

She Figures 2015

Comprehensive Methodology – New Research & Innovation Output Indicators

Project

RTD-B6-PP-00963-2013

Date

May 26, 2015

Submitted to



EUROPEAN COMMISSION
DIRECTORATE-GENERAL FOR RESEARCH & INNOVATION

Directorate B – Innovation Union and European Research Area
B.7 – Science with and for Society

by

Science-Metrix (Lead) with
ICF International, Beyond 20/20, tKorps bvba, and Imprimerie Centrale (subcontractors)

contact

Éric Archambault, Ph.D.
President and CEO | Science-Metrix Inc.
info@science-metrix.com | 1-514-495-6505



Acknowledgements

The production of this methodology for the She Figures 2015 was led by Science-Metrix, with extensive contributions from the Directorate-General for Research and Innovation of the European Commission; EIGE; Eurostat; the Helsinki Group on Gender in Research and Innovation and their Statistical Correspondents; ICF International, KU Leuven and the OECD.

Contents

Acknowledgements.....	i
Contents.....	ii
Tables.....	iii
Figures.....	iii
1 Introduction.....	1
2 Output size, quality, and collaboration in scientific research as well as output size in innovation.....	2
2.1 Ratio of women-to-men scientific authorships.....	4
2.1.1 Definition.....	4
2.1.2 Source of data.....	4
2.1.3 Availability over time.....	5
2.1.4 International availability.....	5
2.1.5 Availability across Fields of Science and Technology (FOS).....	6
2.1.6 Comparability.....	6
2.1.7 Calculation method.....	11
2.2 Ratio of women-to-men international co-authorship rate.....	17
2.2.1 Definition.....	17
2.2.2 Source of data.....	17
2.2.3 Availability over time.....	17
2.2.4 International availability.....	17
2.2.5 Availability across Fields of Science and Technology.....	17
2.2.6 Comparability.....	17
2.2.7 Calculation method.....	17
2.3 Ratio of women-to-men scientific quality/impact.....	20
2.3.1 Definition.....	20
2.3.2 Source of data.....	20
2.3.3 Availability over time.....	21
2.3.4 International availability.....	21
2.3.5 Availability across Fields of Science and Technology.....	21
2.3.6 Comparability.....	21
2.3.7 Calculation method.....	21
2.4 Ratio of women-to-men inventorships.....	24
2.4.1 Definition.....	24
2.4.2 Source of data.....	25
2.4.3 Availability over time.....	25
2.4.4 International availability.....	25
2.4.5 Availability across technological fields (IPC classes).....	26
2.4.6 Comparability.....	26
2.4.7 Calculation method.....	28
3 Gender dimension in research content.....	31
3.1 Proportion of a country's research outputs integrating a gender dimension in its research content (GDRC).....	32
3.1.1 Definition.....	32
3.1.2 Source of data.....	32
3.1.3 Availability over time.....	32
3.1.4 International availability.....	32
3.1.5 Availability across Fields of Science and Technology (FOS).....	32
3.1.6 Comparability.....	32
3.1.7 Calculation method.....	33
Appendix 1: Match between Science-Metrix's classification and the FOS in the Frascati Manual.....	35
Appendix 2: Coverage bias across subfields in the WoS.....	37
Appendix 3: Gender name disambiguation.....	41

Appendix 4: Gender dimension in research content.....	55
Appendix 5: The various queries to build the dataset on Gender Dimension in Research Content	70

Tables

Table 1	Share of papers for which the author-address link and full given name are available in the WoS based on reprint authors, 2002–2013.....	7
Table 2	Share of paper fractions for which the author-address link and full given name are available in the WoS based on all authors, 2008–2013.....	8
Table 3	Percentage of papers for which the given name of the reprint author is available by field in the WoS, 2007–2013.....	9
Table 4	Relationship between the recall and ratio of women-to-men authorships by field in the WoS, 2013.....	9
Table 5	Share of reprint author in first and last position by country in the WoS (2007–2013).....	13
Table 6	Share of EPO patent applications for which the country of affiliation and full given name of all inventors appearing on an application is available in PATSTAT, 2002–2013.....	27
Table 7	Correspondence table between Science-Metrix’s classification and the Field of Science and Technology (FOS) as defined in the Frascati Manual (revised classification of 2007).....	35
Table 8	Number of scientific publications beyond the WoS (2007–2013).....	38
Table 9	Validation methods of the gender assignment procedure for author names in the WoS (2007–2013) by country for those covered in She Figures 2015.....	47
Table 10	Validation results of the gender assignment procedure for author names in the WoS (2007–2013) by country for those covered in She Figures 2015.....	50
Table 11	Validation methods of the gender assignment procedure for inventor names in PATSTAT (2002–2013) by country for those covered in She Figures 2015.....	51
Table 12	Validation results of the gender assignment procedure for inventor names in PATSTAT (2002–2013) by country for those covered in She Figures 2015.....	54
Table 13	WoS journals containing the term <i>gender</i> in their name.....	58
Table 14	WoS journals containing publications classified under the subfield Gender Studies.....	58
Table 15	MeSH terms, the number of papers associated with them in the WoS and the verdict for insertion of the associated publications in the seed dataset.....	59
Table 16	TF-IDF weight of the top items appearing in EIGE draft thesaurus on gender equality terms and of the noun phrases extracted from WoS papers.....	60
Table 17	Items with low TF-IDF weight that will not be included in the keyword query.....	61
Table 18	Distribution of publications in the GDRC dataset across fields of science.....	63
Table 19	Recall of the seed dataset (i.e. gender studies subfield, specialist journals and MeSH terms) and of each of the specialist journals using the keyword-based query.....	66
Table 20	Precision of the GDRC dataset as a whole and for some fields in which relevant papers are less likely.....	67

Figures

Figure 1	Share of EPO patent applications for which the country of affiliation and full given name of all inventors appearing on an application are available in PATSTAT by IPC class, 2002–2013.....	28
Figure 2	Validation of NamSor™ GendRE API using data from the Official Directory of the European Union.....	45
Figure 3	Keyword map of the search expressions used in the query.....	65

1 Introduction

This report presents the comprehensive methodology implemented by the research team towards producing the new research and innovation (R&I) output indicators that have been retained for integration in the She Figures 2015 publication. These indicators consist of primary data sources computed as part of the current project using external data sources (i.e. bibliographic databases covering the peer-reviewed scientific literature to measure sex disparities in scientific production, and patent applications to measure sex inequalities in innovation) and are as follows:

1. Ratio of women-to-men scientific authorships;
2. Ratio of women-to-men international co-authorship rate;
3. Ratio of women-to-men scientific quality/impact;
4. Ratio of women-to-men inventorships;
5. Proportion of a country's scientific production including a Gender Dimension in its Research Content (GDRC).

The process used in selecting these R&I output indicators entailed the following:

- The identification of policy issues surrounding gender in science, research and innovation (e.g. horizontal segregation, vertical segregation, and the funding gap) in the European Union and Associated Countries that could be monitored through indicators; in other words, the new indicators must be of relevance to such policy issues.
- An analysis of the availability of timely, representative and validated time series for each of the suggested new indicators.
- An analysis of international availability and cross-country comparability, at least across the EU Member States and Associated Countries, for each of the suggested new indicators.
- An assessment of the accuracy of the suggested new indicators, especially the indicators that do not rely on international standards and definitions, such as the OECD Frascati and Canberra manuals, and official, internationally comparable data from statistical offices.

Following this exhaustive quality review process, the proposed new indicators were reviewed by Commission officials and the statistical correspondents/experts of the She Figures 2015 project, as well as by members of the Helsinki group. Through interactions between the research team and these experts, iterative rounds of improvement were conducted until the retained indicators satisfactorily fulfilled the above four quality requirements for inclusion in the She Figures 2015 publication.

In this methodological report, the indicators were grouped in the same section if they shared a common rationale. Since the four women-to-men ratios mentioned above (indicators #1 to #4) all share the same rationale, they are all presented in Section 2, while the GDRC indicator (indicator #5) is presented separately in Section 3. Each of these two indicator sections is introduced by presenting the general rationale (i.e. policy relevance) for the inclusion of the indicators they respectively cover. For each indicator, a short methodological note follows presenting its definition, source of data, availability over time, international availability, and comparability, as well as its calculation method. This latter item includes a full account of the computation of confidence intervals and other statistics used in reporting the accuracy of the indicators. Some items requiring deeper explanations to allow a third party to reproduce the data computed for the new R&I output indicators included in She Figures 2015 are presented in appendices. These appendices are introduced wherever appropriate in Sections 2 and 3.

2 Output size, quality, and collaboration in scientific research as well as output size in innovation

Many countries in the EU have established research policies promoting gender equality in research including the Austrian Science Fund (FWF), the Academy of Finland, the German Research Foundation (DFG), the Netherlands Organisation for Scientific Research (NOW), the Research Council of Norway, the Science Foundation Ireland (SFI), the Swedish Research Council, the Swiss National Science Foundation (SNSF), and the UK Research Councils. For example, some of them are planning or have already performed studies and monitoring activities on gender equality in research funding.¹ Despite these initiatives, there still appear to be important gaps in research funding with female researchers lagging behind their male counterparts, and these gaps are even maintained in the more proactive countries in this matter – namely, the Nordic countries.²

To some extent this might be attributable to the concomitant rise in the emphasis countries have placed on promoting research ‘excellence’ with the implementation of numerous programmes specifically designed to support the most outstanding researchers. One such example is the prestigious funding provided through the European Research Council (ERC), which was first instigated by the European Commission under the Seventh Framework Programme to support research excellence throughout Europe. For instance, such programmes might, in their current implementation, prove to be incompatible with initiatives fostering gender equality in research funding. For example, this might be the case if men are at an advantage with respect to some of the key dimensions (see below) currently in use for rating the ‘excellence’ of researchers in grant competitions. For this reason, it appears highly relevant to monitor gender differences in such dimensions.

The reliance on bibliometric statistics for research evaluation purposes in research assessment exercises (RAE) 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, especially in the context of grant competitions targeted at ‘excellence’. In grant competitions focusing more heavily on applied research, such as in the European Commission’s Framework Programmes where the transfer of knowledge from academia to the private sector is a key aspect aimed at fostering economic growth through innovation, the number of patents on which a researcher is listed as an inventor might also prove to be a decisive factor in the funding decision.

Hence, if women are at a disadvantage relative to their male counterparts in terms of the size of their contribution to scientific knowledge, they might very well get stuck in a vicious circle whereby the smaller size of their scientific and technological production would reduce their chances of being funded or the actual amount of funding they manage to secure, which would in turn reduce their capacity to produce more outputs. In addition, there are intrinsic relationships between the size of

¹ European Commission, Directorate-General for Research (2009). *The gender challenge in research funding: Assessing the European national scenes*. EUR 23721 EN: http://ec.europa.eu/research/science-society/document_library/pdf_06/gender-challenge-in-research-funding_en.pdf.

² Louët, S. (2014). *Research funding gap: Her excellence dwarfed by his excellence*. <http://euroscientist.com/2014/06/research-funding-gap-excellence-dwarfed-excellence/>.

a researcher's publication portfolio and other characteristics of such a portfolio that can have a negative effect on the perceived scientific quality/impact of his or her work – that is, a dimension that also carries a non-negligible weight in the peer-review process of grant proposals.

For example, there is ample evidence in the scientific literature demonstrating the presence of a Matthew effect in science – 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'.³ In fact, in a 1999 paper, Katz⁴ 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 citation impacts of a researcher gets larger as his or her pool of papers gets larger, in a similar manner to the phenomenon of 'the rich get richer and the poor get poorer'. Thus, if women lag behind men in terms of production size, they might also trail behind in terms of scientific quality and impact, as these dimensions are typically measured through citation counts – that is, the citation counts of the journals in which they publish for assessing scientific quality (the well-known journal impact factor), and citation counts of their papers for assessing scientific impact.

Similarly, there is ample evidence in the scientific literature of a link between the scientific impact of papers as measured through citation counts and international co-authorships. For instance, Science-Metrix recently showed how the citation impact of papers rises as the number of authors and countries involved on scientific papers increases.⁵ As such, 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 through international partnerships. Indeed, if they trail behind men in this regard, this might amplify, or at least contribute to maintaining, the potential gap in impact and, ultimately, in funding.

Larivière and colleagues⁶ released a 2013 study in which they showed that women still lag behind men in terms of the size and impact of their scientific production as well as the extent to which they are involved in international co-authorships. They suggest that the observed gaps between women and men might very well relate to age differences. They stated:

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.

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

⁴ Katz, J.S. (1999). The self-similar science system. *Research Policy*, 28: 501-517.

⁵ Campbell, D., Côté, G., Haustein, S., Lefebvre, C., and 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.forskningradet.no/servlet/Satellite?blobcol=urldata&blobheader=application%2Fpdf&blobheader_name1=Content-Disposition%3A&blobheadervalue1=+attachment%3B+filename%3D%22SMBibliometricsRCNInterimAnalyticalReport.pdf%22&blobkey=id&blobtable=MungoBlobs&blobwhere=1274503843081&ssbinary=true.

⁶ Larivière, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C. R. (2013), Global gender disparities in science. *Nature*, 504: 211-213.

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. However, they still note some discrepancies in findings across various studies on this matter.

This project builds on previous work done by a number of 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 refined through the computation of confidence intervals accounting for widely recognised biases resulting from the use of some of the most comprehensive databases of peer-reviewed scientific publications. Additionally, the dimensions covered by previous studies have been expanded by adding statistics on inventorships by gender (as measured with patent applications). The new indicators will shed light on some of the conditions inherent to current research and innovation systems that might hinder women from breaking the gender funding gap, informing science policy on this very important issue. The short methodological note for each of these indicators is presented below in the following sequence:

- Ratio of women-to-men scientific authorships (Section 2.1)
- Ratio of women-to-men international co-authorship rate (Section 2.2)
- Ratio of women-to-men scientific quality/impact (Section 2.3)
- Ratio of women-to-men inventorships (Section 2.4)

2.1 Ratio of women-to-men scientific authorships

2.1.1 Definition

This indicator is the ratio of women-to-men authorships, or equivalently, the ratio of the proportion of women authorships (in total authorships) over the equivalent proportion for men. It can be computed at various aggregation levels (e.g. organisations, countries, world regions). A score above 1 indicates that women in a given entity produced a larger share of the entity's scientific publications than men, whereas a score below 1 means the opposite. It is only based on the reprint (i.e. corresponding) author of peer-reviewed scientific publications. The reprint author is used as a proxy to compare the contribution of women relative to that of men when in a leading role (see explanation in Section 2.1.7).

2.1.2 Source of data

This indicator was computed by Science-Metrix using raw bibliographic data derived from the Web of Science (WoS), which is produced by Thomson Reuters. The WoS includes three databases: the Science Citation Index Expanded (SCI Expanded), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). It indexes some 12,000 refereed journals (publishing articles that are peer reviewed prior to publication) and covers various fields of science (e.g. Natural Sciences and Engineering [NSE], Health Sciences [HS] and Social Sciences and Humanities [SSH]).

The WoS was chosen because it includes cited references for each document it incorporates (e.g. articles and chapters, published in a journal or book series), allowing for internal coverage monitoring of the database and analysis of scientific impact based on citations and impact factors.

For instance, Thomson Reuters' monitoring procedure ensures that the most important peer-reviewed journals in their respective fields are indexed. As recently shown,⁷

50% of all citations generated by this collection came from only 300 of the journals. In addition, these 300 top journals produced 30% of all articles published by the total collection.

Because science is not static, the list of key international journals is changing continuously. For this reason, Thomson Reuters is adjusting the coverage of the WoS on a regular basis to reflect the dynamics of the science.

Also, the WoS includes all authors and their institutional affiliations, which allows collaboration rates among various entities (e.g. countries, institutions, and researchers) to be analysed. It also indexes the links between authors and their addresses, a key feature for aggregating sex data by country.

In producing this indicator, as well as other indicators based on the WoS (see subsequent sections), only three document types published in refereed scientific journals – articles, notes, and reviews – were retained, as all have been through the peer-review process prior to being accepted for publication. The peer-review process ensures that the research is of good quality and constitutes an original contribution to scientific knowledge. The terms 'papers', or alternatively 'publications', are used throughout this report when referring to these three document types.

Note that Science-Metrix hosts an in-house version of the WoS in the form of an SQL relational database. This has allowed Science-Metrix to carefully condition the database for the purpose of producing large-scale comparative scientometric analyses. Bibliometrics analysts at Science-Metrix have performed a large number of bibliometric projects with it and thus have in-depth knowledge of its respective strengths and limitations.

2.1.3 Availability over time

This indicator can be computed for 6 out of the 12 years to be covered by the She Figures 2015 report. Although the completeness of this indicator in regard to the time coverage is only at 58%, it should be noted that the years covered (2007 to 2013) form a continuous stretch filling the most recent half of the study period. Thus, the timeliness of the data is good. This indicator cannot be produced prior to 2007 because the full given name of reprint authors on scientific publications – which is essential for attributing a sex to authors – is not available for a sufficiently large proportion of the papers in previous years in order to compute reliable statistics; in other words, the sample size would be too small prior to 2007 (see Section 2.1.6). Note that the yearly data is based on the publication year of papers indexed in the WoS.

2.1.4 International availability

Data can readily be produced for all 41 countries to be covered in the She Figures 2015 publication – that is, for the 28 EU Member States as well as for Albania, Bosnia & Herzegovina, Faroe Islands, Iceland, Israel, Liechtenstein, the Former Yugoslav Republic of Macedonia, Moldova, Montenegro, Norway, Serbia, Switzerland, and Turkey. Thus, the completeness of this indicator in regard to the geographical scope stands at 100%. Note that data by country is obtained by assigning the reprint author to a given country based on affiliation address rather than using the author's nationality. The indicator therefore looks at where the research was produced.

⁷ <http://wokinfo.com/essays/journal-selection-process/>

2.1.5 Availability across Fields of Science and Technology (FOS)

All publications indexed in the WoS were classified into six large domains (Applied Sciences, Arts & Humanities, Economic & Social Sciences, General, Health Sciences and Natural Sciences), then further divided into 22 fields and 176 subfields using Science-Metrix's journal-based classification.⁸ This classification is mutually exclusive (i.e. each article is classified into one and only one set of domain, field and subfield) and was developed for the European Commission within the context of the *Analysis and Regular Update of Bibliometric Indicators* study (RTD 2009_S_158-229751). Using information derived from OECD documentation – that is, the fields of science and technology (FOS) classification in the Frascati Manual⁹ (§Table 3.2) and the revised classification¹⁰ – the subfields in Science-Metrix's classification were matched to their corresponding FOS as defined in the Frascati Manual using their 2007 description. Thus, this indicator has been computed for each of the following six FOS:

- (NS) Natural sciences;
- (ET) Engineering and Technology;
- (MS) Medical sciences;
- (AS) Agricultural sciences;
- (SS) Social sciences; and
- (H) Humanities.

The detailed correspondence table between Science-Metrix's classification and the FOS as defined in the Frascati Manual, as of 2007, is provided in Appendix 1.

2.1.6 Comparability

Comparability becomes an issue when data are being compared across periods, geographical regions and disciplines. Whenever an issue of this type is encountered, efforts are made to eliminate or limit its impact on the data (e.g. limiting the analysis to a subset of countries, or disciplines), and the potential biases that could result from it are clearly stated in Science-Metrix's reports, sometimes in the form of confidence intervals of the estimates. Prior to setting up the bibliometric versions of the WoS, Science-Metrix's senior analysts performed a comprehensive testing of their coverage looking for errors such as:

Bias in the number of documents over time

The proportion of papers for which the full given name and address of the reprint author is available (i.e. the recall) varies over time ranging from a low of 0.2% in 2002 to 83% in 2013 (Table 1). Since these pieces of information are essential for assigning a sex to authors when producing sex disaggregated data, as well as in assigning a country of affiliation to authors in producing country

⁸ Archambault É., Beauchesne, O., and Caruso J. (2011), Towards a multilingual, comprehensive and open scientific journal ontology, in B. Noyons, P. Ngulube, and J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)*, Durban, South Africa, pp 66–77.

⁹ OECD (2002), *Frascati manual*, OECD Publishing, Paris, available at <http://dx.doi.org/10.1787/9789264199040-en>.

¹⁰ OECD (2007), *Working Party of National Experts on Science and Technology Indicators: Revised Field of Science and Technology (FOS) Classification in the Frascati Manual*. Available at <http://www.oecd.org/science/inno/38235147.pdf>.

disaggregated data, this means that samples rather than entire populations of the publications indexed in the WoS are available to compute the ratio of women-to-men authorships. Since the recall increases over time, this implies that the estimated ratios will be more accurate in recent years. In fact, from 2006 to 2007, there is a sharp leap in the recall, going from about 50% to 75% with a slightly increasing recall afterwards. It is considered that the recall prior to 2007 is not adequate to provide reliable statistics. As such these years were not considered in computing this indicator. Still, in the 2007–2013 period, the accuracy of the estimated ratios is likely to increase slightly. To reflect this in the data, margins of error have been computed to construct 90% confidence intervals, thereby reflecting the sampling errors in the computed ratios over time (see sub-section on accuracy in Section 2.1.7).

Table 1 Share of papers for which the author-address link and full given name are available in the WoS based on reprint authors, 2002–2013

Year	Papers with author-address link		Papers with author-address link & full given name	
	No.	% of Total	No.	% of Total
2002	801,100	98.4%	1,870	0.2%
2003	845,349	99.0%	2,752	0.3%
2004	894,491	99.6%	5,473	0.6%
2005	940,316	99.6%	14,121	1.5%
2006	993,880	99.5%	493,702	49.4%
2007	1,044,690	99.6%	789,471	75.2%
2008	1,124,675	99.5%	862,424	76.3%
2009	1,178,893	99.4%	920,601	77.7%
2010	1,219,885	99.5%	972,647	79.3%
2011	1,295,733	99.6%	1,045,437	80.3%
2012	1,294,587	99.6%	1,061,144	81.6%
2013	1,359,734	99.6%	1,133,212	83.0%

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Note that the recall was much lower when the analysis was not limited to the reprint author (i.e. including all authors on a paper); it was around 50% regardless of the year in the 2008–2013 period (Table 2). This is because Thomson Reuters only recently started to index the link between an author and his or her address in the WoS and because the full given name of all authors only recently started to become more systematically available on scientific publications. This is one of the limitations that motivated the choice of the reprint author, instead of all authors, in computing this indicator (see Section 2.1.7 for an exhaustive listing of the many factors that motivated this choice).

Table 2 Share of paper fractions for which the author-address link and full given name are available in the WoS based on all authors, 2008–2013

Year	Papers with author-address link		Papers with author-address link & full given name	
	Sum of paper fractions	% of Total	Sum of paper fractions	% of Total
2008	721,480	64.1%	548,845	48.7%
2009	747,585	63.5%	579,447	49.2%
2010	762,508	62.4%	604,108	49.4%
2011	798,574	61.5%	640,777	49.4%
2012	785,826	60.6%	641,580	49.5%
2013	808,372	59.3%	673,251	49.4%

Note: Paper fractions refer to the fact that each author on a publication is attributed an equal share of the publication. Thus, author fractions are counted.

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Bias in favour of some countries

When counting the absolute number of papers associated with a country, a potential linguistic bias exists in favour of Anglo-Saxon countries because the WoS almost exclusively indexes the peer-reviewed scientific literature published in English-language journals. However, the ratio of women-to-men authorships should not be strongly affected by such a bias. For instance, most scientific literature is currently published in English-language journals, regardless of the country (except in some fields/subfields, see below for comparability across countries in some areas of research). As such, there is no *a priori* reason to think that the ratio of women-to-men authorships would be drastically different within the much smaller realm of scientific literature written in the author's mother tongue. Indeed, both the numerator and denominator in this ratio are likely affected by the same bias, which would therefore cancel out. Of course, this assumes that there is no sex difference in the extent to which women researchers publish in their mother tongue compared to men researchers, but even if this was to be the case, the proportion of the literature authored in the mother tongue likely does not represent a sufficiently large pool of the overall population of scientific papers to reduce the cross-country comparability of this indicator.

Bias in favour of disciplines

The proportion of papers for which the full given name of the reprint author is available (i.e. the recall) varies substantially across fields and subfields based on Science-Metrix's classification (ranging, at the field level, from a low of 64% in Physics & Astronomy to a high of 97% in Communication & Textual Studies in 2013, see Table 3; at the subfield level, it varies from 39% to 100% [data not shown]). Since the full given name is an essential piece of information for assigning a sex to authors when producing sex disaggregated data, this means that some subfields will contribute more, and others less, than they should to the women-to-men ratios computed at higher aggregation levels (e.g. for FOS as defined by the Frascati Manual or for all fields of science [entire database]). This would not be a problem if women and men contributed similarly to the scientific production in each field and subfield. However, this is not the case. In fact, the fields in which women appear to contribute the most are among those for which the recall is highest, and those in which they contribute the least are among those for which the recall is lowest (Table 4, similar finding at the subfield level). Since the former fields are the smaller ones (e.g. Communication & Textual Studies and Public Health & Health Services) and the latter fields are quite large (e.g. Physics &

Astronomy), these important differences in their recall would lead to biases at higher aggregation levels.

