Management | 39,993 | 22.2% | [17.3%, 27.2%] | 23.4% | 21.9% | 23.1% | 22.7% | 22.2% | 21.2% | 20.5% | 20.5% | 23.4% | 24.3% |
| Urban & Regional Planning | 32,300 | 35.3% | [30.9%, 39.7%] | 32.1% | 32.4% | 33.6% | 33.6% | 33.2% | 35.4% | 37.0% | 35.8% | 37.3% | 38.5% |
| **Enabling & Strategic Technologies** | **2,010,215** | **29.1%** | [26.7%, 31.5%] | **25.0%** | **25.8%** | **26.7%** | **27.8%** | **28.7%** | **30.3%** | **30.5%** | **30.6%** | **31.1%** | **31.2%** |
| Bioinformatics | 74,127 | 29.3% | [26.6%, 32%] | 24.3% | 24.9% | 28.7% | 30.7% | 30.7% | 29.0% | 28.9% | 30.1% | 31.2% | 31.9% |
| Biotechnology | 142,769 | 36.9% | [33.7%, 40%] | 33.5% | 34.1% | 35.6% | 36.0% | 36.3% | 37.2% | 37.5% | 38.1% | 38.9% | 38.9% |
| Energy | 600,413 | 25.7% | [23.6%, 27.8%] | 22.2% | 22.5% | 23.5% | 24.6% | 26.2% | 26.0% | 26.1% | 26.6% | 28.1% | 28.4% |
| Materials | 626,453 | 32.7% | [30.5%, 34.8%] | 28.0% | 28.9% | 29.3% | 29.9% | 31.5% | 35.0% | 35.5% | 35.0% | 34.0% | 32.6% |
| Nanoscience & Nanotechnology | 193,382 | 29.3% | [26.7%, 31.8%] | 23.7% | 24.7% | 25.7% | 26.2% | 27.4% | 28.7% | 30.4% | 31.2% | 32.0% | 33.0% |
| Optoelectronics & Photonics | 283,531 | 24.3% | [21.9%, 26.6%] | 21.7% | 22.7% | 23.4% | 25.5% | 24.1% | 25.6% | 23.8% | 24.4% | 25.5% | 27.5% |
| Strategic, Defence & Securit. Stud. | 89,539 | 29.4% | [24.9%, 33.9%] | 26.2% | 28.6% | 28.3% | 31.4% | 29.2% | 29.1% | 28.7% | 30.4% | 30.1% | 31.6% |
| **Engineering** | **2,027,332** | **26.3%** | [23.1%, 29.5%] | **22.9%** | **23.4%** | **24.4%** | **24.7%** | **25.6%** | **27.1%** | **27.2%** | **28.1%** | **28.7%** | **28.2%** |
| Aerospace & Aeronautics | 122,980 | 20.1% | [17.2%, 23%] | 18.8% | 18.2% | 18.6% | 20.2% | 20.2% | 20.2% | 19.7% | 21.0% | 21.9% | 24.7% |
| Automobile Design & Engineering | 21,846 | 22.6% | [15.4%, 29.9%] | 19.8% | 19.5% | 20.1% | 19.1% | 24.4% | 22.4% | 23.5% | 25.3% | 26.8% | 25.9% |
| Biomedical Engineering | 134,074 | 28.6% | [25.5%, 31.7%] | 24.8% | 26.4% | 28.8% | 26.8% | 28.3% | 28.2% | 29.3% | 29.6% | 30.9% | 32.2% |
| Chemical Engineering | 171,715 | 29.5% | [26.3%, 32.7%] | 26.8% | 26.2% | 26.8% | 28.3% | 29.0% | 29.4% | 29.3% | 31.6% | 32.6% | 33.5% |
| Civil Engineering | 137,655 | 27.6% | [23.7%, 31.5%] | 26.7% | 28.1% | 26.1% | 27.2% | 27.1% | 29.1% | 27.7% | 27.0% | 28.0% | 28.5% |
| Electrical & Electronic Engineering | 300,296 | 23.4% | [20.9%, 25.8%] | 20.5% | 20.6% | 20.5% | 21.1% | 21.2% | 24.2% | 24.1% | 24.6% | 25.8% | 27.7% |
| Environmental Engineering | 117,661 | 27.4% | [23%, 31.8%] | 23.3% | 24.4% | 25.1% | 26.5% | 26.6% | 28.6% | 29.0% | 29.1% | 29.4% | 30.0% |
| Geological & Geomatics Engin. | 121,618 | 25.4% | [21.1%, 29.7%] | 22.1% | 23.0% | 23.5% | 24.9% | 24.9% | 25.7% | 25.7% | 27.0% | 26.9% | 27.4% |
| Industrial Engin. & Automation | 341,064 | 24.4% | [22.1%, 26.7%] | 21.9% | 21.9% | 25.0% | 23.8% | 25.9% | 25.6% | 24.0% | 24.5% | 25.0% | 24.5% |
| Mechanical Engin. & Transports | 386,898 | 29.0% | [26.4%, 31.6%] | 23.2% | 24.1% | 24.5% | 24.3% | 24.7% | 30.2% | 31.9% | 32.4% | 32.2% | 27.9% |
| Mining & Metallurgy | 75,432 | 31.2% | [25%, 37.4%] | 25.7% | 27.5% | 30.0% | 32.6% | 31.3% | 32.5% | 31.5% | 34.4% | 31.7% | 32.9% |
| Operations Research | 96,093 | 25.6% | [21.2%, 30%] | 22.3% | 22.7% | 24.8% | 24.3% | 29.3% | 27.0% | 24.2% | 26.1% | 26.9% | 25.3% |
| **Information & Comm. Tech.** | **1,859,347** | **25.5%** | [23.6%, 27.4%] | **22.5%** | **22.9%** | **24.5%** | **26.0%** | **26.6%** | **26.3%** | **26.2%** | **25.6%** | **26.1%** | **26.9%** |
| Artificial Intelligence & Image Processing | 831,711 | 26.6% | [25.3%, 27.9%] | 22.5% | 23.5% | 25.7% | 27.6% | 28.8% | 27.7% | 27.2% | 26.3% | 26.8% | 27.5% |
| Computation Theory & Math. | 93,939 | 21.0% | [17.8%, 24.1%] | 23.8% | 19.2% | 19.0% | 20.3% | 19.4% | 20.6% | 21.9% | 21.9% | 22.0% | 20.6% |
| Computer Hardware & Architecture | 44,888 | 16.1% | [12.5%, 19.6%] | 14.7% | 14.3% | 15.1% | 14.9% | 15.4% | 15.8% | 16.4% | 17.4% | 18.1% | 19.9% |
| Distributed Computing | 34,230 | 19.1% | [14.9%, 23.2%] | 16.9% | 20.7% | 16.9% | 18.0% | 18.5% | 18.7% | 19.9% | 18.8% | 20.5% | 21.9% |
| Information Systems | 92,118 | 27.1% | [24%, 30.2%] | 26.1% | 24.7% | 25.6% | 27.1% | 26.7% | 26.3% | 28.2% | 28.9% | 29.6% | 29.2% |
| Medical Informatics | 35,437 | 37.1% | [32.8%, 41.4%] | 37.4% | 34.6% | 34.4% | 37.7% | 35.9% | 37.5% | 36.5% | 36.9% | 40.1% | 39.3% |
| Networking & Telecomm. | 632,966 | 25.1% | [23.4%, 26.8%] | 21.9% | 22.8% | 24.8% | 25.3% | 25.6% | 26.4% | 26.5% | 25.1% | 25.4% | 26.3% |
| Software Engineering | 94,059 | 23.2% | [20.2%, 26.2%] | 21.5% | 21.0% | 21.7% | 24.4% | 22.6% | 23.3% | 22.7% | 24.8% | 24.5% | 27.9% |

