South Africa and the joint data-backed study of publication patterns of the Global Gender Gap project

I M A Gledhill

Flow Research Unit, School of Mechanical, Industrial and Aeronautical Engineering, University of the Witwatersrand, South Africa

E-mail: Igle.Gledhill@wits.ac.za

Abstract. As the level of participation of women in physics changes, it is of interest to understand whether a gender gap exists in publication, and what trends can be observed. This paper quotes data from the joint data-backed study of publication patterns that was undertaken as a task within the Gender Gap in Science project initiated through the International Science Council. The study was based on metadata available through publication databases, which allow inference of author gender from name strings using services that provide access to databases of names. Five such services were benchmarked. Results have been made available in an interactive online tool, from which the data available from South Africa have been drawn. Within the Astrophysics Data System, the rise in South African publications in the field can be observed, together with the evolution of the proportion of authorships by women. A global result from this project is also very relevant to South Africa: fractional authorships by women in high-impact journals in theoretical physics show average percentages of women near 10%, with little or no tendency to rise since 1999, while top journals in astrophysics and astronomy show steadily rising fractional authorships which have approximately doubled since 1999. The implications for South Africa are explored.

1. Introduction

Peer-reviewed publications are the basis of the body of scientific knowledge and of acknowledgement of contributions to science. In many countries, authorship is also used in the evaluation of individual performance and institutional achievement, and has become a part of hiring and promotion practices.

This is the case in South Africa, and is particularly relevant in building a strong base in basic sciences as the community of practice transforms itself towards a more equitable gender balance. Major projects in South Africa include co-hosting of the SKA[1], particle physics at CERN[2], and the National Institute of Theoretical and Computational Sciences. In the wider context relevant to this study lie theoretical physics, astronomy and astrophysics, and the many domains of physics and applied physics in South Africa.

There is a known gender gap in publication practices, and a known under-representation of women as authors in high-ranked journals. This short paper describes results from the Gender Gap in Science

[1] Square Kilometre Array Telescope

[2] European Organization for Nuclear Research

project [1] in the context of two questions relevant to South Africa. Firstly, how do publication patterns among women in astrophysics and astronomy compare with publication patterns in theoretical physics - what can South Africa learn from global data? Secondly, what data can be obtained from the project's study of bibliometrics for women in physics in South Africa?

A scientometric study of physics, including astronomy [2], has recently been completed in South Africa. The data are drawn from the South African Knowledgebase, which captures the self-identified gender of authors. Evidence of positive change of the community of physics with respect to women is captured. Nevertheless, the discipline lags the other basic sciences in percentages of women.

Of papers authored in physics and astronomy in 2016, 83% were co-authored with at least one international researcher [2]. Collaboration appears to foster authorships by women in physics [3].

Contrasts across subdisciplines may exist. A concept of specific relevance to this paper is the "field-specific ability beliefs hypothesis" put forward by Leslie et al. [4], whose formulation indicates that the extent to which practitioners "believe that success depends on sheer brilliance" is a good predictor of under-representation of women in that field.

1.1. The Gender Gap in Science project

This project, "A Global Approach to the Gender Gap in Mathematical, Computing, and Natural Sciences: How to Measure It, How to Reduce It?", was a 3-year collaboration of eleven scientific unions and global organisations[3]. Three major tasks were undertaken: a global survey, a data-backed study of publication patterns, and the formulation of a databases of initiatives known to have successfully addressed the gender gap in science. The project book [1] provides a complete report on the three tasks and on recommendations arising from the study.

2. The Joint Data-Backed Study of Publication Patterns

The objectives of the study of publication patterns included the provision of insights into the proportions of women as publication authors, and the presence of women as authors in top journals ("top" journals in each field were identified by members of the discipline participating in the study) [3]. While other aspects, such as author drop-out rates of men and women, were covered, these are not included in the present brief study.

In a study of this kind, expertise is required across data science, scientometrics, gender studies and sociology, and in the Gender Gap project domain experts were available to collaborate and to check the methodologies used.

3. Methodology of the Joint Data-Backed Study of Publication Patterns

A full description of the methodology [3] will be found in the project publication [1]. In this study, an instance of *authorship* refers to a one-to-many pair of a publication and an author. A *fractional authorship* of a paper with *n* authors is defined as *1/n*.