Table 3 Percentage of papers for which the given name of the reprint author is available by field in the WoS, 2007–2013

Field	2007	2008	2009	2010	2011	2012	2013
Agriculture, Fisheries & Forestry	58%	62%	64%	66%	68%	69%	71%
Biology	80%	80%	82%	83%	83%	84%	84%
Biomedical Research	86%	87%	87%	88%	89%	89%	90%
Built Environment & Design	79%	81%	81%	82%	84%	84%	85%
Chemistry	81%	82%	83%	84%	85%	86%	88%
Clinical Medicine	79%	79%	81%	83%	84%	84%	85%
Communication & Textual Studies	97%	97%	97%	97%	97%	97%	97%
Earth & Environmental Sciences	67%	67%	67%	70%	70%	72%	73%
Economics & Business	95%	95%	95%	95%	95%	95%	97%
Enabling & Strategic Technologies	63%	65%	67%	70%	72%	75%	77%
Engineering	66%	67%	69%	72%	72%	75%	77%
General Arts, Humanities & Social Sciences	96%	97%	94%	95%	96%	96%	96%
General Science & Technology	90%	92%	94%	94%	96%	96%	97%
Historical Studies	93%	93%	93%	92%	93%	93%	93%
Information & Communication Technologies	87%	88%	88%	88%	87%	88%	91%
Mathematics & Statistics	77%	77%	78%	79%	79%	81%	83%
Philosophy & Theology	95%	92%	93%	95%	95%	96%	96%
Physics & Astronomy	53%	54%	57%	58%	60%	62%	64%
Psychology & Cognitive Sciences	94%	94%	95%	95%	96%	95%	96%
Public Health & Health Services	91%	90%	90%	92%	93%	93%	93%
Social Sciences	95%	96%	96%	96%	96%	96%	96%
Visual & Performing Arts	96%	95%	96%	95%	96%	94%	95%

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Table 4 Relationship between the recall and ratio of women-to-men authorships by field in the WoS, 2013

Field	Recall	Ratio of women to men authorships
Public Health & Health Services	93%	1.3
Communication & Textual Studies	97%	0.8
Psychology & Cognitive Sciences	96%	0.8
Visual & Performing Arts	95%	0.8
Social Sciences	96%	0.7
Historical Studies	93%	0.6
Agriculture, Fisheries & Forestry	71%	0.5
General Arts, Humanities & Social Sciences	96%	0.5
Biology	84%	0.5
Biomedical Research	90%	0.5
Clinical Medicine	85%	0.4
General Science & Technology	97%	0.4
Earth & Environmental Sciences	73%	0.4
Built Environment & Design	85%	0.4
Economics & Business	97%	0.4
Philosophy & Theology	96%	0.4
Chemistry	88%	0.3
Enabling & Strategic Technologies	77%	0.3
Engineering	77%	0.3
Mathematics & Statistics	83%	0.2
Information & Communication Technologies	91%	0.2
Physics & Astronomy	64%	0.2

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

In addition, this issue implies that samples rather than entire populations of the publications indexed in the WoS are available to compute the ratio of women-to-men authorships. Indeed, the proportion of the publications indexed in the WoS that can be used to compute this indicator (i.e. the recall) varies across subfields and often does not approach 100%. As a result, margins of error should be computed to adequately reflect the sampling errors associated with the computed ratios at the subfield level (see below).

Another discipline-related bias exists in the WoS and is in favour of the NSE and HS compared to the SSH. The SSH produce a greater proportion of scientific publications that are not journal articles – especially books. Thus, research in these areas would best be examined using instruments that, unlike traditional bibliometrics, also consider publications in or as books. When counting publications using citation databases (e.g. the WoS) as in this study, a greater portion of the scientific output is omitted in the SSH compared to the NSE and HS. In fact, a comparative analysis of the share of referenced publications that are indexed in the WoS across subfields (based on Science-Metrix's classification) revealed a substantial amount of variability in the proportion of a given subfield's publications that is actually covered in the WoS (ranging from a low of 3% in Art Practice, History & Theory to a high of 90% in Developmental Biology, see Table 8 in Appendix 2). The impact of such variation in the coverage of the scientific literature across subfields is that women-to-men ratios computed at higher aggregation levels (e.g. for FOS or all fields combined) will be biased towards the scores of the subfields that are better covered in the WoS. This must therefore be accounted for in computing aggregated ratios (see below).

Finally, another aspect requiring consideration when performing bibliometric analyses of the SSH is the more local orientation of SSH research. This means that SSH scholars publish somewhat more frequently in a language other than English – and in journals with a national distribution rather than international distribution – than do NSE and HS researchers. Yet because we are producing data on a ratio, this potential bias might cancel out, as mentioned above in the 'Bias in favour of some countries' section.

Because the major citation databases that are suitable for performing bibliometric analyses do not perfectly represent the natural distribution of the scientific literature across scientific subfields, the uninformed or careless use of bibliometrics can lead to erroneous conclusions of the types mentioned above. It is certain that the size of the scientific output of an entity in the SSH should not be compared to the size of its production in other areas. However, this is not the goal that is targeted in producing the ratio of women-to-men authorships. In fact, in this case, such biases in the absolute size of a country's production across disciplines should cancel out since they are present both in the numerator and denominator. Still, the aforementioned biases can have problematic effects on the women-to-men ratios when computed at aggregation levels higher than the subfield one.

As will be seen in Section 2.1.7, this issue has been dealt with diligently by first computing the ratios at the subfield level prior to their aggregation at the FOS level (as well as for all fields combined; i.e. for the whole database) using a weighting scheme of the scores that more adequately reflects their representation within the whole realm of scientific research beyond the WoS (i.e. including the relevant literature not indexed in the WoS). The weighting is obtained by computing the proportion of each subfield in this expanded population using the estimated share of each subfield that is indexed in the WoS (see Appendix 2). Additionally, because the set of publications available to compute the proportions of women and men authorships at the subfield level represent samples rather than entire populations, the random sampling errors of these estimates at the subfield level

have also been aggregated using the above-mentioned weighting schemes to construct the 90% confidence intervals of the aggregated ratios at the FOS level (as well as for all fields combined) (see Section 2.1.7).

2.1.7 Calculation method

Author selection

To compute this indicator, information on the sex and country of authors must first be obtained. The sex is obtained using the name of authors, while the country is obtained using the affiliation address of authors as indicated in scientific publications. For the sex, one must have access to the complete name of an author including his or her full given name (not just the initials) and surname. For the country, one must have access to a link associating each author on a paper with their corresponding affiliation address. As seen above in Section 2.1.3, these two pieces of information are not systematically available, although their inclusion in the WoS has improved over time. In Section 2.1.3, it was also shown that this information is not currently sufficiently available to produce reliable statistics using all authors on a publication, but it is so for the reprint author alone. Thus, the approach implemented for the She Figures 2015 publication relies on the reprint author only.

This approach has advantages over one that would be using all authors. Firstly, the reprint author is usually the author who is in a leading position – that is, the principal investigator. The principal investigator is usually the researcher to whom the project grant was awarded and his or her name may appear in different positions on a publication. In many fields, the team leader/reprint author will appear as the last author on a publication, whereas in other fields he or she will appear as the first author. In other circumstances, the investigator who made the most significant contribution to the publication, which might not be the principal investigator, can appear as the reprint author. In this case, the investigator will often appear as the first author on a publication and he or she may be a graduate student, although this is less common. Indeed, the reprint author name is usually that of an author whose affiliation address is stable, which is most often not the case for graduate students. In fact, in the case of papers involving graduate students as part of their authors, it is more common to observe a graduate student in the first author position and the lead/reprint author in the last position. Finally, in the case of single-author publications, the question is irrelevant and it can be assumed that the author is well established and in some kind of lead position. Consequently, by limiting the analysis to the reprint author, graduate students and other types of contributors who may not end up pursuing a research career are, to some extent, discarded from the analysis. This leaves us with those researchers, women and men, who are more likely to apply for funding and this is the population of interest given the rationale introduced earlier in Section 2. The reprint author is therefore referred to as the ‘lead’ author on a publication.

Other researchers have made use of the first author instead of the reprint author in producing similar statistics on the leading author.¹¹ Although the approach using the corresponding author is imperfect in that graduate students can sometimes appear as reprint authors, this is also the case with first authors and this latter approach is prone to other biases. For example, in some fields, authors are listed alphabetically. In such cases, the first author does not relate at all to a leading position, either as the team leader or as the main contributor. Additionally, the team leader often

¹¹ Larivière, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C.R. (2013). Global gender disparities in science. *Nature*, 504: 211–213.

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 since the share of publications for which it is possible to assign a sex and a country to the first author is smaller than with the reprint author.¹²

Table 5 actually shows the share of reprint authors appearing in first and last positions by country in the WoS from 2007 to 2013. From this analysis, it can be seen that the sum of both shares adds up to at least 95% of the total for all countries covered in the She Figures 2015 publication. In most cases, the numbers add up to more than 100%; indeed, for single-author publications, the reprint author is in both first and last place. Most of the time, the reprint author is in first place (76% compared to 30% for last position on average for countries covered in She Figures 2015). Thus, the reprint author either appears in first or last place, two positions that are generally regarded as reflecting some kind of leadership role.

Still, there are also disadvantages in using the reprint author. Indeed, it is 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) 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. On the other hand, these 'omitted' contributions might represent those of women researchers performing less well based on measures affected by the status/influence of researchers (e.g. the Average of Relative Impact Factors used as a proxy to measure scientific impact/quality and the propensity of researchers to collaborate internationally). In that case, the ratio of women-to-men scientific quality/impact (Section 2.3) and the ratio of women-to-men international co-authorship rate (Section 2.2) might be overestimated to some degree.

¹² This holds true for the last author.

Table 5 Share of reprint author in first and last position by country in the WoS (2007–2013)

Country	All papers with reprint author from the corresponding country	Papers with reprint author in 1 st place		Papers with reprint author in last place	
		No.	Share	No.	Share
Belgium	77,014	54,674	71.0%	24,960	32%
Bulgaria	10,305	8,216	79.7%	2,760	27%
Czech Republic	44,883	34,538	77.0%	12,705	28%
Denmark	53,944	40,400	74.9%	16,760	31%
Germany	446,032	309,472	69.4%	161,177	36%
Estonia	6,071	4,859	80.0%	1,879	31%
Ireland	30,208	19,423	64.3%	12,290	41%
Greece	54,605	34,310	62.8%	17,744	32%
Spain	243,974	157,453	64.5%	77,295	32%
France	310,080	204,008	65.8%	110,080	36%
Croatia	18,044	14,647	81.2%	4,173	23%
Italy	278,543	187,590	67.3%	82,150	29%
Cyprus	2,801	2,026	72.3%	1,160	41%
Latvia	2,165	1,821	84.1%	504	23%
Lithuania	11,138	9,512	85.4%	2,686	24%
Luxembourg	1,809	1,296	71.6%	601	33%
Hungary	26,007	17,769	68.3%	10,326	40%
Malta	618	531	85.9%	251	41%
Netherlands	146,464	103,672	70.8%	44,995	31%
Austria	50,283	33,935	67.5%	18,697	37%
Poland	110,808	89,340	80.6%	33,810	31%
Portugal	48,056	30,882	64.3%	15,912	33%
Romania	34,657	27,471	79.3%	9,965	29%
Slovenia	17,253	13,052	75.7%	6,373	37%
Slovakia	13,820	11,260	81.5%	3,777	27%
Finland	48,056	38,477	80.1%	12,297	26%
Sweden	92,077	68,029	73.9%	30,228	33%
United Kingdom	466,740	335,297	71.8%	186,740	40%
Iceland	2,651	1,946	73.4%	979	37%
Liechtenstein	110	93	84.5%	19	17%
Norway	44,250	36,166	81.7%	11,943	27%
Switzerland	92,630	60,998	65.9%	35,185	38%
Montenegro	487	445	91.4%	115	24%
FYR Macedonia	1,019	835	81.9%	250	25%
Albania	365	330	90.4%	54	15%
Serbia	22,264	17,345	77.9%	4,359	20%
Turkey	140,451	106,311	75.7%	33,335	24%
Bosnia and Herzegovina	1,671	1,513	90.5%	269	16%
Israel	62,546	40,387	64.6%	27,611	44%
Faroe Islands	47	44	93.6%	10	21%
Rep. of Moldova	883	738	83.6%	170	19%

Source: Compiled by Science-Matrix using WoS data (Thomson Reuters)

Formulas

Each reprint author on a peer-reviewed scientific paper is first categorised by sex (see Appendix 2 for details on this procedure) and country based on his or her affiliation address. Because some given names are unisex, a number of authorships end up being unclassified. As such, only the publications for which a sex could be attributed to the reprint author of a given country or region (EU28 or the world) are kept, and this subset of publications constitutes the sample with which the indicator is computed. Thus, the variables required to compute the ratio of women-to-men authorships are as follows:

- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S); and
- (n_{CYS}) Sample size for a given country (C), year (Y) and subfield (S) = PF_{CYS} + PM_{CYS}.

The ratio is then obtained by dividing the number of women authorships (PF_{CYS}) by the number of men authorships (PM_{CYS}), which is equivalent to the ratio of the proportion of women authorships (PF_{CYS}/n_{CYS}) over the same proportion for men (PM_{CYS}/n_{CYS}); indeed, the sample size (n_{CYS}) cancels out in the ratio of the two proportions.

The subfield indices (S) in the above variables is essential since the ratio must first be computed at the subfield level prior to aggregating the results at the FOS level (based on the Frascati Manual), as well as for all fields combined (i.e. the entire database). Indeed, as noted in Section 2.1.6, this is essential in order to account for the coverage biases that prevail in the WoS (some subfields being well represented while others are not). To aggregate the ratios computed at the subfield level up to the desired level (i.e. FOS or all fields combined), an additional variable is required. This variable is a weight that allows adjusting the contribution of each subfield to the aggregated ratio so as to more adequately reflect the representation of each subfield within the whole realm of scientific research – that is, including relevant literature not covered in the WoS. The weight for a given subfield (N_{YS}) is actually obtained by dividing the estimated number of papers in this subfield in the world (i.e. beyond the WoS) over the sum of these estimated numbers across all subfields; it therefore corresponds to the proportion of the world's scientific literature (i.e. beyond the WoS) falling in the given subfield (see Appendix 2 for details on how these weights are estimated). The generalised formula for aggregating the women-to-men ratios of authorships computed at the subfield level over any combination of subfields¹³ for a given country and year is as follows:

$$WMRatioAuthorships_{CY} = \frac{\sum_{S=1}^{tbd} \left(\frac{PF_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}{\sum_{S=1}^{tbd} \left(\frac{PM_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S);
- (n_{CYS}) Sample size for a given country (C), year (Y) and subfield (S) = PF_{CYS} + PM_{CYS}; and
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS).

The aggregation of scores using the weighting described in the above formula assumes that women and men behave in the same manner in the indexed (within WoS) and not-indexed (outside WoS) portion of the scientific literature.

Note that the aggregated numerator (i.e. sum of the weighted proportions of women authorships) and denominator (i.e. sum of the weighted proportions of men authorships) do not add up to 100% when aggregates are based on a subset of the 176 subfields used (i.e. for FOS aggregates). In fact,

¹³ Those combinations are determined by the table matching the subfields in Science-Metrix's classification to the FOS in the Frascati Manual (see Appendix 1).

the only aggregate where the two proportions add up to 100% is for all fields combined. This is because the shares (weights) within all the scientific literature (i.e. beyond the WoS) of the 176 subfields only add up to 100% when they are all considered. The weights have not been re-scaled so that when computing the aggregated ratio within a specific FOS, the weights for the underlying subset of subfields add up to 100%. This has no impact on the computed ratios or on the analysis of trends – using the Compound Annual Growth Rates (CAGR) – in the proportion of women or men authorships, since the re-scaling of the scores purely consists of a linear transformation. However, the weighted proportions of women and men authorships should not be displayed. Indeed, the actual numbers cannot be interpreted since they represent weighted scores whose sum will not add up to 100%.

The CAGR in the proportion of women authorships for a given country was computed by applying the following formula:

$$CAGR_C = (P_E/P_S)^{1/N-1}$$

Where,

- (P_E) Aggregated proportion of women authorships for a country (i.e. the numerator for women in the WMRatioAuthorship_{CY} equation above) in the end year;
- (P_S) Aggregated proportion of women authorships for a country (i.e. the numerator for women in the WMRatioAuthorship_{CY} equation above) in the start year; and
- (N) Number of years in the reference period (i.e. e – s).

When using moving periods instead of yearly data (see sub-section on accuracy below), N in the CAGR formula corresponds to the last year in the end period (e.g. E = 2011–2013) minus the last year in the start period (e.g. S = 2007–2009); in the above example, N would therefore equal 4 (i.e. 2013 – 2009).

Accuracy: As noted in Section 2.1.6, the computed ratios at the subfield level are based on samples rather than entire populations due to the presence of papers for which the sex of the reprint author could not be attributed and to coverage biases across subfields. To account for the random sampling errors at the subfield level, the margins of error of the subfield proportions for women and men have been computed and subsequently aggregated at the desired level to construct 90% confidence intervals of the ratio of women-to-men authorships using the following formula (note that the margins of error computed at the subfield level are also weighted based on the weights described above):

Lower limit of the 90% confidence interval of the ratio of women-to-men authorships aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(\frac{PF_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) - \sum_{S=1}^{tbd} \frac{\sqrt{\frac{N_{CYS} - n_{CYS}}{N_{CYS} - 1}} 1.645 \sqrt{\frac{PF_{CYS}/n_{CYS} (1 - PF_{CYS}/n_{CYS})}{n_{CYS}}}}{PF_{CYS}/n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}{\sum_{S=1}^{tbd} \left(\frac{PM_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{176} N_{YS}} \right) + \sum_{S=1}^{tbd} \frac{\sqrt{\frac{N_{CYS} - n_{CYS}}{N_{CYS} - 1}} 1.645 \sqrt{\frac{PM_{CYS}/n_{CYS} (1 - PM_{CYS}/n_{CYS})}{n_{CYS}}}}{PM_{CYS}/n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}$$

Upper limit of the 90% confidence interval of the ratio of women-to-men authorships aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(\frac{PF_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) + \sum_{S=1}^{tbd} \frac{\sqrt{\frac{N_{CYS} - n_{CYS}}{N_{CYS} - 1}} 1.645 \sqrt{\frac{PF_{CYS}/n_{CYS} (1 - PF_{CYS}/n_{CYS})}{n_{CYS}}}}{PF_{CYS}/n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}{\sum_{S=1}^{tbd} \left(\frac{PM_{CYS}}{n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{176} N_{YS}} \right) - \sum_{S=1}^{tbd} \frac{\sqrt{\frac{N_{CYS} - n_{CYS}}{N_{CYS} - 1}} 1.645 \sqrt{\frac{PM_{CYS}/n_{CYS} (1 - PM_{CYS}/n_{CYS})}{n_{CYS}}}}{PM_{CYS}/n_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S);
- (n_{CYS}) Sample size for a given country (C), year (Y) and subfield (S) = PF_{CYS} + PM_{CYS}
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS); and
- (N_{CYS}) The estimated number of papers in a given subfield and year for a given country (i.e. beyond the WoS) (as for N_{YS}, see Appendix 2 for more details).

The confidence intervals computed with this approach assume that the publications indexed in the WoS represent a random sample of the entire population at the subfield level. Although there might still be some biases in the coverage of the relevant literature *within* individual subfields, these are likely much less pronounced and their effect negligible relative to coverage biases *between* subfields.

However, note that the confidence intervals thus obtained do not account for the accuracy of the tool used in assigning a sex to reprint authors. Therefore, it is assumed that the attribution of a sex to author names is 100% accurate (excluding the unclassified cases, which are discarded in the computation). Manual validation showed that it was indeed highly accurate. For instance, the average accuracy across the countries included in She Figures 2015 is 97%. The lowest accuracies are actually quite high and are observed for Latvia (91%), Iceland (92%), Estonia (93%) and Turkey (93%) (see the AVPN column in Table 10 of Appendix 3).

Because the confidence intervals of some of the smaller countries were sometimes quite large on a yearly basis due to the size of the available samples by subfield, the ratios were computed using three-year moving periods (e.g. 2007–2009, 2008–2010, 2009–2011, 2010–2012, and 2011–2013). This way, the samples used were larger, providing more robust estimates. A similar approach has been used in measuring the Compound Annual Growth Rate (CAGR) in the proportion of women authorships. Thus, in interpreting the CAGR for this indicator, it is important to note that it measures the annual growth of the three-year scores shifting by one year every year instead of the annual growth of the yearly scores shifting by one year every year.

2.2 Ratio of women-to-men international co-authorship rate

2.2.1 Definition

To obtain this indicator, the international co-authorship rate of female researchers is divided by the international co-authorship rate of male researchers, whereby a score above 1 indicates that women publish their publications more frequently through involvement in international teams than men, whereas a score below 1 means the opposite. It is only based on the reprint (i.e. corresponding) author of peer-reviewed scientific publications. The reprint author is used as a proxy to compare the contribution of women relative to that of men when in a leading role (see sub-section on author selection in Section 2.1.7).

2.2.2 Source of data

Same as above (see Section 2.1.2).

2.2.3 Availability over time

Same as above (see Section 2.1.3).

2.2.4 International availability

Same as above (see Section 2.1.4).

2.2.5 Availability across Fields of Science and Technology

Same as above (see Section 2.1.5).

2.2.6 Comparability

Same as above (see Section 2.1.6).

2.2.7 Calculation method

Author selection

Same as above (see Section 2.1.7).

Formulas

The samples used to compute this indicator are the same as those used for the ratio of women-to-men authorships; they are similarly first computed at the subfield level for a given period (i.e. three-

year moving period) and country. Thus it is based on only those publications for which a sex (either male or female) could be attributed to the reprint author; it excludes cases where the sex of the authorship could not be classified (e.g. unisex names).

Using these samples, the international co-authorship rate of women in a given country, period and subfield is computed by dividing the number of women authorships (as reprint author) from said country, period and subfield that involved co-authors from at least two countries by the total number of women authorships (as reprint author) from said country, period and subfield. This rate actually measures how frequently women authors, when in a leading position, are involved in international teams when it comes to publishing their research outputs in a given country, period and subfield. The international co-authorship rate of men in a given country, period and subfield is computed in the same manner. Once computed at the subfield level, these scores are aggregated at the desired level (i.e. FOS level and all fields combined) using the weights described in Section 2.1.7. The formula used to compute the aggregated ratio of women-to-men co-authorship rate for a given country and year is as follows:

$$WMRatioCollaboration_{CY} = \sum_{S=1}^{tbd} \left(\frac{PFC_{CYS}}{PF_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) / \sum_{S=1}^{tbd} \left(\frac{PMC_{CYS}}{PM_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S);
- (PFC_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S) with international co-authors;
- (PMC_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S) with international co-authors; and
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS, see Appendix 2).

The year (Y) in the above variables and formula can be replaced by a period for computing moving ratios (by analogy to moving averages). The aggregation of scores using the weighting described in the above formula assumes that women and men behave in the same manner in the indexed (within WoS) and not-indexed (outside WoS) portion of the scientific literature.

As for the ratio of women-to-men authorships, the aggregated numerator and denominator should not be presented since they are difficult to interpret in their weighted form (only the ratio should be presented; see Section 2.1.7 for more details). They can nevertheless be used to analyse trends using the Compound Annual Growth Rates (CAGR). The CAGR in the international co-authorship rate of women for a given country was computed by applying the following formula:

$$CAGR_C = (P_E/P_S)^{1/N-1}$$

Where,

- (P_E) Aggregated international co-authorship rate of women (i.e. the numerator for women in the WMRatioCollaboration_{CY} equation above) in the end year;
- (P_S) Aggregated international co-authorship rate of women (i.e. the numerator for women in the WMRatioCollaboration_{CY} equation above) in the start year; and
- (N) Number of years in the reference period (i.e. e – s).

When using moving periods instead of yearly data (see accuracy sub-section in Section 2.1.7), N in the CAGR formula corresponds to the last year in the end period (e.g. E = 2011–2013) minus the last year in the start period (e.g. S = 2007–2009); in the above example, N would therefore equal 4 (i.e. 2013 – 2009).