Note: 95% RI = lower and upper limit for a 95% reliability interval.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

# Table XI Proportion of women in health sciences, by field & subfield, 2006–2015

| Domain / Field / Subfield | 2006-2015 Papers | 2006-2015 P women | 2006-2015 95% RI | Proportion (P) of women 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 | 2013 | 2014 | 2015 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **TOTAL (S&E in Scopus)** | **19,113,853** | **31.1%** | [28.4%, 33.9%] | **28.0%** | **28.6%** | **29.5%** | **30.2%** | **30.8%** | **31.5%** | **32.0%** | **32.5%** | **33.1%** | **33.9%** |
| **Health Sciences** | **6,096,375** | **37.0%** | [34.6%, 39.4%] | **33.8%** | **34.5%** | **35.3%** | **35.9%** | **36.7%** | **37.3%** | **38.0%** | **38.5%** | **38.9%** | **39.5%** |
| **Biomedical Research** | **1,167,616** | **38.2%** | [35.7%, 40.7%] | **35.4%** | **36.3%** | **36.8%** | **37.3%** | **38.2%** | **38.6%** | **39.2%** | **39.6%** | **39.8%** | **40.3%** |
| Anatomy & Morphology | 16,606 | 34.2% | [27%, 41.4%] | 27.9% | 32.5% | 33.0% | 35.4% | 34.5% | 35.9% | 35.3% | 35.4% | 34.3% | 36.4% |
| Biochemistry & Molecular Biology | 268,936 | 36.9% | [35.4%, 38.5%] | 34.3% | 34.7% | 35.4% | 36.0% | 36.7% | 37.3% | 38.2% | 38.5% | 39.2% | 39.1% |
| Biophysics | 48,497 | 29.4% | [26.2%, 32.7%] | 27.5% | 27.9% | 28.2% | 28.8% | 29.2% | 29.8% | 30.4% | 30.2% | 30.9% | 31.7% |
| Developmental Biology | 200,169 | 35.0% | [33.8%, 36.1%] | 33.0% | 33.6% | 34.5% | 33.9% | 35.1% | 35.2% | 35.7% | 36.2% | 35.6% | 35.5% |
| Genetics & Heredity | 46,402 | 45.1% | [41.4%, 48.9%] | 41.7% | 44.0% | 45.6% | 44.5% | 45.1% | 44.9% | 45.8% | 45.0% | 47.6% | 47.1% |
| Microbiology | 230,019 | 40.1% | [38%, 42.2%] | 38.0% | 38.4% | 39.2% | 39.4% | 40.1% | 40.2% | 40.8% | 41.1% | 41.3% | 42.1% |
| Microscopy | 12,301 | 27.2% | [17.2%, 37.3%] | 24.8% | 27.7% | 25.1% | 24.7% | 30.4% | 27.2% | 31.3% | 30.0% | 26.6% | 31.1% |
| Mycology & Parasitology | 39,961 | 36.3% | [30.6%, 42%] | 33.2% | 34.7% | 35.0% | 36.2% | 36.7% | 36.8% | 37.1% | 37.9% | 37.5% | 36.5% |
| Nutrition & Dietetics | 86,480 | 48.3% | [44.6%, 51.9%] | 44.9% | 45.1% | 45.0% | 46.9% | 47.5% | 48.9% | 49.4% | 50.4% | 50.8% | 50.8% |
| Physiology | 46,308 | 32.0% | [28.2%, 35.7%] | 29.4% | 30.5% | 30.8% | 31.7% | 31.8% | 31.5% | 32.9% | 33.1% | 34.0% | 35.1% |
| Toxicology | 82,650 | 40.3% | [37.1%, 43.4%] | 37.1% | 38.7% | 38.1% | 39.1% | 39.3% | 40.2% | 41.3% | 41.2% | 42.9% | 43.3% |
| Virology | 89,287 | 40.4% | [37.9%, 42.9%] | 38.2% | 38.7% | 39.5% | 39.2% | 40.3% | 40.5% | 40.9% | 41.7% | 41.9% | 42.5% |
| **Clinical Medicine** | **4,133,290** | **34.1%** | [31.8%, 36.5%] | **30.9%** | **31.5%** | **32.4%** | **33.0%** | **33.7%** | **34.4%** | **35.1%** | **35.8%** | **36.2%** | **36.9%** |
| Allergy | 25,598 | 39.5% | [33.4%, 45.6%] | 36.8% | 36.3% | 36.7% | 38.8% | 39.1% | 39.2% | 41.9% | 40.7% | 42.7% | 42.4% |
| Anesthesiology | 69,163 | 31.8% | [28.1%, 35.6%] | 27.9% | 29.1% | 29.6% | 30.9% | 32.5% | 32.2% | 33.0% | 34.0% | 33.7% | 34.7% |
| Arthritis & Rheumatology | 56,397 | 37.8% | [33.8%, 41.9%] | 35.2% | 35.3% | 36.1% | 35.8% | 38.2% | 38.0% | 39.3% | 39.9% | 39.5% | 39.9% |
| Cardiovascular System & Hematology | 289,530 | 26.9% | [25.5%, 28.3%] | 24.2% | 24.7% | 25.7% | 25.7% | 26.7% | 27.1% | 27.9% | 28.7% | 28.7% | 28.6% |
