Differences in publication behaviour between female and male scientists

Bibliometric analysis of longitudinal data from 1980 to 2005 with regard to gender differences in productivity and involvement, collaboration and citation impact

Abschlussarbeit zur Erlangung des akademischen Grades Master of Arts
im Fach

Bibliotheks- und Informationswissenschaft

eingereicht an der
Philosophischen Fakultät I
Humboldt-Universität zu Berlin

von
Paul Donner B.A.
geboren in Berlin, 25. November 1985

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekan der Philosophischen Fakultät I:
Prof. Michael Seadle, PhD

Gutachter:
1. Dr. Frank Havemann
2. Dipl.-Math. Michael Heinz

Ich danke meinen Betreuern für die Begleitung dieser Arbeit. Ihre Unterstützung und die anregenden Diskussionen waren für diese Abschlussarbeit von höchstem Wert. Zudem gilt mein Dank meiner Familie für all ihre Hilfe und Geduld.

Abstract

The topic of this thesis are differences in publication behaviour between female and male scientists. It is possible to assign a gender to a person based on their given name, and in some countries, in whose languages surnames are also inflected by gender, also based on their surname. For said countries, both methods are compared and evaluated. By way of the surname form method, gender is assigned to the names of authors or scientific publication of a number of countries and the groups of female and male authors are compared for the period of 1980 to 2010 in regards to productivity and involvement, cooperation and citation impact. Suitable metrics for these categories are introduced. Each countriy's results are discussed and subsequently all countries are compared.

Zusammenfassung

Der Gegenstand dieser Abhandlung sind Unterschiede im Publikationsverhalten zwischen Wissenschaftlern weiblichen und männlichen Geschlechts. Eine Bestimmung des Geschlechtes einer Person ist in vielen Fällen über ihren Vornamen, in einigen Ländern, in deren Sprachen auch Nachnamen nach dem grammatikalischen Geschlecht gebeugt werden, auch über ihren Nachnamen, möglich. Für diese Länder werden die beiden genannten Methoden miteinander verglichen und bewertet. Mittels der Geschlechtserkennung über die Nachnamensform werden für eine Reihe von Ländern den Autoren wissenschaftlicher Artikel das Geschlecht zugeordnet und die Gruppen männlicher und weiblicher Autoren eines jeden Landes über den Zeitraum 1980 bis 2010 miteinander verglichen hinsichtlich der Produktivität, der Beteiligung, Maßen der Kooperation und des Zitationserfolges. Geeignete Maße für diese Kategorien werden vorgestellt. Die Resultate der einzelnen Länder werden diskutiert und alle Länder im Anschluss miteinander verglichen.

Inhaltsverzeichnis

1 Introduction

This thesis is meant to enrich the presently existing statistical basis which underlies the discourse on the involvement of women in science. To understand the current situation better, it is imperative to investigate the past development leading up to it. To put results for one discipline of science or one country into a meaningful context, including other disciplines or countries as well is advanteguous. In order to attain this broader perspective, the present study compares several bibliometrical measures describing women's involvement in science in a number of selected countries over a long period of time with the aim to corroborate and extend previous studies' findings. The countries were chosen for whether it is possible to derive a person's gender from their surname in the language of their country with high reliability. This is due to the restrictions of the data source, the Science Citation Index, which for much of the past literature does not include authors' given names, which lend themselves better for gender detection, including many countries for which surname gender detection is impossible. Both methods are compared for the countries of interest in the time period for which given names are available, 2006 to 2010.

Using the gender data information gained from the authors' surnames, it is possible to determine in which proportions female and male first and last authors appear and which shares of contribution in publication-equivalents each group accounts for. From the raw publication counts, measures of productivity per author by gender and overall female involvement, displayed in the share of publications including at least one female author, are derived. When comparing collaboration, mainly possible differences in relative shares of sole authorship, differences in average team sizes of teams led by women and men and average collaborative coefficients are investigated. Finally, impact metrics for female and male authors from selected samples will be compared. Established and newly introduced metrics are used and discussed.

At no point will I try to point out the causes that might have led to the observed situations or are keeping them as they are from the presented findings. This is decidedly not within the scope of this thesis, even though the reasons are ultimately far more important than a mere statistical description. The underlying reasons for gender differences can not be investigated with bibliometric methods. Trying to give reasons as to why women are disadvantaged in many ways in science on just the basis of quantitative analyses of publications can amount to nothing but mere speculation. However, in this area, important and very insightful work is being conducted by means of surveys from social scientists and notable results from this area will be discussed briefly in the appropiate section.

One particular problem that could regrettably not tackled in this thesis is the question whether people of one gender tend to cite works by authors of their own gender more often than would be expected from the given author gender distribution in the citable literature in their field. The available data did not lend itself to the kind of analysis required for this.

Literature review In this chapter previous work is discussed, and both well-established and conflicting results, and, in some cases, methodology on which this study's methodology is based is being pointed out. The reviewed papers are grouped according to their main topic and ordered chronologically.

1.1 Gender in science – overview

A report on the situation of women in science in general is given by Bentley and Adamson (2003). Their review of literature on the topic concludes that women earn less then men, do not get promoted as much and do not publish as much as their male colleagues.

Concerning empirical research not focused on research output, some studies have found gender biases in grant proposal peer review processes, some found none. In a recent meta analysis, Bornmann et al. (2007), reported significant evidence (of small effect) for biases in favour of men but were later, contradicted by Marsh et al. (2009). These latest findings indicate no larger systematic bias, though some sporadic bias might remain.

To provide an overview, first some key phrases that identify important issues in the discourse on gender in science should be explained: *leaky pipeline*, *glass ceiling*, *productivity puzzle* and *citations gap*. The leaky pipeline and the glass ceiling are two closely related metaphors that try to describe the difficulties women face in science education and careers: as the level of education or staff position gets higher, the share of women decreases as more women than men drop out of careers or find it harder to get promoted. In fact, the discrepancy of the share of women enrolled in introductory programmes and those attaining full professorship in most fields of science is staggering. For a detailed literature review that also points out reasons and possible solutions for these problems, consider Blickenstaff (2005).

Cole and Zuckerman (1984) assert that women scientists are far less productive than their male counterparts based on a study of doctorate recipients. The differences increased over time for the cohort. The causes for this remained elusive, hence the term *productivity puzzle*. They found no differences in co-authorship patterns.

Ferber (1986) observed a bias in both genders towards citing references by authors of their own gender in a study of a constrained sample from labour economics literature. Ferber (1988) confirmed these initial results for four other fields: mathematics, financial economics, developmental psychology, and sociology. The observed difference was termed the *citations gap* and found to be stronger in fields with little participation by women:

> The "citations gap" appears to decrease as the proportion of women in the field, and of articles written by women, increases. The larger the proportion

> of women in a field the less invisible they are; first, because there are more articles written by women, in which women are more often cited; second, because men cite them more frequently.

1.2 Reasons for gender inequality

A common assumption in the literature is that women are more burdened by family obligations (pregnancy, staying at home to care for babies, more time spent raising children). Luukkonen-Gronow and Stolte-Heiskanen (1983) found no evidence for this in their survey of a sample of Finnish scientists. Not only is family life not perceived as a hindrance in this cohort, being married is even positively correlated to being productive, for both women and men.

Long (1990) observed small sex differences, consistently in favour of men, in many variables, such as quality of education and mentors and collaboration with the mentor among graduate students. Having children decreased the mentor's collaboration with women, but not with men.

A thorough quantitative analysis into the effect of number and age of children on research productivity was performed by Stack (2004). After controlling for the number of children, gender differences in productivity between women and men remained prominent among the cohort of US PhD holders. The number of young children correlated positively with productivity, i.e. PhDs with children published more. The only negative effect found is that women with preschool children publish less than those with older children. This effect was also confirmed in the social sciences, where other than that no productivity gender differences could be found due to the high number of women in the field.

Fox (2005) conducted a survey of US university faculty from five disciplines regarding gender, family composition, marital status and research productivity. Confirming other studies, the major productivity differences were in the non-publishing and very frequently publishing sectors. Independent of gender, married and cohabiting authors publish at higher rates than those who were never married. Men publish at higher rates in all categories except "never married". Women in subsequent marriage are almost twice as productive as women in their first marriage. Contrary to other studies, Fox found that women with only preschool children are more productive than their childless colleagues and those with elementary or secondary school children. Between various factors considered, the reason for this turned out to be a particularly strong interest in active research among these women.

Ledin et al. (2007) surveyed applicants to two molecular biology grant programmes to reveal the reasons for the puzzling differences in acceptance rates between male and female applicants, even after gender-blinding the proposals. After carefully ruling out any other possibilities, the reasons they identified were fewer support for female scientists by their superiors, especially young group leaders and those with children, and a more general discrimination against women in their department, noted by both men and women.

1.3 Gender and research output

Webster (2001) gives an overview of the activities of female scientists from Poland in the years from 1980 to 1999. Using surname suffixes denoting names as male or female, the author was able to assign a gender to 60 % of the authors, the results being in line with the proportions of women to men at three Polish medical schools. During the time frame, participation of women did not increase, in spite of more women being employed in science. Strong gender preference for disciplines was observed and remained constant over time. Polish women did not engage as actively in international collaboration as their male colleagues, leaning more towards domestic journals.

Prpić (2002) examined differences in scientific productivity among young Croatian researchers through an exhaustive questionnaire. Men are more productive in all investigated figures, except the number of single authored papers. The differences have only increased from an earlier study (Prpić (1990)). Total career output of women amounts to 67.9 % that of men. The author speculates that the observed increases may be due to a transition to a more competitive international scientific publication system that escalated existing differences. Among highly productive scientists, men are overrepresented. Remarkably, disciplines with low women participation display only minute differences in productivity. Among natural scientists, medical scientists, bio-technical scientists and social and humanities scholars sex differences are significant and consistent, even in the early years of their careers. Through regression analyses the author was able to isolate the most important productivity predictor: professional position. This factor has a stronger positive effect for women. Family factors' influences, however, are marginal and affect both sexes equally.

Naldi and Vannini Parenti (2002a,b) performed large scale analyses for their report for the European Commission, which was both an analysis and a feasibility study on using given names for sex identification. They compiled a rather comprehensive first name database and used it to determine the sex of authors of patents and scholarly journal articles from six European countries, a methodology they found to be robust enough to be useful for studies concerned with author gender. Their sample consisted of articles from 157 scientific journals from the year 1995 whose authors came from France, Germany, Italy, Spain, Sweden or the United Kingdom. In it, women made up 22 % of all authors and contributed to 20 % of all items. Participation, however, is not that balanced:

> The items with at least one female author are 45.8 % while the items with at least one male author are 94.7 %. As a consequence 54.2 % of the items have been entirely produced by men and 5.3 % entirely by women.

Regarding disciplines, women's contribution is low in mathematics and engineering and above average in biology, bio-medicine and earth and space sciences. Looking at countries, women's participation is low in Germany and Great Britain, high in Spain and Italy. The authors also noticed that men had a higher likelihood to write publications on their own.

Another large scale comparative study between fourteen European countries' female scientists' output (journal articles and patent applications) was carried out by Frietsch et al. (2009). The authors selected publications in about 300 high-ranking international journals from the sciences published from 1996 to 2005. They chose to include only countries that they were able to assign a gender based on author first names with a success rate of at least 85 % from data of patent publications using country specific lists (lowering the hit rate for people not from that country or not with a name common in it). Those 14 countries' gender assignment rates in the journal article data ranged from 75.6 % (USA) to 94.5 % (Italy). Their results indicate that women's participation is higher in scholarly publication than in patent application material. Women's participation rates are high in Italy, France and Spain, and low in Germany, Austria and Switzerland. Overall contribution shares grew moderately, though much less so for countries with already fairly high rates. Differences across disciplines are expectedly pronounced, with mathematics and physics having low rates of women participation, while biology and bio-medicine ranked high – these shares being relatively similar across countries.

Puuska (2009) analysed publications by Helsinki University research staff (2002-2004) in order to explore relationships between output, gender, and professional position and taking into account all publication venues, not just international journals. Professors are the overall most productive group; PhD holders publish more than non-PhDs. In general, men outperformed women in the respective status groups, except among professors in the natural sciences category. However, no significant difference between genders was observed in conference papers.

1.4 Gender and collaboration

McDowell and Kiholm Smith (1992) found co-author choice in a cohort of PhD holders in economics to be significantly influenced by their gender. Propensity for single authorship was higher among women than men. The authors argue that for their sample, these differences affect promotion decisions to the disadvantage of women.

In a study of authorship patterns in the field of information systems, Cunningham and Dillon (1997) found women authors far more likely to engage in teamwork than men. The set consisted of articles from five journals (1989-1995). Men publish alone nearly twice as often as women. The fraction of male authors in male-only teams as a percentage of all male authors (43 %) was higher than that of females in female-only teams (7 %).

Boschini and Sjögren (2007) further confirmed that co-author seeking behaviour in economics is not neutral in regard to gender. This was shown by modelling voluntary authorship team formation as a random matching process influenced by individuals' preference for co-author gender and team size. The model was then examined against empirical data from three top economics journals (1991-2002). Articles were classified into three sub-fields. In this set of articles, women are twice as likely to coauthor with women than men are. Contrary to prior studies, differences in preference for single authorship were minor. Single authorship, especially among women, decreased over time. Preference for single authorship and presence of women vary vastly among sub-fields. In fields with higher percentages of women authors, woman-woman co-authored articles are more frequent. The tendency of men to co-author with women increases less than expected. Presence of women in the field did not correlate with the fraction of single authored articles. These findings allowed the rejection of assumptions of gender irrelevance and gender neutrality in co-author choice.

The collaboration in the field of distance education was examined by Zawacki-Richter and von Prümmer (2010). They examined articles published in five leading distance education journals between 2000 and 2008. 46 percent of the sample's authors were women. Proportions of male and female authors in sole-authored articles were close (male: 53, female: 47 %). However, men are slightly more often first authors of collaborative articles than women (male: 57, female: 43 %). No difference between genders for preference in team size was observed, but women first authors showed a slight preference for just one co-author and men for three and more co-authors. In this sample, men were slightly more inclined to choose male-only teams than women were for women-only teams. The ratio of women partaking in mixed gender teams was higher than that of men (female: 57, male: 48 %). Furthermore, remarkable associations between gender and research topic, as well as research methods, were observed.

1.5 Gender and impact

A study of the role of gender in the context of scholarly production was conducted by Lutz (1990), who examined literature in sociocultural anthropology, a field with a traditionally high ratio of female scholars, from 1977 to 1986, published in four core journals. The share of articles written by women in that set rose from 27 % to 34 %. Citation analysis for a subset of 446 articles from 1982-86 with single authors was carried out. The main finding is that “Men cite women authors only half as often as women do”. However, it must not be overlooked that in this case there was no proportional overcitation of women by women, rather, men did not cite women as much as their proportion of articles in the cited period would suggest, assuming equal quality.

Ward et al. (1992) looked into citation patterns in a sociology subfield by examining citations in one journal partly specialized in gender topics from 1974 to 1983. Such articles on gender issues were cited more often (18 times vs. 10 times) than articles not concerned with this topic. While overall received citations did not differ by the gender of the author, women were cited more often by women, while no statistically significant dependencies for cited men were observed.

In a study of productivity, collaboration and impact of biochemistry PhD recipients throughout their careers, Long (1992) observed that men publish more frequently than their female colleagues in the beginning of their careers, but women later catch up on the average number of papers per year. A similar pattern shows in the average position in author lists per year. Levels of collaboration, as measured in number of authors in in the papers coauthored, are not different for the two genders. Articles by women consistently receive more citations on average than those by men. All results indicate that differences in publication behaviour mainly arise from a lower publication rate of women, which in turn can be traced to more female non-publishers and fewer women among highly productive authors.

Davenport and Snyder (1993) report on the topic of different citation practice in sociology. They randomly chose 100 articles from a set (first authors male: 72 %, female: 28 %) of 25 journals' publications ranging from 1985 to 1994. The number of female authors cited by males was significantly lower than that of women citers citing women, which was close to the sample's ratio of women. Men undercited women authors compared to the share of the sample population they make up, substantiating the findings by Lutz (1990).

Lewison (2001) sheds light on the numbers, collaboration and perceived relative research quality of female researchers from Iceland, utilizing the custom of children receiving a surname consisting of their father's given name (genitive) plus one of the suffixes -son or -dóttir, depending on the child's gender. The dataset investigated comprised all articles with Icelandic contributions in the SCI from 1980 to 2000. The ratio of female to male authors grew steadily. A citation analysis was undertaken for the subset of articles from clinical medicine (1988-1996, n=555). Male and female authored articles performed very similar, though they were outperformed by the undetermined category (mostly international coauthors).

Hutson (2002) presented findings similar to Lutz' in one of four archaeology journals examined. He stresses the importance of factors other then gender on citations included such as field specialization and regional specialization.

Citation differences were observed in the fields of sociolinguistics and linguistic anthropology by McElhinny et al. (2003) in a set of articles from five journals (1965-2000). The rates at which women were cited by men and women differed considerably. In the entire journal article set 27 % of citations were made to women's articles, men cited them at 22 %, women at 35 %. Differences at journal level were conspicuous. The authors' findings support Lutz' results, stating that differences in self-citation between genders do not explain their differing citation rates of women.

Håkanson (2005) explored citation practices in three library and information science core journals (1980-2000, 1739 articles). She observed a significant deviation from the average towards citing one's own gender's authors for both male and female authors. As authorship share becomes more balanced over time, the shares in references expectedly converge as well. Håkanson estimates the time lag to be about ten years. Shares of received citations were dependent on the citing author's gender, i.e. in a given year male authored articles were cited more often by males than by females and vice-versa. Reece-Evans (2010) continued Håkanson's inquiries for two open access electronic LIS journals (277 articles from 1995-2007). Again, referencing patterns for both and men and women were found to differ from the average due to a preference to cite one's own gender, the preference was more pronounced for men. Articles authored by women received more citations overall.

Sánchez Peñas and Willett (2006) analyzed publications of a cohort of US and UK librarianship and information science university department staff. Articles were categorized into eight classes by topic. After normalization, it was observed that men publish significantly more articles in each category, measured in mean publications per author. In two of the categories, females received more mean citations per publication, men outdid them in the other six. These differences in publication count and received citations were not found to be statistically significant.

Recently, Borrego et al. (2010) report on investigations about a cohort of 731 Spanish PhD holders. Among those with articles in the SCI database, no significant difference in publication count in regard to gender was found. They determined that it was more common for women to have no post-doc output at all than for men, the number of female authors with any post-doc output was lower than that of men. Women's articles received a significantly higher median citation count. When considering the publishing venues' JIFs, women again outperformed men, however, this measure is not generally considered reliable.

Dendrochronology is a relatively small and interdisciplinary field of research in which men outnumber women (an estimate of 70 percent men is given). A sample of 20 comparable male and 20 female scientists from this field was compiled and analyzed for a possible citation bias by Copenheaver et al. (2010). Researchers were compared on the basis of average citations per first-authored article. No significant difference in citation rates between genders was observed. The authors point out that in contrast to other fields, female dendrochronologists are highly productive, with only insignificantly less papers per author than men. Another reason might be less time spent in teaching, since dendrochronology is not commonly taught at undergraduate level. Women were as eager to coauthor as men, lessening a possible sole authorship isolation effect observed to be higher for women in other fields.