Accuracy: As for the ratio of women-to-men authorships, the margins of error at the subfield level have been computed and subsequently aggregated at the desired level to construct 90% confidence intervals of the ratio of women-to-men international co-authorship rate using the following formula (with the same weighting approach as in Section 2.1.7):

Lower limit of the 90% confidence interval of the ratio of women-to-men international co-authorship rate aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(\frac{PFC_{CYS}}{PF_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) - \sum_{S=1}^{tbd} \sqrt{\frac{NF_{CYS} - PFC_{CYS}}{NF_{CYS} - 1}} 1.645 \sqrt{\frac{PFC_{CYS}/PF_{CYS} (1 - PFC_{CYS}/PF_{CYS})}{PF_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}}{\sum_{S=1}^{tbd} \left(\frac{PMC_{CYS}}{PM_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) + \sum_{S=1}^{tbd} \sqrt{\frac{NM_{CYS} - PMC_{CYS}}{NM_{CYS} - 1}} 1.645 \sqrt{\frac{PMC_{CYS}/PM_{CYS} (1 - PMC_{CYS}/PM_{CYS})}{PM_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}}$$

Upper limit of the 90% confidence interval of the ratio of women-to-men international co-authorship rate aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(\frac{PFC_{CYS}}{PF_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) + \sum_{S=1}^{tbd} \sqrt{\frac{NF_{CYS} - PFC_{CYS}}{NF_{CYS} - 1}} 1.645 \sqrt{\frac{PFC_{CYS}/PF_{CYS} (1 - PFC_{CYS}/PF_{CYS})}{PF_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}}{\sum_{S=1}^{tbd} \left(\frac{PMC_{CYS}}{PM_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right) - \sum_{S=1}^{tbd} \sqrt{\frac{NM_{CYS} - PMC_{CYS}}{NM_{CYS} - 1}} 1.645 \sqrt{\frac{PMC_{CYS}/PM_{CYS} (1 - PMC_{CYS}/PM_{CYS})}{PM_{CYS}} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}}$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S);
- (PFC_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S) with international co-authors;
- (PMC_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S) with international co-authors; and
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS, see Appendix 2).
- (NF_{CYS}) The estimated number of women authorships in a given subfield and year for a given country (i.e. beyond the WoS; NF_{CYS} = N_{CYS} * PF_{CYS}/n_{CYS})

Where,

$$(n_{CYS}) \quad \text{Sample size for a given country (C), year (Y) and subfield (S) = } PF_{CYS} + PM_{CYS}$$

- (N_{CYS}) The estimated number of papers in a given subfield and year for a given country (i.e. beyond the WoS) (as for N_{YS}, see Appendix 2 for more details).
- (NM_{CYS}) The estimated number of men authorships in a given subfield and year for a given country (i.e. beyond the WoS; $NM_{CYS} = N_{CYS} * PM_{CYS}/n_{CYS}$)
Where,
 - (n_{CYS}) Sample size for a given country (C), year (Y) and subfield (S) = $PF_{CYS} + PM_{CYS}$
 - (N_{CYS}) The estimated number of papers in a given subfield and year for a given country (i.e. beyond the WoS) (as for N_{YS}, see Appendix 2 for more details).

The confidence intervals computed with this approach assume that the publications indexed in the WoS represent a random sample of the entire population at the subfield level. Although there might still be some biases in the coverage of the relevant literature *within* individual subfields, these are likely much less pronounced and their effect negligible relative to coverage biases *between* subfields.

However, note that the confidence intervals thus obtained do not account for the accuracy of the tool used in assigning a sex to reprint authors. Therefore, it is assumed that the attribution of a sex to author names is 100% accurate (excluding the unclassified cases, which are discarded in the computation). Manual validation showed that it was indeed highly accurate. For instance, the average accuracy across the countries included in She Figures 2015 is 97%. The lowest accuracies are actually quite high and are observed for Latvia (91%), Iceland (92%), Estonia (93%) and Turkey (93%) (see the AVPN column in Table 10 of Appendix 3).

Because the confidence intervals of some of the smaller countries were sometimes quite large on a yearly basis due to the size of the available samples by subfield, the ratios were computed using three-year moving periods (e.g. 2007–2009, 2008–2010, 2009–2011, 2010–2012, and 2011–2013). This way, the samples used were larger, providing more robust estimates. A similar approach has been used in measuring the Compound Annual Growth Rate (CAGR) in the proportion of women authorships. Thus, in interpreting the CAGR for this indicator, it is important to note that it measures the annual growth of the three-year scores shifting by one year every year instead of the annual growth of the yearly scores shifting by one year every year.

2.3 Ratio of women-to-men scientific quality/impact

2.3.1 Definition

The indicator comparing the scientific quality/impact of women and men researchers will consist of the ratio of the Average of Relative Impact Factors (ARIF) of the papers whose reprint author is a woman, over the ARIF of the papers whose reprint author is a man in the corresponding country. A score above 1 indicates that women in a given entity produced publications that were published, on average, in higher impact journals than men, whereas a score below 1 means the opposite. The reprint author is used as a proxy to measure the gap in the quality/impact of research between women and men when in a lead position (see sub-section on author selection in Section 2.1.7).

2.3.2 Source of data

Same as above (see Section 2.1.2).

2.3.3 Availability over time

Although the Average of Relative Citations (ARC) is a more direct measure of the scientific impact of a given entity (e.g. a researcher, an institution, a country) as it is based on the citations received by the actual publications of an entity instead of the publication venue, it does not adequately meet the quality requirement in terms of its availability over time. Indeed, in the case of the ARC, citations are counted in a forward manner from the publication date of a scientific paper as opposed to a backward manner as in the case of the ARIF. Since citations must be counted for at least the publication year and the two following years to obtain accurate data, the ARC could only be provided for the years 2007 (citation window of eight years) through 2011 (citation window of three years; i.e. citations are counted in the publication year of 2011 and in the two following years [2012 and 2013]). Thus, the completeness of this indicator, if it relied on the ARC, would be as low as 42% (5 out of 12 years) and it would not provide timely data as the two most recent years would be unavailable.

Using the ARIF instead obtains the same coverage as the previous indicators (see Section 2.1.3). The ARIF, in addition to providing an indirect measure of scientific impact (it correlates well with the ARC at the country level), provides a good proxy for the scientific quality of the research performed by women relative to men (see definition of the ARIF in Section 2.3.7). Moreover, the ARIF is also an indicator of prestige, as researchers usually prefer to publish in the most highly cited journals.

2.3.4 International availability

Same as above (see Section 2.1.4).

2.3.5 Availability across Fields of Science and Technology

Same as above (see Section 2.1.5).

2.3.6 Comparability

Same as above (see Section 2.1.6).

2.3.7 Calculation method

Author selection

Same as above (see Section 2.1.7).

Definition of ARIF

The ARIF is a measure of the scientific impact of papers produced by a given entity (e.g. a country) based on the impact factors of the journals in which they were published. As such, the ARIF is an *indirect* impact metric reflecting the scientific quality measured by the average citation rate of the publication venue instead of the actual publications. Indeed, the most cited journals (i.e. those with the highest impact factor) also generally have the more restrictive inclusion policy – as they are more cited, more researchers want to publish in them, more papers are submitted for a limited space, and therefore more papers are rejected. Moreover, the ARIF is also an indicator of prestige, as researchers usually prefer to publish in the most highly cited journals.

Thomson Reuters calculates an annual impact factor (IF) for each journal based on the number of citations it received in the previous two years relative to the number of papers it published in the previous two years. Thus, each journal's IF will vary from year to year. The IF of a journal in 2007 is equal to the number of citations to articles published in 2006 (8) and 2005 (15) divided by the number of articles published in 2006 (15) and 2005 (23) (i.e. $IF = \text{numerator [23]}/\text{denominator [38]} = 0.605$). However, as Archambault pointed out,¹⁴ this indicator carries the weight of history and of many choices that were made a long time ago, when their effects had not yet been thoroughly studied. For example, Moed and colleagues have described the effect of the observed asymmetry between the numerator and denominator of the Thomson Reuters' IF:¹⁵

ISI classifies documents into types. In calculating the nominator of the IF, ISI counts citations to all types of documents, while as citable documents in the denominator ISI includes as a standard only normal articles, notes and reviews. However, editorials, letters and several other types are cited rather frequently in a number of journals. When they are cited, these types do contribute to the citation counts in the IF's numerator, but are not included in the denominator. In a sense, the citations to these documents are 'for free'.

In this study, Science-Metrix will therefore compute and use a symmetric IF based on the document types that are used throughout this entire project for producing bibliometric data. More precisely, the IF of publications is calculated by ascribing to them the IF of the journal in which they are published for the year in which they are published using a five-year citation window (publication year and four previous years). This longer citation window is more accurate, especially in the SSH where citations take longer to accumulate. Subsequently, to account for different citation patterns across fields and subfields of science (e.g. there are more citations in biomedical research than mathematics), each paper's IF is divided by the average IF of all papers that were published the same year in the same subfield to obtain a Relative Impact Factor (RIF). The ARIF of a given entity is the average of its RIFs (i.e. if an institution has 20 papers, the ARIF is the average of 20 RIFs: one per paper). When the ARIF is above 1, it means that the entity scores higher than the world average; when it is below 1, it means that, on average, the entity publishes in journals that are not cited as often as the world level.

Formulas

The samples used to compute the ratio of women-to-men scientific quality/impact are the same as those used for the ratio of women-to-men authorships; they are similarly first computed at the subfield level for a given period (i.e. three-year moving period) and country. Thus, it is based on only those publications for which a sex (either male or female) could be attributed to the reprint author; it excludes cases where the sex of the authorship could not be classified (e.g. unisex name).

Sex disaggregated data for the ARIF can be obtained by averaging the RIFs of all papers whose reprint author is a woman in a given country and averaging the RIFs of all papers whose reprint author is a man in the corresponding country.

Using these samples, the ARIF of women in a given country, period and subfield is computed by averaging the RIFs of the papers whose reprint author is a woman for said country, period and

¹⁴ Archambault É. and Larivière V. (2009). History of the journal impact factor: Contingencies and consequences. *Scientometrics*, 79(3): 635–649.

¹⁵ Moed, H. F., Van Leeuwen, T. H. N., and Reedijk, J. (1999). Towards appropriate indicators of journal impact. *Scientometrics*, 46: 575–589.

subfield. The ARIF of men in a given country, period and subfield is computed in the same manner. Once computed at the subfield level, these scores are aggregated at the desired level (i.e. FOS level and all fields combined) using the weights described in Section 2.1.7. The formula used to compute the aggregated ratio of women-to-men scientific quality/impact for a given country and year is as follows:

$$WMRatioQI_{CY} = \sum_{S=1}^{tbd} \left(\frac{\sum_{P=1}^{PF_{CYS}} RIF_{CYSP} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}{PF_{CYS}} \right) / \sum_{S=1}^{tbd} \left(\frac{\sum_{P=1}^{PM_{CYS}} RIF_{CYSP} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}}}{PM_{CYS}} \right)$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (PF_{CYS}) Number of papers by women reprint authors in a given country (C), year (Y) and subfield (S);
- (PM_{CYS}) Number of papers by men reprint authors in a given country (C), year (Y) and subfield (S);
- (RIF_{CYSP}) Relative impact factor of a given paper (P) in a given country (C), year (Y) and subfield (S); and
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS, see Appendix 2).

The year (Y) in the above variables and formula can be replaced by a period for computing moving ratios (by analogy to moving averages). The aggregation of scores using the weighting described in the above formula assumes that women and men behave in the same manner in the indexed (within WoS) and not-indexed (outside WoS) portion of the scientific literature.

As for the ratio of women-to-men authorships, the aggregated numerator and denominator in the above formula should not be presented since they are difficult to interpret in their weighted form (only the ratio should be presented; see Section 2.1.7 for more details). They can nevertheless be used to analyse trends using the Compound Annual Growth Rates (CAGR). The CAGR in the ARIF of women for a given country was computed by applying the following formula:

$$CAGR_C = (P_E/P_S)^{1/N-1}$$

Where,

- (P_E) The aggregated ARIF of women (i.e. the numerator for women in the WMRatioQI_{CY} equation above) in the end year;
- (P_S) The aggregated ARIF of women (i.e. the numerator for women in the WMRatioQI_{CY} equation above) in the start year; and
- (N) Number of years in the reference period (i.e. e – s).

When using moving periods instead of yearly data (see accuracy sub-section in Section 2.1.7), N in the CAGR formula corresponds to the last year in the end period (e.g. E = 2011–2013) minus the last year in the start period (e.g. S = 2007–2009); in the above example, N would therefore equal 4 (i.e. 2013 – 2009).

Accuracy: As for the ratio of women-to-men authorships, the margins of error of the numerator (for women) and denominator (for men) have first been computed at the subfield level. However, since the ARIF is based on an average instead of a proportion, the margins of error could not be computed using the formula for the margin of error for a sample proportion. Instead, 90% confidence intervals have been computed empirically at the subfield level for the ARIF of women and men prior to their aggregation at the FOS level and for all fields combined. A bootstrapping procedure was applied to

achieve this. For the ARIF of women, the approach consisted of the re-sampling, with replacement, of the papers whose reprint author is a woman using a sample size (N) equal to the size of the population being sampled (i.e. PF_{CYS} = the number of papers by women reprint authors in a given country, year and subfield). The re-sampling was achieved multiple times (i.e. 500 iterations). Using the empirical distribution of the ARIF of women thus constructed, it was possible to extract the lower and upper limit of the ARIF of women for a given country, year and subfield at the 90% confidence level. The 90% confidence interval for the ARIF of men was computed in the same manner with $N = PM_{CYS}$ (the number of papers by men reprint authors in a given country, year and subfield). The lower and upper limits thus obtained for the ARIF of women and men were subsequently aggregated at the desired level to construct 90% confidence intervals of the ratio of women-to-men scientific quality/impact using the following formula (with the same weighting approach as in Section 2.1.7):

Lower limit of the 90% confidence interval of the ratio of women-to-men scientific quality/impact aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(LLW_{CYS} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}{\sum_{S=1}^{tbd} \left(ULM_{CYS} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}$$

Upper limit of the 90% confidence interval of the ratio of women-to-men scientific quality/impact aggregated over any combination of subfields for a given country and year

$$\frac{\sum_{S=1}^{tbd} \left(ULW_{CYS} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}{\sum_{S=1}^{tbd} \left(LLM_{CYS} \frac{N_{YS}}{\sum_{S=1}^{tbd} N_{YS}} \right)}$$

Where,

- (tbd) to be determined based on the desired aggregation of subfields;
- (LLW_{CYS}) Lower limit of the ARIF of women obtained through the bootstrapping procedure for a given country (C), year (Y) and subfield (S);
- (ULW_{CYS}) Upper limit of the ARIF of women obtained through the bootstrapping procedure for a given country (C), year (Y) and subfield (S);
- (LLM_{CYS}) Lower limit of the ARIF of men obtained through the bootstrapping procedure for a given country (C), year (Y) and subfield (S);
- (ULM_{CYS}) Upper limit of the ARIF of men obtained through the bootstrapping procedure for a given country (C), year (Y) and subfield (S); and
- (N_{YS}) The estimated number of papers in a given subfield and year in the world (i.e. beyond the WoS, see Appendix 2).

2.4 Ratio of women-to-men inventorships

2.4.1 Definition

This indicator is the ratio of women-to-men inventorships, or equivalently, the ratio of the proportion of women inventorships (in total inventorships) over the equivalent proportion for men. It can be computed at various aggregation levels (e.g. organisations, countries, world regions). A score above 1 indicates that women in a given entity produced a larger share of the entity's inventions (as measured with patent applications) than men, whereas a score below 1 means the opposite. The

absolute number of inventorships used in computing this indicator is based on fractionalised counts of patent applications across their corresponding inventors; for example, if a patent application involves 10 inventors, each inventor is attributed an equal fraction of the inventorships (i.e. 1/10 of the invention).

2.4.2 Source of data

Patents, like scientific papers, contain interesting information that may be analysed to produce results for policy purposes. Just like for papers, document counts can be performed using patent data. Measurement of patent data is often seen as a proxy for measuring innovative activities, as patents provide formal protection for new technological progress often resulting from formal research. Indeed, studies have established statistical significance between patent counts and innovation. As such, patents can be used to characterise countries' inventive performance in terms of new technologies, new processes or new products.

Patent data are provided in a number of national, regional and international databases. Science-Metrix has an in-house version of EPO Worldwide Patent Statistical Database (PATSTAT), which covers patent data from over 150 offices worldwide, including the USPTO, EPO, and JPO. The USPTO covers the United States, the EPO covers Europe, the JPO covers Japan, and so forth. This has allowed Science-Metrix to condition the database for the purpose of producing large-scale comparative technometric analyses.

For this project, the statistics are based on the EPO within PATSTAT as the European market is one of the largest in the world and certainly the most relevant in this project as nearly all countries covered are part of the European Research Area (ERA).

Note that statistics on inventorships can be produced by measuring issued patents or patent applications when working with EPO data. On a conceptual level, if the goal is to get a sense of the inventive/innovative capacity of a given entity (e.g. women in a given country) rather than of 'marketable/innovative outputs', as in this study, then applications are more appropriate. Furthermore, in cases where trends in the inventiveness of entities are to be investigated, as in this study, the capacity to produce timely data is important. In this regard, issued patents have the disadvantage of running behind and only becoming visible years after the innovative activity has taken place. Thus, from a methodological standpoint, applications are still preferable. Consequently, EPO patent applications (kind codes: A1 and A2) were retained in computing the ratio of women-to-men inventorships. These patent applications are later referred to as 'patent applications', 'patents' or 'inventions'.

2.4.3 Availability over time

This indicator can be computed for the 12 years to be covered by the She Figures 2015 publication. Thus, the completeness of this indicator in regard to the time coverage is at 100%. The year of a patent application is obtained from the application date.

2.4.4 International availability

Data can readily be produced for all 41 countries to be covered in the She Figures 2015 publication – that is, for the 28 EU Member States as well as for Albania, Bosnia & Herzegovina, Faroe Islands, Iceland, Israel, Liechtenstein, the Former Yugoslav Republic of Macedonia, Moldova, Montenegro,

Norway, Serbia, Switzerland, and Turkey. Thus, the completeness of this indicator in regard to the geographical scope stands at 100%. Note that data by country is obtained by assigning each inventor on a patent application to a given country based on his or her affiliation address rather than using the inventor's nationality. The indicator therefore looks at where the innovation took place.

2.4.5 Availability across technological fields (IPC classes)

All EPO patent applications are classified based on the International Patent Classification (IPC) of the World Intellectual Property Organization (WIPO) in PATSTAT.¹⁶ This hierarchical classification is divided into eight sections (level 1), which are further divided into classes (level 2), subclasses (level 3), main groups (level 4) and subgroups (lower level). This classification is not mutually exclusive (i.e. each patent application is classified into one or more sections, classes, subclasses, main groups and subgroups). Thus, a given patent application can contribute to the scores of more than one of the eight sections for which this indicator has been computed:

- (A) Human Necessities;
- (B) Performing Operations & Transporting;
- (C) Chemistry & Metallurgy;
- (D) Textiles & Paper;
- (E) Fixed Constructions;
- (F) Mechanical Engineering, Lighting, Heating, Weapons & Blasting;
- (G) Physics;
- (H) Electricity.

2.4.6 Comparability

Comparability becomes an issue when data are being compared across periods, geographical regions and disciplines. Whenever an issue of this type is encountered, efforts are made to eliminate or limit its impact on the data (e.g. limiting the analysis to a subset of countries, disciplines), and the potential biases that could result from it are clearly stated in Science-Metrix's reports, sometimes in the form of confidence intervals of the estimates.

To be used appropriately, technometrics, much like scientometrics, must address important shortcomings in the data. These include: incompleteness, as patenting is only one way of protecting an invention; inconsistency in quality, as the importance and value of patented inventions vary considerably; variations between scientific fields and industries in their propensity to patent inventions; and variations between countries, where propensity to patent varies and so do Intellectual Property (IP) laws. Additional sources of potential distortions include:

- international and sectoral differences in patenting behaviour;
- differences in patenting between large companies and smaller firms;
- the same weight is given to important patents and run-of-the-mill patents;
- the fact that patents only cover a part of the overall trajectory from R&D to innovation.

Prior to setting up the in-house bibliometric versions of PATSTAT, Science-Metrix's senior analysts performed a comprehensive testing of their coverage looking for errors such as:

¹⁶ WIPO (2015). *International Patent Classification*. Version 2015.01, <http://www.wipo.int/classifications/ipc/en/>.

Bias in the number of documents over time

Note that the affiliation address of inventors and their full given name represent essential pieces of information towards assigning a sex to inventors in producing sex disaggregated data, as well as in assigning a country of affiliation to authors in producing country disaggregated data. The proportion of patent applications for which this information is available for all inventors on an application (i.e. the recall) declined slightly over time from a high of 95% in 2002 to a low of 89% in 2013 (Table 6). Even though the recall is not equal to 100%, it is fairly safe to assume that the available samples are sufficiently large to produce highly accurate statistics over the entire time frame of the She Figures publication, even though the margins of error might be slightly more pronounced in recent years. To reflect this in the computed data, margins of error have been computed to construct 90% confidence intervals of the ratio of women-to-men inventorships (see sub-section on accuracy in Section 2.4.7).

Table 6 Share of EPO patent applications for which the country of affiliation and full given name of all inventors appearing on an application is available in PATSTAT, 2002–2013

Year	EPO patent applications with country of affiliation and full given name	EPO patent applications	% of Total
2002	97,684	102,430	95%
2003	98,879	104,405	95%
2004	112,768	119,655	94%
2005	110,552	117,839	94%
2006	118,102	126,347	93%
2007	121,870	130,750	93%
2008	128,250	137,480	93%
2009	119,666	128,225	93%
2010	118,831	127,699	93%
2011	119,260	130,959	91%
2012	122,259	136,688	89%
2013	122,586	137,878	89%

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Bias in favour of some countries

Differences in the extent of use of ‘continuation-in-part’, ‘continuation of’ and ‘divisional’ patent applications are problematic, as they can lead to biases in cross-regional comparisons of the number of inventions. Because these types of applications lead to patent splitting – whereby the full scope of an invention is protected by a set of patents instead of just one – regions in which the use of these types of applications is common among inventors (e.g. the US) are advantaged relative to others in which this behaviour is less widespread (e.g. Canada and Europe). Thus, by counting individual patents, some inventions could be counted multiple times in some countries and less so in others. However, since the current indicator consists of the ratio of the number of women inventions over the same number for men in a given country, it is assumed that such biases should cancel out and should therefore not affect the cross-country comparability of the ratio of women-to-men inventorships. Of course, this assumes that women and men within a country behave in the same manner in their extent of use of ‘continuation-in-part’, ‘continuation of’ and ‘divisional’ patent applications.

Bias in favour of disciplines

The proportion of EPO patent applications for which the affiliation country and full given name is available for all inventors on an application is high and relatively similar across IPC classes; it ranges from 80% (D07) to 100% (B68) among the ~120 IPC classes (Figure 1) compared to 30% to 100% for scientific subfields (see Section 2.1.7). It is therefore concluded that none of the IPC classes should contribute significantly more or less than it should to the women-to-men ratios computed at higher aggregation levels (e.g. for IPC Sections or for all EPO patent applications). As such, the ratios can be computed at the IPC section level without first having to compute them at the IPC class level. Indeed, it is not necessary to compute the ratios at the IPC section level by weighting the scores at the class level according to their representation in PATSTAT, as was done for authorships. Still, since the recall is not equal to 100%, margins of error have been computed to adequately reflect the sampling errors associated with the computed ratios at the IPC section level (see Section 2.4.7).

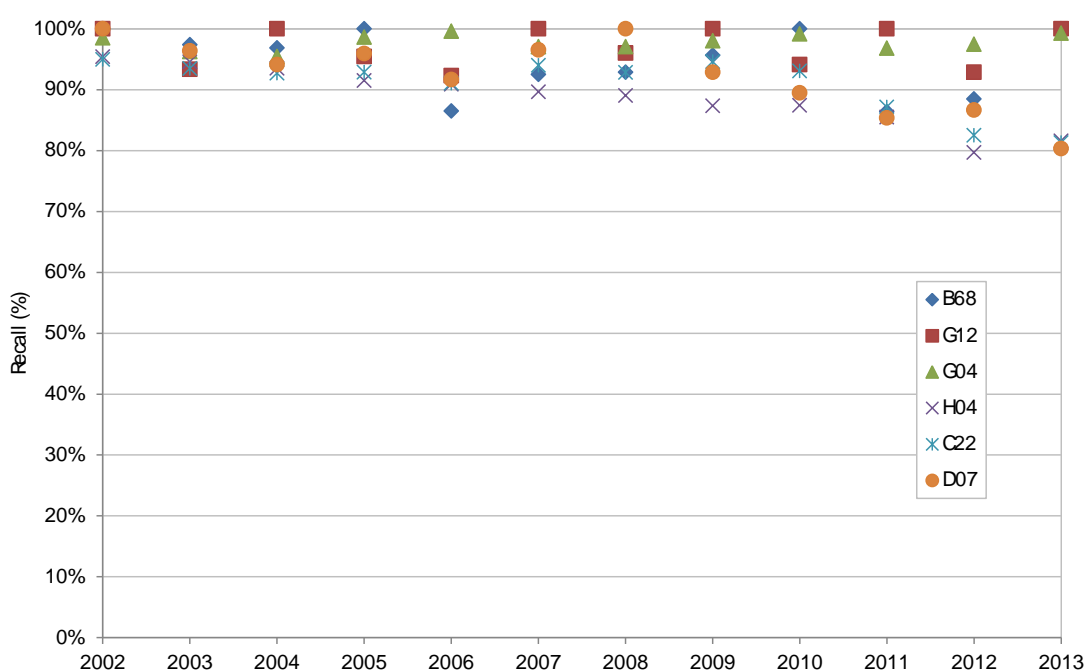


Figure 1 Share of EPO patent applications for which the country of affiliation and full given name of all inventors appearing on an application are available in PATSTAT by IPC class, 2002–2013

Note: The data is displayed for the three IPC classes for which the recall is highest in 2013 and for the three IPC classes for which the recall is smallest in 2013. There are about 120 IPC classes in total.