| Complementary & Alternative Medicine | 20,967 | 42.5% | [36.2%, 48.8%] | 38.0% | 41.5% | 40.7% | 40.0% | 41.4% | 41.9% | 42.9% | 43.7% | 44.4% | 45.0% |
| Dentistry | 105,485 | 32.7% | [29.6%, 35.8%] | 28.0% | 29.1% | 29.2% | 31.6% | 32.3% | 34.0% | 33.8% | 34.5% | 35.9% | 35.6% |
| Dermatology & Venereal Diseases | 91,615 | 42.2% | [38.7%, 45.8%] | 38.1% | 39.3% | 40.3% | 41.1% | 42.5% | 42.9% | 43.9% | 44.1% | 44.5% | 46.0% |
| Emergency & Critical Care Medicine | 60,210 | 28.8% | [25.9%, 31.6%] | 25.4% | 25.8% | 26.2% | 27.4% | 28.0% | 29.8% | 30.7% | 30.8% | 31.8% | 31.6% |
| Endocrinology & Metabolism | 138,133 | 39.7% | [37.3%, 42.1%] | 36.7% | 37.2% | 38.3% | 38.5% | 38.9% | 40.1% | 40.8% | 41.0% | 42.0% | 42.5% |
| Environmental & Occupational Health | 21,320 | 41.1% | [35.4%, 46.9%] | 35.0% | 38.6% | 39.5% | 42.0% | 40.5% | 40.5% | 46.0% | 42.9% | 43.5% | 43.6% |
| Gastroenterology & Hepatology | 143,446 | 30.2% | [27.9%, 32.6%] | 27.3% | 28.1% | 29.0% | 29.3% | 29.9% | 30.2% | 30.7% | 31.7% | 32.5% | 33.1% |
| General & Internal Medicine | 369,652 | 35.5% | [33.5%, 37.6%] | 32.6% | 32.7% | 33.6% | 34.3% | 35.4% | 36.0% | 36.5% | 36.8% | 38.2% | 38.6% |
| General Clinical Medicine | 41,627 | 37.0% | [32.1%, 41.8%] | 32.4% | 33.0% | 34.4% | 35.9% | 35.9% | 39.0% | 38.1% | 39.5% | 40.4% | 39.9% |
| Geriatrics | 26,456 | 45.4% | [42.4%, 48.3%] | 39.8% | 42.1% | 44.0% | 43.7% | 45.1% | 45.5% | 47.3% | 47.1% | 48.6% | 49.1% |
| Immunology | 182,478 | 38.1% | [36.4%, 39.9%] | 35.1% | 35.5% | 36.9% | 37.4% | 38.1% | 38.9% | 39.2% | 39.6% | 39.7% | 40.3% |
| Legal & Forensic Medicine | 19,806 | 35.5% | [29%, 42%] | 32.9% | 33.7% | 35.9% | 34.7% | 34.8% | 35.7% | 35.7% | 36.6% | 36.5% | 37.3% |
| Neurology & Neurosurgery | 449,953 | 34.3% | [32.8%, 35.7%] | 31.9% | 32.3% | 33.3% | 33.6% | 33.9% | 34.4% | 34.9% | 35.5% | 35.8% | 37.0% |
| Nuclear Medicine & Medical Imaging | 185,367 | 27.7% | [25.3%, 30%] | 25.0% | 25.5% | 26.5% | 27.5% | 28.4% | 27.8% | 28.4% | 29.1% | 29.2% | 28.7% |
| Obstetrics & Reproductive Medicine | 138,968 | 44.5% | [41.7%, 47.4%] | 39.5% | 41.3% | 42.3% | 42.9% | 43.8% | 45.0% | 45.9% | 47.4% | 47.6% | 48.7% |
| Oncology & Carcinogenesis | 387,667 | 37.6% | [36.1%, 39%] | 33.8% | 34.6% | 35.7% | 36.2% | 36.6% | 37.3% | 38.1% | 39.3% | 39.6% | 40.7% |
| Ophthalmology & Optometry | 105,921 | 34.3% | [31.6%, 37.1%] | 31.3% | 32.2% | 33.4% | 33.9% | 34.7% | 34.5% | 35.1% | 35.4% | 35.9% | 36.4% |
| Orthopedics | 124,360 | 19.9% | [17.7%, 22.1%] | 16.5% | 17.5% | 18.9% | 19.1% | 19.5% | 20.6% | 20.8% | 21.4% | 21.3% | 21.7% |
| Otorhinolaryngology | 66,710 | 29.5% | [26.3%, 32.7%] | 25.6% | 26.3% | 28.2% | 28.2% | 29.4% | 29.7% | 31.0% | 31.7% | 32.3% | 32.1% |
| Pathology | 47,994 | 38.9% | [35.9%, 42%] | 37.0% | 36.9% | 38.6% | 39.2% | 39.0% | 38.6% | 39.2% | 39.9% | 40.3% | 40.6% |
| Pediatrics | 107,270 | 44.4% | [41.2%, 47.6%] | 41.3% | 41.9% | 43.1% | 42.4% | 43.7% | 45.4% | 45.9% | 46.8% | 46.8% | 45.4% |
| Pharmacology & Pharmacy | 252,191 | 38.2% | [36.1%, 40.3%] | 35.6% | 36.2% | 36.9% | 37.2% | 36.8% | 37.6% | 38.9% | 39.9% | 40.1% | 40.3% |
| Psychiatry | 138,600 | 39.7% | [37.7%, 41.6%] | 35.4% | 35.5% | 37.0% | 38.5% | 39.2% | 40.1% | 41.2% | 41.9% | 42.6% | 43.7% |