Data for physics and astronomy were sourced from the Astrophysics Data System (ADS) [5] and the arXiv [6]. The limitations and rationale is provided by Mihaljević and Santamaría [3]. Geoinformation was extracted from affiliation where possible using the Stanford Named Entity Recogniser NER, GeoNames, and CERMINE [7]. Authorships and author profiles were extracted; gender inference was performed using Gender API, genderize.io, Python gender guesser, and Russian surname suffices [8]. All author names were assigned an identifier of 'female', 'male', or 'unknown'. The method was benchmarked [9] and it was shown that unknown names more likely to be associated with men, and

[3] IMU, the International Mathematical Union; IUPAC, the International Union of Pure and Applied Chemistry; IUPAP, the International Union of Pure and Applied Physics; IAU, the International Astronomical Union; ICIAM, the International Council of Industrial and Applied Mathematics; IUBS, the International Union of Biological Sciences; IUHPST, the International Union of History and Philosophy of Science and Technology; ACM, the Association for Computing Machinery; UNESCO, the United Nations Educational, Scientific and Cultural Organisation; GenderInSITE, Gender in Science, Innovation, Technology and Engineering and OWSD, the Organisation of Women in Science for the Developing World.

therefore that fractions of women in the final analysis are likely to be upper bounds. A public interactive database was built which has been used as the data source for the present paper.

Automated inference of gender is a sensitive and complex subject. The methodology and its limitations were discussed in a critical paper by Mihaljević et al., 2019 [10]. Automatically inferred gender may contrast with internal perception of gender. In the present study, only binary gender was inferred. Statements about societal issues based on publication meta-data may be subject to error. Fairness, Accountability and Transparency in Machine Learning is a critical element of the field [10]. Selection bias may be present due to the datasets that are available; in the present study this primarily affects names from certain regions. The removal of authors classified as "unknown" from the dataset may cause exclusion bias favouring groups for whom gender inference is easier.

Usage of the research may change the system being observed. One or another group may become objectified, to the detriment of the scientific field. The logical misuse of publication data to justify the lack of women professors or grant holders is not unknown [10].

Given these limitations, why should publication pattern analysis be attempted? The answer is at least partially that physics is experiencing a well-defined problem in gender balance, and the gender gap needs to be understood and remedied based on the evidence available. This short paper addresses two useful aspects emerging from the data-backed study of publications in the Gender Gap project, concerning a global gap between astronomy and physics, and specific data available on the publication patterns of women in South Africa.

4. Global findings

The database is publicly available [11]. From the ADS, data from 1970 to March 2018 were drawn, with selection of publications in "Astro" only. From 777 270 documents, 181 172 author profiles were extracted. Gender could be assigned to 93 608 of which 15% could be classified as female and 60% as male. From the arXiv, data from 1991 to July 2019 were drawn, with selection of the theoretical physics subfield only (noting that High Energy Physics has its own platform, not on arXiv). Of 1 667 512 documents, 458 485 author profiles were extracted. Gender could be assigned to 281 602 of which 17% could be classified as female and 83% as male. The first result that is evident is that the fraction of female authorships in astronomy/astrophysics contrasts with other physics subfields in arXiv (figure 1).

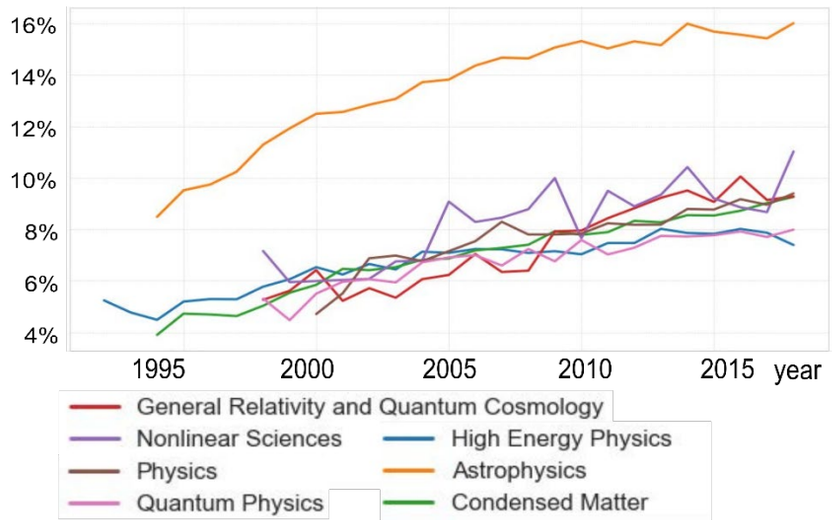

Figure 1. Global fractions of authorships by women per physics subfield and year in publications indexed in the arXiv. Reproduced under CC-BY 4.0 [3].