Aksnes et al. (2011) performed a citation analysis on four Norwegian universities. They found small differences in favour of men in both the field and journal normalized citation indexes, which for the latter might likely be due to publication in more cited journals. In the field index' case, women are affected by their low production rates, whereas men profit from the cumulative advantage of publishing more often. Gender differences were still observed when only looking at men and women of equal academic status, even on professor level. From this it could be concluded that overall differences could not be attributed to the small number of women in important positions.

Lewison and Markusova (2011) conducted a productivity and citation impact study of female Russian scientists, once more by determining their sex from the name endings, for publications in 1985, 1995 and 2005. They analyzed the presence of female authors in major scientific fields and determined citation scores, taking into consideration international collaboration, language preference, team structure and article type by applying multiple regression analysis. Their results confirm similar earlier studies in that women lag behind in output and their citation scores are lower then men's, though a part of that difference can be traced to the mentioned factors, as their multiple regression indicated.

1.6 Names in bibliometrics

Working with personal names, especially in bibliometrics, is subject to considerable caveats. Names are ambiguous and name forms are ambiguous. As for name forms, a person's name might be entered into a database with more than one form of heading – a problem caused by the lack of universal authority control in academic publishing, which has recently begun to be addressed. A name might be abbreviated in different ways, transcribed differently or contain spelling errors. Marx (2011) gives an account of this issue for databases used in bibliometrics.

More than one person might share the same name (homonymity). For example, Aksnes (2008) reports that 14 % of all Norwegian research personell share their name with one or more people in the surname plus given name initials form as used in the SCI. The problem is worse for East Asian countries. In Japanese, Cornell (1982) reports, the lack of middle names, the possibility that names of different persons, written with different kanji characters, are correctly transliterated to the same latin alphabet strings and the syllabic language system make it difficult to distinguish scholars with the same surname.

Xu and Nicolson (1992) remark that in China, unlike in the West, there are far more given names than surnames. Abbreviating given names to initals thus leads to very high occurance of accidentally and unneccessarily homonymic strings, cf. Sprouse (2007). Furthermore, accurate transcription of Chinese is difficult per se and simply not possible in plain ASCII, as tonal marks are omitted, amplifying the issues. These two problem areas, use of few surnames and difficulties in transcription, also hold true for Korean and Vietnamese names.

Personal names may change, which is a problem that pertains more to women, who, in much of the World, traditionally take on their husband's surname at marriage. Pellack and Kappmeyer (2011) studied women author names' versions in bibliographies and online databases. They found that the correct name form appeared in 82 % of all cases.

1.7 Summary

Gender differences are widespread and diverse but not universal. They have been investigated in studies of hiring processes, researcher promotion, tenure decisions, and awarding of grants with methods of quantitative social science (surveys, statistics of grant application outcome, faculty composition, earnings etc.) on one hand and studies of output, impact and collaboration as reflected in published works analyzed with bibliometric methods on the other.

There are structural inequalities throughout most sciences: Women are underrepresented in scholarship in general. Women drop out of academic careers far more often then men. They advance to higher academic ranks slower and much more rarely then their male counterparts. They are not on par with men in other venues of academic recognition to various degrees (prestigious institutions, scholarships, awards, professorships); a problem of course both caused and amplified by the previously mentioned differences. Women are paid less, even for similar work. Why?

Besides the long since ruled out notion of different mental capabilities, one further oft-adduced cause for smaller involvement of women in science is the burden of childbirth and childcare. Several studies have shown convincingly that family obligations are not only not neccessarily a burden, but a possible indicator for higher productivity and any disadvantage is shared by parents, regardless of gender. For possible causes, that leaves two areas, first, certain socio-psychological factors, proposed to affect men and women differently, such as commitment, motivation or interest, and second, true systematic biases or discrimination against women.

The methodology used and described in the following chapters is based in part on that by Webster (2001), Lewison (2001) and Lewison and Markusova (2011) (surname methodology), Naldi and Vannini Parenti (2002a,b) (given name methodology and the contribution and participation metrics).

2 Data and methods

2.1 Data

Bibliograpic items with address fields matching a given country name, publication year matching the range 1980 to 2010 (plus one earlier and later year for early and late publications) and having the document types “Article”, “Letter” or “Note” were downloaded from the Science Citation Index (SCI) in Thomson Reuter’s Web of Science database in October 2011, as were additional items in journal volumes required to calculate mean expected citation rates for the impact analyses. All files for a given country were concatenated and all double quote characters were removed.

After importing the data for a country into `R` (R Development Core Team (2011)), items with more than 100 authors were excluded, because they would very likely conspicuously skew collaboration measures in years in which they occur and introduce outliers, especially for countries with comparatively few publishing scientists. At the same time, they carry only very little information content about each author that participated in them due to the small share of contribution of each group author. Furthermore, items for which the surname method was unable to find the gender of any author were also excluded. The number of excluded items are listed in each country’s data section.

2.2 Methods

2.2.1 Personal name gender assignment

Personal names identify people. Names, personal names, are ubiquituous. The usage of names is culture. They are dependent on location and time. For this study mostly present European names are of interest and it is assumed that they consist of a *given name* and a *surname*. Other terms are not appropriate or not exact enough: as there are no hereditary family names in Iceland, that term is not used. First name, forename, last name and christian name are not used because these terms derive from the position of a name part in a complex name and cultural tradition respectively and are not broadly appropriate.

Very often, particular given names are gender specific. In a number of languages, the gender of a person is also apparent from the name form of the surname. In both cases, many names are not indicative of gender: unisex given names, surnames that by rule are not inflected by gender or that did not originate in the country with that language. These are exceptions and since in the general case, the gender indicating properties of names work rather well, they are used in this study to investigate sets of scientific publications with regards to gender of the authors under the given simplifications and caveats.

Thus, for the purpose of this study the concept of gender is treated in a simplified manner, as a basic biological distinction: an unambiguous dichotomy of female and male. This way of assessing gender is not reflecting the finer details of this concept in biology and society, but it is a neccessary simplification.

Gender assignment is carried out by a `Perl` program (`ug.pl`) that analyzes the surnames (in conjunction with module `Gdm.pm`) and calls an external `C` program (called `gender.c`) to perform given name gender identification. The program reads the preprocessed data from SCI for one country and takes arguments specifying which analyses to perform[1]. Both surname and given name gender association are always country specific. The program outputs text files for the surname and given name processes, as applicable, which are structured to mirror the input file in such a way that each line of input (one article description) results in one line of output (author gender results). The output is a list of tab-separated one-letter codes denoting the result, taking the values `f` or `m` for unequivocal male or female names, respectively, `u` for a name found to be from the country in question but of unknown gender (either because the surname suffix is ambiguous or the given name is unisex or not in the database), `n` for authors not from the country, `i` for authors who could not be matched to an address with certainty, and finally, `l` for a lack of a proper given name (not applicable to surname gender identification), which just means that only given name initials were available.

Given name gender identification

The given name gender identification was done entirely with the `gender.c` program (version 1.2, 2008-11-30) as described in Michael (2007). It was called with the given names and country names as arguments, if full given names were available, from within `ug.pl` and the returned result string was mapped to the aforementioned letter codes. Results *is mostly male* and *is mostly female* are mapped to `m` and `f` respectively.

[1] Perl and R code for this thesis is available under `http://amor.cms.hu-berlin.de/~donnerpa/thesis/`

Country	Detection rate	
	all names	unique names
Selected countries		
Azerbaijan	0.720	0.657
Bulgaria	0.875	0.716
Czech Republic	0.984	0.734
Greece	0.833	0.534
Iceland	0.914	0.850
Kazakhstan	0.693	0.652
Latvia	0.967	0.930
Lithuania	0.929	0.749
Macedonia	0.918	0.834
Poland	0.973	0.518
Russia	0.938	0.735
Slovakia	0.971	0.769
Uzbekistan	0.554	0.549

Tabelle 2.1: Given name gender detection rates, 2006-2010

Surname gender identification

Information on gender specific surname suffixes was gleaned mostly from Brendler and Brendler (2007) and complemented with information from the Wikipedia article “Family name”. The gender specific suffix rules were translated into `Perl` regular expressions and the surnames of authors which were found to have an address from the country in question were matched against those. Determining the correct address, and thus country, for each author was irksomely difficult, as the source data did for the most part not directly associate a name to an address. This the reason for the `i` code.

In some cases the same sets of suffix regexes could be used for more than one country:

- Czech and Slovak names (and those from Czechoslovakia)
- Russian, Azerbaijan, Kazakh and Uzbek names - due to russification of the other Soviet Republics (as well as those from Belarus, Kyrgyzstan, Tajikistan and Ukraine, which were eventually not used)

The results of this simple method are improved by

1. Considering double-barrelled names a special case. Women may choose to keep their birth name and use it as one part of a double-barrelled name. If only the

birth name is inflected in the female form and it is the first part of the double-barrelled name, the whole name would incorrectly be classified as male. Hence, in names with a hyphen, the first part is checked if it is a female form name. This proved especially effective for Poland, where this practice is seemingly quite common.

2. First building a buffer of all distinct full names that were found to be from the particular country for certain. In a second run, all names that no address could be associated with (`i` code) were looked up in the buffer. If they were found in it, the author was considered to be from the country with reasonable surety, because the person (or one with the same name) had lived there at one point.

3. Several minor tweaks to the regexes:

   - Greece: Added -eli, -elis, -alis, which are probably transcription inconsistencies of -elli, -ellis, -allis.

   - Latvia: Some Russian suffix form were included.

   - Lithuania: Modified some expressions to avoid conflicts with Slavic name suffixes and identify them correctly, since many authors from neighbouring countries work in Lithuania. Added -ienie, which seems to be an occasional misspelling or transcription inconsistency of -iene.

   - Russia/Russian group: Some transcription inconsistencies were allowed for.

Table 2.2 is a listing of the countries considered for surname gender identification. The second column gives the range of publication years taken into account for the calculation, even if there were only few articles in a year. That is the reason why the range often begins before the country really existed with that name. The third column is the ratio of all female and male names to all names (female, male and unknown gender) as a quotient. This is the criterium by which countries were selected to be included in the analysis, which means, in other words, the proportion of names to which a gender can not be associated with certainty should be as small as possible. Countries with a $\frac{f+m}{f+m+u}$ ratio greater than 0.70 were selected. An exception is Poland, where according to Webster (2001, p. 188), the gender ratios of the identified and the true population are similar. The actual detection rates for Poland in the analysis are much higher because items for which no gender could be assigned to any name are excluded. Belarus and Ukraine have considerable Russian minorities, but the detection ratio is too low to make it useful. Kyrgyzstan and Tajikistan, on the other hand, do not produce enough literature to make meaningful statistical studies possible. The fourth column lists the detection ratio for unique full names.

Country	Years	Detection rate	
		all names	unique names
Selected countries			
Azerbaijan	1989-2010	0.897	0.752
Bulgaria	1980-2010	0.836	0.603
Czech Republic	1989-2010	0.835	0.841
Czechoslovakia	1980-1995	0.844	0.836
Greece	1980-2010	0.833	0.758
Iceland	1980-2010	0.799	0.460
Latvia	1981-2010	0.827	0.786
Lithuania	1989-2010	0.757	0.660
Macedonia	1991-2010	0.868	0.784
Kazakhstan	1988-2010	0.856	0.819
Poland	1980-2010	0.388	0.330
Russia	1989-2010	0.725	0.624
Slovakia	1989-2010	0.847	0.852
USSR (Russian SFSR)	1980-1992	0.748	0.609
Uzbekistan	1989-2010	0.884	0.828
Discarded countries			
Belarus	2007-2010	0.466	0.466
Kyrgyzstan	2006-2010	0.896	0.835
Tajikistan	2006-2010	0.908	0.876
Ukraine	2009	0.392	0.410

Tabelle 2.2: Initial surname gender detection rates

Other European languages with surname gender differences that could not be further investigated are Irish and Scottish Gaelic, in which the different forms are not prevalent in written text meant for international audiences, and Sorbian, whose speakers are not directly associated with a country of their own, since the Sorbs are a small ethnic group in eastern Germany, and thus cannot be searched for in the SCI.

In the Slavonian and Baltic languages and in Greek, the male and female name forms are the declensed forms of the root word according to grammatical gender, which is identical to the actual gender of the person, and word class (noun or adjective). Additionally, Lithuanian female surnames are also indicative of their bearer's marital status. A very similar practice used to be common in Poland in the past, but has fallen out of use almost completely. In Icelandic on the other hand, there are no hereditary family surnames, instead the surname is a patronymic, formed of the given name of the father suffixed by -son or -dóttir, for sons and daughters respectively. Occasionally a matronymic is used, but that makes no difference for determining the gender. This is not to be confused with fixed surnames that derived from former true patronymics. The practice of differently suffixed patronymics was once very common throughout Europe (especially in Scandinavian and Gaelic language regions) and is extant in some other languages of the world, for example Arabic (Nasab name part), Habesha, Igbo.

It has been mentioned in section 1.6 that the use of personal names in bibliometrics studies is quite problematic, some minor inconsistencies were addressed in section 2.2.1. For this study, fuzziness coming from transcription is a particular problem. All names in the SCI database are transcribed to plain ASCII. That means that a lot of languages must be transliterated into a system with fewer symbols than the language requires. Consider for example that Czech and Slovak surnames that end in -á are unambiguously female (these are names whose basic form is an adjective with a male ending of -ý). However, names ending in -a are never inflected and are always gender-ambiguous. In SCI, the letter á is transliterated to a. Thus a very common female ending is irreversibly confounded with an ambiguous one.

One other method of identifying names' genders that is not reliant on suffixes is possible: construction of two exclusively male and female name dictionaries per country based on census data. This was experimentally attempted for the Czech Republic because the necessary data was freely available. Name data for 2007 was downloaded[2], names unique to the male and female name files were filtered with the UNIX `comm` tool and the names transliterated to plain ASCII (`translit_cs.pl`) and saved as two text files. The files were used to check the names and identify the gender of authors for the Czech publication items. This method performs virtually equally as well as the suffix surname gender detection method for names from the Czech Republic, as table 2.3 illustrates. However, since it was not clearly better and data was only available for this one country, this method was not used in the analysis but is mentioned here as a viable alternative. The errors that occur when using this method are of different nature: there are names which will be assigned to be unambiguously male or female, which just at the time happened to have only people of one gender bearing that name, which can easily change through marriage or birth if the name is in fact gender-ambiguous.

[2] `http://www.mvcr.cz/clanek/cetnost-jmen-a-prijmeni-722752.aspx`

Year	Detection rate		Difference
	Suffix method	Names dictionaries method	
1994	0.881	0.860	-0.022
1995	0.880	0.851	-0.029
1996	0.863	0.852	-0.011
1997	0.862	0.851	-0.011
1998	0.853	0.854	0.002
1999	0.857	0.853	-0.004
2000	0.855	0.849	-0.006
2001	0.853	0.846	-0.007
2002	0.858	0.850	-0.008
2003	0.846	0.857	0.010
2004	0.853	0.849	-0.004
2005	0.850	0.845	-0.006
2006	0.856	0.863	0.006
2007	0.851	0.877	0.027
2008	0.851	0.840	-0.011
2009	0.846	0.838	-0.007
2010	0.848	0.841	-0.007

Tabelle 2.3: Suffix method and names dictionaries methods surname gender detection rates for names from the Czech Republic

2.2.2 Bibliometric measures

In order to assess the involvement of women in science over time, a variety of metrics is computed for each year and each country, or, in some cases for overlapping periods of time when data is scarce. Since there are several metrics for all of the three areas, productivity, collaboration and impact, they are to a certain degree redundant. However, each one has a specific purpose and they are intended to complement each other. This study is concerned primarily with the relative performance of male and female scientists, hence only the figures and ratios of these two groups are reported, not those for authors of unknown gender and authors from other countries. For each country there will be a summary of the results and full sets of figures in the appendix which are essential for a good overview of the situation in one country. All functions were written in and all analyses performed with the `R` statistics programming language (v. 2.14). The level of significance α used throughout all statistical tests is 0.05. All productivity and collaboration metrics are computed for years with at least 50 publications.

Productivity

The first two metrics are the distributions of the gender of first authors of all articles and last authors of articles with two or more authors, both expressed as the ratio of female to male plus female authors. The first one might at first seem as a crude way to measure the general gender distribution in the light of the more sophisticated metrics introduced later, but it provides insight into not just how many authors of any gender were involved, but rather how many were involved at this particular position in the author list. The first author is generally regarded as the most important participant unless an author team opted for alphabetical name ordering. The last author in a team publication is oftentimes the senior researcher or head of the working group overseeing the project.

The next two measures, participation and contribution, follow the methodology of Naldi and Vannini Parenti (2002b, p. 26). Participation, here, refers to the share of articles which have at least one author of the given gender in the author team. A two sample proportion test under the null hypothesis that the two proportions are equal is performed between the male and female participation counts. If the null hypothesis was rejected, i.e. the two proportions are significantly different, the difference (female minus male proportion) is indicated under the chart. No datapoint in the small lower chart for a year indicates no significant difference. If there were no differences found in any year, the charts are omitted.

The contribution is computed by dividing each article (1 unit of contribution or publication-equivalent) into n equal shares, where n is the number of authors. The contribution of a time period, in this study one year, is the sum of all shares by gender.

For example, an article with three authors, two women and one man, has $\frac{2}{3}$ units of female and $\frac{1}{3}$ unit of male contribution. Thus the sum of all the sums of contributions for all genders (including those of unkown gender and from different countries) in a year is equal to the number of all published articles in that year. The relative share of female to female plus male contribution is reported (Women's contribution).

It is insightful to set women's contribution in relation to the share of women among publishing female and male authors, this is the female contribution relative to the share of publishing female authors, or, relative publishing women's contribution. In dividing women's contribution by the share of women authors among female plus male authors in a year, a figure is derived that is greater than 1 when a given share of female contribution is created by a smaller share of authors than would be expected. A figure smaller than 1 indicates that the contribution is smaller than the share would suggest. Non-publishing scientists are not represented in this figure at all.

In order to provide more visual context, the plots for share of female first and last authors, women's contribution and relative publishing women's contribution are drawn such as that all other countries' values are indicated by grey lines while the country in question has a black line. As a consequence, these plots have a range of time from 1980 to 2010 on the x-axis in all cases, which is not true for all other plots.

The final productivity metric is average productivity per author, or, the mean number of publication-equivalents of all female and all male authors in a year. For this purpose, a list of all authors, based on the author name strings, their genders, and sum of shares of contribution is computed for each year. The average of the sums of shares of a gender is that gender's contribution per author. The geometric mean is used for the average, because the sum of shares does not follow a normal distribution. An exact Wilcoxon rank sum test (otherwise known as Mann-Whitney U test) is performed on the sets of female and male values in a given year under the null hypothesis that the distributions are equal. The `wilcox.exact()` function from the `exactRankTests` package, Hothorn and Hornik (2011), is used because the distributions often include ties. If the null hypothesis was rejected, the approximate effect size of the difference between the female and male samples is indicated below the chart, calculated with a slightly optimized implementation of the method given in Sachs and Hedderich (2006, p. 460-462).