| Respiratory System | 94,701 | 26.6% | [24.3%, 28.8%] | 23.5% | 24.8% | 25.0% | 25.8% | 25.6% | 26.7% | 27.4% | 28.0% | 29.0% | 29.5% |
| Sport Sciences | 54,592 | 28.9% | [25.7%, 32.1%] | 26.7% | 27.0% | 27.9% | 28.8% | 29.1% | 29.4% | 29.2% | 29.8% | 28.9% | 30.6% |
| Surgery | 152,811 | 24.9% | [22.4%, 27.4%] | 21.9% | 22.2% | 23.0% | 23.1% | 25.2% | 25.3% | 26.1% | 26.8% | 26.4% | 27.5% |
| Tropical Medicine | 44,812 | 37.2% | [32.3%, 42.1%] | 35.4% | 35.0% | 36.0% | 36.2% | 37.1% | 37.0% | 37.0% | 38.3% | 38.5% | 40.0% |
| Urology & Nephrology | 119,490 | 25.8% | [23.7%, 27.8%] | 22.4% | 23.3% | 23.2% | 24.6% | 25.8% | 26.4% | 27.2% | 28.2% | 28.1% | 29.2% |
| **Psychology & Cognitive Sciences** | **331,013** | **42.4%** | [40%, 44.8%] | **40.2%** | **40.0%** | **41.3%** | **41.6%** | **41.4%** | **42.5%** | **42.5%** | **43.3%** | **44.0%** | **46.2%** |
| Behavioral Science & Comparative Psychol. | 31,858 | 41.5% | [38.2%, 44.8%] | 38.4% | 39.5% | 40.4% | 40.9% | 42.1% | 40.3% | 42.0% | 42.7% | 44.3% | 44.0% |
| Clinical Psychology | 37,486 | 48.2% | [46.1%, 50.3%] | 43.5% | 43.6% | 46.6% | 46.7% | 46.8% | 49.0% | 50.2% | 51.0% | 50.8% | 51.6% |
| Developmental & Child Psychology | 45,828 | 57.6% | [55.2%, 60%] | 54.5% | 54.6% | 55.9% | 57.0% | 56.4% | 57.8% | 57.7% | 58.9% | 60.4% | 60.7% |
| Experimental Psychology | 94,321 | 36.4% | [34.8%, 38%] | 33.5% | 33.8% | 34.9% | 35.2% | 36.0% | 36.4% | 36.5% | 37.0% | 37.5% | 40.7% |
| General Psychology & Cognitive Sciences | 9,427 | 51.5% | [47.3%, 55.7%] | 49.9% | 51.6% | 51.2% | 49.0% | 51.5% | 52.7% | 51.5% | 50.8% | 52.4% | 54.0% |
| Human Factors | 36,946 | 31.7% | [28.2%, 35.2%] | 32.2% | 30.8% | 32.5% | 31.9% | 30.3% | 31.6% | 32.0% | 30.9% | 32.3% | 33.7% |
| Psychoanalysis | 11,125 | 39.7% | [36.7%, 42.8%] | 39.8% | 35.2% | 39.5% | 39.3% | 37.2% | 41.2% | 40.6% | 39.9% | 42.8% | 41.9% |
| Social Psychology | 64,022 | 42.8% | [40.3%, 45.3%] | 40.7% | 39.6% | 40.5% | 41.6% | 42.3% | 43.6% | 43.2% | 44.5% | 44.9% | 45.2% |
| **Public Health & Health Services** | **464,456** | **55.2%** | [52.6%, 57.8%] | **53.1%** | **53.8%** | **54.1%** | **54.6%** | **55.1%** | **55.0%** | **55.9%** | **56.3%** | **56.5%** | **56.5%** |
| Epidemiology | 32,261 | 44.6% | [40.6%, 48.7%] | 41.2% | 42.6% | 43.6% | 43.2% | 43.8% | 44.9% | 46.3% | 46.4% | 47.3% | 46.7% |
| Gerontology | 26,394 | 52.9% | [49.4%, 56.4%] | 51.4% | 51.8% | 51.8% | 51.4% | 53.5% | 53.5% | 53.1% | 53.1% | 54.2% | 54.4% |
| Health Policy & Services | 42,346 | 41.4% | [38.8%, 44.1%] | 37.5% | 38.5% | 39.7% | 39.9% | 40.8% | 40.9% | 42.3% | 42.8% | 44.8% | 45.5% |
| Nursing | 106,541 | 73.7% | [71.7%, 75.7%] | 75.1% | 74.9% | 73.7% | 74.4% | 74.7% | 74.0% | 73.8% | 73.2% | 72.5% | 71.5% |
| Public Health | 136,885 | 53.4% | [51.5%, 55.3%] | 50.5% | 51.3% | 51.5% | 52.9% | 53.0% | 53.5% | 53.9% | 54.6% | 54.7% | 55.9% |
| Rehabilitation | 63,402 | 52.1% | [48.8%, 55.3%] | 50.0% | 51.8% | 51.3% | 51.6% | 53.1% | 51.4% | 51.7% | 53.0% | 53.5% | 51.9% |
| Speech-Language Pathology & Audiology | 23,524 | 46.9% | [42.6%, 51.1%] | 40.0% | 41.3% | 44.5% | 44.5% | 44.0% | 43.8% | 53.4% | 53.2% | 54.5% | 55.8% |
| Substance Abuse | 33,104 | 44.8% | [42.4%, 47.3%] | 41.3% | 41.9% | 43.1% | 43.3% | 45.1% | 45.4% | 46.0% | 46.5% | 46.5% | 47.0% |