From "top" journals identified by participants in the project, the percentages of authorships by women in astronomy and astrophysics, and in theoretical physics, have been derived and are shown in figures 2 and 3.

It is of interest to note that in the Global Survey of Scientists, to which there were 32 346 respondents from 159 countries, of whom there were 7 570 in physics and 2 597 in astronomy/astrophysics, the

question was asked "During the last five years, how many articles have you submitted to journals that are top-ranked in your field?". A model indicated that the self-perceptions of men and women of their rate of submissions to highly ranked journals are similar [3].

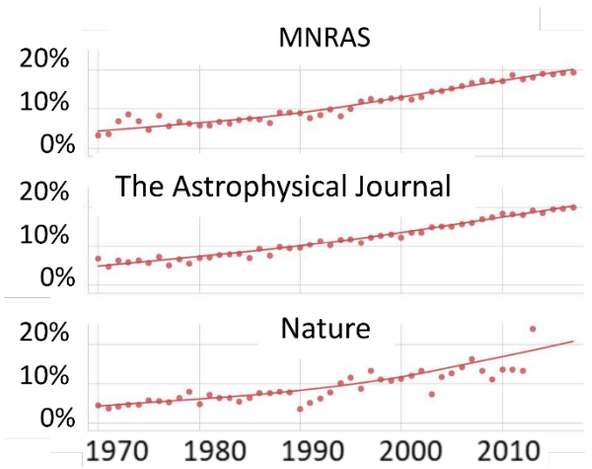

Figure 2. Percentage of fractional authorships from women in three top astronomy and astrophysics journals per year since 1970. MNRAS: Monthly Notices of the Royal Astronomical Society. The journals Astronomy and Astrophysics, The Astronomical Journal, and Science follow similar trends. Reproduced under CC-BY 4.0 [3].

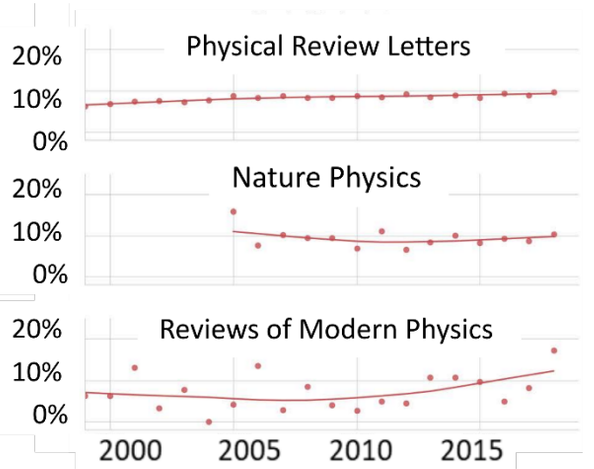

Figure 3. Percentage of fractional authorships from women in top theoretical physics journals per year since 1999. Physical Review D, Journal of Physics A, and Journal of Mathematical Physics show similar trends. Reproduced under CC-BY 4.0 [3].

5. South Africa

Within the Gender Gap publication database [11], the available data relevant to physics are from ADS only. The last year shown in each study does not include all papers published in that year.

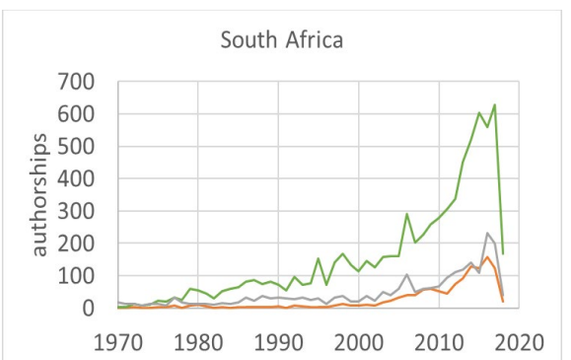

Figure 4. Chronology of absolute numbers of authorships extracted from ADS, [11].