Collaboration

The first measure of collaboration considered is one of an absence of collaboration: the rate of sole authorship (articles with a single author). The absolute number of sole authored articles in a year is divided by the number of articles that have the same gender as a first author, which yields the ratio of sole authorship. A chi-squared test on the 2×2 contingency table of female and male sole authored and team authored article counts is performed under the null hypothesis of independence in respect to gender. If the null hypothesis is rejected, the difference of female to male relative share is plotted.

To answer if the team size of collaboratively authored articles differs in respect to the gender of the primary author, the average team size (geometric mean) of the number of authors for articles with two or more authors is calculated. An exact Wilcoxon rank sum test under the null hypothesis that the values of the female and male team sizes have equal distributions is performed and if the null hypothesis is rejected, the approximate effect size between the female and male values is computed and indicated below the main chart. A positive value would indicate that the average team size of male-headed teams is larger, and a negative value that female-headed teams are larger. The unit of the effect size can be regarded as number of authors in this case.

Another way of measuring the collaboration of authors is comparing the average collaborative coefficient (CC) (Ajiferuke et al. (1988)) of all authors by gender. The arithmetic mean of the CC of all male and female authors respectively, as identified by name string, is computed and an exact Wilcoxon rank sum test under the null hypothesis of equality between the female and male values is performed. The approximate effect size is computed and plotted if the null hypothesis was rejected.

Further insight into collaboration differentials can be gained by analyzing the gender composition of author teams. Does the gender structure of the rest of the team depend on the gender of the first author? To try to answer this question, all articles with two or more authors for which the gender of all authors is known (which also means all authors are from the same country) are categorized into 10 disjoint groups by the relative numbers of author genders represented in the author collective, minus the first author, and by the gender of the first author. The categories for the columns are *only male*, *more male than female*, *equal numbers of male and female*, *more female* and *only female*. An example of a count table obtained from this method is given in table 2.4.

	male only	more male	male eq. female	more female	female only
female	32	90	23	24	123
male	218	41	21	45	21

Tabelle 2.4: Team structure example table

If both rows contain at least 20 observations a rank correlation between the *female* and *male* rows is performed. Kendall's τ (tau) statistic was chosen over Spearman's ρ because ρ assumes equidistance of the divisions of the scale. A τ value of close to 1 would indicate that the ranks of the categories of the rest team gender composition between male-led and female-led teams are very similar, i.e. team structure in terms of gender proportions are independent of the gender of the primary author. Values around -1 would point to inverse rank ordering of categories: very dissimilar structures. Another way of describing this approach is that assuming if only the relative proportions of female and male authors in the pool of available scientists and the number of possible combinations which lead to the relative proportions in the columns groups are relevant to the choice of author group members by first author, the τ-statistic would be close to 1.

Impact

For the two citation impact metrics, the citation window is from the point of publication to October 2011, when the article descriptions were downloaded.

The first metric of impact is the proportion of female and male first authored articles that remained uncited. A chi-squared test under the null hypothesis of independence is performed on the 2×2 contingency table of article counts of female and male first authored articles that remained uncited and those that were cited at least once. The difference of the female to male proportion is calculated and plotted if the null hypothesis was rejected.

The last measure is mean relative citation rate for male and female first-authored articles respectively, similar to those proposed by Schubert and Braun (1986), done only for a sample of all articles, since the mean expected citation rate is calculated from sets of articles with the same properties as those considered for the citation analysis and no direct access to all those figures was readily available. To keep the number of article descriptions to be downloaded for the mean expected citation rate (MECR) reasonable, the analysis was restricted to articles published in those journals which together account for at least 20 % of all publications for each country. The intersection of the sets of top journals across all countries was a list of 199 journals. For these, article descriptions for the three article types “Article”, “Letter” and “Note”, for the years 1980 to 2005, were downloaded. Year, journal and document type specific mean expected citation rates (MECR) were calculated, which are normalization divisors by which then the actual citation count of every publication that appeared in one of these journals was divided by. This is the article specific normalized citation score, normalized for year of publication, type of publication and journal (and thus discipline). The target was to have at least 10 observations in both the male and female groups. For countries with comparably low output, this meant the observations of several years had to be culled, resulting in running means.

The impact in respect to the gender of the first author is the arithmetic mean of all normalized citation scores divided by the number of authors involved in the article for all female and male first-authored articles respectively, within the stated restrictions. Viewed over the entire period of time for which data was available, this gives a non-representative sample of at least 20 % for each country, consisting of publications that appeared in popular journals. An exact Wilcoxon rank sum test of the female and male distributions under the null hypothesis of equality is performed and the approximate effect size computed if the null hypothesis was rejected.

A more accurate impact metric would be the yearly average normalized citation scores by gender calculated as means over all individuals' summed up fractional citation scores within a given period of time, but for this metric, all citation counts have to be at hand, not just a sample.

3 Results

3.1 Azerbaijan

3.1.1 Data

The dataset for publications with contributors from Azerbaijan comprises 3484 items from 1992 to 2010. 10 items were excluded due to having more than 100 authors; 545 items were excluded because no usable author gender attributes could be determined. The number of articles increased from 123 in 1992 to 397 in 2010.

Yearly surname detection rates (1992-2010) varied between 0.9 and 0.968, with an arithmetic mean across all years of 0.934. A comparison of the results to the given name gender identification method is not useful because it yielded very poor results for Azerbaijan names (0.720). The high mean detection rate of the surname method in conjunction with the results of comparing the two methods for Russian names, which use the same rules, indicates that the surname method works well enough to use for Azerbaijan.

3.1.2 Productivity

Azerbaijan's share of female primary authors is the lowest among the countries compared and does not increase notably, remaining between about 10 and 20 %, these figures being subject to much variation due to the low overall number of publications. The share of female last authors in author teams shows medium values, increasing from below 10 % to above 35 %. Women's contribution remains comparatively low, however it increases from about 10 to about 25 %. Likewise, the contribution of women relative to their presence among publishing scientists is comparatively low. Men's productivity per author is higher than women's throughout all years, with modest effect sizes. The average values for both genders decrease only slightly over time. Participation rates for men remained virtually constant, while those for women increased from below 20 to above 40 %.

3.1.3 Collaboration

Sole authorship rates for men were higher for all years expect 2010. The differences were not statistically significant. The team size with regard to the gender of the first author showed no clear differences, the only significant result being a higher value for men in 2010. Female authors' average collaborative coefficient values were above those of men throughout all years, with modest to high effect sizes. For the four years for which enough data was available, the team composition with respect to the gender of the first author shows relatively low τ values (between 0.11 and 0.40), indicating a low correlation.

3.1.4 Citation analysis

No significant difference in shares of uncited papers could be observed between the groups. Due to the scarcity of data, no citation analysis could be carried out.

3.2 Bulgaria

3.2.1 Data

The dataset for publications with contributors from Bulgaria comprises 41229 items from 1980 to 2010. 455 items were excluded due to having more than 100 authors; 2354 items were excluded because no usable author gender attributes could be determined. The number of articles increased from 1017 in 1980 to 1829 in 2010.

Yearly surname detection rates (1980-2010) varied between 0.908 and 0.977, mean: 0.94, whereas the given name detection rates (2006-2010) lied between 0.864 and 0.886, mean: 0.876. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0854 and -0.0363 and a mean difference of -0.0531. This suggests that the surname method yields moderately accurate results for Bulgaria, the given name method calculates approximately 5 % higher values for the f:f+m ratio. Since the surname method strongly outperforms the given name method in detection rate, its results are likely to be more accurate.

3.2.2 Productivity

Bulgaria's share of female first authors is continuously very high, rising from approximately 30 to above 50 %. The share of female last authors in group authored articles is high as well, increasing from below 30 %, after a phase of stagnation from 1980 to 1991, to 40 % in 2010. Women's contribution shows the highest value for almost all years among the countries investigated, increasing from 29 % in 1980 to 47 % in 2010. Female contribution relative to the share of women among publishing scientists is likewise comparatively very high, with values at circa 85 % from 1980 to 1992 and values around 90 % since 1995. Male productivity per author outperforms womens' throughout all years. The difference between groups slowly decreased, while both groups' values also decreased. Male participation dropped slightly over the observation period, while female values rose from 45 to 65 %.

3.2.3 Collaboration

The share of sole authorship was continuously higher for male authors, the difference remaining nearly constant, while the values for both groups decreased. Women-led teams are generally of larger size than male-led teams, the difference in average team size is becoming larger over time. Female authors' average collaborative coefficient was also higher than male authors', but the difference is small. The relative composition of teams by gender in dependence of the primary author's gender show moderately high values (around 0.5) throughout all years, indicating moderate positive correlation.

3.2.4 Citation analysis

Differences in the shares of papers remaining uncited are small and relatively balanced between gender groups, with four statistically significant higher values for the female group between 1982 and 1992, and four for the male group between 1999 and 2005. Similarly, for the impact by gender of first author, no clear continuous difference emerges (9510 papers considered). These findings indicate that there are no distinct differences in impact due to author gender.

3.3 Czech Republic

3.3.1 Data

The dataset for publications with contributors from the Czech Republic comprises 66893 items from 1990 to 2010. 1071 items were excluded due to having more than 100 authors; 5549, were excluded because no usable author gender attributes could be determined. Articles from 1994 to 2010 were analyzed and plotted (66174 items). The number of articles increased from 2301 in 1994 to 7101 in 2010.

Yearly surname detection rates (1994-2010) varied between 0.846 and 0.881, mean: 0.857, whereas the given name detection rates (2006-2010) lied between 0.982 and 0.987, mean: 0.984. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between 0.00073 and 0.0117 and a mean difference of 0.00604. This suggests that the surname method yields very accurate results for the names from the Czech Republic.

3.3.2 Productivity

The share of female first authors steadily rose from below 20 to about 30 %, while the share of female last authors in team-authored papers increased from 15 to 20 % between 1994 and 2000 and has remained on that level since. Women's contribution increased from 16 to 27 %. Meanwhile, the female contribution relative to the share of women among active authors remained commensurate between 1994 and 2005 at a comparatively very low value of about 0.75 and even decreased to 0.71 from 2005 to 2010. The male group had higher productivity per author throughout all years. Female participation increased from 26 to 48 %, while male participation slowly decreased from 96 to 92 %. The values for all productivity metrics place the Czech Republic in the range of relatively low female involvement.

3.3.3 Collaboration

Shares of male sole authorship were higher than those of female authors in all years with small effect sizes between -0.07 and -0.13. Team size by gender of first author was higher for female-led teams throughout all years, as was the average collaborative coefficient, the differences being substantive enough to be statistically significant in most years. Correlation of team structure by gender of first author was around 0.5 for all years, indicating a moderate positive correlation. The results indicate that author gender is not generally a factor in citation success for Czech authors.

3.3.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years, the male group's value being higher at significant levels in 2003 and 2005. The figures for impact by gender of first author are also very close, with one significant difference for the male group in 1999 (8329 papers considered).

3.4 Czechoslovakia

3.4.1 Data

No distinction is made between the political entities referred to as Czechoslovakia (Czechoslovak Socialist Republic, Czechoslovak Federative Republic/Czech and Slovak Federative Republic) in the time period investigated. The dataset for publications with contributors from Czechoslovakia comprises 47564 items from 1980 to 1994. 22 items were excluded due to having more than 100 authors; 3781 were excluded because no usable author gender attributes could be determined. Articles from 1980 to 1993 were analyzed and plotted (47419 items). The number of articles were 3300 in 1980 and 3069 in 1993, when numerous articles were labelled to have originated in its successor states. The highest number of publications, 3543, was reached in 1989.

Yearly surname detection rates (1980-1993) varied between 0.866 and 0.885, mean: 0.874. Very high accuracy of the surname detection method is assumed for Czechoslovakia from the results for the Czech Republic and Slovakia.

3.4.2 Productivity

In the period of 1980 to 1993 the share of female first authors only slightly increased from 19 to 22 %, that of female last authors in groups remained relatively stable between 15 and 20 %. The share of women's contribution grew slightly as well, from 18 to 21 %. The contribution of women relative to the share of women among publishing scientists dropped from 0.8 to 0.75 in 1980 to 1982 and recovered to 0.8 in 1993. Productivity per author was continuously higher for the male group. Participation rates for both groups remained nearly at the same level, women: 32, men 91 %. All measures of productivity show only small changes in the time period of 1980 to 1993.

3.4.3 Collaboration

The share of sole authorship is higher for articles written by male authors, with the difference not changing over the years. Average team sizes for teams led by male and female authors were nearly equal in all years, with only very small effect size values in favour of female-led teams in two years. The average collaborative coefficient was slightly, but significantly, higher in all years for female authors. Correlation of relative team structure with the gender of first author shows comparatively low values in Czechoslovakia, indicating that the gender composition is influenced more by the relative ratios of genders of the authors available than by preference for any gender of coauthors among the authors. All results indicate only minor differences between genders for collaboration.

3.4.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years. The figures for impact by gender of first author are also very close, with significantly different values in three years for the male group, in 1981, 1987 and 1991 (16804 papers considered). These findings point to only minute if any differences in impact between female and male authors.

3.5 Greece

3.5.1 Data

The dataset for publications with contributors from Greece comprises 110073 items from 1980 to 2010. 1028 items were excluded due to having more than 100 authors; 6849 were excluded because no usable author gender attributes could be determined. The number of articles increased from 808 in 1980 to 8542 in 2010.

Yearly surname detection rates varied between 0.854 and 0.88, mean: 0.872, whereas the given name detection rates (2006-2010) lied between 0.824 and 0.84, mean: 0.833. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.116 and -0.095 and a mean difference of -0.11. This suggests that the surname method yields very poor results for Greece.

The cause of these differences is a shortcoming in the performance of the correct identification of names as female in the surname method. While the given name method has ca. 99 % rate of agreement with the given name method for male names, the rate for female names is only 95 %. This means a large number of female names are falsely classified as male because they do not have a female, but a male surname suffix. For Greece in the years when both methods can be used, the methods had directly opposing results in 1119 cases (not unique names). In 809 cases (72 %) the surname result was *male*, while the given name method returned *female*.

From these results it has to be concluded that the a surname based gender assignment is not possible within acceptable error rates for names from Greece.

3.6 Iceland

3.6.1 Data

The dataset for publications with contributors from Iceland comprises 5989 items from 1980 to 2010. 32 items were excluded due to having more than 100 authors; 1132 were excluded because no usable author gender attributes could be determined. The number of articles increased from 50 in 1980 to 568 in 2010. Yearly surname detection rates varied between 0.742 and 0.923, mean: 0.846, whereas the given name detection rates (2006-2010) lied between 0.895 and 0.941, mean: 0.92. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0194 and 0.0309 and a mean difference of 0.00113. This suggests that the surname method yields very good results for Iceland.

3.6.2 Productivity

The share of female first authors increased after a mostly stagnant phase in the first half of the 1980s (with, alas, very few data) from about 5 % to 43 % in 2010, a remarkable development. The share of female last authors in group authored articles increased from below 10 % throughout the 1980s to around 25 % from 2005 onwards. Women's contribution increased from very low values of about 5 % in the mid-1980s to 30 % from 2005 onwards. Women's contribution in relation to their relative presence among active authors, due to data scarcity, varied very strongly around 0.8. Men's productivity per author is slightly above women's, the difference being significant for many years from 1994 onwards, evincing small to medium size effect sizes. Participation rates of men slowly decrease from well above 95 % to 83 %, while that of women increased from below 10 % throughout much of the 1980s to 47 % in 2010.

3.6.3 Collaboration

The shares of sole authorship for male and female authors are not statistically significantly different, but men's are higher from the mid-1990s onwards, when that of women is very low. Team sizes for male and female-led teams are generally similar, with female-led teams being significantly larger in 1995, 1999 and 2004. Average collaborative coefficients are also similar for both groups, with that of female authors being a little higher in several years between 1994 and 2005. Too few data was available to compute meaningful values for the correlation between relative team structure and gender of first author.

3.6.4 Citation analysis

No significant differences between male and female first-authored papers could be observed in the shares of uncited papers and the impact measured in average normalized citation counts (442 papers considered).

3.7 Kazakhstan

3.7.1 Data

The dataset for publications with contributors from Kazakhstan comprises 2892 items from 1992 to 2010. 141 items were excluded due to having more than 100 authors; 614 were excluded because no usable author gender attributes could be determined.

The number of articles increased from 137 (1992) to 199 (2010), with large variations. Yearly surname detection rates varied between 0.863 and 0.958, mean: 0.904. The given name detection rate at 0.693 was too low to be used to compare the surname detection rate against. Good performance of the surname method is assumed by way of the results for Russia which is using the same rules.

3.7.2 Productivity

The share of female first authors varies strongly, but does not show any clear trend. The share of female last authors in groups increased from below 20 to 35 % throughout the observation period. The share of women's contribution did not show an obvious trend either. Women's contribution relative to the share of women among publishing scientists, in spite of large fluctuation, reveals rather high values for many years. Productivity per author was very close, with that of the male authors being significantly higher with moderate effect sizes in some years. The participation rate for men did not change notably, while that of women increased from below 40 to above 50 %.

3.7.3 Collaboration

No significant differences between male and female groups appear in both the share of sole authorship and the average team sizes by first author gender. Average collaborative coefficients are very similar in values as well, with those of female authors being significantly higher in 1994 and 2010. Correlation of relative team structure and gender of first author was not computed because not enough data was available.

3.7.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years. No citation impact analysis was performed due to data scarcity.

3.8 Latvia

3.8.1 Data

The dataset for publications with contributors from Latvia comprises 4034 items from 1991 to 2010. 3 items were excluded due to having more than 100 authors; 723 were excluded because no usable author gender attributes could be determined. The number of articles increased from 54 (1991) to 320 (2010).

Yearly surname detection rates varied between 0.775 and 0.907, mean: 0.85, whereas the given name detection rates (2006-2010) lied between 0.953 and 0.98, mean: 0.966. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0675 and 0.0139 and a mean difference of -0.0207. This suggests that the surname method yields good results for Latvia.

3.8.2 Productivity

The share of female first authors increased from around 30 % in the mid-1990s to 50 % (2007). Female last author share in group authored articles also increased strongly, from below 20 to above 35 %. Women's contribution increased from 27 % in 1992 to 39 % in 2010, with a maximum of 43 % (2004). Over the same time span, the female contribution relative to female presence among publishing scientists increased distinctly from below 0.7 to circa 0.8. Throughout all years, productivity per author is higher for men. Men's participation stayed nearly the same, women's increased from approximately 40 to 60 %, a relatively small difference compared to other countries.