Note: 95% RI = lower and upper limit for a 95% reliability interval.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

# Table XII Proportion of women in economics & social sciences, and in general journals, by field & subfield, 2006–2015

| | 2006-2015 | | | Proportion (P) of women | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Domain / Field / Subfield** | **Papers** | **P women** | **95% RI** | **2006** | **2007** | **2008** | **2009** | **2010** | **2011** | **2012** | **2013** | **2014** | **2015** |
| **TOTAL (S&E in Scopus)** | **19,113,853** | **31.1%** | [28.4%, 33.9%] | **28.0%** | **28.6%** | **29.5%** | **30.2%** | **30.8%** | **31.5%** | **32.0%** | **32.5%** | **33.1%** | **33.9%** |
| **Economic & Social Sciences** | **1,089,594** | **36.4%** | [33%, 39.7%] | **32.8%** | **33.3%** | **34.0%** | **35.3%** | **36.4%** | **36.7%** | **36.9%** | **37.8%** | **38.2%** | **39.6%** |
| **Economics & Business** | **469,931** | **29.1%** | [25.5%, 32.8%] | **24.8%** | **25.7%** | **26.6%** | **28.7%** | **29.3%** | **30.1%** | **29.7%** | **30.6%** | **31.0%** | **31.8%** |
| Accounting | 11,332 | 30.0% | [25.3%, 34.8%] | 27.7% | 29.0% | 27.8% | 28.5% | 28.3% | 30.3% | 32.3% | 32.0% | 29.9% | 31.8% |
| Agricultural Economics & Policy | 14,870 | 27.0% | [20.8%, 33.2%] | 22.0% | 24.7% | 26.2% | 26.4% | 26.7% | 27.2% | 28.4% | 28.9% | 29.1% | 28.5% |
| Business & Management | 132,907 | 34.4% | [31.7%, 37.1%] | 29.3% | 30.4% | 31.5% | 33.7% | 35.1% | 36.4% | 35.2% | 35.8% | 35.8% | 36.8% |
| Development Studies | 16,030 | 34.6% | [28.2%, 40.9%] | 31.3% | 33.0% | 32.4% | 34.4% | 34.4% | 33.5% | 34.8% | 34.8% | 36.8% | 39.1% |
| Econometrics | 6,309 | 17.2% | [8.6%, 25.9%] | 13.5% | 14.0% | 15.2% | 15.4% | 16.5% | 18.9% | 19.9% | 18.5% | 20.9% | 17.4% |
| Economic Theory | 8,567 | 15.4% | [9.2%, 21.5%] | 14.5% | 13.7% | 15.8% | 13.5% | 16.4% | 16.4% | 14.3% | 16.5% | 16.4% | 16.2% |
| Economics | 124,290 | 24.1% | [21.5%, 26.6%] | 19.2% | 20.2% | 21.2% | 23.1% | 23.2% | 22.7% | 24.4% | 27.3% | 27.9% | 27.1% |
| Finance | 32,239 | 20.9% | [16.3%, 25.5%] | 18.9% | 18.6% | 19.1% | 21.4% | 21.3% | 21.0% | 20.9% | 21.3% | 22.1% | 23.3% |
| Industrial Relations | 7,556 | 37.4% | [31.7%, 43%] | 31.3% | 35.3% | 37.5% | 35.2% | 37.2% | 37.1% | 39.2% | 37.9% | 38.9% | 41.8% |
| Logistics & Transportation | 61,968 | 26.3% | [22.2%, 30.3%] | 22.2% | 24.6% | 25.3% | 27.6% | 26.2% | 27.4% | 25.5% | 26.8% | 27.2% | 28.6% |
| Marketing | 34,306 | 35.9% | [32%, 39.8%] | 33.1% | 31.5% | 32.1% | 33.9% | 35.0% | 36.6% | 37.7% | 38.2% | 38.9% | 39.2% |
| Sport, Leisure & Tourism | 19,557 | 39.5% | [34.1%, 44.8%] | 39.1% | 36.4% | 38.4% | 39.7% | 40.5% | 39.1% | 39.8% | 39.5% | 39.7% | 40.3% |
| **Social Sciences** | **619,663** | **41.9%** | [38.8%, 45%] | **38.8%** | **39.5%** | **40.2%** | **40.9%** | **41.7%** | **41.9%** | **42.2%** | **42.8%** | **43.4%** | **45.1%** |
| Criminology | 31,223 | 44.5% | [42.2%, 46.8%] | 40.6% | 42.5% | 42.7% | 43.6% | 43.2% | 43.7% | 43.9% | 45.2% | 48.0% | 48.6% |
| Cultural Studies | 23,574 | 38.9% | [33.1%, 44.6%] | 37.4% | 37.9% | 38.1% | 40.1% | 38.1% | 38.3% | 38.8% | 40.1% | 38.7% | 39.7% |
| Demography | 8,351 | 44.2% | [37.2%, 51.3%] | 44.5% | 41.1% | 42.8% | 43.0% | 41.5% | 44.9% | 44.2% | 45.3% | 47.0% | 46.1% |
| Education | 233,846 | 45.7% | [43.7%, 47.8%] | 41.6% | 43.3% | 44.3% | 45.1% | 46.0% | 45.1% | 46.0% | 46.1% | 47.2% | 49.5% |
| Family Studies | 9,898 | 59.8% | [56.4%, 63.1%] | 55.6% | 58.7% | 57.4% | 57.4% | 59.8% | 61.6% | 59.8% | 61.7% | 61.9% | 62.3% |
| Gender Studies | 8,213 | 75.9% | [70.4%, 81.5%] | 76.9% | 78.4% | 77.7% | 77.0% | 78.7% | 75.0% | 76.0% | 75.0% | 73.5% | 73.1% |
| Geography | 48,207 | 34.8% | [30.8%, 38.7%] | 31.9% | 31.2% | 32.3% | 33.7% | 34.2% | 34.5% | 36.1% | 36.2% | 37.1% | 38.1% |
| Information & Library Sciences | 44,702 | 47.3% | [43.3%, 51.2%] | 45.7% | 45.5% | 45.9% | 46.5% | 47.7% | 48.6% | 46.5% | 48.0% | 48.2% | 50.7% |
| International Relations | 24,756 | 27.2% | [22.9%, 31.4%] | 22.1% | 22.4% | 23.7% | 26.3% | 26.1% | 27.9% | 28.2% | 29.9% | 30.0% | 30.0% |
| Law | 39,048 | 33.4% | [30.2%, 36.6%] | 30.6% | 30.9% | 32.5% | 31.1% | 33.5% | 34.1% | 33.9% | 35.4% | 35.5% | 35.6% |
| Political Science & Public Administration | 67,157 | 32.0% | [29.4%, 34.6%] | 29.4% | 29.7% | 29.6% | 30.3% | 30.7% | 32.3% | 32.7% | 33.3% | 34.3% | 33.7% |
| Science Studies | 19,809 | 29.7% | [24.4%, 35%] | 25.5% | 28.2% | 28.2% | 30.7% | 29.9% | 30.1% | 31.6% | 28.6% | 30.0% | 32.3% |
| Social Sciences Methods | 10,076 | 39.7% | [34%, 45.4%] | 36.6% | 35.5% | 38.8% | 38.7% | 38.3% | 38.9% | 39.9% | 43.7% | 40.5% | 43.8% |
| Social Work | 19,042 | 60.1% | [56.5%, 63.7%] | 58.0% | 58.2% | 56.1% | 56.8% | 60.0% | 61.1% | 60.4% | 61.5% | 63.0% | 63.8% |
| Sociology | 31,762 | 42.0% | [38.7%, 45.3%] | 39.5% | 38.0% | 39.9% | 40.0% | 42.5% | 42.1% | 42.5% | 43.5% | 43.6% | 45.0% |
| **General Science** | **330,571** | **36.0%** | [34.4%, 37.6%] | **34.1%** | **33.9%** | **34.2%** | **33.8%** | **34.9%** | **35.3%** | **36.5%** | **37.2%** | **37.9%** | **37.1%** |