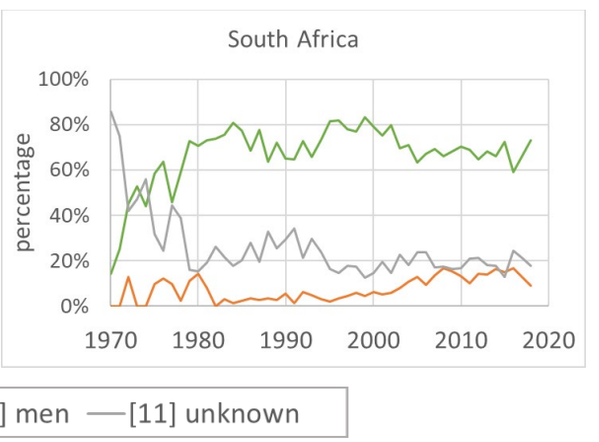

Figure 5. Chronology of percentages of authorships extracted from ADS, [11].

Figure 4 shows that South Africa experienced steady growth in the number of papers accessible through ADS, and figure 5 indicates that the percentage of authorships identified within this study as those of women rose from 6% in 2000 to 13% in 2017.

In figures 6 and 7, the same data [11] are compared with authorships identified from a recent South African scientometric study of physics [2]. These author profiles are based on self-identification of gender in the South African Knowledgebase. Only binary gender is available in the results [2]. It is observed that the number of authorships assigned through the Gender Gap publication pattern study [11] is between ⅛ and ⅓ of those available to the scientometric study. Figure 7 shows the percentages of authorships by gender in the two studies. The fraction of authors identified as women is very similar in the two studies. Trends with time are similar for the percentage of authors who are men. The automated gender assignment benchmark study [9] indicated that unknown names are more likely to be those of men, and it is not impossible that this is supported by figure 7.

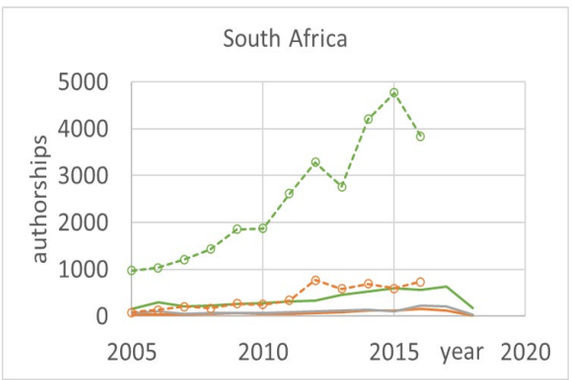

Figure 6. Numbers of auhorships, [2], [11]

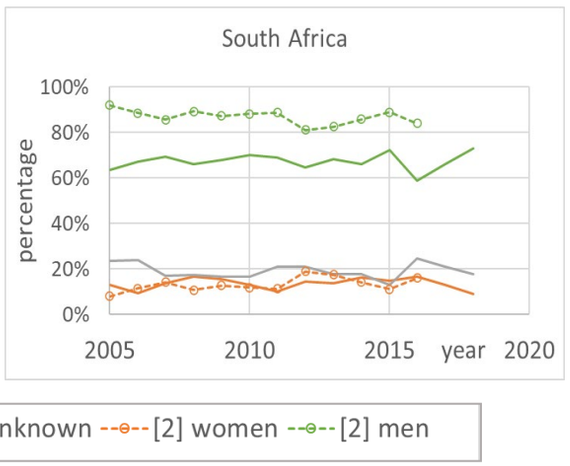

Figure 7. Percentages of authorships, [2], [11]

Profiles for two additional countries were drawn for comparison with South Africa. Egypt was chosen as an African country in which significant publication in physics is taking place. The data for Egypt show a surge in publications after 2010, and the percentage of authorships by women dropped from 9% (2000) to 7% (2017). Brazil was chosen as a country perhaps close to South Africa in science infrastructure. The percentage of authorships by women was 15% in 2000 and 17% in 2017, a slow rise.

6. Conclusions and recommendations

In global trends within publication patterns of women, the difference in authorships between astrophysics/astronomy (≈20%) and theoretical physics (≈10%) is both striking and relevant in the context of South Africa, which has significant success and investment in both these fields. In the database of publication patterns, top journals in astronomy and astrophysics exhibit rising percentages of female authorships, while percentages in some of the top theoretical physics journals appear to be stagnating or rising more slowly. Understanding the successes visible in astronomy may help theoretical physics.