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

2.4.7 Calculation method

To compute this indicator, each inventor on an EPO patent application is first assigned a sex based on his or her name (i.e. full given name plus surname) using a gender name disambiguation algorithm (see Appendix 2 for details on this procedure). Because some given names are unisex, and because a sex cannot always be attributed to all names (i.e. those where only the initials of the given name are available), a number of inventorships end up being unclassified. Given the high proportion of inventors for whom sufficient information was available to assign a sex to inventors in PATSTAT (see Section 2.4.6), each EPO patent application for which a sex could not be attributed

to all of its inventors was simply discarded from further analyses. Each inventorship on a patent application (i.e. only those with a known sex remain) is then attributed a fraction of the invention. If there are 10 inventors on a paper, then each of them is attributed a 10th of the invention.

In turn, the fractionalised inventorships of women associated with a given country, year and IPC section are summed to obtain the numerator. Similarly, the fractionalised inventorships of men associated with the corresponding country, year and IPC section are summed to obtain the denominator. The ratio is then obtained by dividing the sum of the fractionalised inventorships of women by that of men for a given country, year and IPC section. This indicator is equivalent to the ratio of the proportion of female inventorships over the proportion of male inventorships:

$$WMRatioInventorships_{CYI} = \frac{WI_{CYI}}{TI_{CYI}} \bigg/ \frac{MI_{CYI}}{TI_{CYI}} = \frac{WI_{CYI}}{MI_{CYI}}$$

Where,

- (WI_{CYI}) Sum of fractionalised inventorships for women in a given country (C), year (Y) and IPC section (I);
- (MI_{CYI}) Sum of fractionalised inventorships for men in a given country (C), year (Y) and IPC section (I); and
- (TI_{CYI}) Sum of fractionalised inventorships across women and men in a given country (C), year (Y) and IPC section (I).

The CAGR in the proportion of women inventorships for a given country and IPC section was computed by applying the following formula:

$$CAGR_{CI} = (P_E/P_S)^{1/N-1}$$

Where,

- (P_E) Proportion of women inventorships for a country (C) and IPC section (I) (i.e. the numerator for women in the $WMRatioInventorships_{CYI}$ equation above) in the end year;
- (P_S) Proportion of women inventorships for a country (C) and IPC section (I) (i.e. the numerator for women in the $WMRatioInventorships_{CYI}$ equation above) in the start year; and
- (N) Number of years in the reference period (i.e. $e - s$).

When using moving periods instead of yearly data (see accuracy sub-section below), N in the CAGR formula corresponds to the last year in the end period (e.g. $E = 2010-2013$) minus the last year in the start period (e.g. $S = 2002-2005$); in the above example, N would therefore equal 8 (i.e. $2013 - 2005$).

Accuracy: As noted in Section 2.4.6, the computed ratios are based on samples (rather large ones since the recall is very high) rather than entire populations due to the presence of patent applications for which the sex of all inventors could not be attributed. To account for the random sampling errors, the margins of error in the proportions of women inventorships and men inventorships have been computed and subsequently used to construct 90% confidence intervals of the ratio of women-to-men inventorships using the following formula:

Lower limit of the 90% confidence interval of the ratio of women-to-men inventorships for a given country, year and IPC section

$$\frac{\frac{WI_{CYI}}{TI_{CYI}} - \sqrt{\frac{NCYI - TI_{CYI}}{NCYI - 1}} \cdot 1.645 \sqrt{\frac{WI_{CYI}/TI_{CYI} (1 - WI_{CYI}/TI_{CYI})}{TI_{CYI}}}{WI_{CYI}/TI_{CYI}}$$

$$\frac{\frac{MI_{CYI}}{TI_{CYI}} + \sqrt{\frac{NCYI - TI_{CYI}}{NCYI - 1}} \cdot 1.645 \sqrt{\frac{MI_{CYI}/TI_{CYI} (1 - MI_{CYI}/TI_{CYI})}{TI_{CYI}}}{MI_{CYI}/TI_{CYI}}$$

Upper limit of the 90% confidence interval of the ratio of women-to-men authorships aggregated over any combination of subfields for a given country and year

$$\frac{\frac{WI_{CYI}}{TI_{CYI}} + \sqrt{\frac{NCYI - TI_{CYI}}{NCYI - 1}} Z_{\alpha/2} \sqrt{\frac{WI_{CYI}/TI_{CYI} (1 - WI_{CYI}/TI_{CYI})}{TI_{CYI}}}{WI_{CYI}/TI_{CYI}}$$

$$\frac{\frac{MI_{CYI}}{TI_{CYI}} - \sqrt{\frac{NCYI - TI_{CYI}}{NCYI - 1}} Z_{\alpha/2} \sqrt{\frac{MI_{CYI}/TI_{CYI} (1 - MI_{CYI}/TI_{CYI})}{TI_{CYI}}}{MI_{CYI}/TI_{CYI}}$$

Where,

- (WI_{CYI}) Sum of fractionalised inventorships for women in a given country (C), year (Y) and IPC section (I);
- (MI_{CYI}) Sum of fractionalised inventorships for men in a given country (C), year (Y) and IPC section (I); and
- (TI_{CYI}) Sum of fractionalised inventorships across women and men in a given country (C), year (Y) and IPC section (I); and
- $(NCYI)$ Sum of fractionalised inventorships (including unclassified ones; i.e. no sex attributed) in PATSTAT in a given country (C), year (Y) and IPC section (I).

Note that the confidence intervals thus obtained do not account for the accuracy of the tool used in assigning a sex to inventors. Therefore, it is assumed that the attribution of a sex to inventor names is 100% accurate (excluding the unclassified cases, which are discarded in the computation). Manual validation showed that it was indeed highly accurate. For instance, the average accuracy across the countries included in She Figures 2015 is 97%. The accuracy is smaller than 90% for only one country: Montenegro (80%) (see the AVPN column in Table 12 of Appendix 3).

Because the confidence intervals of some of the smaller countries were sometimes quite large on a yearly basis due to the size of the available samples by IPC section, the ratios were computed using four-year moving periods (e.g. 2002–2005, 2003–2006, and so on). This way, the samples used were larger, providing more robust estimates. A similar approach has been used in measuring the Compound Annual Growth Rate (CAGR) in the proportion of women inventorships. Thus, in interpreting the CAGR for this indicator, it is important to note that it measures the annual growth of the three-year scores shifting by one year every year instead of the annual growth of the yearly scores shifting by one year every year.

3 Gender dimension in research content

Within the context of Horizon 2020 (H2020) – that is, the European Commission’s Eighth Framework Programme for Research and Technological Development – activities towards achieving gender equality are being implemented along three main objectives:¹⁷

- fostering gender balance in research teams;
- ensuring gender balance in decision-making; and
- integrating gender analysis in research and innovation (R&I) content.

There is, in fact, a legal basis in place to ensure the attainment of these objectives, as laid out in three core documents on H2020 covering its regulation,¹⁸ the rules of participation¹⁹ and the specific programme implementing it.²⁰ For example, it is now mandatory for H2020 participants to specify in their grant proposals how they intend to integrate a gender dimension into the subject matter of their projects.

As such, it becomes highly relevant to start monitoring trends in the extent to which researchers in different countries incorporate such aspects into their research content. For instance, in the future it will be interesting to investigate whether the actions taken in H2020 translated into a significant increase in the integration of the gender dimension in H2020 R&I content relative to the Seventh Framework Programme (FP7), and whether these actions actually led to an increase in the extent to which the gender dimension appears in the research content of the scientific outputs produced by the participating countries.

A new indicator is proposed to monitor the extent to which researchers integrate a gender dimension into their research content:²¹

- Proportion of a country’s research outputs integrating a gender dimension in their subject matter.

¹⁷ European Commission. (2014). *Gender equality in Horizon 2020*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/gender/h2020-hi-guide-gender_en.pdf

¹⁸ European Commission. (2013). Establishing Horizon 2020 – the Framework Programme for Research and Innovation (2014–2020) and repealing Decision No 1982/2006/EC. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/fp/h2020-eu-establact_en.pdf

¹⁹ European Commission. (2013). Laying down the rules for participation and dissemination in ‘Horizon 2020 – the Framework Programme for Research and Innovation (2014–2020)’ and repealing Regulation (EC) No 1906/2006. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/rules_participation/h2020-rules-participation_en.pdf

²⁰ European Commission. (2013). Council decision of 3 December 2013 establishing the specific programme implementing Horizon 2020 – the Framework Programme for Research and Innovation (2014–2020) and repealing Decisions 2006/971/EC, 2006/972/EC, 2006/973/EC, 2006/974/EC and 2006/975/EC. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/legal_basis/sp/h2020-sp_en.pdf

²¹ There were originally two indicators proposed in this area; however, discussions at the second Steering Group Meeting and first Plenary Meeting led to the decision to discontinue work on the FP7 variable due to its focus solely on project titles.

3.1 Proportion of a country's research outputs integrating a gender dimension in its research content (GDRC)

3.1.1 Definition

This indicator simply consists of a country's number of peer-reviewed scientific papers (those with at least one author from said country; the analysis is not restricted to the reprint author in this case) in which a gender dimension has been identified in their research content divided by the total number of peer-reviewed scientific papers of the corresponding country. Note that the concept of the Gender Dimension in Research Content (GDRC) covers both the sex (biological characteristics of both women and men) and gender (social/cultural factors of both women and men) dimension (see Appendix 4).

3.1.2 Source of data

This indicator will be computed by Science-Metrix using raw bibliographic data derived from the Web of Science (WoS) produced by Thomson Reuters (see Section 2.1.2).

3.1.3 Availability over time

The indicator can be computed for the 12 years to be covered by the She Figures 2015 publication (2002 to 2013). Thus, the completeness of this indicator in regard to the time coverage stands at 100%. Note that the yearly data is based on the publication year of papers indexed in the WoS.

3.1.4 International availability

Data can readily be produced for all 41 countries to be covered in the She Figures 2015 publication – that is, for the 28 EU Member States, as well as for Albania, Bosnia & Herzegovina, Faroe Islands, Iceland, Israel, Liechtenstein, the Former Yugoslav Republic of Macedonia, Moldova, Montenegro, Norway, Serbia, Switzerland, and Turkey. Thus, the completeness of this indicator in regard to the geographical scope stands at 100%. Note that data by country is based on author addresses.

3.1.5 Availability across Fields of Science and Technology (FOS)

Same as above (see Section 2.1.5).

3.1.6 Comparability

Contrary to the indicators on authorships, international co-authorships and scientific quality/impact, most of the comparability issues discussed above in Section 2.1.6 do not apply here. This is because there is no need to assign a sex to authors in the present case.

Bias in the number of documents over time

Unless there was to be a drastic change in the terminology used to describe GDRC-related content in scientific publications (see Appendix 4 on the methods used to create the GDRC dataset), there is no reason to expect any bias over time. An analysis of trends in the computed indicator indicates that such an issue is not present (data not shown).

Bias in favour of some countries

Same as above (see Section 2.1.5).

Bias in favour of disciplines

The component of the disciplinary bias related to the proportion of papers for which the full given name of the reprint author is available (i.e. the recall) does not apply here. However, the component related to the coverage bias in the WoS *does* apply. Thus, because the set of publications available to compute the proportion of a country's publications integrating a gender dimension in its research content represent samples rather than entire populations, the random sampling error of this proportion was computed at the FOS level (as well as for all fields combined) to obtain a 90% confidence interval (see Section 2.1.7).

3.1.7 Calculation method

Formula

The method used in identifying the peer-reviewed scientific publications with a gender dimension in their research content is described in detail in Appendix 4 and 5. Once the GDRC dataset is completed, the computation of the proportion of a country's publications integrating a gender dimension in its research content for a given year and FOS is straightforward:

$$\text{ProportionGDRC}_{CYF} = \text{GDRCP}_{CYF} / \text{TP}_{CYF}$$

Where,

- (GDRCP_{CYF}) Number of GDRC papers in a given country (C), year (Y) and FOS (F);
- (TP_{CYF}) Total number of papers in a given country (C), year (Y) and FOS (F) in WoS.

The CAGR in this indicator was computed by applying the following formula:

$$\text{CAGR}_{CYF} = (P_E / P_S)^{1/N-1}$$

Where,

- (P_E) ProportionGDRC_{CYF} in the end year;
- (P_S) ProportionGDRC_{CYF} in the start year; and
- (N) Number of years in the reference period (i.e. e – s).

When using moving periods instead of yearly data (see accuracy sub-section below), N in the CAGR formula corresponds to the last year in the end period (e.g. E = 2010–2013) minus the last year in the start period (e.g. S = 2002–2005); in the above example, N would therefore equal 8 (i.e. 2013 – 2005).

Accuracy: As noted in Section 3.1.6, the computed proportions of GDRC papers are based on samples rather than entire populations due to the presence of coverage biases across subfields in the WoS. To account for the random sampling errors, the margin of error in a country's proportion of publications integrating a gender dimension in their research content has been computed and subsequently used to construct 90% confidence intervals for this indicator using the following formula:

Lower limit of the 90% confidence interval of the proportion of a country's publications including a GDRC for a given year and FOS

$$\frac{\text{GDRCP}_{CYF}}{\text{TP}_{CYF}} + \frac{\sqrt{\frac{N_{CYF} - \text{TP}_{CYF}}{N_{CYF} - 1}} \cdot 1.645 \cdot \sqrt{\frac{\text{GDRCP}_{CYF}/\text{TP}_{CYF} (1 - \text{GDRCP}_{CYF}/\text{TP}_{CYF})}{\text{TP}_{CYF}}}}{\text{GDRCP}_{CYF}/\text{TP}_{CYF}}$$

Upper limit of the 90% confidence interval of the proportion of a country's publications including a GDRC for a given year and FOS

$$\frac{\text{GDRCP}_{CYF}}{\text{TP}_{CYF}} - \frac{\sqrt{\frac{N_{CYF} - \text{TP}_{CYF}}{N_{CYF} - 1}} \cdot 1.645 \cdot \sqrt{\frac{\text{GDRCP}_{CYF}/\text{TP}_{CYF} (1 - \text{GDRCP}_{CYF}/\text{TP}_{CYF})}{\text{TP}_{CYF}}}}{\text{GDRCP}_{CYF}/\text{TP}_{CYF}}$$

Where,

- (GDRCP_{CYF}) Number of GDRC papers in a given country (C), year (Y) and FOS (F);
- (TP_{CYF}) Total number of papers in a given country (C), year (Y) and FOS (F) in WoS; and
- (N_{CYF}) The estimated number of papers in a given FOS and year for a given country (i.e. beyond the WoS) (N_{CYF} equals the sum of N_{YS} (see Appendix 2) across the subfields found in the given FOS).

Note that the confidence intervals thus obtained do not account for the accuracy and recall of the GDRC dataset, which are respectively estimated at: 97% and 58% (see Appendix 4 and 5).

Appendix 1: Match between Science-Matrix's classification and the FOS in the Frascati Manual

Table 7 Correspondence table between Science-Matrix's classification and the Field of Science and Technology (FOS) as defined in the Frascati Manual (revised classification of 2007)

Domain	Field	Subfield	FOS	FOS_code
Applied Sciences	Agriculture, Fisheries & Forestry	Agronomy & Agriculture	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Agriculture, Fisheries & Forestry	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Fisheries	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Food Science	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Forestry	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Horticulture	Agricultural Sciences	AS
Applied Sciences	Agriculture, Fisheries & Forestry	Veterinary Sciences	Agricultural Sciences	AS
Applied Sciences	Built Environment & Design	Architecture	Humanities	H
Applied Sciences	Built Environment & Design	Building & Construction	Engineering and Technology	ET
Applied Sciences	Built Environment & Design	Design Practice & Management	Engineering and Technology	ET
Applied Sciences	Built Environment & Design	Urban & Regional Planning	Social Sciences	SS
Applied Sciences	Enabling & Strategic Technologies	Bioinformatics	Natural Sciences	NS
Applied Sciences	Enabling & Strategic Technologies	Biotechnology	Engineering and Technology	ET
Applied Sciences	Enabling & Strategic Technologies	Energy	Engineering and Technology	ET
Applied Sciences	Enabling & Strategic Technologies	Materials	Engineering and Technology	ET
Applied Sciences	Enabling & Strategic Technologies	Nanoscience & Nanotechnology	Engineering and Technology	ET
Applied Sciences	Enabling & Strategic Technologies	Optoelectronics & Photonics	Engineering and Technology	ET
Applied Sciences	Enabling & Strategic Technologies	Strategic, Defence & Security Studies	Unknown	Unknown
Applied Sciences	Engineering	Aerospace & Aeronautics	Engineering and Technology	ET
Applied Sciences	Engineering	Automobile Design & Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Biomedical Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Chemical Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Civil Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Electrical & Electronic Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Environmental Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Geological & Geomatics Engineering	Engineering and Technology	ET
Applied Sciences	Engineering	Industrial Engineering & Automation	Engineering and Technology	ET
Applied Sciences	Engineering	Mechanical Engineering & Transports	Engineering and Technology	ET
Applied Sciences	Engineering	Mining & Metallurgy	Engineering and Technology	ET
Applied Sciences	Engineering	Operations Research	Engineering and Technology	ET
Applied Sciences	Information & Communication Technologies	Artificial Intelligence & Image Processing	Natural Sciences	NS
Applied Sciences	Information & Communication Technologies	Computation Theory & Mathematics	Natural Sciences	NS
Applied Sciences	Information & Communication Technologies	Computer Hardware & Architecture	Engineering and Technology	ET
Applied Sciences	Information & Communication Technologies	Distributed Computing	Natural Sciences	NS
Applied Sciences	Information & Communication Technologies	Information Systems	Natural Sciences	NS
Applied Sciences	Information & Communication Technologies	Medical Informatics	Natural Sciences	NS
Applied Sciences	Information & Communication Technologies	Networking & Telecommunications	Engineering and Technology	ET
Applied Sciences	Information & Communication Technologies	Software Engineering	Natural Sciences	NS
Arts & Humanities	Communication & Textual Studies	Communication & Media Studies	Social Sciences	SS
Arts & Humanities	Communication & Textual Studies	Languages & Linguistics	Humanities	H
Arts & Humanities	Communication & Textual Studies	Literary Studies	Humanities	H
Arts & Humanities	Historical Studies	Anthropology	Social Sciences	SS
Arts & Humanities	Historical Studies	Archaeology	Humanities	H
Arts & Humanities	Historical Studies	Classics	Humanities	H
Arts & Humanities	Historical Studies	History	Humanities	H
Arts & Humanities	Historical Studies	History of Science, Technology & Medicine	Humanities	H
Arts & Humanities	Historical Studies	History of Social Sciences	Social Sciences	SS
Arts & Humanities	Historical Studies	zzHistorical Studies - Unclassified	Humanities	H
Arts & Humanities	Philosophy & Theology	Applied Ethics	Social Sciences	SS
Arts & Humanities	Philosophy & Theology	Philosophy	Humanities	H
Arts & Humanities	Philosophy & Theology	Religions & Theology	Humanities	H
Arts & Humanities	Visual & Performing Arts	Art Practice, History & Theory	Humanities	H
Arts & Humanities	Visual & Performing Arts	Drama & Theater	Humanities	H
Arts & Humanities	Visual & Performing Arts	Folklore	Humanities	H
Arts & Humanities	Visual & Performing Arts	Music	Humanities	H
Economic & Social Sciences	Economics & Business	Accounting	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Agricultural Economics & Policy	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Business & Management	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Development Studies	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Econometrics	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Economic Theory	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Economics	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Finance	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Industrial Relations	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Logistics & Transportation	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Marketing	Social Sciences	SS
Economic & Social Sciences	Economics & Business	Sport, Leisure & Tourism	Social Sciences	SS
Economic & Social Sciences	Economics & Business	zzEconomics & Business - Unclassified	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Criminology	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Cultural Studies	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Demography	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Education	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Family Studies	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Gender Studies	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Geography	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Information & Library Sciences	Social Sciences	SS
Economic & Social Sciences	Social Sciences	International Relations	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Law	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Political Science & Public Administration	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Science Studies	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Social Sciences Methods	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Social Work	Social Sciences	SS
Economic & Social Sciences	Social Sciences	Sociology	Social Sciences	SS
General	General Arts, Humanities & Social Sciences	General Arts, Humanities & Social Sciences	Unknown	Unknown
General	General Science & Technology	General Science & Technology	Unknown	Unknown
Health Sciences	Biomedical Research	Anatomy & Morphology	Medical Sciences	MS

Source: Compiled by Science-Matrix using the Frascati Manual and Science-Matrix classification

Table 7 Continued

Domain	Field	Subfield	FOS	FOS_code
Health Sciences	Biomedical Research	Biochemistry & Molecular Biology	Natural Sciences	NS
Health Sciences	Biomedical Research	Biophysics	Natural Sciences	NS
Health Sciences	Biomedical Research	Developmental Biology	Natural Sciences	NS
Health Sciences	Biomedical Research	Genetics & Heredity	Natural Sciences	NS
Health Sciences	Biomedical Research	Microbiology	Natural Sciences	NS
Health Sciences	Biomedical Research	Microscopy	Natural Sciences	NS
Health Sciences	Biomedical Research	Mycology & Parasitology	Medical Sciences	M5
Health Sciences	Biomedical Research	Nutrition & Dietetics	Medical Sciences	M5
Health Sciences	Biomedical Research	Physiology	Medical Sciences	M5
Health Sciences	Biomedical Research	Toxicology	Medical Sciences	M5
Health Sciences	Biomedical Research	Virology	Natural Sciences	NS
Health Sciences	Biomedical Research	zzBiomedical Research - Unclassified	Medical Sciences	M5
Health Sciences	Clinical Medicine	Allergy	Medical Sciences	M5
Health Sciences	Clinical Medicine	Anesthesiology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Arthritis & Rheumatology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Cardiovascular System & Hematology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Complementary & Alternative Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Dentistry	Medical Sciences	M5
Health Sciences	Clinical Medicine	Dermatology & Venereal Diseases	Medical Sciences	M5
Health Sciences	Clinical Medicine	Emergency & Critical Care Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Endocrinology & Metabolism	Medical Sciences	M5
Health Sciences	Clinical Medicine	Environmental & Occupational Health	Medical Sciences	M5
Health Sciences	Clinical Medicine	Gastroenterology & Hepatology	Medical Sciences	M5
Health Sciences	Clinical Medicine	General & Internal Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	General Clinical Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Geriatrics	Medical Sciences	M5
Health Sciences	Clinical Medicine	Immunology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Legal & Forensic Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Neurology & Neurosurgery	Medical Sciences	M5
Health Sciences	Clinical Medicine	Nuclear Medicine & Medical Imaging	Medical Sciences	M5
Health Sciences	Clinical Medicine	Obstetrics & Reproductive Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Oncology & Carcinogenesis	Medical Sciences	M5
Health Sciences	Clinical Medicine	Ophthalmology & Optometry	Medical Sciences	M5
Health Sciences	Clinical Medicine	Orthopedics	Medical Sciences	M5
Health Sciences	Clinical Medicine	Otorhinolaryngology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Pathology	Medical Sciences	M5
Health Sciences	Clinical Medicine	Pediatrics	Medical Sciences	M5
Health Sciences	Clinical Medicine	Pharmacology & Pharmacy	Medical Sciences	M5
Health Sciences	Clinical Medicine	Psychiatry	Medical Sciences	M5
Health Sciences	Clinical Medicine	Respiratory System	Medical Sciences	M5
Health Sciences	Clinical Medicine	Sport Sciences	Medical Sciences	M5
Health Sciences	Clinical Medicine	Surgery	Medical Sciences	M5
Health Sciences	Clinical Medicine	Tropical Medicine	Medical Sciences	M5
Health Sciences	Clinical Medicine	Urology & Nephrology	Medical Sciences	M5
Health Sciences	Psychology & Cognitive Sciences	Behavioral Science & Comparative Psychology	Natural Sciences	NS
Health Sciences	Psychology & Cognitive Sciences	Clinical Psychology	Medical Sciences	M5
Health Sciences	Psychology & Cognitive Sciences	Developmental & Child Psychology	Social Sciences	SS
Health Sciences	Psychology & Cognitive Sciences	Experimental Psychology	Medical Sciences	M5
Health Sciences	Psychology & Cognitive Sciences	General Psychology & Cognitive Sciences	Social Sciences	SS
Health Sciences	Psychology & Cognitive Sciences	Human Factors	Social Sciences	SS
Health Sciences	Psychology & Cognitive Sciences	Psychoanalysis	Social Sciences	SS
Health Sciences	Psychology & Cognitive Sciences	Social Psychology	Social Sciences	SS
Health Sciences	Public Health & Health Services	Epidemiology	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Gerontology	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Health Policy & Services	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Nursing	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Public Health	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Rehabilitation	Medical Sciences	M5
Health Sciences	Public Health & Health Services	Speech-Language Pathology & Audiology	Social Sciences	SS
Health Sciences	Public Health & Health Services	Substance Abuse	Medical Sciences	M5
Natural Sciences	Biology	Ecology	Natural Sciences	NS
Natural Sciences	Biology	Entomology	Natural Sciences	NS
Natural Sciences	Biology	Evolutionary Biology	Natural Sciences	NS
Natural Sciences	Biology	Marine Biology & Hydrobiology	Natural Sciences	NS
Natural Sciences	Biology	Ornithology	Natural Sciences	NS
Natural Sciences	Biology	Plant Biology & Botany	Natural Sciences	NS
Natural Sciences	Biology	Zoology	Natural Sciences	NS
Natural Sciences	Chemistry	Analytical Chemistry	Natural Sciences	NS
Natural Sciences	Chemistry	General Chemistry	Natural Sciences	NS
Natural Sciences	Chemistry	Inorganic & Nuclear Chemistry	Natural Sciences	NS
Natural Sciences	Chemistry	Medicinal & Biomolecular Chemistry	Medical Sciences	M5
Natural Sciences	Chemistry	Organic Chemistry	Natural Sciences	NS
Natural Sciences	Chemistry	Physical Chemistry	Natural Sciences	NS
Natural Sciences	Chemistry	Polymers	Natural Sciences	NS
Natural Sciences	Chemistry	zzChemistry - Unclassified	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Environmental Sciences	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Geochemistry & Geophysics	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Geology	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Meteorology & Atmospheric Sciences	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Oceanography	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	Paleontology	Natural Sciences	NS
Natural Sciences	Earth & Environmental Sciences	zzEarth & Environmental Sciences - Unclassified	Natural Sciences	NS
Natural Sciences	Mathematics & Statistics	Applied Mathematics	Natural Sciences	NS
Natural Sciences	Mathematics & Statistics	General Mathematics	Natural Sciences	NS
Natural Sciences	Mathematics & Statistics	Numerical & Computational Mathematics	Natural Sciences	NS
Natural Sciences	Mathematics & Statistics	Statistics & Probability	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Acoustics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Applied Physics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Astronomy & Astrophysics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Chemical Physics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Fluids & Plasmas	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	General Physics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Mathematical Physics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Nuclear & Particles Physics	Natural Sciences	NS
Natural Sciences	Physics & Astronomy	Optics	Natural Sciences	NS
UNKNOWN	UNKNOWN	UNKNOWN	Unknown	Unknown