Note: 95% RI = lower and upper limit for a 95% reliability interval.
Source: Compiled by Science-Metrix using Scopus (Elsevier)

# Appendix B – Data underlying report figures

Table XIII Underlying data for Figure 1

| Year | Web of Science | | | | | Scopus | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | papers | authorships | authors/ paper | first name available | | papers | authorships | authors/ paper | first name available | |
| | | | | n | % of all | | | | n | % of all |
| 1996 | 727,425 | 2,622,045 | 3.60 | 6,505 | 0.2% | 923,659 | 3,201,125 | 3.47 | 1,610,328 | 50.3% |
| 1997 | 741,617 | 2,745,184 | 3.70 | 6,999 | 0.3% | 949,662 | 3,362,015 | 3.54 | 1,710,665 | 50.9% |
| 1998 | 748,830 | 2,782,111 | 3.72 | 9,969 | 0.4% | 954,944 | 3,429,186 | 3.59 | 1,791,302 | 52.2% |
| 1999 | 762,186 | 2,876,608 | 3.77 | 8,363 | 0.3% | 965,222 | 3,509,623 | 3.64 | 1,895,528 | 54.0% |
| 2000 | 777,781 | 2,985,247 | 3.84 | 10,947 | 0.4% | 1,013,189 | 3,732,268 | 3.68 | 1,829,671 | 49.0% |
| 2001 | 781,105 | 3,046,479 | 3.90 | 11,693 | 0.4% | 1,034,607 | 3,848,014 | 3.72 | 1,705,746 | 44.3% |
| 2002 | 800,229 | 3,188,680 | 3.98 | 14,039 | 0.4% | 1,083,671 | 4,076,342 | 3.76 | 2,749,549 | 67.5% |
| 2003 | 838,694 | 3,429,742 | 4.09 | 18,319 | 0.5% | 1,163,320 | 4,428,525 | 3.81 | 3,059,138 | 69.1% |
| 2004 | 882,561 | 3,748,041 | 4.25 | 28,605 | 0.8% | 1,296,211 | 5,012,165 | 3.87 | 3,561,718 | 71.1% |
| 2005 | 927,253 | 4,029,205 | 4.35 | 48,989 | 1.2% | 1,484,047 | 5,777,781 | 3.89 | 4,196,847 | 72.6% |
| 2006 | 980,477 | 4,295,038 | 4.38 | 1,654,664 | 38.5% | 1,586,737 | 6,262,054 | 3.95 | 4,584,025 | 73.2% |
| 2007 | 1,050,083 | 4,586,885 | 4.37 | 3,282,938 | 71.6% | 1,687,677 | 6,758,090 | 4.00 | 5,023,032 | 74.3% |
| 2008 | 1,129,441 | 4,952,147 | 4.38 | 3,614,497 | 73.0% | 1,788,987 | 7,180,347 | 4.01 | 5,458,109 | 76.0% |
| 2009 | 1,183,706 | 5,283,497 | 4.46 | 3,938,012 | 74.5% | 1,902,144 | 7,716,151 | 4.06 | 5,968,948 | 77.4% |
| 2010 | 1,226,929 | 5,691,863 | 4.64 | 4,281,031 | 75.2% | 2,011,013 | 8,328,408 | 4.14 | 6,492,141 | 78.0% |
| 2011 | 1,308,110 | 6,408,631 | 4.90 | 4,742,711 | 74.0% | 2,156,203 | 9,217,743 | 4.27 | 7,178,292 | 77.9% |
| 2012 | 1,375,335 | 7,177,501 | 5.22 | 5,225,136 | 72.8% | 2,232,912 | 10,090,486 | 4.52 | 7,732,517 | 76.6% |
| 2013 | 1,451,327 | 7,531,312 | 5.19 | 5,734,421 | 76.1% | 2,314,550 | 10,462,374 | 4.52 | 8,332,036 | 79.6% |
| 2014 | 1,490,237 | 7,839,751 | 5.26 | 6,097,782 | 77.8% | 2,318,257 | 10,787,377 | 4.65 | 8,647,846 | 80.2% |
| 2015 | 1,455,361 | 7,903,934 | 5.43 | 6,186,711 | 78.3% | 1,643,360 | 8,132,251 | 4.95 | 6,540,139 | 80.4% |

Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics) and Scopus (Elsevier)

## Table XIV Underlying data for Figure 2

| Year | Link to address | First name recorded | Link to address + first name recorded |
|---|---|---|---|
| 2001 | 0.2% | 0.4% | 0.1% |
| 2002 | 0.3% | 0.4% | 0.2% |
| 2003 | 0.4% | 0.5% | 0.3% |
| 2004 | 0.6% | 0.8% | 0.4% |
| 2005 | 0.7% | 1.2% | 0.5% |
| 2006 | 0.9% | 38.5% | 0.6% |
| 2007 | 11.4% | 71.6% | 7.9% |
| 2008 | 92.2% | 73.0% | 67.2% |
| 2009 | 92.3% | 74.5% | 68.4% |
| 2010 | 93.0% | 75.2% | 69.8% |
| 2011 | 93.9% | 74.0% | 69.2% |
| 2012 | 94.4% | 72.8% | 68.4% |
| 2013 | 96.4% | 76.1% | 73.3% |
| 2014 | 97.0% | 77.8% | 75.5% |
| 2015 | 97.5% | 78.3% | 76.4% |

Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics)

## Table XV Underlying data for Figure 3

| | Scopus | Web of Science | |
|---|---|---|---|
| | all authors | corresp. author | all authors |
| 1996 | 100.0% | 69.4% | 0.2% |
| 1997 | 100.0% | 71.5% | 0.2% |
| 1998 | 100.0% | 97.7% | 0.3% |
| 1999 | 100.0% | 97.9% | 0.2% |
| 2000 | 100.0% | 98.2% | 0.2% |
| 2001 | 100.0% | 98.3% | 0.3% |
| 2002 | 100.0% | 98.4% | 0.3% |
| 2003 | 100.0% | 99.1% | 0.4% |
| 2004 | 100.0% | 99.7% | 0.6% |
| 2005 | 100.0% | 99.7% | 0.7% |
| 2006 | 100.0% | 99.6% | 1.0% |
| 2007 | 100.0% | 99.7% | 11.9% |
| 2008 | 100.0% | 99.7% | 88.5% |
| 2009 | 100.0% | 99.7% | 89.5% |
| 2010 | 100.0% | 99.8% | 90.3% |
| 2011 | 100.0% | 99.8% | 90.9% |
| 2012 | 100.0% | 99.8% | 91.3% |
| 2013 | 100.0% | 99.9% | 92.1% |
| 2014 | 100.0% | 99.9% | 92.7% |
| 2015 | 100.0% | 99.9% | 94.6% |

Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics) and Scopus (Elsevier)

Table XVI Underlying data for Figure 4

| Year | corresp. author | all authors |
|---|---|---|
| 2001 | 0.5% | 0.1% |
| 2002 | 0.5% | 0.2% |
| 2003 | 0.7% | 0.3% |
| 2004 | 0.9% | 0.4% |
| 2005 | 1.6% | 0.5% |
| 2006 | 42.4% | 0.6% |
| 2007 | 75.5% | 7.9% |
| 2008 | 76.7% | 67.2% |
| 2009 | 78.1% | 68.4% |
| 2010 | 79.7% | 69.8% |
| 2011 | 80.7% | 69.2% |
| 2012 | 82.1% | 68.4% |
| 2013 | 83.5% | 73.3% |
| 2014 | 84.6% | 75.5% |
| 2015 | 85.4% | 76.4% |

Source: Compiled by Science-Metrix using the WoS (Clarivate Analytics)

Table XVII Underlying data for Figure 5

| | Proportion of women | |
|---|---|---|
| | not weighted | weighted |
| 2006 | 27.1% | 28.1% |
| 2007 | 27.6% | 28.7% |
| 2008 | 28.4% | 29.6% |
| 2009 | 29.0% | 30.3% |
| 2010 | 29.6% | 30.9% |
| 2011 | 30.3% | 31.6% |
| 2012 | 30.7% | 32.1% |
| 2013 | 31.3% | 32.7% |
| 2014 | 31.9% | 33.3% |
| 2015 | 32.9% | 34.0% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

Table XVIII Underlying data for Figure 6

| Year | point estimate | measurement error | sampling error |
|---|---|---|---|
| 2006 | 28.0% | ±1.4% | ±1.8% |
| 2007 | 28.6% | ±1.3% | ±1.7% |
| 2008 | 29.5% | ±1.3% | ±1.7% |
| 2009 | 30.2% | ±1.3% | ±1.6% |
| 2010 | 30.8% | ±1.3% | ±1.5% |
| 2011 | 31.5% | ±1.2% | ±1.4% |
| 2012 | 32.0% | ±1.2% | ±1.4% |
| 2013 | 32.5% | ±1.2% | ±1.3% |
| 2014 | 33.1% | ±1.2% | ±1.3% |
| 2015 | 33.9% | ±1.4% | ±1.4% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