A starting hypothesis is that there is a higher percentage of women in astronomy and astrophysics than in theoretical physics. As an illustrative example, data from the USA for 2017 [12] indicate that women earned 40% of astronomy doctorates, and 20% of physics doctorates. The Gender Gap project was not designed to measure percentages of women across the disciplines studied, and the data from the scientometric study do not provide gender disaggregation in subfields. If the fraction is indeed higher, this would contribute to explaining the lower publication rate in physics.

However, we should then ask why physics is less attractive to women, or loses women faster, than astronomy [3]. In terms of culture or bias, is the field-specific ability belief hypothesis [4] that "success depends on sheer brilliance" useful in explaining lower numbers, and stagnating publication rates, within theoretical physics? It is recommended that testing whether such beliefs are present may provide useful insight, and may lead to cultural changes which attract and retain women in theoretical physics.

Collaboration, in general, fosters higher publication rates by women [3]. Both astronomy and high energy physics are highly collaborative fields. However, the latter has its own preprint platform and is not well-represented on the arXiv database, and papers with more than 100 authors may not have been captured in the South African scientometric study due to rules governing the data collection process. Collaboration may be assisting women to publish, but the accessible data do not allow the effect to be definitively seen in these studies for South Africa.

South Africa experienced significant growth in the percentages of female authorships illustrated in publications accessible through ADS: the percentage of authorships identified within the publication pattern study as those of women rose from 6% in 2000 to 13% in 2017. The percentages of authorships by women in the publication pattern analysis using ADS [11] are similar to the percentages of authorships by women in the South African scientometric study of physics [2], although the two source datasets each have limitations.

Astronomy has run active development programmes in Southern Africa, and both the physics community and university departments have specific initiatives targeted at closing the gender gap. The outcomes of an intense effort to build an equitable physics culture may be evident in the fact that South Africa has a much greater rate of increase in physics publication by women than the two countries with which comparisons were made, Egypt and Brazil.

Acknowledgments

Images are reproduced from the final report which appeared under a CC-BY 4.0 International License [3]. The funding for this project was provided by the ISC and the eleven participating Unions and organisations[3]. The African Workshop was supported by the African Institute for Mathematical Sciences and the ISC Regional Office for Africa. The author thanks Helena Mihaljević, Lucía Santamaría and Christian Steinfeldt for their impactful research and for providing Open Access to the project data.

References

[1] Roy M-F *et al*. 2020 *A Global Approach to the Gender Gap in Mathematical, Computing, and Natural Sciences: How to Measure It, How to Reduce It?* (Berlin: IMU)

[2] Mouton J et al. 2020 A scientometric assessment of Physics in South Africa (Pretoria: Department of Science and Innovation, Republic of South Africa)

[3] Mihaljević H and Santamaría L 2020 Measuring and analysing the Gender Gap in science through the Joint Data-Backed Study of Publication Patterns in Roy M-F *et al*. [1] (Berlin: IMU) pp 85–152

[4] Leslie S-J, Cimpian A, Meyer M and Freeland T 2015 Expectations of brilliance underlie gender distributions across academic disciplines *Science* **347** pp 262–265

[5] Smithsonian Astrophysical Observatory-NASA https://ui.adsabs.harvard.edu

[6] arXiv https://arxiv.org

[7] Stanford Named Entity Recogniser NER https://nlp.stanford.edu/software/CRF-NER.html
GeoNames http://www.geonames.org/
CERMINE https://github.com/CeON/CERMINE

[8] Gender API https://gender-api.com
genderize.io https://genderize.io
Python gender guesser https://github.com/lead-ratings/gender-guesser
Wikipedia List of Surnames in Russia https://en.wikipedia.org/wiki/List_of_surnames_in_Russia

[9] Santamaría L and Mihaljević-Brandt H 2018 Comparison and benchmark of name-to-gender inference services *PeerJ Computer Science* **4**:e156; DOI 10.7717/peerj-cs.p156 pp 1–29

[10] Mihaljević H, Tullney M, Santamaría L and Steinfeldt C 2019 Reflections on Gender Analyses of Bibliographic Corpora *Frontiers in Big Data* **2** 29 pp1–6

[11] Gender Gap in Publications https://gender-publication-gap.f4.htw-berlin.de/

[12] Porter A M and Ivie R 2019 *Women in Physics and Astronomy, 2019* (American Institute of Physics, College Park MD USA)