3.8.3 Collaboration

While the share of sole authorship of male authors is higher than women's in all years, the only statistically significant difference is in 2007. In many years, average team sizes of female-led teams are significantly higher than those of male-led teams. Significantly higher values for female authors occur in the average collaborative coefficient for most years. Correlation between relative team composition and gender of first author shows small to large positive values, with no clear trend.

3.8.4 Citation analysis

The impact metrics show no distinct differences between male and female first authored papers in regards to uncitedness and average citation impact by gender of first author. The first author gender citation analysis covered 698 papers.

3.9 Lithuania

3.9.1 Data

The dataset for publications with contributors from Lithuania comprises 11091 items from 1993 to 2010. 33 items were excluded due to having more than 100 authors; 895 were excluded because no usable author gender attributes could be determined. The number of articles increased from 161 (1993) to 1461 (2010).

Yearly surname detection rates varied between 0.832 and 0.942, mean: 0.9, whereas the given name detection rates (2006-2010) lied between 0.922 and 0.946, mean: 0.931.

Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0682 and -0.016 and a mean difference of -0.0487. This suggests that the surname method yields moderate results for Lithuania, with a small bias of the surname method towards underestimating the proportion of female authors.

3.9.2 Productivity

Female first author share increased from values around 20 % in the 1990s to 37 % (2009). The share of female last authors in group authored papers likewise increased, from just 7 % in 1993 to 27 % in 2004, from that point on the values did not continue to increase. Women's contribution rose from 14 % in 1993 to 32 % in 2010, with only minor gains since 2003, while the female contribution in relation to women's relative presence among publishing scientists remained at a level of around 0.8, despite large variation. Throughout all years, male productivity per author outperformed that of women. Female participation increased from 21 to 54 %, while male participation decreased somewhat, from 91 to 87 %.

3.9.3 Collaboration

Shares of sole authorship are significantly higher for male authored papers from 2002 onwards. Average team size by gender of first author is significantly higher for female-led teams for 2007 to 2010. Average collaborative coefficients are for the most part very close, with values for female authors occasionally being significantly higher. As for the correlation between gender of the primary author and the gender composition of the rest of the team, the values for Lithuania are relatively high compared to other countries, indicating a somewhat stronger association.

3.9.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years for publications with male and female first authors respectively. The figures for impact by gender of first author reveal no significantly different values between male first-authored and female first-authored articles (824 papers considered).

3.10 Macedonia

3.10.1 Data

The dataset for publications with contributors from Macedonia comprises 1674 items from 1995 to 2010. No items had to be excluded due to very high numbers of authors. 212 items were excluded because no usable author gender attributes could be determined. The number of articles increased from 53 (1995) to 198 (2010).

Yearly surname detection rates varied between 0.758 and 0.94, mean: 0.89, whereas the given name detection rates (2006-2010) lied between 0.908 and 0.943, mean: 0.917. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0632 and 0.00633 and a mean difference of -0.0377. This suggests that the surname method yields good results for Macedonia

3.10.2 Productivity

Female first author shares are very high, but fluctuate strongly between 33 and 63 %, due to the low number of papers in each year. The share of female last authors in group authored papers increased from 8 % in 1995 to 39 % in 2010. Women's contribution rose from 33 to 43 %, while the female contribution relative to the female presence among publishing scientists showed a pronounced drop from 1998 (0.88) to 2004 (0.62) with a subsequent recovery to a value of 0.88 in 2008. Throughout all years, male productivity per author outperformed that of women, significantly so from 2003 to 2006 and in 2009. Female participation increased only modestly, being already at a comparatively high level (40 % in 1995), while male participation remained nearly unchanged.

3.10.3 Collaboration

Shares of sole authorship are higher for male authored papers, but the differences were not significant. Average team size by gender of first author is for most years very close for the two groups, with both groups having a single significantly higher value in one year. Average collaborative coefficients are for the most part very close, with values for female authors occasionally being significantly higher (1996, 2003, 2004). Due to too few items, no correlation between first author gender and gender composition of the other team members was calculated.

3.10.4 Citation analysis

The shares of uncited papers between papers with male and female first authors were close for all years, with no significant differences. No citation analysis was performed because not enough data was available.

3.11 Poland

3.11.1 Data

The dataset for publications with contributors from Poland comprises 144731 items from 1980 to 2010. 1794 items were excluded due to having more than 100 authors; 91043 were excluded because no usable author gender attributes could be determined. The number of articles increased from 3105 (1980) to 10830 (2010) with a low of 2500 in 1982.

Yearly surname detection rates varied between 0.507 and 0.613, mean: 0.559, whereas the given name detection rates (2006-2010) lied between 0.973 and 0.985, mean: 0.98. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0159 and 0.00538 and a mean difference of -0.00334. This suggests that the surname method yields surprisingly good results for Poland in spite of the low detection rate. The reason is that there is no inherent bias of the surname method to determine any one gender better than the other or incorrectly find names to be of the opposite gender if the suffixes that are to be used are carefully chosen to avoid very ambiguous ones. However, the percentage of names found to be of unknown gender is therefore high, leading to a high rate of excluded articles.

3.11.2 Productivity

The share of female first authors decreased from 1980 (31 %) to 1991 (25 %) and increased from that point to 45 % in 2010. The pattern is similar for the share of female last authors in group authored papers, decreasing from 31 to 21 % and then rising up to 34 %. Women's contribution also shows a similar development, a slow drop from 31 to 25 % and a recovery to 41 %. On the other hand, the female contribution relative to the share of women among publishing scientists remained relatively stable at comparatively medium-large values of between 0.80 and 0.85 throughout the three decades of observation. Productivity per author is continuously higher for men, the difference, as expressed in effect size, remaining at similar values. Participation once more echoes the drop in women's involvement seen in the other metrics up to the beginning of the 1990s (from 39 to 32 %, then rising to 54 % in 2010).

3.11.3 Collaboration

The share of sole authorship among articles first-authored by men is continuously much higher than that of women. Average team sizes in respect to the gender of the first author were approximately equal until 1998, when female-led teams become significantly larger in average. Women also showed higher average collaborative coeffients then men, the difference, as expressed in effect size, remained at relatively constant small levels. Correlation between gender of first author and composition of the rest of the team showed modest positive values, which are likely very slowly decreasing over time.

3.11.4 Citation analysis

No clear differences in the shares of uncited papers are apparent, though the share of male-authored ones is higher by small margins from 1995 onwards. For the most part, the impact by gender of first author is not different between articles with male and female primary authors, except in 1991, 1992 and 2004, when the male groups show small but significantly higher values (17492 papers considered).

3.12 Slovakia

3.12.1 Data

The dataset for publications with contributors from Slovakia comprises 25709 items from 1991 to 2010. 602 items were excluded due to having more than 100 authors; 2808 were excluded because no usable author gender attributes could be determined. Articles from 1994 to 2010 were analyzed and plotted. The number of articles increased from 1396 (1994) to 2063 (2010).

Yearly surname detection rates varied between 0.854 and 0.904, mean: 0.89, whereas the given name detection rates (2006-2010) lied between 0.97 and 0.975, mean: 0.972. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.00992 and 0.0308 and a mean difference of 0.0102. This suggests that the surname method yields very good results for Slovakia.

3.12.2 Productivity

From 1994 to 2010 Slovakia's share of female first authors increases from 30 to 41 %, while the share of female last authors in group authored publications does not increase notably, remaining between 25 and 30 % until 2006, and staying consistently at around 30 % from then on. Women's contribution increased from 28 to 36 %. However, the contribution of women relative to the share of women among publishing scientists decreased sligthly from 0.86 to 0.83. Productivity per author was continuously significantly higher for male authors in all years with small to moderate effect sizes. While male participation remained approximately commensurate, female participation increased from 41 to 55 %.

3.12.3 Collaboration

The share of solely authored articles by men is higher than that of women in all years. Team size by gender of first author is significantly higher for female-led teams starting from 2000. The average collaborative coefficients of female authors were slightly higher then male authors' for all years except 2006. Values for correlation between gender of first author and team composition are medium-scale.

3.12.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years, with no statistically significant differences. The figures for impact by gender of first author are also very close, with one significant difference in favour of the male group in 1996 (4648 papers considered).

3.13 Uzbekistan

3.13.1 Data

The dataset for publications with contributors from Uzbekistan comprises 4981 items from 1991 to 2010. 20 items were excluded due to having more than 100 authors; 743 were excluded because no usable author gender attributes could be determined. The number of articles increased from 89 (1991) to 273 (2010).

Yearly surname detection rates varied between 0.861 and 0.947, mean: 0.913. Given name detection rates were to low to be meaningful in a comparison, but in the light of very high surname detection rates and good results with similar names for Russian data, the surname method is used with confidence.

3.13.2 Productivity

The share of female first authors decreased from around 25 % in the early 1990s to under 20 %, with large variations due to few data. The share of female last authors in group authored publications does not increase or decrease notably, remaining between 15 and 25 %. Women's contribution decreased from above 20 to approximately 15 %. Female contribution relative to the presence of women among publishing scientists is comparatively high at times, but does not reveal any clear trend over time. Productivity per author was higher for male authors in all years with small effect sizes. Participation remained approximately commensurate for both female and male authors. These figures portend relatively little involvement of women.

3.13.3 Collaboration

Shares of sole-authored articles are not significantly different. Team size by gender of first author was significantly higher for female-led teams only in 2006. Collaborative coefficient of female and male authors was very close in all years except for a higher value for female authors in 2006. Values for correlation between gender of first author and team composition are high compared to other countries, indicating a moderately strong association.

3.13.4 Citation analysis

The shares of uncited papers were nearly equal or very close for all years, with statistically significant differences in favour of female first-authored papers in 2001 and 2003, in other words, significantly higher shares of articles by women primary authors than male authors from these two years remained uncited. Impact analysis in respect to first author gender was not executed due to a lack of sufficient data.

3.14 Russian SFSR and Russian Federation

3.14.1 Data

The data from Soviet Russian times had to be cleaned of items that were from the USSR but do not have any contributors with addresses in the Russian SFSR, such as those from the Baltic states and others. To only get items with Russia-based authors it is not sufficient to simply keep just those which have “Russia” in its address field because in many cases the political entity , i.e. union republic, within the USSR is not specified, though it often is. If it is not specified, the authors are assumed to be from Russia. However, if the name of a union republic other than the Russian SFSR appears and “Russia” does not appear in the same field, it can be safely concluded that no Russian authors contributed and the item can be excluded from the final dataset.

The selection was performed as follows in GNU bash (version 4.1.5) and awk (v. 3.1.7):

```
awk -F’    ’ ’toupper($22) !~ /KAZAKHSTAN|UKRAINE|BEL(O|A)RUS|
    UZBEKISTAN|GEORGIA|AZERBAIJAN|LITHUANIA|MOLDAVIA|LATVIA|KYRG|
    TAJIKISTAN|ARMENIA|TURKMENISTAN|ESTONIA/’ all.compl.txt > all.
    excl.non.russia.txt
awk -F’    ’ ’toupper($22) ~ /KAZAKHSTAN|UKRAINE|BEL(O|A)RUS|
    UZBEKISTAN|GEORGIA|AZERBAIJAN|LITHUANIA|MOLDAVIA|LATVIA|KYRG|
    TAJIKISTAN|ARMENIA|TURKMENISTAN|ESTONIA/’ all.compl.txt > all.
    non.russia.txt
awk -F’    ’ ’toupper($22) ~ /RUSSIA/’ all.non.russia.txt > all.
    non.russia.with.also.russia.txt
cat all.excl.non.russia.txt all.non.russia.with.also.russia.txt >
    all.txt
wc -l *
     403447 all.complete.txt
     335435 all.excl.non.russia.txt
      68012 all.non.russia.txt
       8079 all.non.russia.with.also.russia.txt
     343514 all.txt
    1158487 total
```

Note that there is a single tab between the single quotes in the awk commands.

The dataset for publications with contributors from Soviet Russia and Russia comprises 674290 items from 1980 to 2010. 3476 items were excluded due to having more than 100 authors; 85269 were excluded because no usable author gender attributes could be determined. The number of articles increased from 20260 (1980) to 22527 (2010) with a minimum of 17574 (2006).

Yearly surname detection rates varied between 0.787 and 0.825, mean: 0.803, whereas the given name detection rates (2006-2010) lied between 0.941 and 0.95 %, mean: 0.944. Comparing the surname method's and given name method's results for the fraction of female names to female plus male names shows differences between -0.0326 and -0.0107 and a mean difference of -0.0187. This suggests that the surname method yields very good results for Russian names.

3.14.2 Productivity

The share of female first authors very slowly decreases from 25 to 21 % from 1982 to 1997 and after that increases to 27 % in 2007. The share of female last authors in group authored publications does not increase notably, being at 25 % in 1982, dropping to 19 % in 1994 and slowly recovering to 26 % in 2010. Women's contribution likewise slowly dropped from 25 to 20.5 % from 1980 to 1994 and increased to 28 % in 2010. Productivity per author was continuously significantly higher for male authors in all years with moderate effect sizes. While male participation did not change, female participation remained nearly commensurate as well from 1980 to 1989 (at 39 %), dropped and recovered throughout the 1990s with a low of 32 % (1994) and reached 47 % in 2010.

3.14.3 Collaboration

The share of solely-authored articles by men was higher than that of women in all years, most markedly throughout the 1990s. Team size by gender of first author is significantly higher for female-led teams, but with only small effect sizes. Average collaborative coefficients of female authors were slightly higher then male authors'. Values for correlation between gender of first author and team composition were comparatively low, indicating weak association.

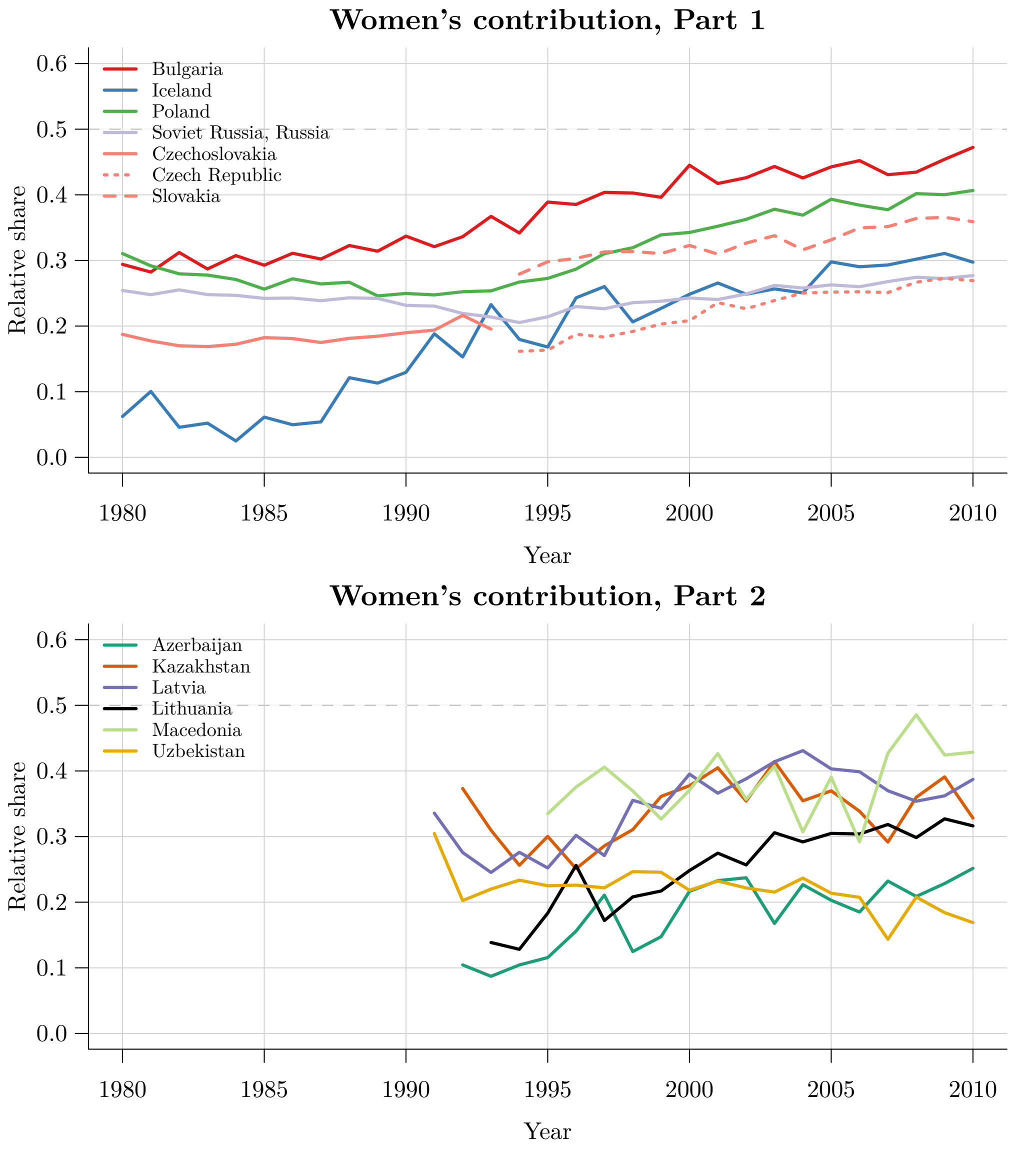
3.14.4 Citation analysis

The shares of female uncitedness were significantly higher than those of men for most years, with the difference getting smaller over time until the values equalized after 2000. In all years the effect sizes were very small. Male first-authored articles reached higher average citation scores with small effect sizes in all years (140546 papers considered).

3.15 Comparison

After having discussed the individual countries in detail, selected metrics of women's involvement will now be presented together for all countries. Not statistically significant values do not appear in the figures, so that gaps in the timelines occur.

Women's contribution is a robust metric of relative productivity. In the overview charts, figure 3.15, part 1 shows those countries with data dating back to 1980 and part 2 those with data from after 1990. Countries with comparatively strong involvement of women are Bulgaria, Poland, Macedonia and Latvia. On the other hand, the Czech Republic, Russia, Azerbaijan and Uzbekistan have little female involvement. While for most countries the contribution of women has increased since the beginning of observations, this is not the case for Uzbekistan. Moreover, there have been phases of stagnation or slow decrease of the figures of women's contribution and subsequent increase in a number of countries: the 1980s in Czechoslovakia, the period form 1995 to 2006 in Macedonia and the first half of the 1980s in Iceland show very little change, while longer phases of decline and subsequent slow rebound occured in Poland and Soviet Russia and Russia. There is a pronounced difference in the relative shares of women's contribution between Slovakia and the Czech Republic.