Source:

Compiled by Science-Metrix using the Frascati Manual and Science-Metrix classification

Appendix 2: Coverage bias across subfields in the WoS

Of relevance to Section 2.1 to 2.3

A comparative analysis of the share of referenced publications that are indexed in the WoS across subfields (based on Science-Metrix's classification) revealed a substantial amount of variability in the proportion of a given subfield's publications that is actually covered in the WoS (ranging from a low of 3% in Art Practice, History & Theory to a high of 90% in Developmental Biology, Table 8). The impact of such variation in the coverage of the scientific literature across subfields is that women-to-men ratios (see Sections 2.1 to 2.3) computed at higher aggregation levels (e.g. for FOS or all fields combined) will be biased towards the scores of the subfields that are better covered in the WoS.

This bias must therefore be accounted for in computing aggregated ratios. This is achieved using a weighting scheme of the ratios computed for individual subfields, which more adequately reflects the representation of subfields within the whole realm of scientific research beyond the WoS (i.e. including the relevant literature not indexed in the WoS). The weighting is obtained by first computing the number of publications in each subfield and year in the world beyond the WoS (denoted N_{YS} in the below formula [same variable as found in the formulas of Sections 2.1 to 2.3; 3rd data column in Table 8) using the estimated share of the corresponding subfield and year that is indexed in the WoS (denoted S_{YS} in the below formula; 2nd data column in Table 8) and the actual number of papers in the WoS in the corresponding subfield and year (denoted WP_{YS} in the below formula 1st data column in Table 8). For a given subfield and year, the number of publications in each subfield in this expanded population is obtained using the following formula:

$$N_{YS} = WP_{YS} * (1/S_{YS})$$

The number of publications by subfield in this expanded population can then be used to compute the weight for a given subfield and year. This weight corresponds to the proportion of the subfield of interest in the expanded population for the given year as expressed in the following formula, which appears as a component of the formulas presented in Sections 2.1 to 2.3 (tbd = to be determined based on the subfields that are to be aggregated, see Section 2.1.7 for an explanation):

$$Weight_{YS} = \frac{N_{YS}}{\sum_{S=1}^{TBD} N_{YS}}$$

See Section 2.1.7 for information on how these weights are used in computing the aggregated ratios at the FOS level (as well as for all fields combined).

Additionally, because the set of publications available to compute the proportions of women and men authorships at the subfield level represent samples rather than entire populations, the random sampling errors of these estimates at the subfield level have also been aggregated using the above-mentioned weighting schemes to construct the 90% confidence intervals of the aggregated ratios. To achieve this, it is necessary to compute the number of publications in each country, year and subfield beyond the WoS (denoted N_{CYS} in the below formula [same variable as found in the formulas of Sections 2.1 to 2.3). This is done using the estimated share of the corresponding subfield and year, which is indexed in the WoS (denoted S_{YS} in the below formula; 2nd data column in Table 8; no country specific values are used) and the actual number of papers in the WoS in the corresponding country, subfield and year (denoted WP_{CYS} in the formula below):

$$N_{CYS} = WP_{CYS} * (1/S_{YS})$$

Table 8 Number of scientific publications beyond the WoS (2007–2013)

Subfield	No. of papers in WoS	Share of cited references indexed in WoS	Estimated no. of papers in the world (N_{VS} in the indicator formulas for Section 2.1, 2.2 and 2.3)
Developmental Biology	97,862	90%	108,687
Immunology	107,173	89%	120,011
Biochemistry & Molecular Biology	176,630	89%	198,626
Endocrinology & Metabolism	80,393	85%	94,541
Oncology & Carcinogenesis	194,483	85%	229,770
Biophysics	34,690	84%	41,065
Virology	55,613	84%	66,021
Neurology & Neurosurgery	238,135	84%	283,461
Nanoscience & Nanotechnology	94,847	83%	113,944
Gastroenterology & Hepatology	68,663	83%	82,607
Pathology	31,741	83%	38,229
Physiology	32,570	83%	39,399
Cardiovascular System & Hematology	148,897	83%	180,480
Urology & Nephrology	64,236	82%	78,738
Genetics & Heredity	31,069	82%	38,103
Arthritis & Rheumatology	30,473	81%	37,393
Pharmacology & Pharmacy	103,053	81%	126,500
General Science & Technology	152,668	81%	188,400
Microbiology	149,097	81%	184,182
Allergy	14,225	81%	17,600
Organic Chemistry	174,641	80%	217,405
Bioinformatics	39,330	80%	49,159
Respiratory System	43,948	80%	54,960
Physical Chemistry	88,349	80%	110,995
Ophthalmology & Optometry	50,845	79%	64,132
Medicinal & Biomolecular Chemistry	81,086	79%	102,941
Biomedical Engineering	49,634	79%	63,136
General Chemistry	113,964	78%	145,397
Analytical Chemistry	133,694	78%	171,200
Surgery	86,446	78%	110,795
Obstetrics & Reproductive Medicine	73,437	78%	94,179
Nuclear Medicine & Medical Imaging	83,040	78%	106,513
Chemical Physics	133,789	78%	172,301
Astronomy & Astrophysics	74,348	77%	96,385
General Clinical Medicine	25,291	77%	32,835
Biotechnology	89,077	77%	115,748
Anesthesiology	29,274	77%	38,067
Toxicology	51,212	76%	67,045
Polymers	113,659	76%	148,900
Psychiatry	72,371	75%	96,814
Orthopedics	63,362	75%	84,871
Inorganic & Nuclear Chemistry	117,577	75%	157,777
Nutrition & Dietetics	48,612	74%	65,288
Emergency & Critical Care Medicine	30,329	74%	40,847
Applied Physics	276,944	74%	374,656
Dermatology & Venereal Diseases	40,772	74%	55,234
Epidemiology	20,155	74%	27,308
Otorhinolaryngology	33,439	73%	45,546
General Physics	128,199	73%	174,850
Geriatrics	13,591	73%	18,591
Fluids & Plasmas	140,707	73%	192,665
Behavioral Science & Comparative Psychology	21,549	73%	29,639
Pediatrics	48,633	72%	67,423
Plant Biology & Botany	114,236	72%	158,938
Experimental Psychology	55,362	71%	77,657
Meteorology & Atmospheric Sciences	91,184	71%	127,938
Optoelectronics & Photonics	41,852	71%	58,940
Materials	190,629	71%	268,582
Optics	73,430	71%	103,507
Food Science	66,680	70%	94,822
Chemical Engineering	79,692	70%	113,490
General & Internal Medicine	114,191	70%	163,926
Veterinary Sciences	60,941	70%	87,535
Dentistry	55,810	69%	80,471
Microscopy	7,245	69%	10,471
Gerontology	13,193	69%	19,230
Substance Abuse	19,132	68%	27,991
Sport Sciences	30,659	68%	45,248
Anatomy & Morphology	9,207	67%	13,680
Nuclear & Particles Physics	116,882	67%	175,275
Mycology & Parasitology	26,117	66%	39,442
Evolutionary Biology	52,248	66%	79,665
Tropical Medicine	24,082	65%	37,210
Dairy & Animal Science	60,827	65%	94,123
Complementary & Alternative Medicine	9,413	64%	14,646
Environmental Sciences	74,440	64%	116,710
Clinical Psychology	21,176	63%	33,444
Developmental & Child Psychology	24,956	63%	39,799
Rehabilitation	29,355	63%	46,939
Oceanography	16,790	63%	26,851
Energy	179,186	62%	287,269
Geochemistry & Geophysics	71,405	61%	116,465
Marine Biology & Hydrobiology	56,426	61%	92,164
Fisheries	34,881	61%	57,486
Ecology	71,446	60%	118,495
Environmental & Occupational Health	11,112	60%	18,434
Environmental Engineering	53,013	59%	89,723

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Table 8 Continued

Subfield	No. of papers in WoS	Share of cited references Indexed in WoS	Estimated no. of papers in the world (N_{VS} in the indicator formulas for Section 2.1, 2.2 and 2.3)
Legal & Forensic Medicine	8,093	59%	13,761
Entomology	37,995	58%	65,598
Mathematical Physics	23,584	58%	40,786
Finance	12,540	58%	21,764
Acoustics	27,117	57%	47,373
Applied Mathematics	60,346	57%	106,673
Econometrics	4,077	57%	7,216
Mechanical Engineering & Transports	82,204	56%	145,650
Horticulture	11,164	56%	19,858
Public Health	73,788	56%	131,453
Health Policy & Services	22,859	56%	40,754
Agronomy & Agriculture	76,002	56%	136,190
Social Psychology	37,075	55%	67,536
Ornithology	8,100	54%	15,068
Economic Theory	5,451	54%	10,161
Medical Informatics	13,215	54%	24,638
Numerical & Computational Mathematics	46,491	53%	87,205
Statistics & Probability	41,358	53%	77,771
Family Studies	4,239	53%	8,073
Geological & Geomatics Engineering	28,366	52%	54,715
Strategic, Defence & Security Studies	37,345	52%	72,035
Operations Research	34,770	51%	67,530
Marketing	12,109	51%	23,607
Geology	17,101	51%	33,685
Business & Management	32,891	50%	65,769
Forestry	26,487	50%	53,187
Industrial Engineering & Automation	55,328	49%	112,243
Speech-Language Pathology & Audiology	8,842	48%	18,386
Electrical & Electronic Engineering	47,211	48%	98,683
Nursing	39,689	48%	83,055
General Psychology & Cognitive Sciences	5,013	46%	10,843
Accounting	3,751	46%	8,163
Mining & Metallurgy	25,655	45%	56,397
Economics	57,609	45%	128,108
Paleontology	26,816	45%	60,072
Artificial Intelligence & Image Processing	60,213	44%	136,275
General Mathematics	121,621	44%	278,002
Human Factors	7,148	43%	16,477
Agricultural Economics & Policy	8,458	43%	19,640
Design Practice & Management	7,911	43%	18,392
Networking & Telecommunications	96,335	43%	224,650
Automobile Design & Engineering	4,914	43%	11,477
Civil Engineering	31,439	43%	73,427
Building & Construction	23,745	42%	55,893
Aerospace & Aeronautics	19,131	42%	45,911
Industrial Relations	4,794	41%	11,746
Criminology	14,647	40%	36,873
Social Sciences Methods	5,519	39%	14,010
Computation Theory & Mathematics	33,241	39%	85,830
Logistics & Transportation	20,672	38%	53,723
Social Work	7,993	36%	22,255
Demography	4,205	36%	11,738
Applied Ethics	11,111	36%	31,145
Information Systems	17,861	35%	50,401
Zoology	29,009	35%	81,992
Geography	21,017	35%	60,434
Urban & Regional Planning	12,556	33%	37,706
Science Studies	9,341	33%	28,145
Sport, Leisure & Tourism	6,180	33%	18,837
Education	57,433	32%	181,087
Anthropology	13,068	30%	43,278
Information & Library Sciences	12,182	29%	41,342
Development Studies	6,662	29%	22,757
Sociology	14,677	29%	51,414
Software Engineering	18,271	28%	65,116
Computer Hardware & Architecture	7,602	27%	27,761
Distributed Computing	6,843	27%	25,644
Psychoanalysis	4,029	25%	16,010
General Arts, Humanities & Social Sciences	9,131	25%	36,474
Political Science & Public Administration	31,673	23%	136,528
Communication & Media Studies	12,031	23%	52,442
Gender Studies	4,112	23%	18,243
Archaeology	12,213	21%	59,328
Languages & Linguistics	23,450	16%	149,294
International Relations	9,989	16%	64,001
Law	16,556	15%	109,938
Philosophy	20,074	15%	136,817
History of Social Sciences	3,682	13%	29,112
Music	5,137	9%	58,243
Religions & Theology	12,254	7%	166,138
History of Science, Technology & Medicine	4,918	7%	70,547
Cultural Studies	12,468	7%	189,027
Architecture	2,777	6%	45,571
Folklore	1,467	5%	27,803
Classics	4,182	5%	80,440
Drama & Theater	2,461	5%	53,588
History	25,744	4%	681,479
Literary Studies	33,823	4%	927,800
Art Practice, History & Theory	4,552	3%	155,117

Source: Compiled by Science-Metrix using WoS data (Thomson Reuters)

Of relevance to Section 3.1

For the proportion of a country's publications including a Gender Dimension in its Research Content (GDRC), the estimated number of papers in a given country, subfield and year (i.e. N_{CYS}) were also used in computing the 90% confidence interval of this indicator. The estimated number of papers for a given country and year were summed across the relevant subfields to obtain the estimated number of papers for a given country, year and FOS (subfield aggregates matching the FOS as defined in the Frascati Manual, see Appendix 1) in the entire scientific literature (i.e. beyond the WoS).

Appendix 3: Gender name disambiguation

To produce the indicators measuring various dimensions of the scientific (see Sections 2.1 to 2.3) and technological (see Section 2.4) productions of women and men, the names of authors on peer-reviewed scientific publications – those indexed in the Web of Science (WoS) – and of inventors on European patent applications – those indexed in PATSTAT – must be attributed a sex.

The first step in performing this task consisted of conditioning the names of authors and inventors as they appear in the WoS and PATSTAT so as to match the format required by the gender name disambiguation algorithm that has been used in this study (see later in this section). Note that the required format is as follows:

- one column for the given name (without initials); and
- one column for the surname.

Conditioning of author names in the WoS

In the WoS, the names of authors are already parsed into two columns, one for the given name and the other one for the surname. However, the given name of authors is not reported in a unified fashion. Sometimes, given names are simply the initials of the authors, whereas in other instances they represent complete names. Intermediate cases mixing initials with full names exist. Furthermore, special characters are sometimes introduced in this field such as: '()', '-', '.' and ' '. For instance, 'John H.' could appear in many different forms such as:

- John
- John Harold
- John-Harold
- John H.
- J.H.
- J.
- J. Harold
- J. (Harold)

As a first step towards conditioning the given names to ensure a proper match in the gender name disambiguation algorithm, all entries consisting of initials only were identified and removed from the list. Because the format of initials in the database most often follows the format in which each initial is capitalised and separated from other initials using a period – for example, A.G.H. – they can easily be identified by:

- Counting the number of '.' in the string.
- Calculating the total length of the string.
- If the total length of the string is smaller than three times the number of '.', then the string consists of initials only and is discarded.

Once the given names consisting of initials only were discarded, all special characters were replaced by a hyphen. Subsequent to this, each string was divided into its constituent parts (i.e. the substrings, which are separated by a '-'). Any segments of length 0 (replaced special character) or 1 (remaining initials in a given name) were then discarded prior to re-merging all segments of a string using, again, a hyphen. For example, this step allowed removing the initial in 'Louis-Philippe-C' to obtain 'Louis-Philippe'.

Subsequently, all given names of length 2 were excluded if they were in the form of two vowels or two consonants, as well as if they were both capitalised (manual validations whereby the actual papers were downloaded confirmed that two uppercase letters most often represent initials). Since

the given names that satisfied the mentioned rule could not be pronounced, they were removed and considered as cases of two initials that could not be removed in the first step because they followed a different format (e.g. AB).

In the case of given names ranging in length from three to seven characters, those that were fully capitalised were also removed as they consisted, again, of initials only (e.g. ATD and SKP). Once this process was completed, nearly 334,000 unique given names were left in the WoS for the 2002–2013 period, only considering the reprint author.

Once in this format, the given names were ready for gender attribution using the gender name disambiguation algorithm. In the WoS, there were about 2,028,000 unique combinations (i.e. given name/surname) to be treated using the algorithm once the cleaning of given names was completed for the 2002–2013 period for reprint authors only.

Conditioning of inventor names in PATSTAT-EPO

In PATSTAT, the names of inventors are always linked to their addresses and the country code has been extracted in a separate column for EPO patent applications (kind codes: A1 and A2). However, the names of inventors have not been parsed into two columns to separate the given name from the surname. Fortunately, most of the time the format is uniform with the surname coming first and being separated from the given name by a comma.

In cases where no comma is present in the string, the order of appearance is not reliable for identifying the given name. As such, these names (about 26,000 unique names for 2002–2013) have been treated as 'Unclassified' in performing the gender attribution.

In cases where multiple commas are present, only the substring delimited by the first and second comma is kept as the given name. The remaining portion often consists of the names of the companies in which inventors are located. Note that these formats are the exception rather than the rule.

Subsequently, an intensive cleaning was performed on all the given names for company abbreviations (e.g. ab, ag, as, bv, c/o, corp, gmbh, inc, ltd, limited), academic-related words (e.g. university, college, school, engineering), industry-related keywords and abbreviations (e.g. indust*²², technolog*, pharma*, optic*, genetic*), and so forth, to identify cases in which the name of the inventor appears along with some other useless information. In these cases, the names of the inventors usually appear at the start of the string. Therefore, only the first segment is retained.

At this point, an approach similar to the one used for authors in the WoS was used. Briefly, all special characters were replaced by a hyphen. Subsequent to this, each string was divided into its constituent parts (i.e. the segments that are separated by '-'). Any segments of length 0 (replaced special character) or 1 (remaining initials in a given name) were then discarded. Prior to re-merging all segments of a string using, again, a hyphen, a visual scan was performed to identify recurring substrings (e.g. a person's title) that are not specific to an individual's name (e.g. Prof., Dr., Ing.). These terms were also removed and the remaining segments of a given name were reassembled. Once this process was completed, there were about 1,800,000 unique combinations (about 188,000 unique given names) for the 2002–2013 period.

²² 'word*' represents an umbrella term that starts with the 'word'.

Gender name disambiguation algorithm

In many countries, including most European ones, given names are gender-specific such that by searching a name in repositories such as the lists of frequently occurring women's and men's names prepared by the US Census Bureau²³ or various databases of baby boys' and girls' names by country/ethnicity,^{24,25} it is feasible to attribute a gender to a given name with a high level of accuracy. This step involves matching a given name against the corresponding entry in the selected repository and of assigning the matched sex. Usually, the assignment is only performed in cases where the given name can be attributed unambiguously, as some names are unisex and may be found among both women and men. Larivière et al., in their *Nature* paper on gender disparities in science, applied the following rule regarding unisex attribution:²⁶

In cases where a name was used for both genders, it was only attributed to a specific gender when it was used at least ten times more frequently for one gender than the other.

In performing such matches, it is preferable to account for the geographic location (i.e. country), or at least the ethnicity, of the origin of a person's (i.e. author or inventor) name, since a given name might be associated with a different gender depending on the cultural context from which it originates. For example, Andrea is likely a man in Italy, but a woman in the United Kingdom, though Andrea in the Italian context might be present in the United Kingdom due to migration. Thus, two approaches can be used to set the reference context against which to attribute the appropriate sex. The first one involves using *a priori* information on the actual geographic location of an individual. However, this method is prone to errors in the case of migrants and their descendants. The second one is more reliable and involves matching the surname of the person against other databases listing family names by countries of origin.

In other cases, gender is implicit in the grammar.²⁷ For example, in Iceland the surname consists of the father's given name and ends with *-sson* for a man and *-sdóttir* for a woman. In Russia, men's family names typically end in *-ov*, *-ev* or *-in*, whereas women's surnames usually end in *-ova*, *-eva* or *-ina*. In Lithuania, women's given names usually end with: *-a*, *-e* or *-ia* and men's given names usually end in *-s*, *-as*, *-is*, *-ys*, *-us* or *-ius*. Country-specific rules can therefore also be applied in determining the gender of an individual or in validating the name repositories to be used.

Finally, in some countries, it is extremely difficult to establish the gender of given names with accuracy. This is the case for China and the Republic of Korea.

In their 2013 paper on gender disparities in science, Larivière et al. applied a mixed-method approach combining automated matches against name repositories by gender, rules for countries in which gender is inherent to the grammar, and manual validations for problematic countries (e.g.

²³ https://www.census.gov/genealogy/www/data/1990surnames/names_files.html

²⁴ <http://www.babynology.com/>

²⁵ <http://www.20000-names.com/>

²⁶ Larivière, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C.R. (2013). Global gender disparities in science. *Nature*, 504: 211–213.

²⁷ <http://www.w3.org/International/questions/qa-personal-names>

China).²⁸ Yet their approach relied on the manual consolidation of a set of repositories covering the most frequent names in a restrained number of countries.

More recently, NamSor™ – a European designer of name recognition software committed to promoting diversity and equal opportunity – released GendRE API, a free API to extract gender from personal names. Since the launch of their API, NamSor has already enhanced it 13 times and the size of the database used in the background had doubled from a collection of about 400,000 names in March 2014 to 800,000 names covering all countries in the world by July 2014. Their algorithm makes use of sociolinguistics to (1) recognise the origin of the given name/surname combination to (2) subsequently infer whether the name sounds male or female in the corresponding cultural context.²⁹

NamSor have implemented a rigorous protocol to assess the quality of their tool, demonstrating that it is capable of achieving 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. For instance, based on their quality assessment, precision is systematically above 95% and recall is above 99% for the countries covered in *She Figures 2015*. Their validation procedure relies on the use of directories listing names along with their geographic location (i.e. country) and titles (Mr. for men and Ms. for women). Using the title of individual, they can validate whether or not their algorithm attributes the correct sex.

Interestingly, as noted in an article published on their website on 9 September 2014, they recently used data from the Official Directory of the European Union to validate their GendRE API, as well as to study the gender gap in the European Union. In particular, they report data on vertical segregation as it relates to the positions being occupied by women relative to men.³⁰ Based on this dataset, their API achieves nearly 100% accuracy (see Figure 2 below).

²⁸ Larivière, V., Ni, C., Gingras, Y., Cronin, B. and Sugimoto, C.R. (2013). Global gender disparities in science. *Nature*, 504: 211–213.