Table XIX Underlying data for Figure 7

| Year | point estimate | measurement error | sampling error |
|---|---|---|---|
| 2006 | 25.7% | ±0.33% | ±0.41% |
| 2007 | 26.1% | ±0.33% | ±0.39% |
| 2008 | 26.9% | ±0.33% | ±0.40% |
| 2009 | 27.4% | ±0.33% | ±0.39% |
| 2010 | 28.0% | ±0.33% | ±0.38% |
| 2011 | 28.8% | ±0.32% | ±0.37% |
| 2012 | 29.3% | ±0.31% | ±0.36% |
| 2013 | 29.9% | ±0.31% | ±0.35% |
| 2014 | 30.7% | ±0.31% | ±0.35% |
| 2015 | 32.4% | ±0.36% | ±0.40% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

Table XX Underlying data for Figure 8

| Field | point estimate | measurement error | sampling error |
|---|---|---|---|
| Public Hlth & Hlth Serv | 55.2% | ±1.34% | ±1.23% |
| Psychol & Cognitive Sci | 42.4% | ±1.32% | ±1.10% |
| Social Sciences | 41.9% | ±1.90% | ±1.23% |
| Biomedical Research | 38.2% | ±1.13% | ±1.36% |
| General Sci & Tech | 36.0% | ±0.85% | ±0.74% |
| Agric, Fisheries & Forestry | 35.4% | ±1.88% | ±2.86% |
| Clinical Medicine | 34.1% | ±1.05% | ±1.27% |
| Biology | 33.0% | ±1.49% | ±1.86% |
| Chemistry | 32.5% | ±1.20% | ±1.38% |
| Economics & Business | 29.1% | ±2.07% | ±1.57% |
| Enabling & Strategic Tech | 29.1% | ±1.06% | ±1.34% |
| Earth & Environ Sci | 28.1% | ±1.55% | ±2.15% |
| Built Envir & Design | 26.9% | ±2.40% | ±2.63% |
| Engineering | 26.3% | ±1.44% | ±1.74% |
| ICT | 25.5% | ±0.97% | ±0.91% |
| Mathematics & Statistics | 22.6% | ±2.10% | ±2.50% |
| Physics & Astronomy | 22.2% | ±1.27% | ±1.93% |
| **TOTAL (all S&E fields)** | **31.1%** | **±1.28%** | **±1.50%** |

Source: Compiled by Science-Metrix using Scopus (Elsevier)

## Table XXI Underlying data for Figure 9

| Country | point estimate | measurement error | sampling error |
|---|---|---|---|
| Thailand | 45.9% | ±5.85% | ±3.13% |
| Serbia | 45.7% | ±2.06% | ±3.34% |
| Rep. of Korea | 44.3% | ±4.07% | ±3.29% |
| Croatia | 43.9% | ±2.85% | ±2.98% |
| Argentina | 43.5% | ±1.36% | ±3.09% |
| Romania | 43.3% | ±1.84% | ±2.59% |
| China | 42.8% | ±1.64% | ±1.57% |
| Brazil | 37.9% | ±1.21% | ±1.00% |
| Portugal | 37.0% | ±1.38% | ±2.68% |
| Finland | 36.2% | ±1.72% | ±2.24% |
| Poland | 35.4% | ±0.51% | ±1.49% |
| Slovenia | 35.2% | ±3.08% | ±3.32% |
| Tunisia | 34.9% | ±3.95% | ±4.80% |
| Slovakia | 34.7% | ±1.98% | ±4.43% |
| Malaysia | 33.6% | ±3.13% | ±2.67% |
| South Africa | 33.5% | ±3.52% | ±4.46% |
| Spain | 32.7% | ±0.73% | ±1.41% |
| Australia | 31.9% | ±1.00% | ±1.38% |
| Turkey | 31.9% | ±1.22% | ±1.26% |
| New Zealand | 31.8% | ±2.29% | ±3.33% |
| Italy | 31.5% | ±0.41% | ±1.12% |
| Sweden | 30.0% | ±1.13% | ±1.64% |
| Israel | 29.8% | ±2.39% | ±2.00% |
| India | 29.7% | ±1.06% | ±1.44% |
| Ukraine | 29.5% | ±2.95% | ±7.68% |
| Canada | 29.3% | ±0.89% | ±1.12% |
| Mexico | 29.3% | ±1.58% | ±2.39% |
| Singapore | 29.1% | ±4.60% | ±4.38% |
| Ireland | 28.9% | ±1.73% | ±3.14% |
| Norway | 28.8% | ±2.17% | ±2.23% |
| France | 28.8% | ±0.74% | ±1.18% |
| Russia | 28.8% | ±1.77% | ±4.03% |
| Egypt | 28.6% | ±3.38% | ±3.47% |
| United States | 28.5% | ±0.33% | ±0.38% |
| Nigeria | 28.3% | ±7.27% | ±7.71% |
| United Kingdom | 27.8% | ±0.61% | ±0.98% |
| Denmark | 27.5% | ±1.63% | ±2.00% |
| Netherlands | 27.1% | ±1.13% | ±1.60% |
| Chile | 26.7% | ±1.93% | ±3.00% |
| Belgium | 26.2% | ±1.29% | ±2.21% |
| Hungary | 25.2% | ±1.55% | ±2.51% |
| Czech Republic | 25.1% | ±1.01% | ±2.42% |
| Greece | 24.5% | ±1.44% | ±2.58% |
| Pakistan | 24.1% | ±1.75% | ±2.18% |
| Iran | 22.2% | ±1.67% | ±1.67% |
| Switzerland | 22.1% | ±1.08% | ±1.57% |
| Austria | 21.7% | ±1.12% | ±2.04% |
| Germany | 21.1% | ±0.51% | ±0.78% |
| Saudi Arabia | 15.2% | ±3.55% | ±2.71% |
| Japan | 13.5% | ±0.44% | ±0.37% |

Source: Compiled by Science-Metrix using Scopus (Elsevier)