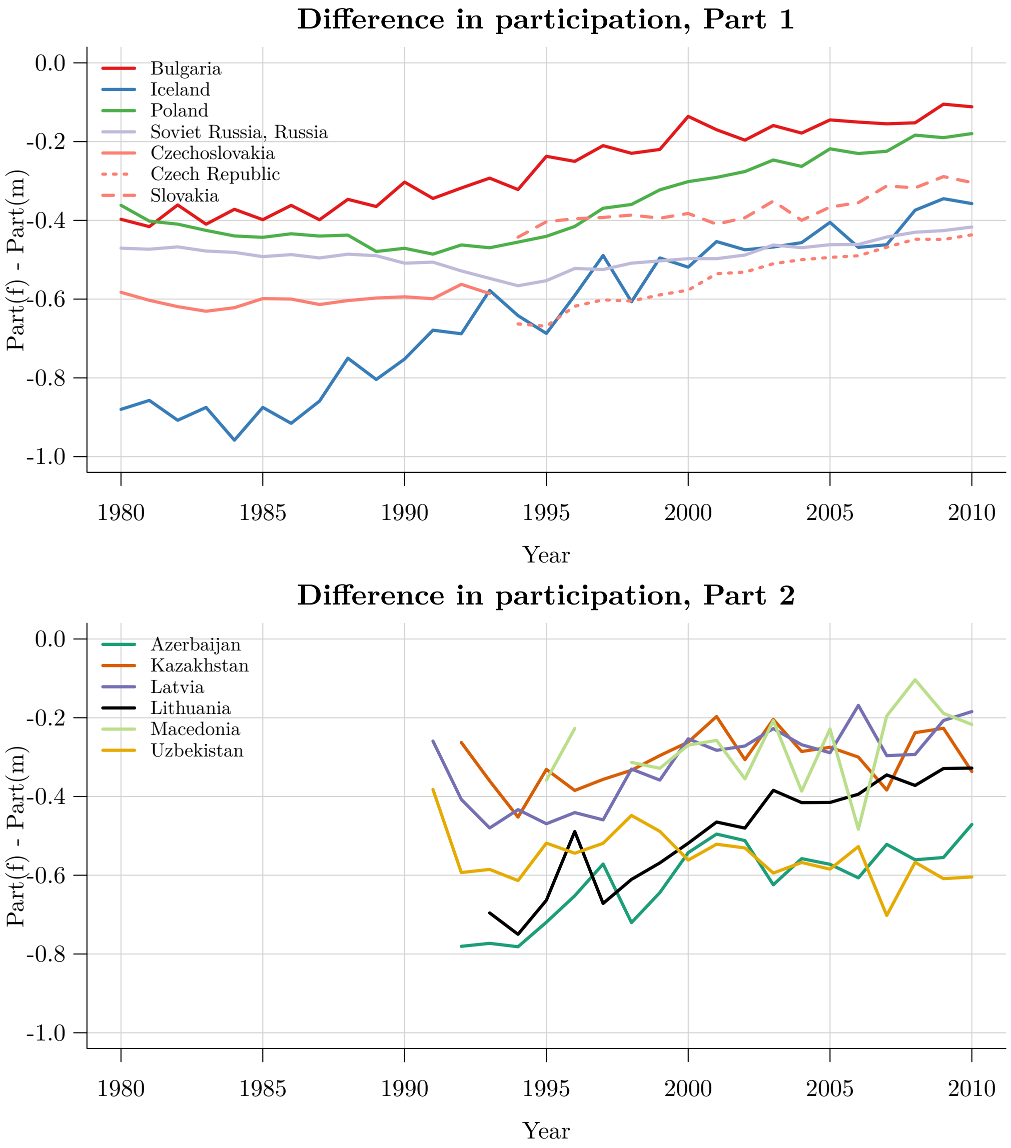
Women’s contribution, Part 1

Women’s contribution, Part 2

Abbildung 3.1: Comparison of women’s contribution

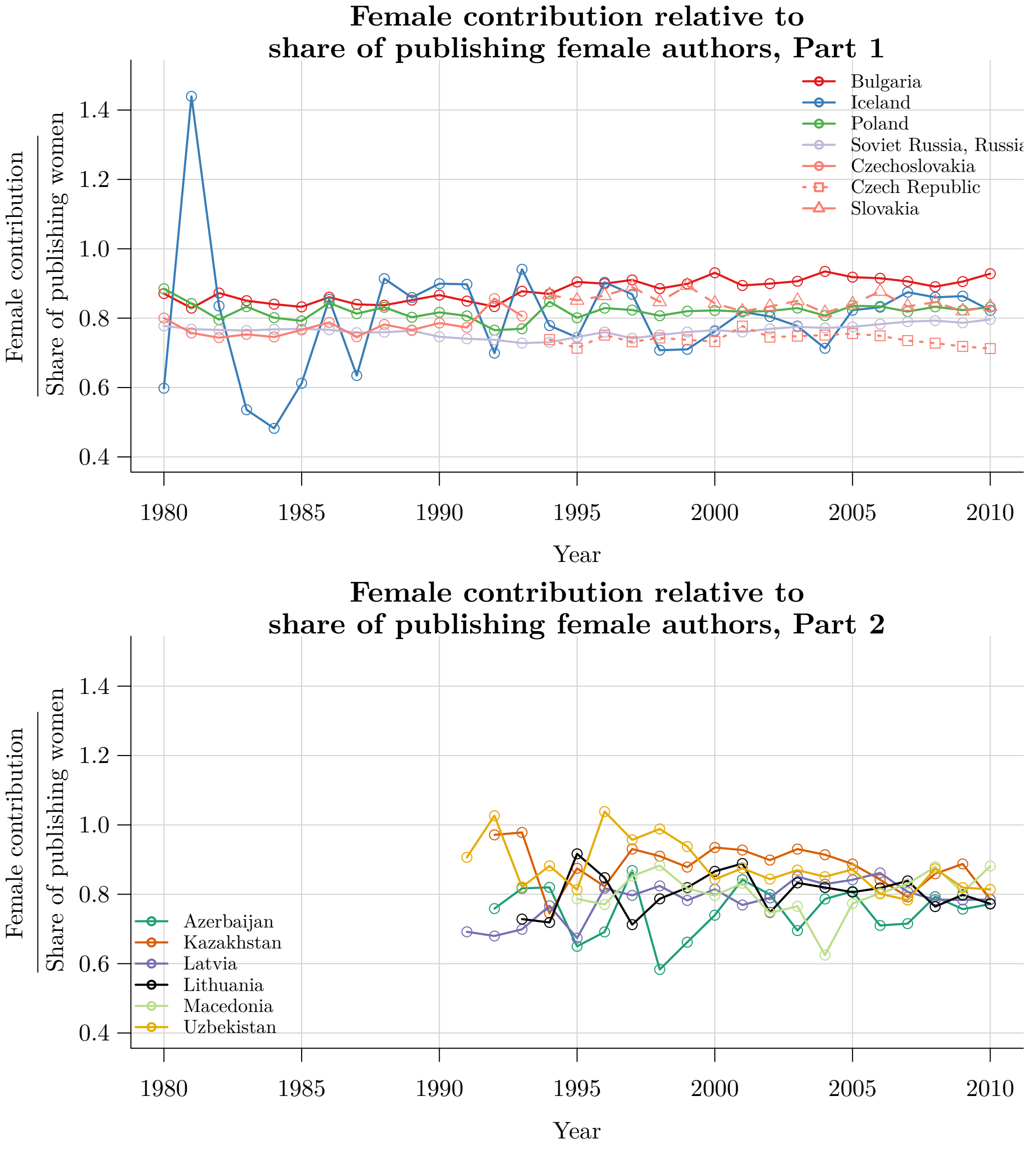
A similar pictures emerges when reviewing all countries' values for participation, here as difference between female and male value of participation (figure 3.15). A value of 0 would indicate that among all papers, the number with at least one female author and the number with at least one male author were equal, while a negative value indicates less papers with at least one female author then with at least one male author, and a positive value vice-versa, though none such values did occur. The overall trend is continuous diminishing of the difference. Once more the 1980s were marked by stagnation in Czechoslovakia, continuing to some degree in Slovakia as far as 2004. Iceland's and Bulgaria's values were not increasing for the first half of the 1980s. The differences in Soviet Russia and Russia and Poland increased until the first half of the 1990s and decreased from that point on. The differences in Uzbekistan and Kazakhstan did not decrease when regarding the entire period for which data exists. Azerbaijan, Latvia, Lithuania, the Czech Republic and Macedonia showed clear trends towards a decrease in participation difference throughout the entire observation period.

The extent of the higher average productivity of male authors is small but very stable for most countries. Small decrease over long periods are apparent in Azerbaijan, Bulgaria and Soviet Russia/Russia.



Comparison of difference in participation

Comparing all countries for the female contribution relative to the share of women among actively publishing male and female scientists shows once more high figures occurring in Bulgaria while the other countries are very close together for the most part, at values around 0.8. Soviet Russia/Russia and the Czech Republic's figures are comparatively low.

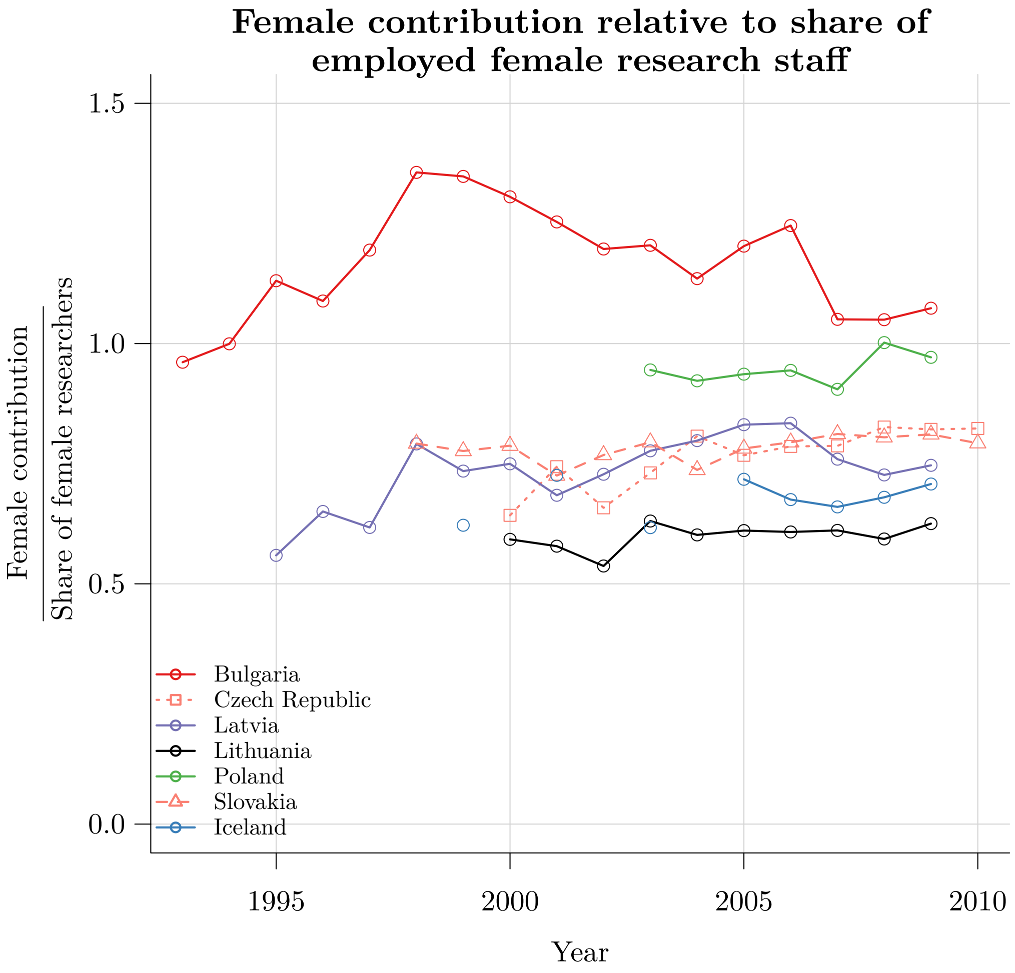


Comparison of women's contribution relative to the share of women among publishing authors

Another similar metric is the female contribution in relation not to publishing women scientists, but the actual relative shares of female scientists from external workforce statistics. The figures for shares of women among scientists in the higher education in full time equivalent employees was obtained from Eurostat[1] as available. This includes figures for Bulgaria, the Czech Republic, Latvia, Lithuania, Poland, Slovakia and Iceland, starting from 1993. Dividing the figure for women's contribution by the share of

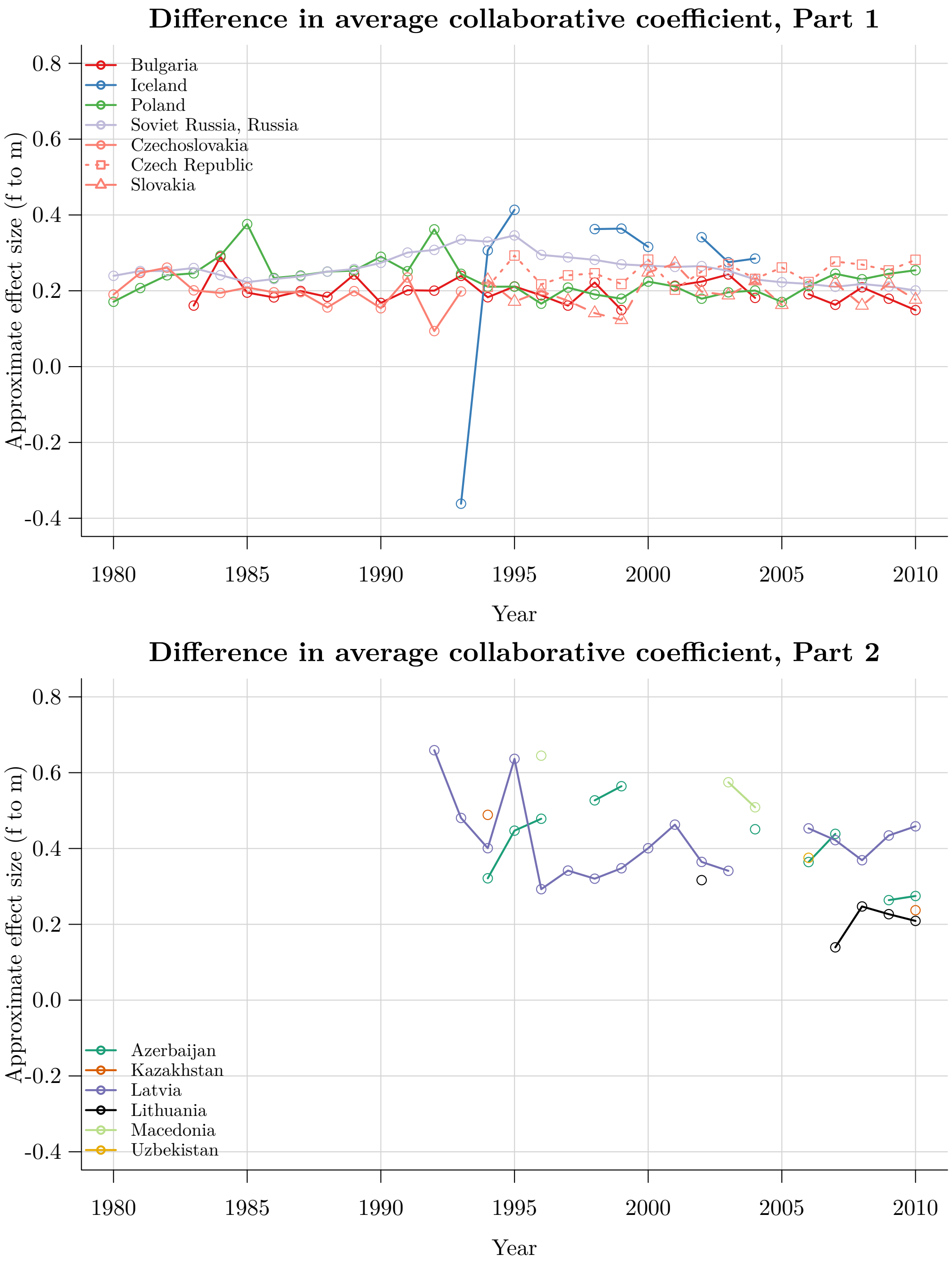
[1] `http://epp.eurostat.ec.europa.eu/portal/page/portal/product_details/dataset?p_product_code=RD_P_FEMRES`

female employees, in the same way as by female publishing scientist, yields an indicator that when below 1 indicates that the contribution of women is lower than would be expected from their share of workers, or when above 1 higher than the share of workers suggests, with the improvement over the aforementioned similar indicator being that non-publishing scientists are also included. The two countries whose women scientist make a publication contribution as high as to be expected are Bulgaria, whose women are much more productive than those of any other country investigated, and Poland, which has values close to 1 in all years. The Czech Republic, Slovakia, Lavia, Lithuania and Iceland fall between 0.5 and 0.8 with no clear upward or downward trends.



Comparison of women's contribution relative to the share of women in the workforce

Viewing all the values for the average collaborative coefficients of authors across all countries (figure 3.15), expressed in approximate effect size, together, reveals that the countries' author populations are much alike in this particular metric. With one exception, the value from Iceland in 1993 which can be considered an outlier, the yearly averages of female authors are higher, resulting in positive effect size values. The differences appear to be very stable across time and countries, remaining at typically between 0.2 and 0.6.



Comparison of difference in average collaborative coefficient

Male authors have generally higher rates of sole authorship, the difference between male and female groups stays almost level in all countries except Soviet Russia/Russia over time in spite of major decrease of overall sole authorship rates. In Russia, the male sole authorship rose distincly from 1989 to 1995 and returned to former levels subsequently until up to 2006, which led to an evident temporary increase in difference to the female group.

The general trend in team sizes is for female-led team to be of larger average size, with the differences becoming more pronounced over time. This is the case for Bulgaria, the Czech Republic, Lithuania, Poland, Slovakia and Soviet Russia/Russia.

Only for Russia are there substantive, though very small, differences in the shares of papers that remained uncited and the average impact of papers with female and male first authors respectively.

4 Discussion

The methodological basis of past studies could be extended by incorporating further countries with surname forms different by gender, improving the details of the surname suffixes used and closely evaluating surname-based gender recognition methods against given name gender recognition. This allowed for approximate assessment of quality of surname gender recognition for several countries and fine-tuning of surname suffix lists, and in the case of Greece, dismissing the usefulness of surname-based gender recognition entirely for the present data.

One main obstacle for studies such as this one is gathering data of sufficient quality (both given name and surname gender assignment is comparatively straightforward), especially with respect to proper text encoding and detailed field structure. In general, using given names for gender assignment is preferable when it is possible, however, the results can be checked against and improved with surname method-derived results for the countries covered in this study. In cases of missing given names, gender-ambigiuous given names of given names not contained in the used database the surname method is very useful. In the present study, this was evident for the countries Uzbekistan, Azerbaijan and Kazakhstan, whose given names were not very well represented in the given name database used, so that the surname gender assignment method outperformed the given name method in these cases.

The findings of Lewison (2001) for Iceland can be partly confirmed and expanded. The suspected levelling off of women's involvement in science, i.e. relative productivity, did not occur, but female contribution continued to increase, albeit slower (measured in contribution instead of overall f:f+m authorships). Lewison found little difference in the numbers of citations to male and female authored papers in a subset comprising of clinical medicine papers. The present study's citation analysis, covering all subject areas, did not reveal any statistically significant advantages in average normalized citation counts for publications with male first authors.