²⁹ <http://namesorts.com/api/>

³⁰ <http://namesorts.com/2014/09/09/whats-the-gender-gap-in-the-european-union-whoiswho/>

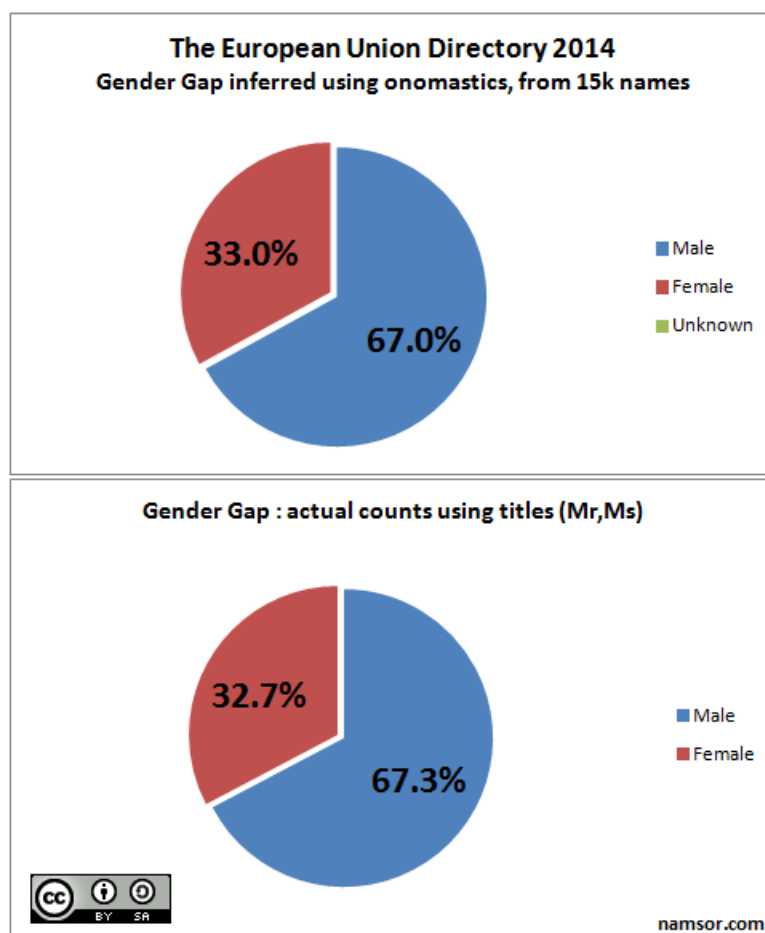


Figure 2 Validation of NamSor™ GendRE API using data from the Official Directory of the European Union

Source: NamSor™ (<http://namesorts.com/2014/09/09/whats-the-gender-gap-in-the-european-union-whoiswho/>)

Science-Metrix's analysts have applied the GenDRE API to all author and inventor names found in the WoS (conditioned names for the reprint authors of scientific papers in the WoS) and PATSTAT (conditioned names for EPO applications only). The attribution of sex to author and inventor names thus obtained was then validated to confirm the accuracy of the tool in the current context. The validation involved one or many of the following procedures depending on the availability of data by country:

- Application of sex assignment rules in the cases of countries for which the sex is encoded in the spelling (e.g. in Iceland the surname consists of the father's given name and ends with –sson for a man and –sdóttir for a woman);
- Assignment of sex using lists providing the names of men and women along with their frequency of occurrence in the population (e.g. the list provided by the US Census Bureau). Note that the following rule was applied: 'In cases where a name is used for both sexes, it is only attributed to a specific sex when it is used at least ten times more frequently for one sex than the other'; and
- Online searches were performed for specific names in the Nordic countries and a manual validation was performed for the Balkans by an analyst of Serbian origin.

Table 9 describes the validation procedure applied to each country for author names in the WoS while Table 10 presents the results of this validation procedure by country. The accuracy of assignments for validated paper/name combinations (see AVPNF column; see Table 10 notes) indicates that the tool used in attributing a sex to authors was highly accurate, with the assignments

being correct across countries included in She Figures 2015 in 97% of the cases, on average. The lowest accuracies are actually quite high and are observed for Latvia (91%), Iceland (92%), Estonia (93%) and Turkey (93%).

Table 11 describes the validation procedure applied to each country for inventor names in PATSTAT, while Table 12 presents the results of this validation procedure by country. The accuracy of assignments for validated patent application/name combinations (see AVPNF column; see Table 12 notes) indicates that the tool used in attributing a sex to inventors was highly accurate with the assignments being correct across countries included in She Figures 2015 in 97% of the cases, on average. The accuracy is smaller than 90% for only one country: Montenegro (80%).

Table 9 Validation methods of the gender assignment procedure for author names in the WoS (2007–2013) by country for those covered in She Figures 2015

Country	Method	Source	Description of rule
Belgium	Name lists	Belgian census data (http://statbel.fgov.be/fr/modules/publications/statistiques/population/prenoms_de_la_population_toale.jsp)	N/A
Bulgaria	Naming rule	Larivière, V. et al. Supplementary Information to: Nature 504, 211–213; 2013)	Male last names finish in -ev, -ov while female last names finish in -eva, -ova, -ska
Czech Republic	Name lists	Slovakian name list from the Ministry of Interior of the Slovak Republic (http://www.minv.sk/?tlacove-spravy-8&sprava=najcastejsim-menom-ktorym-rodicia-pomenovali-svoje-dieta-v-roku-2013-je-jakub)	N/A
Denmark	Name lists	Names of Newborn Children - Statistics Denmark (http://www.dst.dk/en/Statistik/emner/navne/navne-til-nyfoedte.aspx), List of Norwegian names (from validated Norway file)	N/A
Germany	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names (http://www.beliebte-vornamen.de/jahrgang/j2013/top500-2013), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm	N/A
Estonia	Name lists	Estonian given names - Wikipedia (http://en.wiktionary.org/wiki/Category:Estonian_given_names)	N/A
Ireland	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Name.org.uk: http://www.babynames.org.uk/irish-boy-names.htm	N/A
Greece	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Name.org.uk: http://www.babynames.org.uk/greek-girl-baby-names.htm , http://www.babynames.org.uk/greek-boy-baby-names.htm ; Greek names (Wikipedia and Wiktionary): http://en.wikipedia.org/wiki/Category:Greek_masculine_given_names ; http://en.wikipedia.org/wiki/Category:Greek_feminine_given_names , http://en.wiktionary.org/wiki/Category:en:Greek_female_given_names	N/A
Spain	Name lists	Instituto nacional de estadística - Names with a frequency greater than or equal to 20, and their average ages (http://www.ine.es/en/daco/daco42/nombayapel/nombayapel_en.htm)	N/A
France	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), French names (wikipedia.org/wiki/Category:French_feminine_given_names and wikipedia.org/wiki/Category:French_masculine_given_names)	N/A
Croatia	Name lists	Serbian name list (which was hand validated), Croatian name list (http://en.wikipedia.org/wiki/Croatian_name)	N/A
Italy	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Italian names (http://www.studentsoftheworld.info/penpals/stats.php?Pays=ITA)	N/A
Cyprus	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Name.org.uk: http://www.babynames.org.uk/greek-girl-baby-names.htm , http://www.babynames.org.uk/greek-boy-baby-names.htm ; Greek names (Wikipedia and Wiktionary): http://en.wikipedia.org/wiki/Category:Greek_masculine_given_names ; http://en.wikipedia.org/wiki/Category:Greek_feminine_given_names , http://en.wiktionary.org/wiki/Category:en:Greek_female_given_names	N/A
Latvia	Name lists	Latvian names: wikipedia (wikipedia.org/wiki/Category:Latvian_female_given_names and wikipedia.org/wiki/Category:Latvian_male_given_names)	N/A
Lithuania	Naming rules + Name lists	Naming rules: Larivière, V. et al. Supplementary Information to: Nature 504, 211–213; 2013, Statistics Lithuania (http://www.stat.gov.lt/en/homejsessionid=A48C1E3581D231EB3D3250C829ECCB24?p_id=3&p_p_lifecycle=0&p_p_state=maximized&p_p_mode=view&p_p_col_pos=1&p_p_col_count=2&_3_struts_action=%2Fsearch%2Fsearch&_3_redirect=%2Fen%2Fhome&_3_keywords=names&_3_groupId=0)	When assigning the genders the naming rules (MALE given name ending: -as, -is, -ys, -us; FEMALE given name ending: -a, -e, -ia) for Lithuania were applied.
Luxembourg	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm , Baby Name Wizard: http://www.babynamewizard.com/name-list/austrian-boys-names-most-popular-names-for-boys-in-austria ; http://www.babynamewizard.com/name-list/austrian-girls-names-most-popular-names-for-girls-in-austria	N/A
Hungary	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Name Wizard: http://www.babynamewizard.com/name-list/hungarian-girls-names-most-popular-names-for-girls-in-hungary ; http://www.babynamewizard.com/name-list/hungarian-boys-names-most-popular-names-for-boys-in-hungary , and Students of the World: http://www.studentsoftheworld.info/penpals/stats.php?Pays=HUN	N/A
Malta	Name lists	France name list (validated names from the France file)	N/A

Note: N/A = Not applicable.

Source: Compiled by Science-Metrix from various sources

Table 9 Continued

Country	Method	Source	Description of rule
Netherlands	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Dutch names (Wikipedia and Wiktionary): http://en.wiktionary.org/wiki/Category:Dutch_female_given_names ; http://en.wikipedia.org/wiki/Category:Dutch_masculine_given_names ; http://en.wiktionary.org/wiki/Category:Dutch_male_given_names ; http://en.wikipedia.org/wiki/Category:Dutch_feminine_given_names , Common names in Netherlands (http://web.archive.org/web/20080203090920/http://www.meertens.knaw.nl/voornamen/modern.html)	N/A
Austria	Name lists	Germany name list + US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names (http://www.beliebte-vornamen.de/jahrgang/j2013/top500-2013), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm)	N/A
Poland	Name lists	Polish name list - Polish Ministry of the Interior (https://www.msw.gov.pl/pl/aktualnosci/11689.Lena-i-Jakub-to-najpopularniejsze-imiona-mijajacego-roku.html?search=853621)	N/A
Portugal	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Portuguese names: http://en.wikipedia.org/wiki/Category:Portuguese_masculine_given_names ; http://en.wikipedia.org/wiki/Category:Portuguese_feminine_given_names	N/A
Romania	Name lists	Romanian names: (Wikipedia): http://ro.wikipedia.org/wiki/List%C4%83_de_prenume_rom%C3%A2ne%C8%99ti	N/A
Slovenia	Name lists	Slovenian name list from Statistics Slovenia (http://www.stat.si/ImenaRojstva/en/FirstNames/ExpandNames)	N/A
Slovakia	Name lists	Slovakian name list from the Ministry of Interior of the Slovak Republic (http://www.minv.sk/?tlacove-spravny-8&sprava=najcastejsim-menom-ktorym-rodicia-pomenovali-svoje-dieta-v-roku-2013-je-jakub)	N/A
Finland	Name lists + manual validation	Nordic name list (from validated Norway file), Nordic names Wiki (http://www.nordicnames.de), List of common finnish names - Finnish population register center (http://www.vrk.fi/default.aspx?id=279 and http://verkkopalvelu.vrk.fi/Nimipalvelu/default.asp?L=3)	N/A
Sweden	Name lists + manual validation	Nordic name list (from Norway validated file), Nordic Names website (http://www.nordicnames.de)	N/A
United Kingdom	Name lists	UK Census (http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcn%3A77-318125)	N/A
Iceland	Naming rules + Name lists	Naming rules: Larivière, V. et al. Supplementary Information to: Nature 504, 211-213; 2013 and A Guide to Names and Naming Practices (2006) by the UK Government Name lists: US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Names for 2012 (Iceland) http://www.nordicnames.de/wiki/icelandic_Statistics_of_2012 , Behind the name website: http://www.behindthename.com/ , and Research Gate: http://www.researchgate.net/	In order to assign a gender to a name, the naming rules (MALE last name ending: -son and -sson, and FEMALE last name endings: -dottir or -sdottir) for Iceland were applied, and then first names were checked using the Baby Names for 2012 (Iceland) website and Behind the name website; and if conflicting results were produced through either of these means, the profile of the researcher in questions was found (primarily using Research Gate) in order to verify the gender.
Liechtenstein	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm , Baby Name Wizard: http://www.babynamewizard.com/name-list/austrian-boys-names-most-popular-names-for-boys-in-austria- ; http://www.babynamewizard.com/name-list/austrian-girls-names-most-popular-names-for-girls-in-austria-	N/A
Norway	Name lists + manual validation	Statistics Norway (http://www.ssb.no/en/befolkning/statistikker/navn), US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Nordic Names website (http://www.nordicnames.de), validated name lists from Sweden and Finland	N/A
Switzerland	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), French names (wikipedia.org/wiki/Category:French_feminine_given_names and wikipedia.org/wiki/Category:French_masculine_given_names); Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm , Baby Name Wizard: http://www.babynamewizard.com/name-list/austrian-boys-names-most-popular-names-for-boys-in-austria- ; http://www.babynamewizard.com/name-list/austrian-girls-names-most-popular-names-for-girls-in-austria-	N/A
Montenegro	Name lists	Serbian name list (which was hand validated)	N/A
FYR Macedonia	Naming Rules	Naming rules Larivière, V. et al. Supplementary Information to: Nature 504, 211-213; 2013 and A Guide to Names and Naming Practices (2006) by the UK Government	N/A
Albania	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Albanian names: http://www.aboutnames.ch/albanian.htm	N/A
Serbia	Manual validation	Native speaker (due to incomplete coverage with available name lists)	N/A

Note: N/A = Not applicable.
 Source: Compiled by Science-Metrix from various sources

Table 9 Continued

Country	Method	Source	Description of rule
Turkey	Name lists	Turkey statistical office, list of 30 popular names (http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=16054&as_sfid=AAAAAVpx%2FByBO74BwkrMTiV3HRmtviiry9%2BcVdUSKnsfFzidzaPhcFe%2FySouflHPjIuv2duHitdOxQ0Be5DydNQyhxMc490u6vHQFRBsMCFzySw3A%3D%3D&as_fid=pAGvS7Y00QKhwmYjlxor), Turkish names - Wikipedia (http://en.wikipedia.org/wiki/Category:Turkish_masculine_given_names , http://en.wikipedia.org/wiki/Category:Turkish_feminine_given_names , http://en.wikipedia.org/wiki/Category:Turkish_unisex_given_names)	N/A
Bosnia and Herzegovina	Name lists	Serbian name list (which was hand validated), Bosnian name list (http://en.wikipedia.org/wiki/Category:Bosnian_given_names), Bosnian census data (http://www.fzs.ba/Dem/ProcPrist/z100.pdf and http://www.fzs.ba/Dem/ProcPrist/m100.pdf)	N/A
Israel	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Jewish names: http://jewishnames.netzah.org/	N/A
Faroe Islands	Name lists + manual validation	Denmark name list (from validated Denmark file)+ Nordic Names website (http://www.nordicnames.de)	N/A
Rep. of Moldova	Name lists	Romanian name list (http://ro.wikipedia.org/wiki/List%C4%83_de_prenume_rom%C3%A2ne%C8%99ti)	N/A

Note: N/A = Not applicable.

Source: Compiled by Science-Metrix from various sources

Table 10 Validation results of the gender assignment procedure for author names in the WoS (2007–2013) by country for those covered in She Figures 2015

Country	% VN	AVN	AVNF	AVNM	% VPN	AVPN	AVPNF	AVPNM
Belgium	52%	91%	97%	97%	94%	95%	98%	98%
Bulgaria	78%	98%	98%	97%	93%	99%	99%	98%
Czech Republic	28%	98%	98%	97%	91%	97%	100%	95%
Denmark	15%	93%	93%	97%	74%	97%	98%	98%
Germany	8%	92%	93%	97%	80%	99%	97%	100%
Estonia	42%	90%	89%	92%	82%	93%	92%	94%
Ireland	26%	91%	92%	100%	74%	97%	97%	100%
Greece	23%	94%	96%	96%	83%	97%	100%	96%
Spain	36%	93%	96%	95%	94%	98%	99%	99%
France	9%	96%	98%	99%	87%	97%	98%	99%
Croatia	18%	93%	91%	96%	86%	98%	98%	98%
Italy	13%	92%	92%	97%	92%	99%	99%	99%
Cyprus	39%	94%	93%	96%	79%	99%	99%	99%
Latvia	69%	92%	94%	89%	79%	91%	95%	88%
Lithuania	72%	94%	90%	98%	86%	96%	92%	99%
Luxembourg	64%	96%	96%	100%	82%	96%	96%	100%
Hungary	33%	94%	96%	99%	62%	96%	98%	97%
Malta	79%	94%	93%	98%	92%	96%	95%	99%
Netherlands	19%	83%	80%	94%	73%	98%	94%	100%
Austria	22%	94%	95%	99%	82%	99%	98%	100%
Poland	9%	98%	99%	96%	95%	99%	100%	98%
Portugal	38%	80%	73%	94%	82%	98%	91%	100%
Romania	22%	75%	64%	89%	80%	98%	95%	99%
Slovenia	19%	95%	97%	93%	85%	97%	99%	96%
Slovakia	46%	98%	99%	97%	93%	94%	99%	91%
Finland	27%	94%	93%	96%	84%	96%	97%	95%
Sweden	13%	92%	92%	97%	80%	98%	98%	98%
United Kingdom	14%	92%	89%	95%	82%	97%	93%	99%
Iceland	53%	92%	92%	96%	67%	92%	90%	94%
Liechtenstein	71%	95%	92%	96%	90%	96%	81%	98%
Norway	23%	94%	93%	97%	76%	97%	96%	98%
Switzerland	14%	93%	97%	90%	76%	97%	97%	97%
Montenegro	66%	100%	99%	100%	88%	100%	99%	100%
FYR Macedonia	57%	96%	97%	94%	76%	97%	98%	95%
Albania	46%	96%	99%	94%	65%	97%	100%	96%
Serbia	15%	93%	93%	94%	90%	99%	99%	98%
Turkey	6%	92%	94%	98%	63%	93%	94%	98%
Bosnia and Herzegovina	47%	97%	98%	97%	83%	98%	98%	97%
Israel	28%	90%	91%	94%	77%	94%	92%	96%
Faroe Islands	62%	97%	100%	96%	69%	98%	100%	97%
Rep. of Moldova	71%	98%	100%	97%	76%	99%	100%	99%

Note: % VN = Percentage of validated names; AVN = Accuracy of assignments for validated names; AVNF = Accuracy for validated female names; AVNM = Accuracy for validated male names; % VPN = Percentage of validated paper/name combinations; AVPN = Accuracy of assignments for validated paper/name combinations; AVPNF = Accuracy for validated female paper/name combinations; and AVPNM = Accuracy for validated male paper/name combinations. Paper/name combinations represent the sum across all names in a country of all the papers on which these names appear. It is not based on a distinct count of the papers on which they appear. This is because the accuracy was measured for all author names (not just the reprint author) on a paper where the full given name is available. The goal was to account for common names in measuring the accuracy. Indeed, if a name is very common and it was assigned the wrong sex, this could greatly impact the indicators that were computed with this information. The AVPN column therefore provides a better assessment of the accuracy of the gender assignment procedure.

Source: Computed by Science-Metrix using WoS data (Thomson Reuters) and other sources (see Table 9)

Table 11 Continued

Country	Method	Source	Description of rule
Netherlands	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Dutch names (Wikipedia and Wiktionary): http://en.wiktionary.org/wiki/Category:Dutch_female_given_names ; http://en.wikipedia.org/wiki/Category:Dutch_masculine_given_names ; http://en.wiktionary.org/wiki/Category:Dutch_male_given_names ; http://en.wikipedia.org/wiki/Category:Dutch_feminine_given_names , Common names in Netherlands (http://web.archive.org/web/20080203090920/http://www.meertens.knaw.nl/voornamen/modern.html)	N/A
Austria	Name lists	Germany name list + US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names (http://www.beliebte-vornamen.de/jahrgang/j2013/top500-2013), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm)	N/A
Poland	Name lists	Polish name list - Polish Ministry of the Interior (https://www.msw.gov.pl/pl/aktualnosci/11689.Lena-i-Jakub-to-najpopularniejsze-imiona-mijajacego-roku.html?search=853621)	N/A
Portugal	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Portuguese names: http://en.wikipedia.org/wiki/Category:Portuguese_masculine_given_names ; http://en.wikipedia.org/wiki/Category:Portuguese_feminine_given_names	N/A
Romania	Name lists	Romanian names: (Wikipedia): http://ro.wikipedia.org/wiki/List%C4%83_de_prenume_rom%C3%A2ne%C8%99ti	N/A
Slovenia	Name lists	Slovenian name list from Statistics Slovenia (http://www.stat.si/ImenaRojstva/en/FirstNames/ExpandNames)	N/A
Slovakia	Name lists	Slovakian name list from the Ministry of Interior of the Slovak Republic (http://www.minv.sk/?tlacove-spravny-8&sprava=najcastejsim-menom-ktorym-rodicia-pomenovali-svoje-dieta-v-roku-2013-je-jakub)	N/A
Finland	Name lists + manual validation	Nordic name list (from validated Norway file), Nordic names Wiki (http://www.nordicnames.de), List of common finnish names - Finnish population register center (http://www.vrk.fi/default.aspx?id=279 and http://verkkopalvelu.vrk.fi/Nimipalvelu/default.asp?L=3)	N/A
Sweden	Name lists + manual validation	Nordic name list (from Norway validated file), Nordic Names website (http://www.nordicnames.de)	N/A
United Kingdom	Name lists	UK Census (http://www.ons.gov.uk/ons/publications/re-reference-tables.html?edition=tcn%3A77-318125)	N/A
Iceland	Naming rules + Name lists	Naming rules: Larivière, V. et al. Supplementary Information to: Nature 504, 211-213; 2013 and A Guide to Names and Naming Practices (2006) by the UK Government Name lists: US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Baby Names for 2012 (Iceland) http://www.nordicnames.de/wiki/icelandic_Statistics_of_2012 , Behind the name website: http://www.behindthename.com/ , and Research Gate: http://www.researchgate.net/	In order to assign a gender to a name, the naming rules (MALE last name ending: -son and -sson, and FEMALE last name endings: -dottir or -sdottir) for Iceland were applied, and then first names were checked using the Baby Names for 2012 (Iceland) website and Behind the name website; and if conflicting results were produced through either of these means, the profile of the researcher in questions was found (primarily using Research Gate) in order to verify the gender.
Liechtenstein	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm , Baby Name Wizard: http://www.babynamewizard.com/name-list/austrian-boys-names-most-popular-names-for-boys-in-austria- ; http://www.babynamewizard.com/name-list/austrian-girls-names-most-popular-names-for-girls-in-austria-	N/A
Norway	Name lists + manual validation	Statistics Norway (http://www.ssb.no/en/befolkning/statistikker/navn), US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Nordic Names website (http://www.nordicnames.de), validated name lists from Sweden and Finland	N/A
Switzerland	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), French names (wikipedia.org/wiki/Category:French_feminine_given_names and wikipedia.org/wiki/Category:French_masculine_given_names); Top German Names for Boys and Girls: http://german.about.com/library/blname_topDE.htm , Baby Name Wizard: http://www.babynamewizard.com/name-list/austrian-boys-names-most-popular-names-for-boys-in-austria- ; http://www.babynamewizard.com/name-list/austrian-girls-names-most-popular-names-for-girls-in-austria-	N/A
Montenegro	Name lists	Serbian name list (which was hand validated)	N/A
FYR Macedonia	Naming Rules	Naming rules Larivière, V. et al. Supplementary Information to: Nature 504, 211-213; 2013 and A Guide to Names and Naming Practices (2006) by the UK Government	N/A
Albania	Name lists + manual validation	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Albanian names: http://www.aboutnames.ch/albanian.htm	N/A
Serbia	Manual validation	Native speaker (due to incomplete coverage with available name lists)	N/A

Note: N/A = Not applicable.
 Source: Compiled by Science-Metrix from various sources

Table 11 Continued

Country	Method	Source	Description of rule
Turkey	Name lists	Turkey statistical office, list of 30 popular names (http://www.turkstat.gov.tr/PreHaberBultenleri.do?id=16054&as_sfid=AAAAAVpx%2FByBO74BwkrMTiV3HRmtviiry9%2BcVdUSKnsfFzidzaPhcFe%2FySouflHPjIuv2duHitdOxQ0Be5DydNQyhXMc490u6vHQFRBsMCFZySw3A%3D%3D&as_fid=pAGvS7Y00QKhwmYjlxor), Turkish names - Wikipedia (http://en.wikipedia.org/wiki/Category:Turkish_masculine_given_names , http://en.wikipedia.org/wiki/Category:Turkish_feminine_given_names , http://en.wikipedia.org/wiki/Category:Turkish_unisex_given_names)	N/A
Bosnia and Herzegovina	Name lists	Serbian name list (which was hand validated), Bosnian name list (http://en.wikipedia.org/wiki/Category:Bosnian_given_names), Bosnian census data (http://www.fzs.ba/Dem/ProcPrist/z100.pdf and http://www.fzs.ba/Dem/ProcPrist/m100.pdf)	N/A
Israel	Name lists	US Census (https://www.census.gov/topics/population/genealogy/data/1990_census/1990_census_namefiles.html), Jewish names: http://jewishnames.netzah.org/	N/A
Faroe Islands	Name lists + manual validation	Denmark name list (from validated Denmark file)+ Nordic Names website (http://www.nordicnames.de)	N/A
Rep. of Moldova	Name lists	Romanian name list (http://ro.wikipedia.org/wiki/List%C4%83_de_preume_rom%C3%A2ne%C8%99ti)	N/A

Note: N/A = Not applicable.