The output of Polish female scientists has been observed by Webster (2001) to have suffered a substantial drop from 1980 onwards, especially in the first half of the 1990s and a subsequent recovery. The present study can further substantiate these findings and shows that the increase (as observed in both share of female first authors and women's contribution) continued unimpeded until 2010. More than in other countries, the difference in participation in Poland has shrunk substantially.

Although average sole authorship rates of both women and men generally tend to dwindle over time, in the same year and country, women are less likely to be sole authors than men, as measured in the relative shares of publications with one author among all publications with female or male primary authors, in most countries investigated. In contrast, McDowell and Kiholm Smith (1992) report that among economics PhD holders, men were more likely to be sole authors.

Average team sizes of author teams led by men and women both increase with time, but female-led teams were observed to grow faster in countries where the difference is statistically significant. Furthermore, the average collaborative coefficients of women are overwhelmingly often higher then their male colleagues, though mostly only by small margins. Substantial correlation between the relative gender composition of author teams and the primary author's gender was found, which is in accordance to findings by McDowell and Kiholm Smith (1992) and Boschini and Sjögren (2007), but this particular topic demands further in-depth investigation.

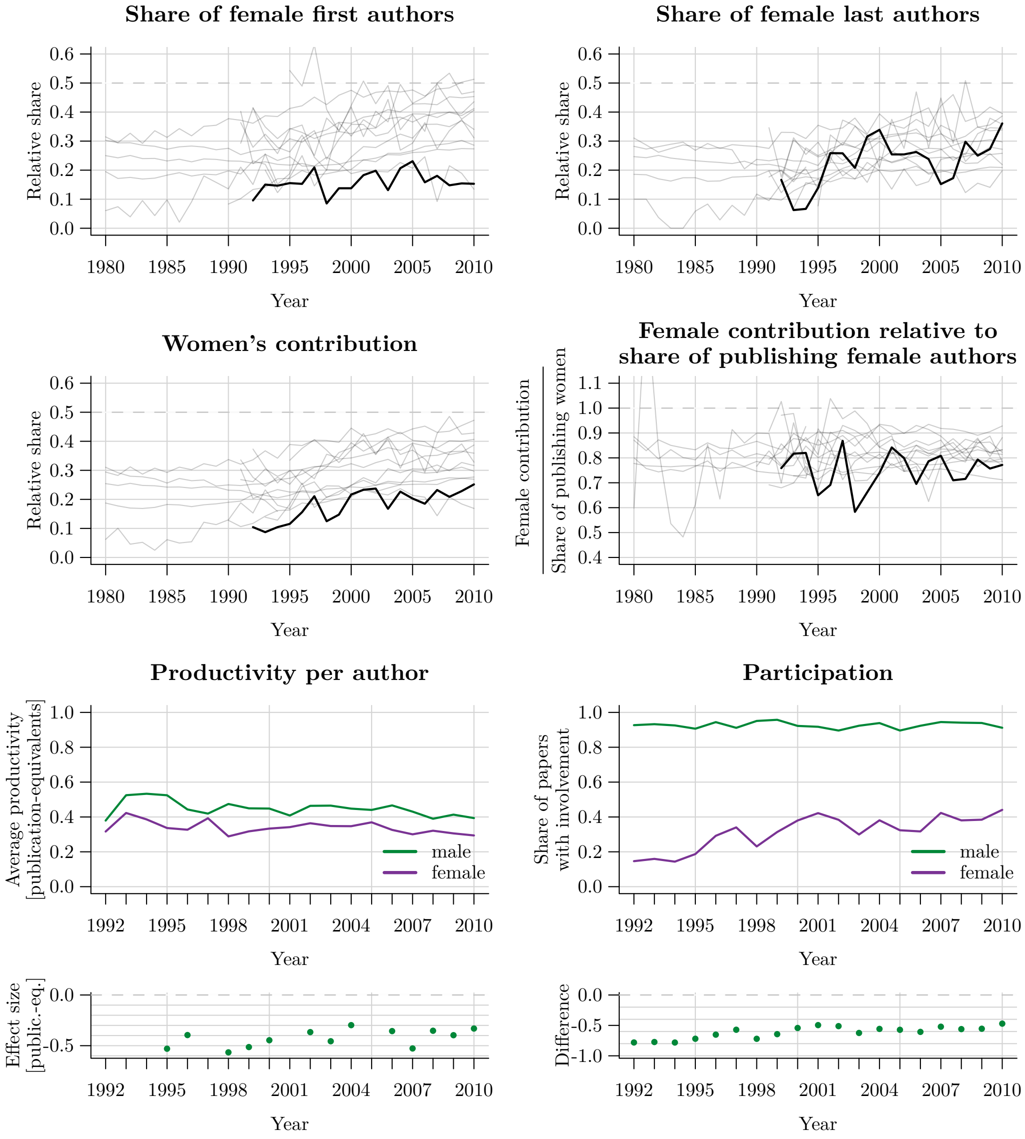
Evidence for differences in impact as measured in normalized citation counts is scarce. Only for Soviet Russia and Russia could consistent statistically significant differences between publications first-authored by women and men be found at all. These differences were minute in value. Likewise, no consistent differences in uncitedness could be found between publications with male and female first authors other than for Soviet Russia and Russia, where the values of the differences were also only very small.

In spite of using a different methodology and data, the results of the present study are in line with those of Lewison and Markusova (2011), in that female Russian scientists are not cited as much as their male colleagues. Lewison and Markusova analysed only papers in which all authors are from Russia, with any publication type and did so seperately for each major area of science, whereas in the present study all articles with Russian primary authors were considered, if the gender of the primary author could be determined, with the publication type being limited to articles, letters and notes, thus excluding reviews. The method of gender assignment also differed, with this study having a specific set of female and male suffixes as opposed to denoting all names without female suffixes (with exceptions) as male. Furthermore, the method for obtaining a male and female citation scores was different. Nonetheless the findings presented here confirm Lewison and Markusova's primary result: women are not as highly cited as men but the difference is very small. In addition, the data presented confirms these differences as being long-lasting and not being subject to change over the time observed, i.e. the higher average impact of publications with male first authors is very consistent over 25 years, as is the difference expressed in effect size to the average impact of female first-authored publications. Similar differences did not occur in the figures for any other country in this study.