Source: Compiled by Science-Metrix from various sources

Table 12 Validation results of the gender assignment procedure for inventor names in PATSTAT (2002–2013) by country for those covered in She Figures 2015

Country	% VN	AVN	AVNF	AVNM	% VPN	AVPN	AVPNF	AVPNM
Belgium	58%	96%	98%	98%	87%	93%	98%	98%
Bulgaria	89%	99%	99%	99%	91%	99%	99%	99%
Czech Republic	49%	98%	100%	96%	85%	97%	100%	97%
Denmark	15%	95%	95%	98%	59%	98%	99%	98%
Germany	10%	91%	92%	99%	60%	98%	93%	100%
Estonia	64%	91%	91%	91%	73%	92%	90%	92%
Ireland	38%	92%	93%	98%	77%	96%	92%	100%
Greece	28%	93%	96%	94%	52%	98%	95%	98%
Spain	74%	92%	90%	95%	93%	98%	96%	98%
France	32%	91%	92%	98%	80%	96%	95%	99%
Croatia	57%	100%	100%	100%	83%	99%	100%	99%
Italy	18%	95%	95%	98%	72%	99%	97%	99%
Cyprus	82%	99%	100%	100%	88%	99%	100%	100%
Latvia	64%	97%	98%	98%	76%	95%	99%	94%
Lithuania	91%	99%	98%	99%	92%	99%	99%	100%
Luxembourg	51%	95%	89%	100%	67%	96%	91%	100%
Hungary	31%	93%	94%	98%	51%	99%	97%	100%
Malta	69%	100%	100%	100%	66%	100%	100%	100%
Netherlands	10%	91%	91%	98%	48%	95%	89%	97%
Austria	34%	100%	97%	100%	78%	95%	94%	98%
Poland	17%	97%	100%	94%	39%	98%	100%	97%
Portugal	13%	97%	100%	100%	27%	100%	100%	100%
Romania	46%	98%	100%	98%	71%	100%	100%	100%
Slovenia	53%	98%	100%	97%	85%	97%	100%	95%
Slovakia	70%	96%	97%	96%	89%	95%	99%	94%
Finland	26%	95%	93%	96%	53%	94%	95%	94%
Sweden	26%	94%	91%	97%	77%	98%	97%	98%
United Kingdom	9%	90%	92%	98%	61%	96%	94%	100%
Iceland	76%	93%	88%	97%	81%	96%	93%	97%
Liechtenstein	52%	96%	100%	98%	61%	98%	100%	99%
Norway	37%	95%	93%	97%	75%	97%	97%	97%
Switzerland	21%	92%	93%	99%	60%	97%	91%	100%
Montenegro	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
FYR Macedonia	100%	89%	100%	86%	100%	80%	100%	75%
Albania	100%	100%	N/A	100%	100%	100%	N/A	100%
Serbia	100%	100%	100%	100%	100%	100%	100%	100%
Turkey	39%	93%	91%	98%	74%	97%	96%	99%
Bosnia and Herzegovina	94%	100%	100%	100%	96%	100%	100%	100%
Israel	32%	90%	94%	92%	66%	93%	93%	95%
Faroe Islands	100%	100%	100%	100%	100%	100%	100%	100%
Rep. of Moldova	66%	95%	100%	94%	65%	96%	100%	96%

Note: % VN = Percentage of validated names; AVN = Accuracy of assignments for validated names; AVNF = Accuracy for validated female names; AVNM = Accuracy for validated male names; % VPN = Percentage of validated paper/name combinations; AVPN = Accuracy of assignments for validated patent application/name combinations; AVPNF = Accuracy for validated female patent application/name combinations; and AVPNM = Accuracy for validated male patent application/name combinations. Patent application/name combinations represent the sum across all names in a country of all the patent applications on which these names appear. It is not based on a distinct count of the patent applications on which they appear. This is because the accuracy was measured for all inventor names on a patent application where the full given name is available. The goal was to account for common names in measuring the accuracy. Indeed, if a name is very common and it was assigned the wrong sex, this could greatly impact the indicators that were computed with this information. The AVPN column therefore provides a better assessment of the accuracy of the gender assignment procedure.

Source: Computed by Science-Metrix using WoS data (Thomson Reuters) and other sources (see Table 11)

Appendix 4: Gender dimension in research content

In order to assess the current status of the gender dimension in research materials, a method to identify peer-reviewed research documents (i.e. mostly journal articles, reviews and notes) in which a gender dimension is present has been developed and is described below. This work was realised in collaboration with the Commission officials and statistical correspondents/experts of the She Figures 2015 project, as well as with members of the Helsinki group. These experts provided fruitful comments and advice throughout the creation process. In a similar manner, some of the experts expressed concerns about the scope of the topics that should be included in the gender dimension. To address these concerns, H2020 documentation was consulted and the work presented in this document tries to reflect, to the greatest extent possible, the definitions established in those documents.

First, let us consider how the gender dimension in research is defined. As stated in the glossary of the document *Gender Equality in Horizon 2020*³¹

Sex refers to biological qualities characteristic of women [females] and men [males] in terms of reproductive organs and functions based on chromosomal complement and physiology. As such, sex is globally understood as the classification of living things as male and female, and intersexed.

Gender – a socio-cultural process – refers to cultural and social attitudes that together shape and sanction ‘feminine’ and ‘masculine’ behaviours, products, technologies, environments, and knowledge. [...]

Sex/gender analysis: is an umbrella term for the entire research cycle that includes the integration of sex/gender issues from the setting of research priorities through developing methodologies, gathering and analysing data to evaluating and reporting results and transferring them to markets.

Gender dimension in research: is a concept regrouping the various elements concerning biological characteristics and social/cultural factors of both women and men into the development of research policies, programmes and projects.

Moreover, extracts from section IV of the same document in which the gender dimension is discussed state:

Gender dimension: ‘*the gender dimension is explicitly integrated into several topics across all the sections of the Work Programme*’ (...) ‘*a topic is considered gender relevant when it and/or its findings affect individuals of groups of persons. In these cases, gender issues should be integrated at various stages of the action and when relevant, specific studies can be included. These topics are flagged to ease access for applicants. This should not however prevent applicants to a non-flagged topic from including a gender dimension in their proposal if they find it relevant*’. [...]

Gender dimension: for flagged topics, evaluators shall check *how sex and/or gender analysis is taken into account in the project’s content* (as requested in the application form).

Hence, based on the excerpts and definitions provided above, the gender dimension in research content includes both the concepts of sex and gender as well as the concept of sex/gender analysis in humans. As such, in addition to research outputs focused on a well-defined gender topic (e.g. feminism, gender pay gap, gender equality, LGBT), research content in which a distinction or a comparison is made between men and women either in the title, abstract, or author keywords of scientific publications were deemed relevant.

³¹ European Commission. (2014). *Gender equality in Horizon 2020*. Retrieved from http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/gender/h2020-hi-guide-gender_en.pdf.

Exclusion: As discussed with Commission officials and the statistical correspondents/experts in the steering group meeting 2, research outputs studying the animal kingdom (e.g. feminisation of fish populations) as well as other non-human biological entities, such as plants, were not to be considered for the construction of the dataset on the gender dimension in research content. Moreover, scientific papers investigating medical conditions specific to one gender (e.g. menopause, erectile dysfunction) were also not to be considered pertinent to the dataset as the inclusion of those would result in the inclusion of a very large portion of scientific publications in the medical fields.

The methodology employed to construct the query for the retrieval of scientific publications in which a gender dimension is addressed is a two-step process that comprises

1. the identification of an initial set of highly relevant papers (i.e. the 'seed'); and
2. the extraction of a gender-specific terminology through an analysis of the textual content present in the seed; these terms are then used towards expanding the seed to obtain the final dataset.

Creation of the seed dataset

Access to bibliographic data on peer-reviewed scientific publications is essential for producing bibliometric data on specific subjects such as the Gender Dimension in Research Content (GDRC). The Web of Science, produced by Thomson Reuters, was selected for this purpose. It includes three databases: the Science Citation Index Expanded (SCI Expanded), the Social Sciences Citation Index (SSCI), and the Arts & Humanities Citation Index (A&HCI). It indexes some 12,000 refereed journals (i.e. articles that are peer reviewed prior to publication) and covers all fields of science as defined in the Frascati Manual.

The Web of Science (WoS) was chosen because it provides cited references for each document it includes (e.g. articles or chapters published in a journal or book series), allowing for internal coverage monitoring of the database and analysis of scientific impact based on citations and impact factors. For instance, Thomson Reuters' monitoring procedure ensures that the most important peer-reviewed journals in their respective fields are indexed. As recently shown, '50% of all citations generated by this collection came from only 300 of the journals. In addition, these 300 top journals produced 30% of all articles published by the total collection.'³² Because science is not static, the list of key international journals is changing continuously. For this reason, Thomson Reuters is adjusting the coverage of the WoS on a regular basis to reflect the dynamics of the science.

Furthermore, the WoS includes names of all authors and their institutional affiliations, which allows collaboration rates among various entities (e.g. countries, institutions, and researchers) to be analysed. It also indexes the links between authors and their addresses, a key feature for aggregating gender data by country.

In producing this indicator, as well as other indicators based on the WoS (see preparatory paper), only three document types published in refereed scientific journals were retained – articles, notes, and reviews. This was because all have been through the peer-review process prior to being accepted for publication. The peer-review process ensures that the research is of good quality and

³² Testa, J. (2012). The Thomson Reuters journal selection process [webpage]. Retrieved from <http://wokinfo.com/essays/journal-selection-process/>.

constitutes an original contribution to scientific knowledge. The terms ‘papers’ and ‘publications/articles’ are used throughout this report in referring to these three document types.

Note that Science-Metrix hosts an in house version of the WoS in the form of an SQL relational database. This has allowed Science-Metrix to carefully condition the database for the purpose of producing large-scale comparative scientometric analyses. Bibliometrics analysts at Science-Metrix have performed a large number of bibliometric projects with it and thus have in-depth knowledge of its respective strengths and limitations.

The Science-Metrix in-house version of the WoS indexes over 25 million publications, which are classified into six large domains:³³ Applied Sciences, Arts & Humanities, Economic & Social Sciences, General, Health Sciences and Natural Sciences. These are divided into 22 fields, which themselves are further split into 176 subfields. The classification³⁴ is mutually exclusive (i.e. each article is classified into one and only one set of domain, field and subfield) and is based on

- the classification used by the NSF in the Science & Engineering Indicators, originally designed by CHI Research and now maintained by the Patent Board;
- the journal classification developed by the Institute for Scientific Information (ISI), which is now part of Thomson Reuters; and
- the Excellence in Research for Australia’s Ranked Journal List by the Australia Research Council.

The first step in identifying scientific publications relevant to the gender dimension in research content was to identify the field(s) or subfield(s) related to gender research. One subfield was found to be directly related to the subject: the Gender Studies subfield under the field of Social Sciences. Note that it is understood that the coverage of this subfield is more limited than the set of relevant literature to be extracted in this study; this document later shows how relevant publications falling outside the core set of gender studies were retrieved. The subfield of Gender Studies contains 6,023 publications (for the period 2002–2013) all discussing a gender-related topic. A validation check (title and abstract reading) of a randomly selected sample of 100 articles has enabled us to confirm the pertinence of the publications in this subfield. These publications constitute the base of the seed dataset.

In the subsequent step, a search for journal names containing the term *gender* was executed in the WoS. The results can be found in Table 13. The scope of each journal was evaluated either by accessing their website or, when that was not possible, by examining the publications contained in these journals. Based on the scope of the journals, a verdict determining if the publications published in each journal were pertinent to the gender dimension in research content was assigned. All the publications contained in the relevant journals (2,150 for 2002–2013) were then added to the seed dataset.

³³ Note that the subfields in Science-Metrix’s classification were later reorganised to allow for producing data at the Field of Science (FOS) level using the Frascati Manual definitions.

³⁴ This classification was developed within the context of the following contract performed for the European Commission: Analysis and Regular Update of Bibliometric Indicators, RTD 2009_S_158-229751. Archambault É., Beauchesne, O., and Caruso J. (2011). Towards a multilingual, comprehensive and open scientific journal ontology, in B. Noyons, P. Ngulube, and J. Leta (Eds.), *Proceedings of the 13th International Conference of the International Society for Scientometrics and Informetrics (ISSI)* Durban, South Africa, pp 66–77.

Table 13 WoS journals containing the term *gender* in their name

Journal name	Papers (2002–2013)	Verdict
FOCUS ON GENDER: PARENT AND CHILD CONTRIBUTIONS TO THE SOCIALIZATION OF EMOTIONAL COMPETENCE	7	ok
PEDIATRIC GENDER ASSIGNMENT: A CRITICAL REAPPRAISAL	19	ok
GENDER AND LANGUAGE	54	ok
POLITICS & GENDER	92	ok
JOURNAL OF WOMENS HEALTH & GENDER-BASED MEDICINE	121	ok
INDIAN JOURNAL OF GENDER STUDIES	147	ok
JOURNAL OF GENDER STUDIES	234	ok
GENDER MEDICINE	266	ok
GENDER PLACE AND CULTURE	333	NO (also on race, ethnicity, sexuality, etc.)
GENDER WORK AND ORGANIZATION	346	ok
GENDER & SOCIETY	394	ok
GENDER AND EDUCATION	468	ok

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

Next, the journals that published articles classified in the subfield of Gender Studies (and which were different from those stated above) were retrieved. These journals are listed in Table 14.

Table 14 WoS journals containing publications classified under the subfield Gender Studies

Journal name	Papers (2002–2013)	Verdict
FEMINISTISCHE STUDIEN	101	ok
FEMINIST THEORY	101	ok
NOUVELLES QUESTIONS FEMINISTES	141	ok
INTERNATIONAL FEMINIST JOURNAL OF POLITICS	163	ok
ASIAN JOURNAL OF WOMENS STUDIES	163	ok
HYPATIA-A JOURNAL OF FEMINIST PHILOSOPHY	177	ok
FRONTIERS-A JOURNAL OF WOMEN STUDIES	179	ok
FEMINIST STUDIES	213	ok
MEN AND MASCULINITIES	215	ok
FEMINIST REVIEW	230	ok
SEXUALITIES	232	ok
AUSTRALIAN FEMINIST STUDIES	237	ok
EUROPEAN JOURNAL OF WOMENS STUDIES	256	ok
FEMINISM & PSYCHOLOGY	408	ok
SIGNS	504	ok
JOURNAL OF SEX RESEARCH	510	ok
WOMENS STUDIES INTERNATIONAL FORUM	604	ok

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

The scope of these journals was assessed in the same manner as previously described in order to evaluate the relevancy of their content. The 3,700 articles published between 2002 and 2013 in appropriate journals were subsequently added to the seed dataset.

The last phase in building the seed dataset involved the use of Medline's controlled vocabulary (MeSH terms) to identify gender-related scientific articles indexed in the WoS and in Medline (a bibliographic database of life sciences and biomedical information compiled by the United States National Library of Medicine (NLM)). Medical subject headings (MeSH) are pre-defined terms created by the NLM that cover all aspects of medicine and health care. They are assigned to each publication upon their inclusion in Medline by subject experts. A search for MeSH terms containing every variant of *gender*, *femin**, *women* and *men* allowed Science-Metrix to identify 18 possible relevant MeSH terms (see Table 15). A MeSH term is classified as a 'major' term for a given publication if it represents a major focus of the study. To enable the extraction of publications using MeSH terms in the WoS, all publications in Medline were matched to their corresponding entry in the WoS using an algorithm developed at Science-Metrix. The pertinence of the papers fetched by each MeSH term was evaluated by selecting a random sample of papers in which the title and abstract were analysed. The only MeSH terms that were not satisfactory were Feminization and Pregnant Women.

Table 15 MeSH terms, the number of papers associated with them in the WoS and the verdict for insertion of the associated publications in the seed dataset

Mesh descriptor term	Papers (2002-2013)	Verdict
Feminine Hygiene Products	9	NO
Sexual and Gender Disorders	17	ok
Dentists, Women	29	ok
Femininity	51	ok
Feminization	68	NO, medical explanation of feminization (most of the time not in human)
Transgendered Persons	113	ok
Feminism	207	ok
Physicians, Women	323	ok
Men's health	390	ok
Women's Rights	403	ok
Women's Health Services	573	ok
Women, Working	627	ok
Pregnant Women	694	NO, as decided in SGM2
Battered Women	991	ok
Women	1,680	ok
Gender Identity	2,045	ok
Women's Health	4,480	ok

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

The articles captured by the selected MeSH terms were also added to the seed dataset. In the end, the seed dataset comprised 17,900 distinct publications. A publication may have been retrieved multiple times by the different techniques (subfield, journals or MeSH terms) but was only counted once.

Creation of the final dataset

In this phase, the seed dataset was expanded using a query searching for gender-related terminology in the title, abstract and author keywords of the publications indexed in the WoS. Highly relevant terms were identified using the term frequency–inverse document frequency (TF-IDF) statistic. The TF-IDF statistic determines the importance of a given expression (a term or set of terms) in a specific set of documents (i.e. the seed dataset) relative to a reference collection of documents (the WoS). The relevance of an expression increases proportionally to the number of times it appears in the seed dataset but is offset by the frequency of the word in the reference collection. This operation increases the detection of rare and specific expressions. Two lists of expressions have been tested using the TF-IDF weight. The first list consists of the European Institute for Gender Equality (EIGE) draft thesaurus on gender equality terms (which contains more than 600 terms) sent to Science-Metrix by EIGE. The second list consists of a set of about 10 million noun phrases (i.e. scientific expressions) extracted from the titles, abstracts and author keywords of publications in the WoS.

Once the TF-IDF weight has been computed for each expression, the terms are placed in descending order of their weight (i.e. relevance to the seed dataset). Table 16 shows the top 30 expressions for both lists and their respective TF-IDF weight. The top keywords from both lists feature some similarities as shown by the bolded keywords in Table 16. Other EIGE keywords appear further down (lower weight) in the WoS list and vice versa (data not shown).

Table 16 TF-IDF weight of the top items appearing in EIGE draft thesaurus on gender equality terms and of the noun phrases extracted from WoS papers

EIGE keywords		WoS extracted noun phrases	
Keyword	Weight	Keyword	Weight
Violence against women	14.36	Sex guilt	16.35
Masculinities	14.35	Gender identity disorder	16.30
Hegemonic masculinity	14.11	Gid	15.88
Feminism	14.07	Ipv	15.76
Femininities	13.77	Gender nonconformity	15.63
Gender role	13.63	Gender identity	15.57
Intimate partner violence	13.62	woman movement	15.50
Gender equality	13.62	Hegemonic masculinity	15.49
Heteronormativity	13.57	Abuse woman	15.49
women empowerment	13.47	Batter woman	15.48
Transgender	13.44	Identity disorder	15.47
Women right	13.42	Woman disability	15.42
Gender relation	13.41	Gender mainstream	15.41
Patriarchy	13.32	Woman health initiative	15.41
Gender mainstreaming	13.29	Transgender	15.36
Transsexual	13.27	Woman physician	15.34
Men health	13.26	Violence against woman	15.34
Domestic violence	13.25	Sexualisation	15.27
Sexualities	13.24	Transsexual	15.26
Gender equity	13.23	Wisewoman	15.25
Gender roles	13.23	Masculinity	15.21
Gender studies	13.12	Partner violence	15.16
Reproductive right	13.04	Woman surgeon	15.15
Queer	12.96	Against woman	15.14
Sexual violence	12.94	Woman study	15.11
Homophobia	12.78	Pokunyokwan	15.04
Gender stereotype	12.72	Feminism	15.04
Feminist studies	12.69	Fa afafine	14.97
Motherhood	12.68	Femininity	14.96
Physical violence	12.64	Feminist	14.95
Sexualities	12.62	Intimate partner violence	14.95

Note: Only the 30 expressions with the highest TF-IDF weight are shown. The items in bold are common to both lists.

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

In total, a TF-IDF weight has been calculated for 640 EIGE keywords (gender specific keywords) and for more than 150,000 scientific expressions found in the WoS (a weight was calculated only for those expressions that were present in the seed dataset, i.e. about 150,000 out of about 10 million in the whole of WoS). Of the 150,000 expressions, only a small fraction was really relevant to GDRC. Nevertheless, the wide range of tested expressions made it possible to detect unsuspected search terms.

The most relevant keywords (i.e. those with the highest TF-IDF weight, some of which are shown in Table 17) are the most promising expressions to identify gender-related publications. As one goes down the list, there is a threshold (not necessarily well defined) at which point the expressions are not specific enough to be used in the search query aimed at expanding the seed dataset (see Table 17); examples of such expressions include the keywords highlighted in red.

Table 17 Items with low TF-IDF weight that will not be included in the keyword query

EIGE keywords		WoS extracted noun phrases	
Keyword	Weight	Keyword	Weight
Transsexual	12.59	Intimate partner	14.89
Demographic change	7.62	Risk taking	10.22
Witness	7.62	Awakening response	10.22
Spending	7.60	Problematise	10.22
Career planning	7.59	Source care	10.22
Migrant worker	7.54	Meaning practice	10.22
Families	7.52	Low libido	10.22
Risk group	7.52	Qualitative method	10.22

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

Each of the promising expressions was tested manually by retrieving the publications in which a given expression appeared in their titles, abstracts or author keywords. The rationale is that when a specific expression is found in the title, abstract or author keywords of a publication, it usually defines an aspect that is, at least partly, the subject of the reported study. Note that in searching publications in which an expression appears, the search term was modified using wildcards so that all wording starting with the same root would be retrieved. For example, for the word femininities the search term was transformed into femininit* so that publications containing femininity OR femininities were both retrieved.

For each search term, the title and abstract of a random sample of the retrieved publications were read and their pertinence to GDRC was assessed by a seasoned analyst. When the majority of the retrieved publications were judged relevant, the search expression was included in the keyword-based query. In some cases, the search terms had to be combined with other search conditions to filter out undesired publications. Either the exclusion of publications classified in a specific field or subfield was necessary or the search had to be tailored to specific fields and subfields. For example, for the keyword *gender role*, numerous articles classified in the field of biology were retrieved by the query. However, most of them concerned the role of a male or female animal when mating or searching for food. Since these subjects were not in the scope of the study, papers from the Biology field were filtered out when using the *gender role* keyword.

Given some of the concerns expressed by the consulted experts that the approach using the TF-IDF weight could induce a bias towards the seed dataset (Gender Studies subfield + specialist journals + MeSH terms), the selection of relevant search expressions was re-iterated using the TF-IDF weight. In this round, the set of publications forming the corpus upon which to measure the relevance of the search expressions was made up of all publications including the *gender* term alone in their titles, abstracts or author keywords (instead of the seed dataset). This ensured that the keywords returned by the second TF-IDF procedure were not only keywords found in the seed dataset, but also expressions that may co-occur with the *gender* term in any of the scientific publications indexed in the WoS. The operations described above were then performed on the promising keywords in order to identify those to include in the final query. By the end of the process around 220 search

expressions were included in the query. For a complete list of these expressions, along with the restrictions applied to them, please refer to Appendix 5.

The last step in the query was to delete articles about the animal kingdom that were not filtered by the field/subfield exclusion criteria described above. Publications containing in their title or abstract the names of animals mostly used in research were not included in the dataset. The remaining 205,000 publications returned by the query were then added to the 17,000 publications found in the seed dataset. Thus, the final SGDRC dataset consisted of some 212,600 distinct publications including a gender dimension in their research content. Thus, the keyword-based query substantially increased the size of the seed dataset, mostly by expanding its coverage of fields that were not at all or only slightly present in the seed dataset (i.e. beyond the core realm of gender studies). It is important to note that the specificity of the keywords employed in the query directly affect the scope and the range of the final dataset. Using only highly specific keywords would result in the retrieval of a small number of very specialised papers while missing other important papers (false negatives). Conversely, using overly broad and unspecific keywords would concomitantly capture a large number of important though undesirable publications (false positives). The selection of the expressions to be included in the query has been carefully thought out and tested to keep the highest degree of precision (low percentage of false positives) in regard to the definition of GDRC, as per the H2020 documentation.