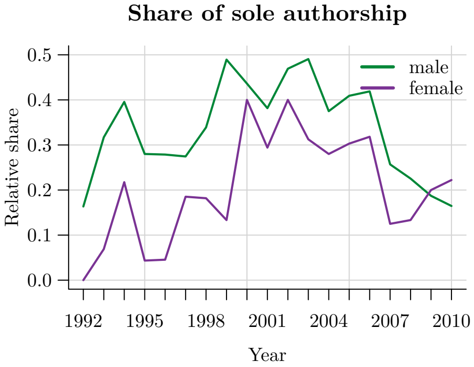
Figures

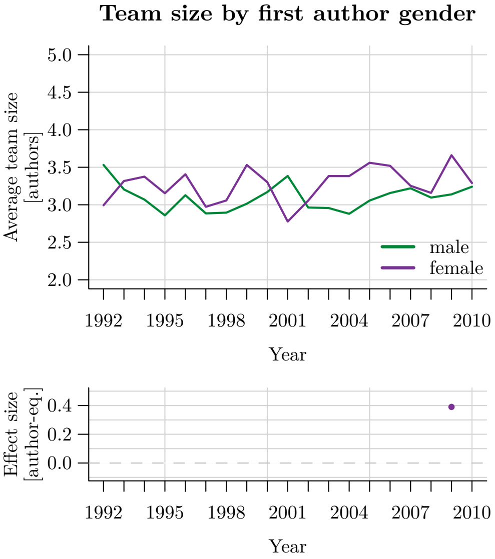
1 Azerbaijan

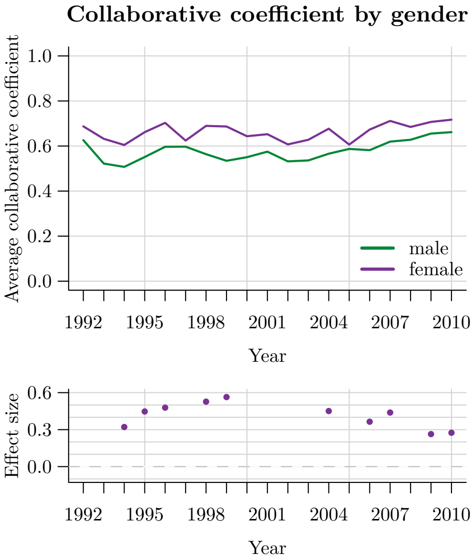
1.1 Productivity

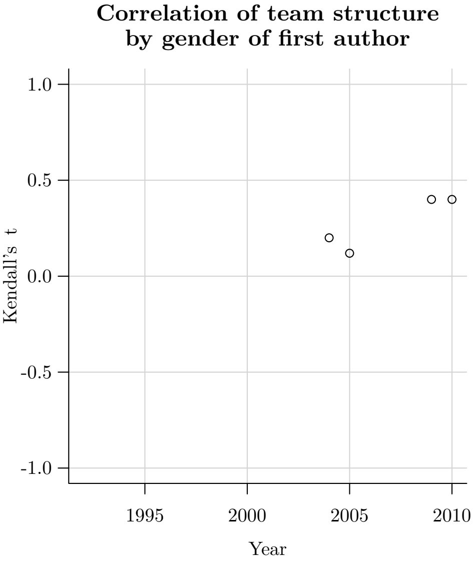


1.2 Collaboration

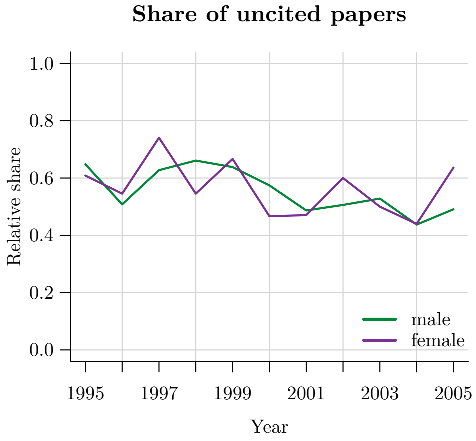

Share of sole authorship


Team size by first author gender


Collaborative coefficient by gender

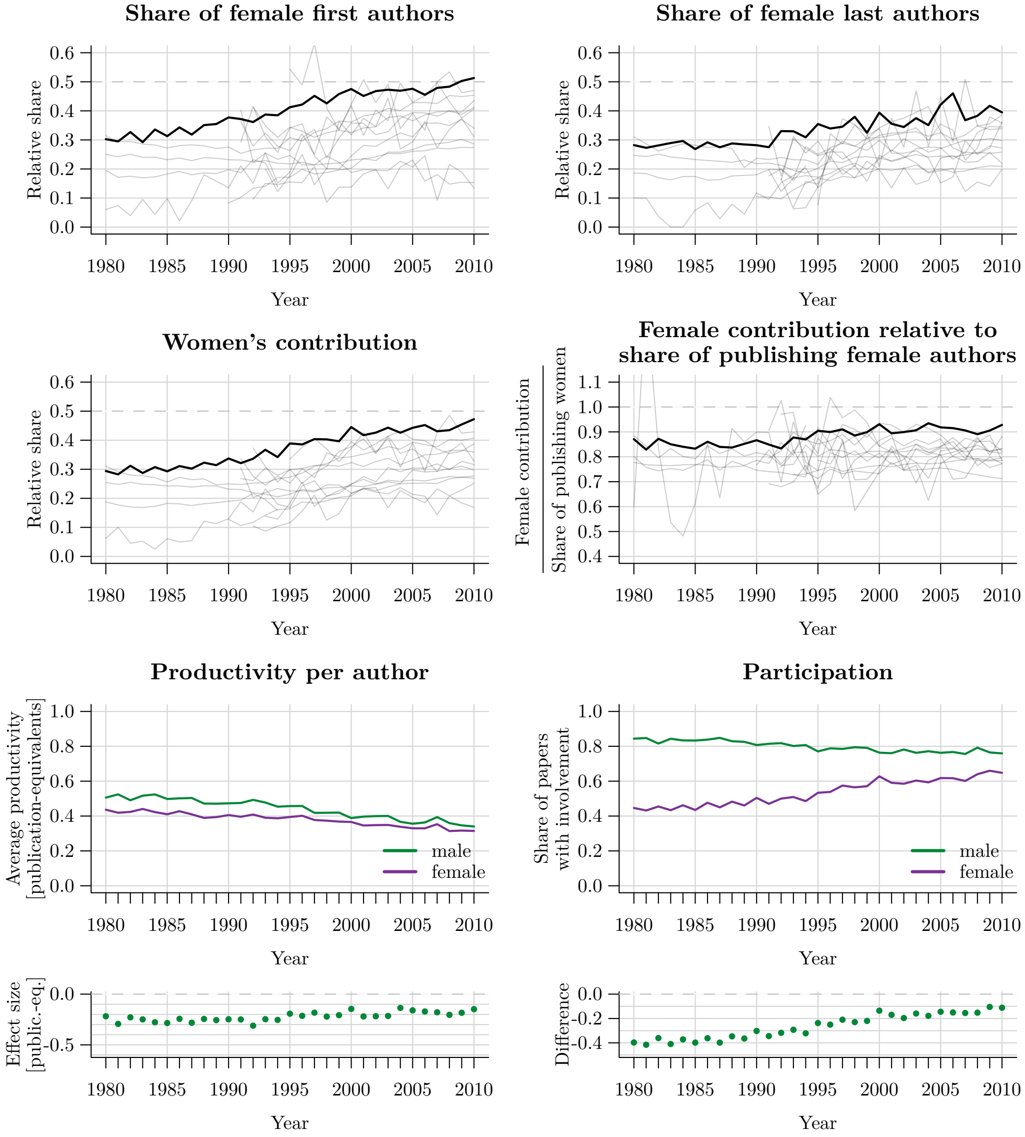

Correlation of team structure by gender of first author

1.3 Citation analysis

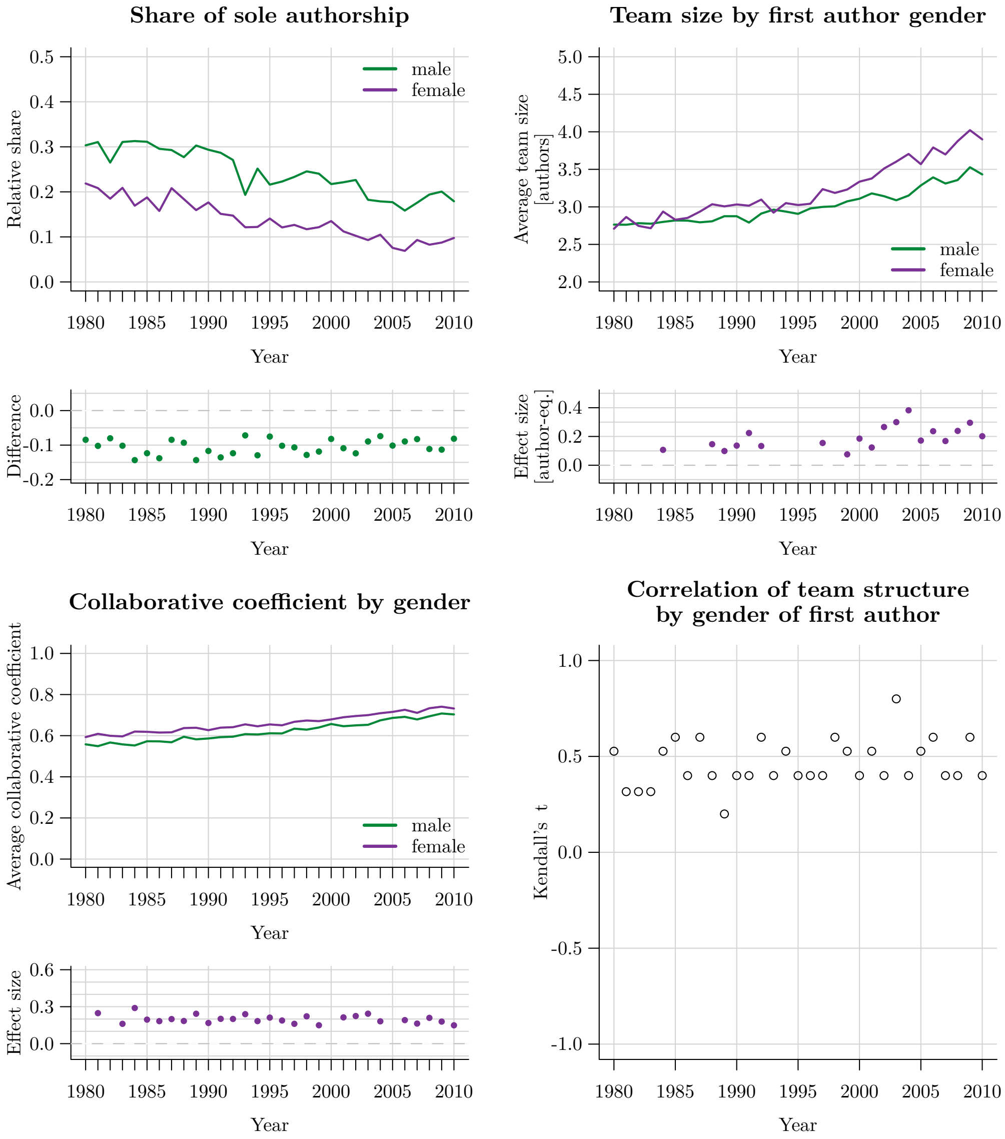


2 Bulgaria

2.1 Productivity



2.2 Collaboration



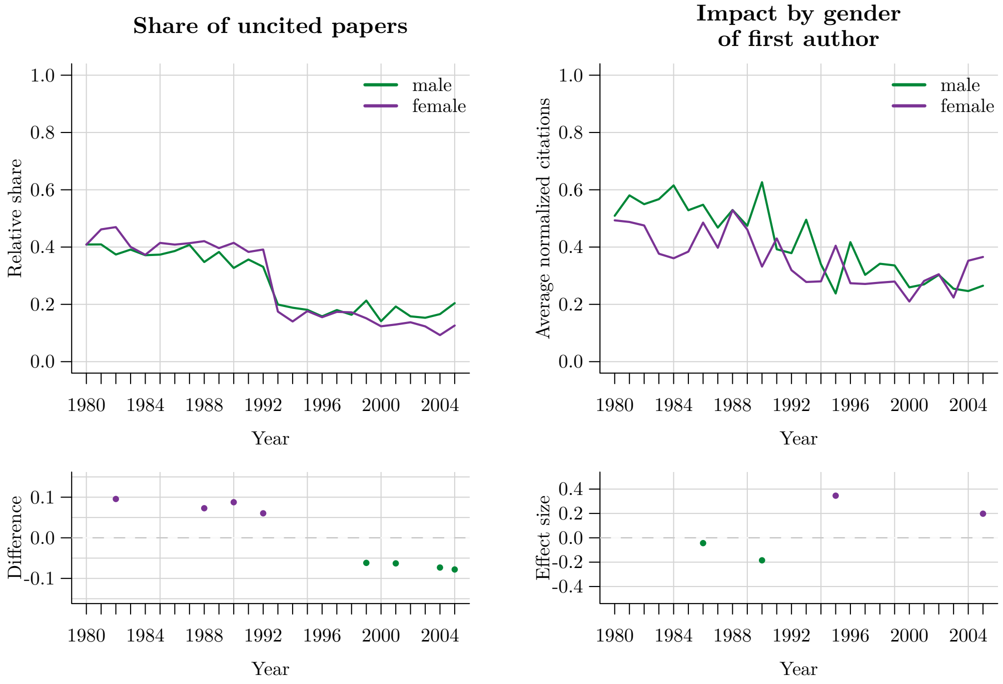
Share of sole authorship

Team size by first author gender

Collaborative coefficient by gender

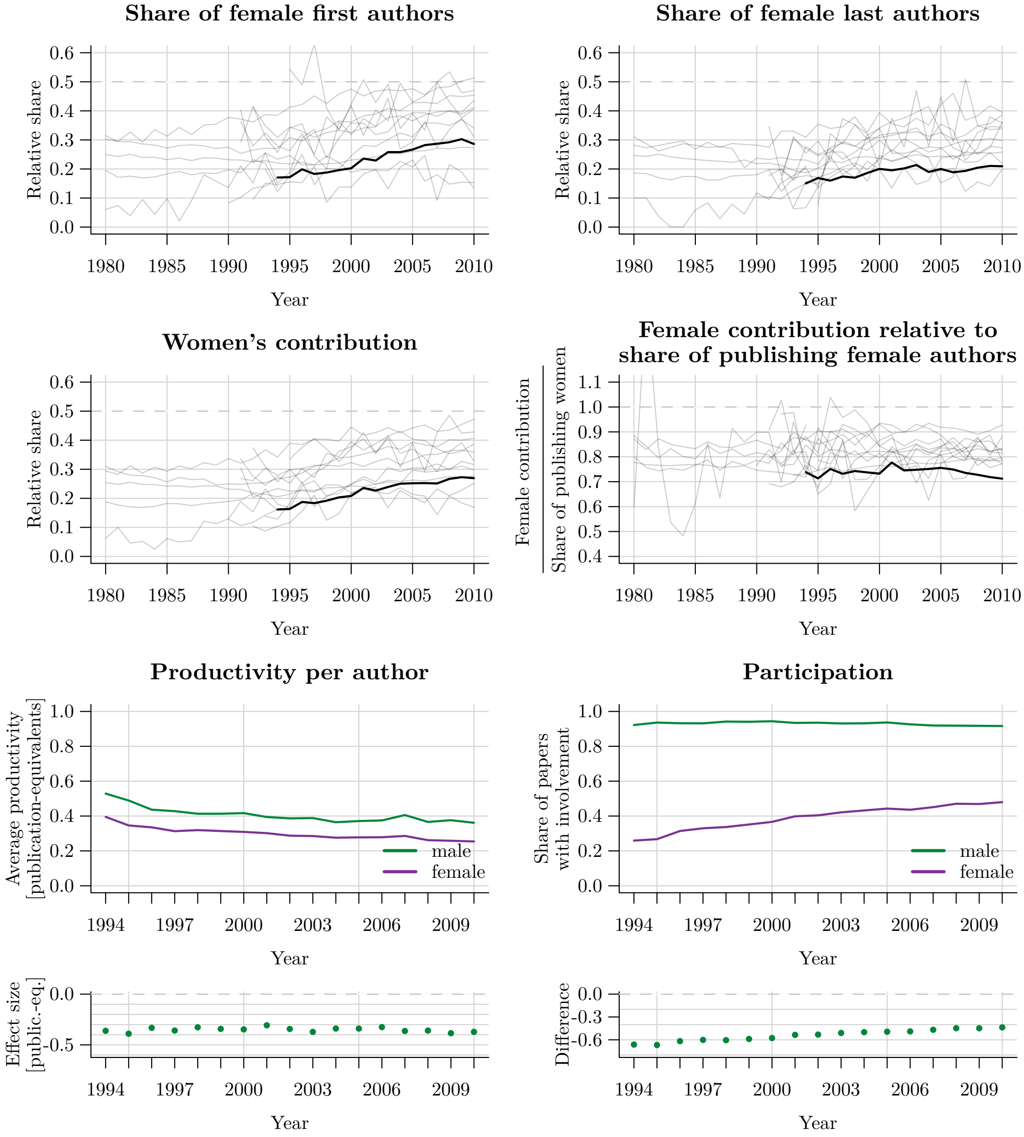
Correlation of team structure by gender of first author

2.3 Citation analysis



3 Czech Republic

3.1 Productivity



Share of female first authors

Share of female last authors

Women's contribution

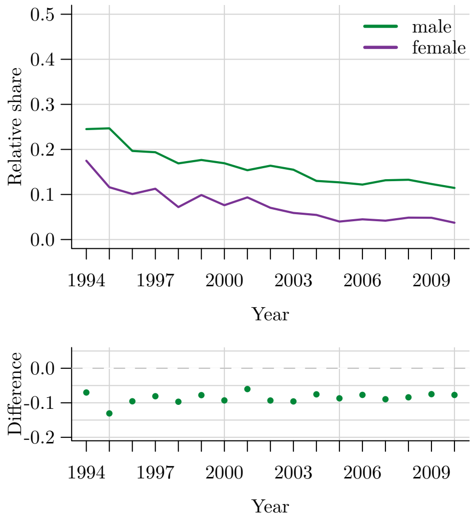
Female contribution relative to share of publishing female authors