Moreover, consulted experts have provided insightful feedback that has been included, in as much as possible, throughout the creation process. For example, one of the statistical correspondents provided information on a database developed by Charité Berlin that indexes sex- and gender-related literature in the biomedical field as a possible source of gender-specific keywords. Their methodology to identify relevant literature (which is described in a paper by Oertelt-Prigione et al.³⁵) also makes use of keyword-based query terms followed by manual validation. Only 10 highly relevant keywords were used in their study – namely, ‘sex difference(s)’, ‘sexual difference(s)’, ‘gender difference(s)’, ‘sexually dimorph(ic)’, ‘sexual dimorphism’, ‘sex dependent’, ‘sex based’, ‘gender based’, and ‘gender dependent’. Although the exclusion and acceptance criteria are not exactly the same as in this work, it was interesting to see that most of the terms were already included in the terms previously selected except for sexual dimorphism and sexually dimorph*. Most of the articles that were using these terms in their titles or abstracts were essentially about animals. In the case of Charité Berlin, this is not a problem since their database was intended to cover animal studies, as opposed to the present study that focuses on humans. These terms were therefore not kept in this study. The exclusion of such terms, even if they retrieve a few relevant papers out of their total, is not critical since the approach implemented in this study includes a lot of redundancy; that is, the inclusion of a wide variety of search terms that often co-occur in publications. As such, if a keyword is omitted, a relevant publication including this keyword will likely be captured by one of the selected search expressions.

Content of the GDRC dataset

The final dataset contains 212,600 publications distributed across all the 22 fields and 176 subfields found in the entire database. The distribution by field is shown in Table 18. One can see that the dataset is dominated by the field of Clinical Medicine. Some statistical correspondents

³⁵ Oertelt-Prigione, S., Parol, R., Krohn, S., Preissner, R., and Regitz-Zagrosek, V. (2010). Analysis of sex and gender-specific research reveals a common increase in publications and marked differences between disciplines. *BMC Medicine*, 8(70), doi: 10.1186/1741-7015-8-70

expressed concerns about the fact that Clinical Medicine may be overrepresented. However, this is not unexpected since it is common practice in medicine to design experiments, and/or present results, for distinct sex groups. Since the H2020 definition of GDRC introduced earlier in this Appendix covers the analysis of sex disaggregated data, all those papers shall be included. It is important to note that Clinical Medicine is a very large field. In fact, the 91,000 articles from this field that fall in the GDRC dataset represent only 2.9% of the whole field in the WoS. Yet to prevent a skew towards the Health Sciences in analysing and interpreting the new indicator on GDRC, the indicator will be presented globally (all fields combined) as well as for each of the six main Fields of Science (FOS) as defined in the Frascati Manual. Note that the approach used to produce the new bibliometric indicators by FOS to be included in the She Figures 2015 will be provided in another document.

Table 18 Distribution of publications in the GDRC dataset across fields of science

Field	Papers	Share of field	Share of dataset
Total	212,673	N/A	100.0%
Clinical Medicine	91,162	2.9%	42.9%
Public Health & Health Services	33,546	10.5%	15.8%
Social Sciences	28,674	8.8%	13.5%
Psychology & Cognitive Sciences	20,961	7.9%	9.9%
Biomedical Research	10,685	0.9%	5.0%
Economics & Business	6,371	2.6%	3.0%
Communication & Textual Studies	5,363	5.2%	2.5%
Historical Studies	4,811	5.1%	2.3%
Philosophy & Theology	1,630	2.6%	0.8%
General Science & Technology	1,169	0.6%	0.5%
Information & Communication Technologies	1,108	0.2%	0.5%
General Arts, Humanities & Social Sciences	922	6.5%	0.4%
Enabling & Strategic Technologies	874	0.1%	0.4%
Visual & Performing Arts	715	3.4%	0.3%
Earth & Environmental Sciences	615	0.1%	0.3%
Engineering	537	0.1%	0.3%
Built Environment & Design	411	0.6%	0.2%
Biology	402	0.1%	0.2%
Chemistry	382	0.0%	0.2%
Mathematics & Statistics	267	0.1%	0.1%
Physics & Astronomy	234	0.0%	0.1%
Agriculture, Fisheries & Forestry	129	0.0%	0.1%
UNKNOWN	1,705	2.1%	0.8%

Source: Compiled by Science-Matrix from WoS data (Thomson Reuters)

The fields that are present in high proportion relative to their size (column *Share of field* in Table) are mostly from the Social Sciences and Humanities domain. Apart from Clinical Medicine and Biomedical Research, the most represented fields (column *Share of dataset* in Table 18) are Public Health & Health Services, Social Sciences, and Psychology & Cognitive Sciences, which is where one would expect most of the gender-related publications to be published. It is also interesting to note that the keyword query did catch a small number of articles from fields in which the presence of gender-related topics is less likely, such as Information & Communication Technologies, Earth & Environmental Sciences or Chemistry, just to name a few.

To get a sense of the topics covered by the GDRC dataset, a keyword map based on the co-occurrence pattern of the searched expressions has been produced (Figure 3). Keywords are placed closer to one another as their co-occurrence in the abstract or the title of publications increases.

The size of the labels and the bubbles is proportional to the occurrence of a given keyword in the GDRC dataset. The search expressions are clustered based on their co-occurrence pattern and the various clusters identified are represented by the colour of the bubbles on the map. The five identified clusters cover topics revolving mainly around gender equality (red), gender associated violence (indigo), sexuality and sex health (green), sex/gender identity (fuchsia), and sex differences (yellow).

Recall and precision

The quality of the dataset was assessed by looking at two parameters often used in information retrieval. These parameters are the recall (i.e. the percentage of false negatives, or relevant papers that were not retrieved, must be reduced) and precision (i.e. the percentage of false positives, or irrelevant papers that were accidentally retrieved) of a dataset. When building a dataset on a specific research area, these parameters must be optimised to improve the resulting dataset. However, the gains for one parameter are often offset by the losses for another. Thus, a balance must be achieved between recall and precision.

The recall of the seed dataset was measured by taking the intersection of the publications in the seed dataset with those retrieved by the keyword-based query over the size of the seed dataset (i.e. the percentage of the seed retrieved with the keyword-based query). This facilitates assessing how well the selected search expressions capture the core literature related to Gender Studies and gender-related MeSH terms. The recall was also assessed at the journal level; it was measured as the percentage of articles captured by the keyword query that are present in the journals that were included in the seed dataset. The results of the various recall measures are presented in Table 19.

Table 19 Recall of the seed dataset (i.e. gender studies subfield, specialist journals and MeSH terms) and of each of the specialist journals using the keyword-based query

Dataset/Journal	Articles from KW query	Total articles	Recall
Seed dataset	10,453	17,885	58.4%
GENDER & SOCIETY	375	394	95.2%
MEN AND MASCULINITIES	202	215	94.0%
GENDER AND LANGUAGE	50	54	92.6%
JOURNAL OF GENDER STUDIES	215	234	91.9%
EUROPEAN JOURNAL OF WOMENS STUDIES	235	256	91.8%
GENDER WORK AND ORGANIZATION	313	346	90.5%
SEXUALITIES	208	232	89.7%
GENDER AND EDUCATION	405	468	86.5%
NOUVELLES QUESTIONS FEMINISTES	116	141	82.3%
ASIAN JOURNAL OF WOMENS STUDIES	132	163	81.0%
WOMENS STUDIES INTERNATIONAL FORUM	484	604	80.1%
FEMINIST REVIEW	180	230	78.3%
FEMINIST THEORY	79	101	78.2%
FEMINISTISCHE STUDIEN	79	101	78.2%
INTERNATIONAL FEMINIST JOURNAL OF POLITICS	127	163	77.9%
POLITICS & GENDER	66	92	71.7%
FOCUS ON GENDER: PARENT AND CHILD CONTRIBUTIONS TO THE	5	7	71.4%
INDIAN JOURNAL OF GENDER STUDIES	102	147	69.4%
FEMINISM & PSYCHOLOGY	266	408	65.2%
GENDER MEDICINE	171	266	64.3%
HYPATIA-A JOURNAL OF FEMINIST PHILOSOPHY	110	177	62.1%
JOURNAL OF SEX RESEARCH	315	510	61.8%
SIGNS	241	504	47.8%
FEMINIST STUDIES	97	213	45.5%
AUSTRALIAN FEMINIST STUDIES	106	237	44.7%
FRONTIERS-A JOURNAL OF WOMEN STUDIES	52	179	29.1%
JOURNAL OF WOMENS HEALTH & GENDER-BASED MEDICINE	34	121	28.1%
PEDIATRIC GENDER ASSIGNMENT: A CRITICAL REAPPRAISAL	5	19	26.3%

Source: Compiled by Science-Matrix from WoS data (Thomson Reuters)

The recall of the seed dataset is near 60%, which is around what is expected for a dataset in the Social Sciences and Humanities (SSH). Indeed, it is difficult to achieve higher recall for subjects related to the SSH, since the expressions used in this domain are usually less specific to a particular

area of research than the expressions used in the Natural Sciences and Engineering or Health Sciences domains. At this point, the process of adding more keywords (likely more generic terms) to increase the recall would be detrimental to the dataset since it would lead to the concomitant retrieval of false positives, thereby reducing the precision of the dataset. The recall of the specialist journals is also satisfactory, with the majority of them having a recall above the 60% mark.

Finally, the precision (i.e. one minus the percentage of false positives, irrelevant papers that were accidentally retrieved) of the dataset was assessed by taking a random sample of 100 articles. The titles and abstracts of those publications were examined and if the subject(s) of a paper did not relate to the above definition of GDRC, it was considered as a false positive. The precision was measured for the dataset as a whole as well as for the fields in which GDRC papers are less likely to be found, such as in Information & Communication Technologies, Agriculture, Fisheries & Forestry, Engineering, and Earth & Environmental Science. Table 20 presents the findings of this assessment.

Table 20 Precision of the GDRC dataset as a whole and for some fields in which relevant papers are less likely

Field	Precision	Sample size
Whole dataset	97%	100
Agriculture, Fisheries & Forestry	70%	50
Information & Communication Technologies	78%	50
Earth & Environmental Sciences	74%	50
Engineering	84%	50
Mathematics & Statistics	60%	50
Physics & Astronomy	80%	50

Source: Compiled by Science-Metrix from WoS data (Thomson Reuters)

The dataset as a whole has an excellent precision of 97%. The precision decreases a little for unusual fields, but since they are almost all above 70% and they do not represent a large proportion of the final dataset, this should not be a cause for concern.

Examples of author keywords, titles and abstracts (delineated by the //xxx// separator) from relevant articles classified under the unexpected fields are shown below with the search terms they contain highlighted in grey.

Field: Earth & Environmental Sciences; subfield: Environmental Sciences

Climate change //xxx// Cooking //xxx// Gender equality //xxx// Global health //xxx// Improved stoves //xxx// Indoor air pollution //xxx// Neglected diseases //xxx// Resources-sociology //xxx// Respiratory diseases //xxx// Social norms //xxx// Soot
A smoke-free kitchen: initiating community based co-production for cleaner cooking and cuts in carbon emissions
Cooking over open fire with solid fuels results in incomplete combustion and indoor air pollution (IAP) causing respiratory and other diseases leading to nearly two million premature deaths per year. In urban areas, IAP interacts with outdoor pollutants in toxic chemical mixtures affecting also other citizens and damaging regional air quality in terms of 'brown clouds'. Deaths result mainly in women, children and infants, who are directly exposed to smoke in unventilated kitchens, thus reflecting differentiated and unequal impacts across population groups. Despite the heavy health burden and discomfort, IAP has only recently been recognised as associated with neglected diseases. In search of synergies between adaptation and mitigation, we seek gender sensitive social innovations to halt smoke, soot and early death while reducing deforestation and carbon emissions. Using transition arenas as a participatory method for experiments and social learning we engaged with local entrepreneurs and peasant farmers in subSaharan Africa to initiate co-production of efficient flue-piped stoves that save energy, labour and lives. Findings indicate that successful design, production and adoption of improved cooking stoves is possible, but the structural challenges of poverty, inequality and distrust may inhibit further diffusion and more profound processes of social learning. Insights from local studies must therefore be contextualised into broader understandings, as attempted here, while local adoption must be combined with wider initiatives and government policies into complex micro-to-macro solutions that provide forceful effects against IAP and its drivers. (C) 2012 Elsevier Ltd. All rights reserved.

Field: Physics & Astronomy; subfield: General Physics

Gender differences in conceptual understanding of Newtonian mechanics: a UK cross-institution comparison. We present the results of a combined study from three UK universities where we investigate the existence and persistence of a performance gender gap in conceptual understanding of Newtonian mechanics. Using the Force Concept Inventory, we find that students at all three universities exhibit a statistically significant gender gap, with males outperforming females. This gap is narrowed but not eliminated after instruction, using a variety of instructional approaches. Furthermore, we find that before instruction the quartile with the lowest performance on the diagnostic instrument comprises a disproportionately high fraction (similar to 50%) of the total female cohort. The majority of these students remain in the lowest-performing quartile post-instruction. Analysis of responses to individual items shows that male students outperform female students on practically all items on the instrument. Comparing the performance of the same group of students on end-of-course examinations, we find no statistically significant gender gaps.

Field: Built Environment & Design; subfield: Architecture

The Woman/Architect Distinction. The complicated links between individual life stories and theories of gender are central to feminist discussion. Since the end of the 1970s, feminist discourse has debated the tensions between feminism's collective category, woman, and the diversity of women. Feminist theory has worked to undo the monolithic category of Woman. This paper argues that research on women in architectural practice should draw on this long-standing debate, particularly when researchers use life-story interview texts as the corner-stone for gender studies. Researchers who use gender as a category of analysis must negotiate the complexity of lived subjectivity narrated in the interview testimonial. The apparent refusal of women interviewees to self-identify as women architects or explain career trajectories primarily through the prism of gender invites us to produce more subtle theories of identity, lived subjectivity and mechanisms of gender identification in feminist architectural research. The term, woman architect, invites us to think about the hyphenated nature of identity.

Field: Earth & Environmental Sciences; subfield: Meteorology & Atmospheric Sciences

gender mainstreaming //xxx// integrated water resource management //xxx// water sector reform //xxx// gender policies //xxx// gender equity //xxx// strategic gender needs Mainstreaming gender in integrated water resources management: the case of Zimbabwe. Zimbabwe embarked on a water sector reform programme in 1995. Two goals of the water reform were to broaden women's access to water and to enhance their participation in water management. This paper analyses how gender was addressed at the national and institutional levels and in the water reform process, paying particular attention on how strategic gender needs were addressed in the process and the resultant policies and Acts. The results show that although the government of Zimbabwe has made considerable progress in mainstreaming gender at the ministerial level, departments which are involved in the actual implementation of water programmes do not have clear gender policies. Therefore although gender equity was one of the main goals of the water reform, most poor women and men were not involved in the consultations. Consequently neither the new Water Act nor the Zimbabwe National Water Authority (ZINWA) Act addresses gender in explicit terms. Strategic gender needs are not addressed at all. It is recommended that all institutions in the water sector, including NGOs, should have clear gender policies, include a gender perspective in their organisation culture and practices and address strategic gender needs through training, education and supporting productive use of water. (C) 2003 Published by Elsevier Ltd.

Field: Engineering; subfield: Geological & Geomatics Engineering

education & training //xxx// European Union //xxx// local government //xxx// social impact Gender mainstreaming within local planning authorities. This paper discusses the extent to which EU-derived gender mainstreaming (GM) requirements are being adopted with reference to Royal Town Planning Institute research on the situation in UK local planning authorities (LPAs). Firstly, the problem of a lack of gender perspective on planning policy is summarised. Barriers to progress are explained and the role of enablers is identified. International and European factors that resulted in GM becoming an integral component of the UK planning system are explained. The second part of the paper outlines the current extent of GM in LPAs. Although generic equalities policy is widespread, it is primarily concerned with personnel matters and there is little understanding of the impact of gender considerations on planning policy. At best there is an assumption that only policies related to women's traditional roles are affected. At worst some LPAs do not consider gender to be of any relevance to the planning process, particularly in departments where there is little social awareness and a technical and quantitative approach predominates. The methodological steps required to achieve GM that need to be applied to the planning process are summarised. Little will change unless central government gives high-level guidance on mainstreaming, and resource allocation and awareness training is increased. Cultural change is needed within the profession to enable planners to take gender seriously.

Appendix 5: The various queries to build the dataset on Gender Dimension in Research Content

Phase 1. Field and subfield whose papers are included in the seed dataset.

Subfield: Gender Studies

Phase 2. Below is the list of specialist journals whose papers are included in the seed dataset:

Focus on Gender: Parent and Child Contributions to the Socialization of Emotional Competence

Genders

Gender Medicine

Gender & Society

Gender and Language

Pediatric Gender Assignment: A Critical Reappraisal

Politics & Gender

Gender, Work & Organization

Indian Journal of Gender Studies

Journal of Women's Health & Gender-Based Medicine

Gender and Education

Gender, Place & Culture

Journal of Gender Studies

Feminist Theory

Frontiers – A Journal of Women Studies

Australian Feminist Studies

Hypatia – A Journal of Feminist Philosophy

Feminist Studies

Feministische studien

Asian Journal of Women's Studies

International Feminist Journal of Politics

Men and Masculinities

Nouvelles questions féministes

Feminist Review

European Journal of Women's Studies

Feminism & Psychology

Signs

Women's Studies International Forum

Phase 3. Major MeSH terms included on the seed dataset:

Battered women
 Dentists, Women
 Femininity
 Feminism
 Gender identity
 Health services for transgendered persons
 Physicians, Women
 Pregnant women
 Sexual and gender disorders
 Transgendered persons
 Women
 Women's health
 Women's health services
 Women's rights
 Women, Working
 Men's health

Phase 4 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS:

abuse woman OR	feminist OR
abusive relationship OR	forced marriage OR
against woman OR	gay man OR
american woman OR	gender development OR
autogynephilia OR	gender difference* OR
batter woman* OR	gender dysphoria OR
bisexual men OR	gender equalit* OR
domestic violence OR	gender equit* OR
ecofeminism OR	gender gap OR
equality between men and women OR	gender identit* OR
equality polic* OR	gender identity disorder OR
female physician OR	gender ideology OR
female sexual dysfunction OR	gender inequalit* OR
femicide OR	gender inequit* OR
feminism OR	gender issue* OR

gender justice OR
 gender mainstream* OR
 gender neutral OR
 gender nonconformit* OR
 gender pay gap OR
 gender perspective* OR
 gender polic* OR
 gender power OR
 gender quota OR
 gender relation OR
 gender segregation* OR
 gender sensitive OR
 gender stereotype OR
 gender stud* OR
 gender varian* OR
 gender-based violence OR
 glass ceiling OR
 glass cliff OR
 hegemonic masculinit* OR
 heteronormativit* OR
 heterosexism OR
 heterosexualit* OR
 homophobia OR
 intimate partner violence OR
 lesbian* OR
 LGBT OR
 machismo OR
 male dominate OR
 manhood OR
 maternal employment OR
 maternity leave OR
 men health OR
 men women equalit* OR
 occupational segregation* OR
 partner violence OR
 patriarchal OR
 patriarchy OR
 postfeminist OR
 reproductive right* OR
 self-objectification OR
 sex equalit* OR
 sex equit* OR
 sex gap* OR
 sex gender OR
 sex guilt OR
 sex segregation* OR
 sexism OR
 sexist OR
 sexist belief OR
 sexual division of labour OR
 sexualit* OR
 spousal violence OR
 transgender OR
 transsexual OR
 transsexualism OR
 violence against women OR
 wife abuse OR
 wife beat OR
 wisewoman OR
 woman autonom* OR
 woman bod* OR
 woman experience OR
 woman health OR
 woman movement OR
 woman perception OR
 womanhood OR
 women autonom* OR
 women empowerment OR

women health OR

women status OR

women Leadership OR

women stud* OR

women men equality OR

working mother OR

women perception OR

working woman

women right OR

Phase 5 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS in combination with another set of noun phrases to limit false positives:

((empowerment OR

equal opportunit* OR

parental leave OR

physical violence OR

pornograph* OR

psychological abuse OR

psychological violence OR

sexual abuse)

AND

gender)

OR

((prostitut* OR

sexual abuse OR

stalking)

AND

(men OR

women))

Phase 6 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the field of *Biology* (to avoid false positives):

sexual harassment OR

motherhood OR

gender discrimination OR

fatherhood OR

gender analysis OR

sexualisation

Phase 7 (expansion of seed dataset using noun phrases): the following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry, Biology* and *Biomedical Research* (to avoid false positives).

male identity OR

cross-sex OR

sex role OR

androgynous OR

feminisation

Phase 8 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry* and *Biology* (to avoid false positives):

gender role OR

sexual coercion OR

sexual identity OR

biological sex

Phase 9 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the field of *Agriculture, Fisheries & Forestry* (to avoid false positives):

sexual orientation OR

tomboy

Phase 10 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the field of *Biomedical Research* (to avoid false positives):

gender awareness OR

male sexual dysfunction

Phase 11 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the field of *Clinical Medicine* (to avoid false positives):

gender norm*

Phase 12 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Biomedical Research* and *Clinical Medicine* (to avoid false positives):

breadwinner

Phase 13 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Clinical Medicine* and *Enabling & Strategic Technologies* (to avoid false positives):

domestic work

Phase 14 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Biology* and *Biomedical Research* (to avoid false positives):

gender role

Phase 15 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Biomedical Research*, *Clinical Medicine* and *Public Health & Health Services* (to avoid false positives):

housework

Phase 16 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry, Biology, Biomedical Research* and *Earth & Environmental Sciences* (to avoid false positives):

intersexuality

Phase 17 (expansion of seed dataset using noun phrases). The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Biology, Biomedical Research, Earth & Environmental Sciences, Physics & Astronomy, and Mathematics & Statistics* (to avoid false positives):

queer

Phase 18. The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the field of *Biology* and the subfields of Behavioral Science & Comparative Psychology, Mycology & Parasitology, Biotechnology, Dairy & Animal Science, Fisheries, Forestry, Food Science, Horticulture, and Veterinary Sciences (to avoid false positives):

gender assignment OR	gender discourse* OR
gender attitude OR	gender disparit* OR
gender atypical OR	gender dissatisfaction* OR
gender base* OR	gender diversit* OR
gender behavior* OR	gender division* OR
gender belief OR	gender egalitarianism OR
gender bias OR	gender empowerment OR
gender change* OR	gender health OR
gender class* OR	gender hierarch* OR
gender composition OR	gender image* OR
gender concordance OR	gender imbalance OR
gender conformit* OR	gender impact* OR
gender consciousness OR	gender medicine* OR
gender differential* OR	gender non-conformit* OR
gender differentiation* OR	gender order OR
gender disaggregat* OR	gender organization OR

gender parit* OR	gender responsive OR
gender perception* OR	gender schema* OR
gender performance* OR	gender socialization OR
gender politic* OR	gender specific* OR
gender practice OR	gender stratification* OR
gender process* OR	gender subjectivit* OR
gender ratio OR	gender type* OR
gender reassignment OR	gender typicalit* OR
gender regime* OR	gender violence* OR
gender representation OR	gender work
gender research* OR	

Phase 19. The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry* and *Biology*, as well as the subfield of Behavioral Science & Comparative Psychology (to avoid false positives):

sex behavior OR	sex politi* OR
sex bias OR	sex power* OR
sex change OR	sex ratio OR
sex composition OR	sex reassignment OR
sex development OR	sex represent* OR
sex difference* OR	sex research* OR
sex different* OR	sex specifi* OR
sex disaggregat* OR	sex stereotype* OR
sex diver* OR	sex stratification OR
sex divis* OR	sex violen* OR
sex empower* OR	sex wage gap OR
sex health* OR	sex work* OR
sex identi* OR	female identit* OR
sex ideolog* OR	male behavior OR
sex inequalit* OR	female behavior OR
sex inequit* OR	feminine behavior OR
sex medicine OR	masculine behavior OR
sex norm* OR	maternal behavior OR
sex percept* OR	paternal behavior OR
sex perform* OR	masculinit* OR

feminit*

Phase 20. The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry* and *Biology*, as well as the subfields of Behavioral Science & Comparative Psychology, Mycology & Parasitology, General Science & Technology, Biotechnology and Bioinformatics (to avoid false positives):

gender*

Phase 21. The following noun phrases were searched for in the title, abstract, and author keywords of papers in the WoS except in the fields of *Agriculture, Fisheries & Forestry* and *Biology*, as well as the subfields of Behavioral Science & Comparative Psychology, Mycology & Parasitology, General Science & Technology, Biotechnology and Bioinformatics (to avoid false positives):

gender*

General exclusions: Articles with the following terms in their title, abstract, and author keywords were removed from the dataset (this list could be expanded to less common animals):

animal*	bird*
rat	trout*
rats	nematode*
mice	elegans
mouse	primate*
drosophil*	dog*
goat*	horse*
bear	calf
bears	calves
macaque*	pig
monkey*	pigs