Productivity per author

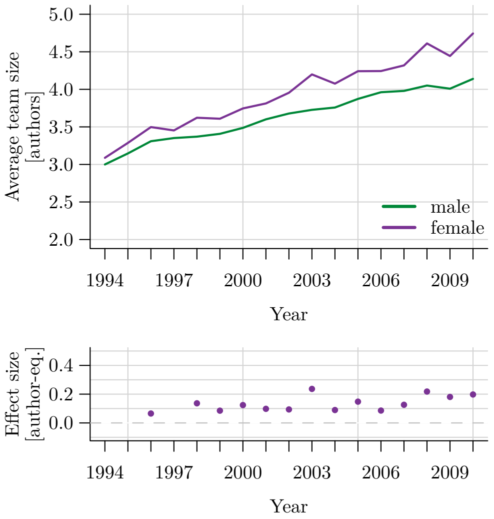
Participation

3.2 Collaboration

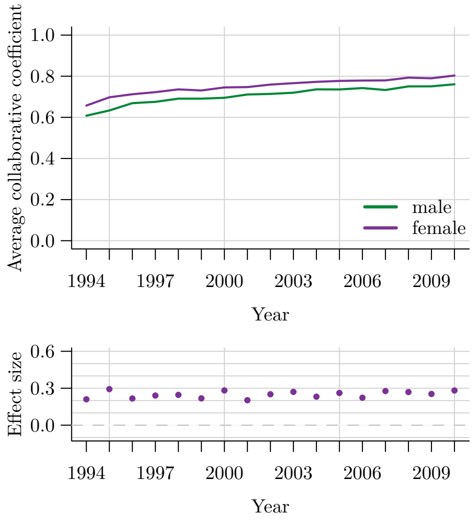
Share of sole authorship



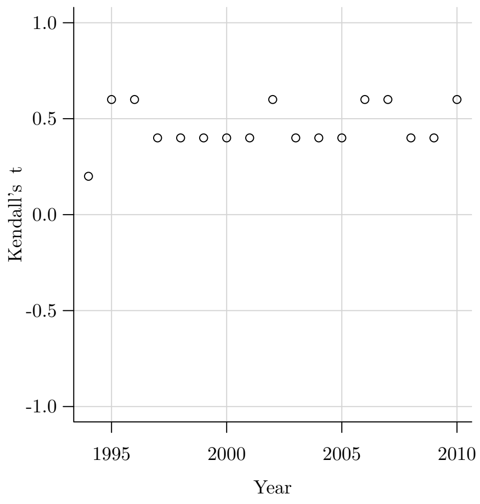
Team size by first author gender



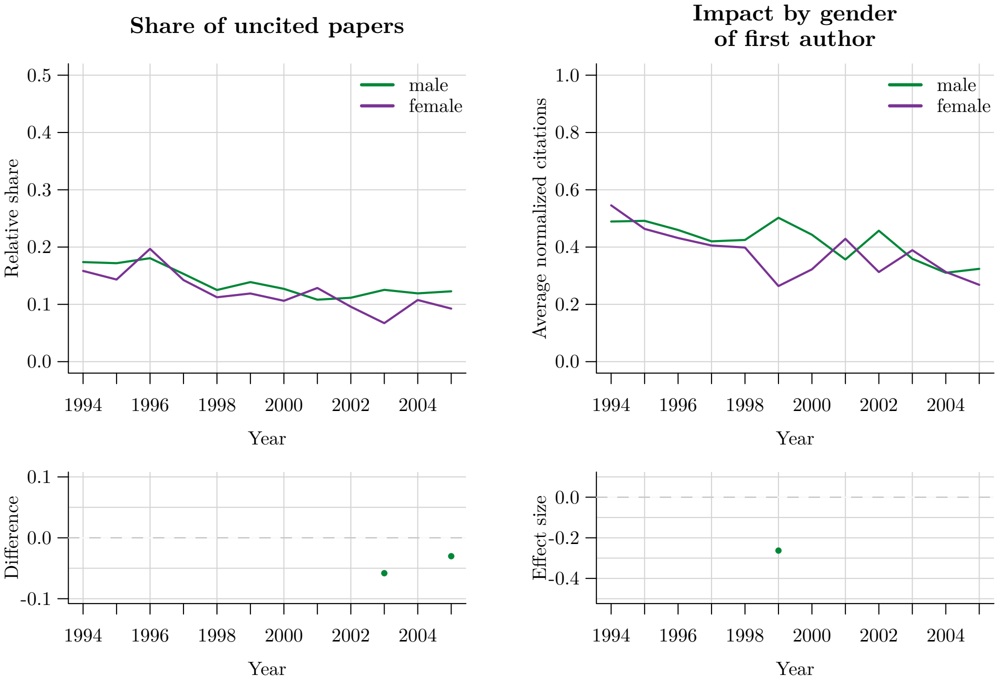
Collaborative coefficient by gender



Correlation of team structure by gender of first author

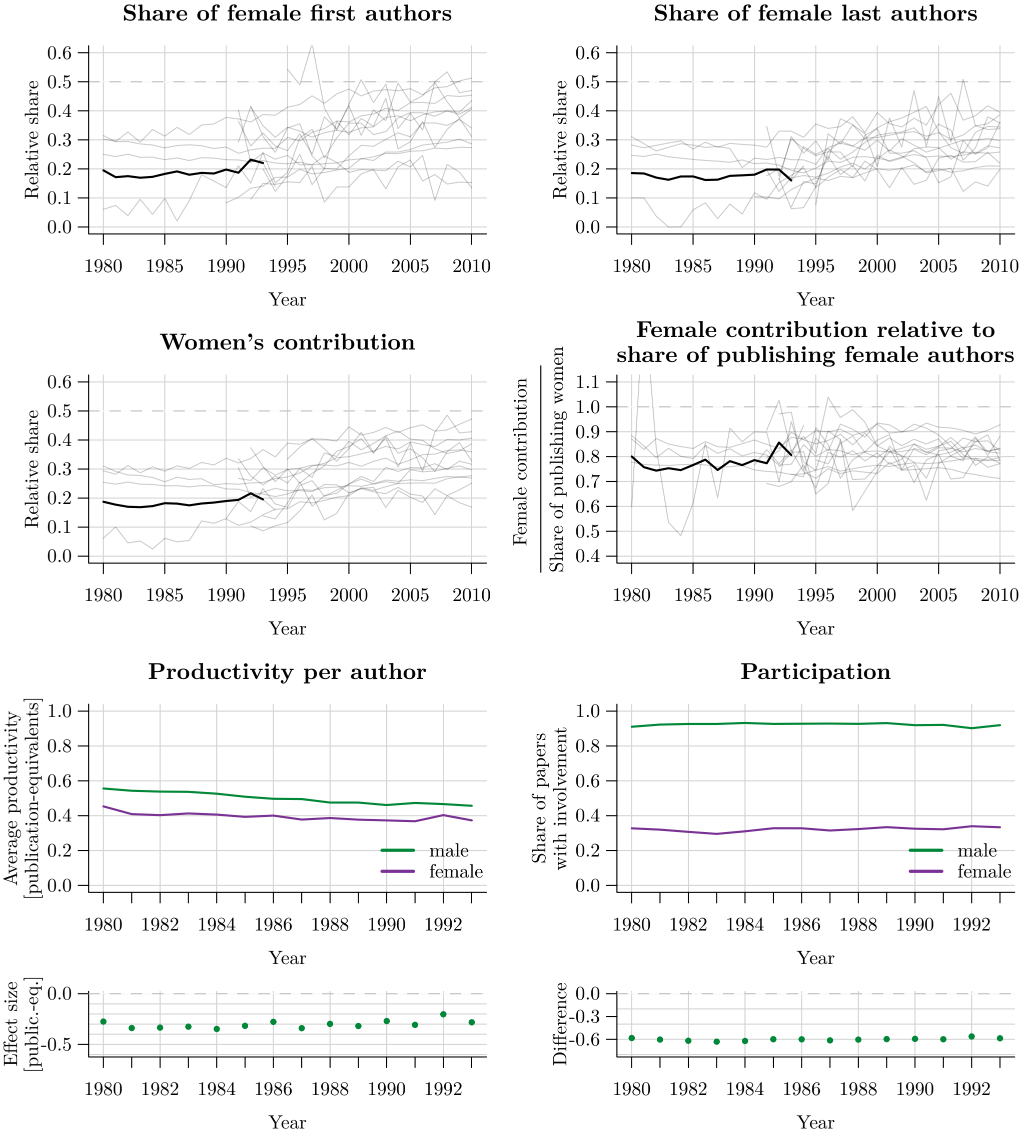


3.3 Citation analysis

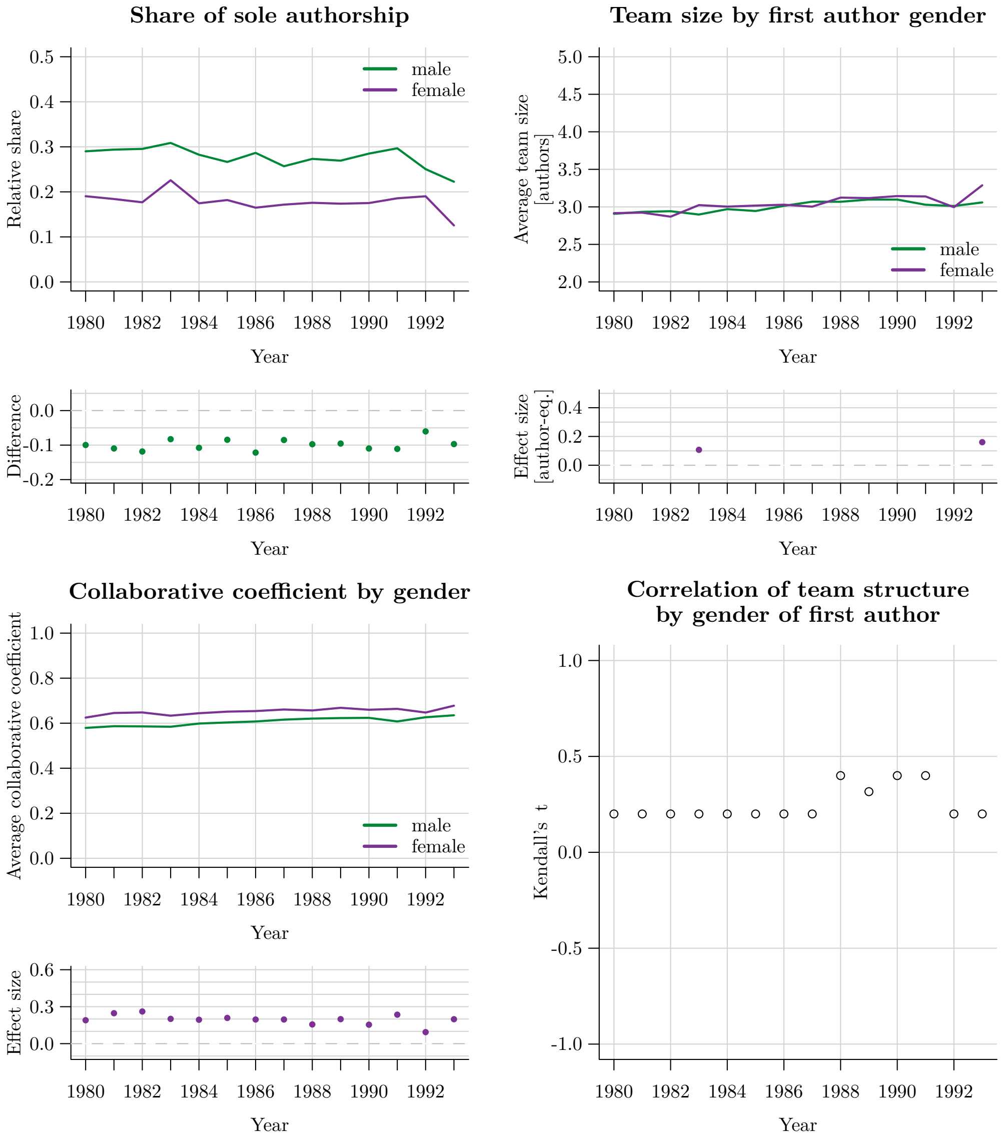


4 Czechoslovakia

4.1 Productivity



4.2 Collaboration



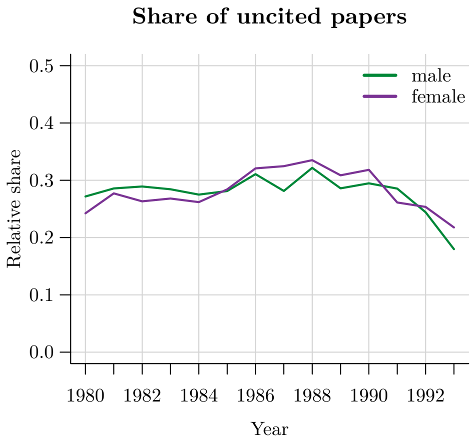
Share of sole authorship

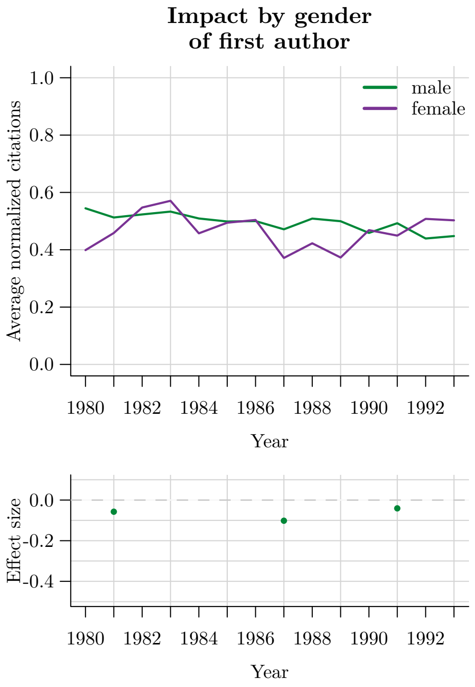
Team size by first author gender

Collaborative coefficient by gender

Correlation of team structure by gender of first author

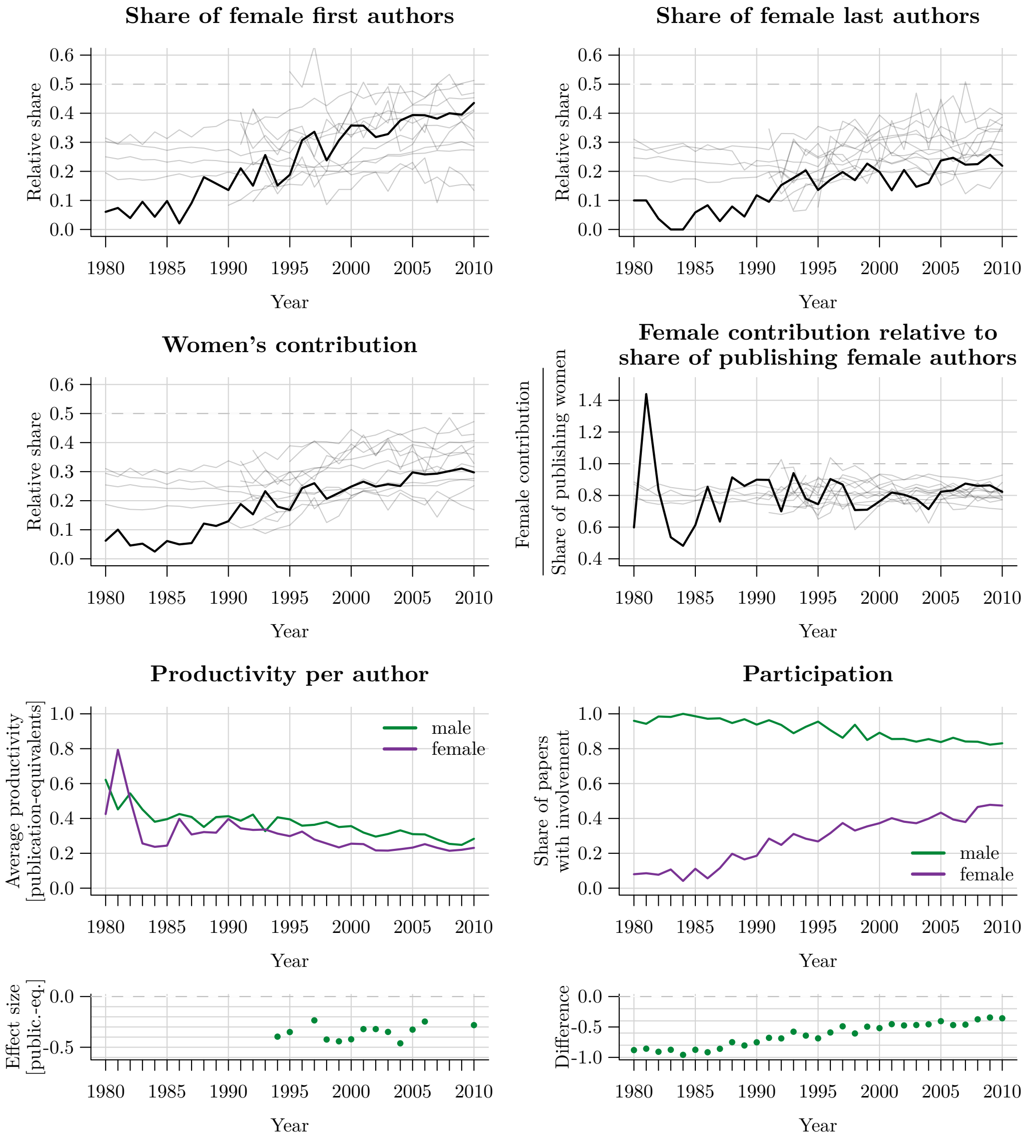
4.3 Citation analysis


Share of uncited papers


Impact by gender of first author

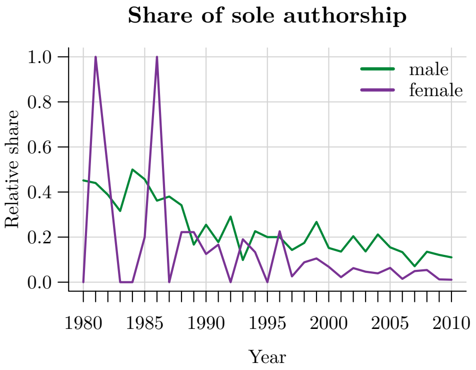
5 Iceland

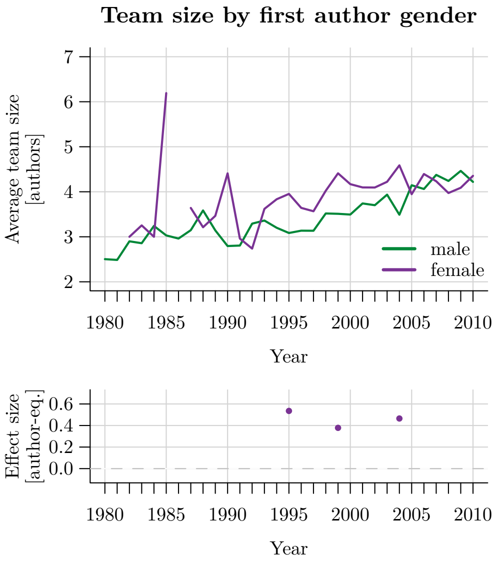
5.1 Productivity

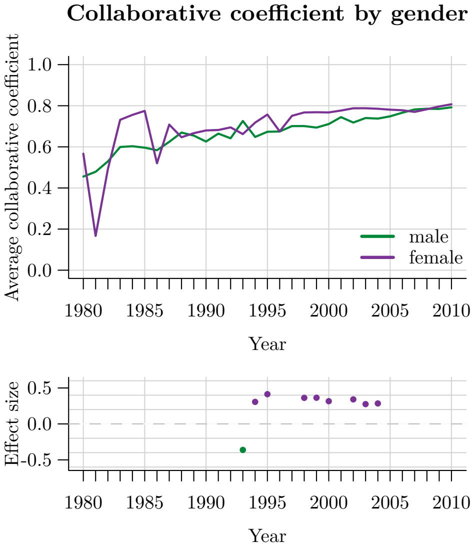


o

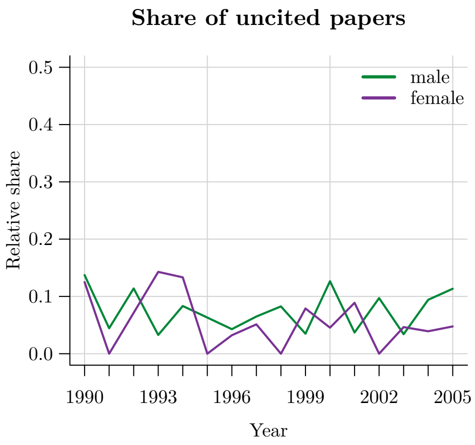
5.2 Collaboration

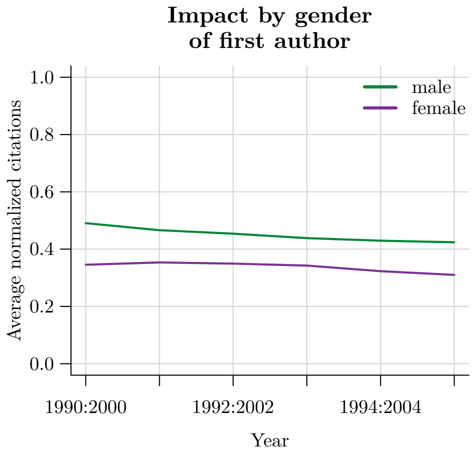






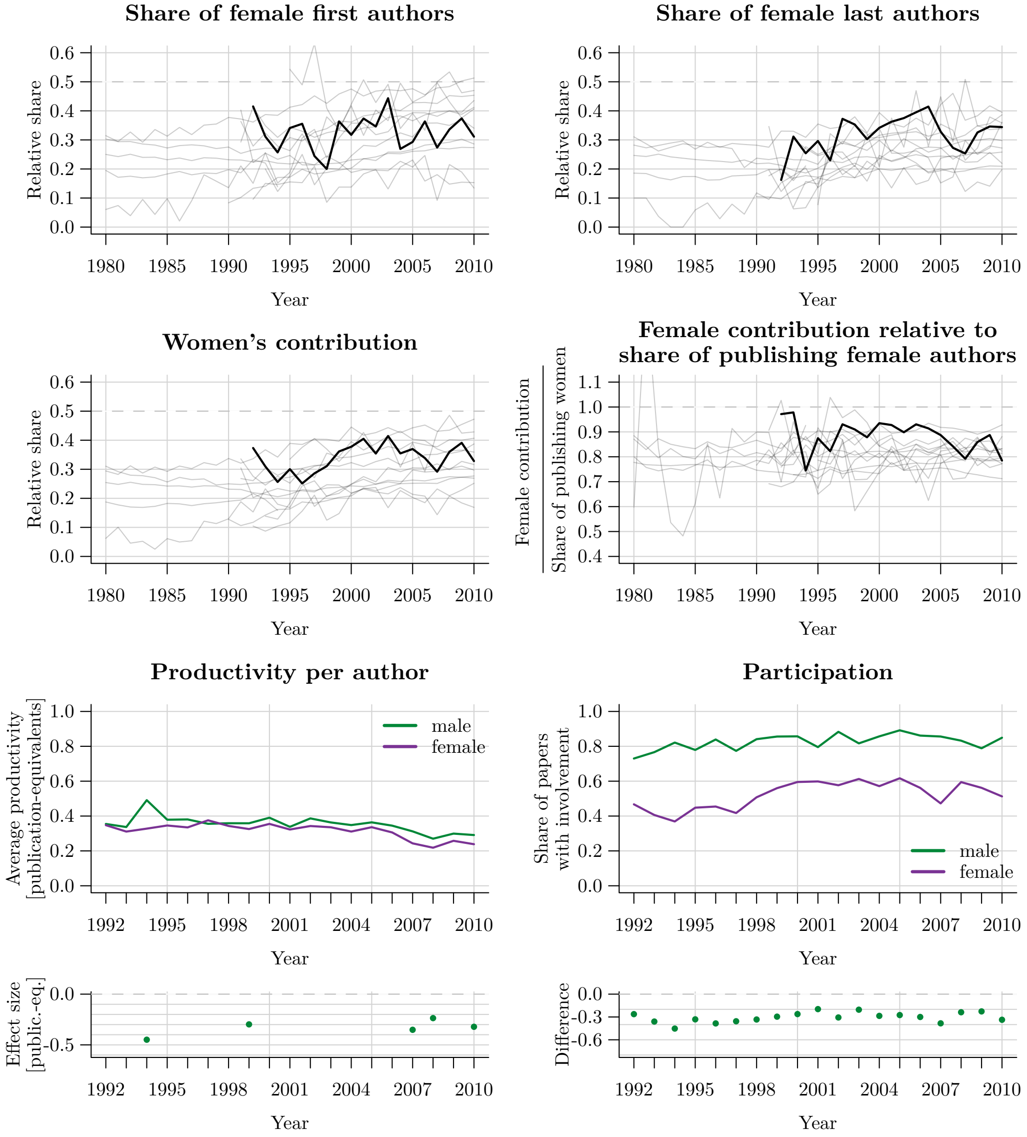
5.3 Citation analysis





6 Kazakhstan

6.1 Productivity



Share of female first authors

Share of female last authors

Women's contribution

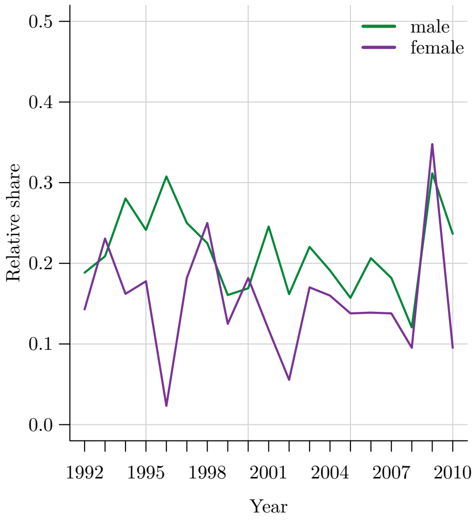
Female contribution relative to share of publishing female authors

Productivity per author

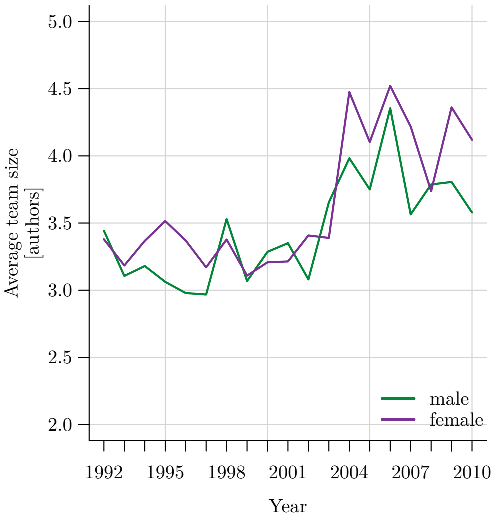
Participation

6.2 Collaboration

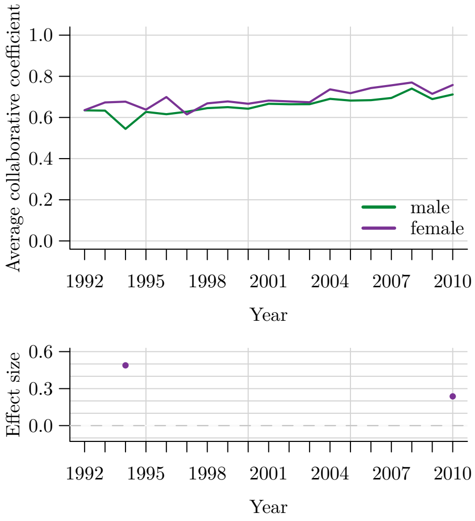
Share of sole authorship



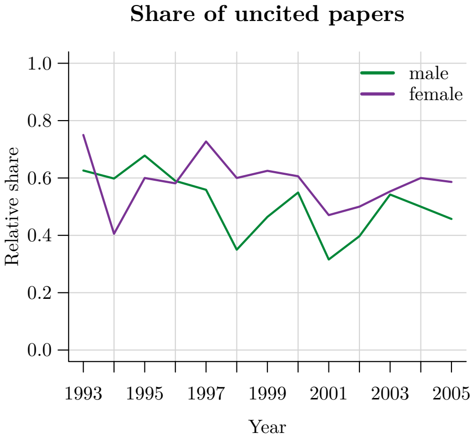
Team size by first author gender



Collaborative coefficient by gender

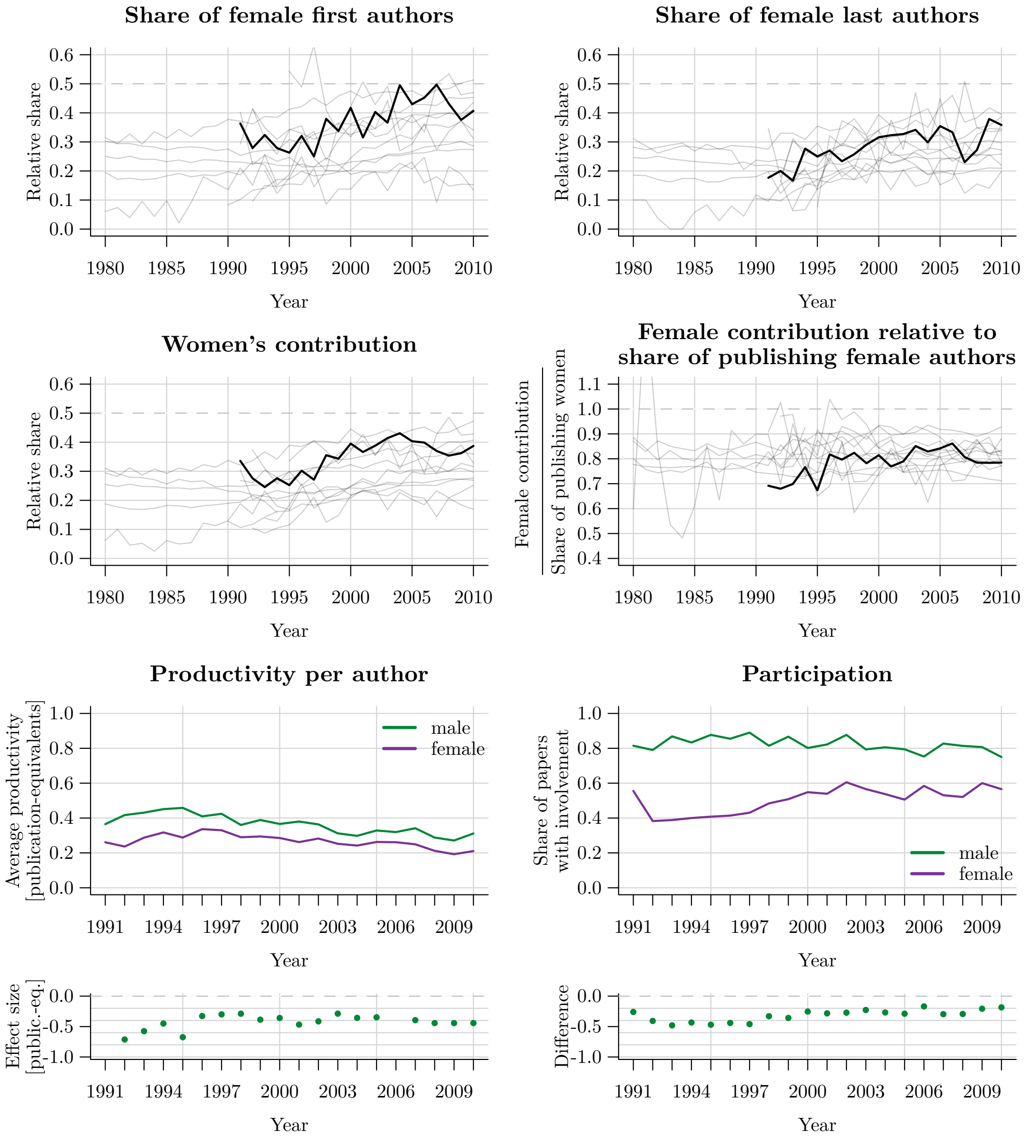


6.3 Citation analysis

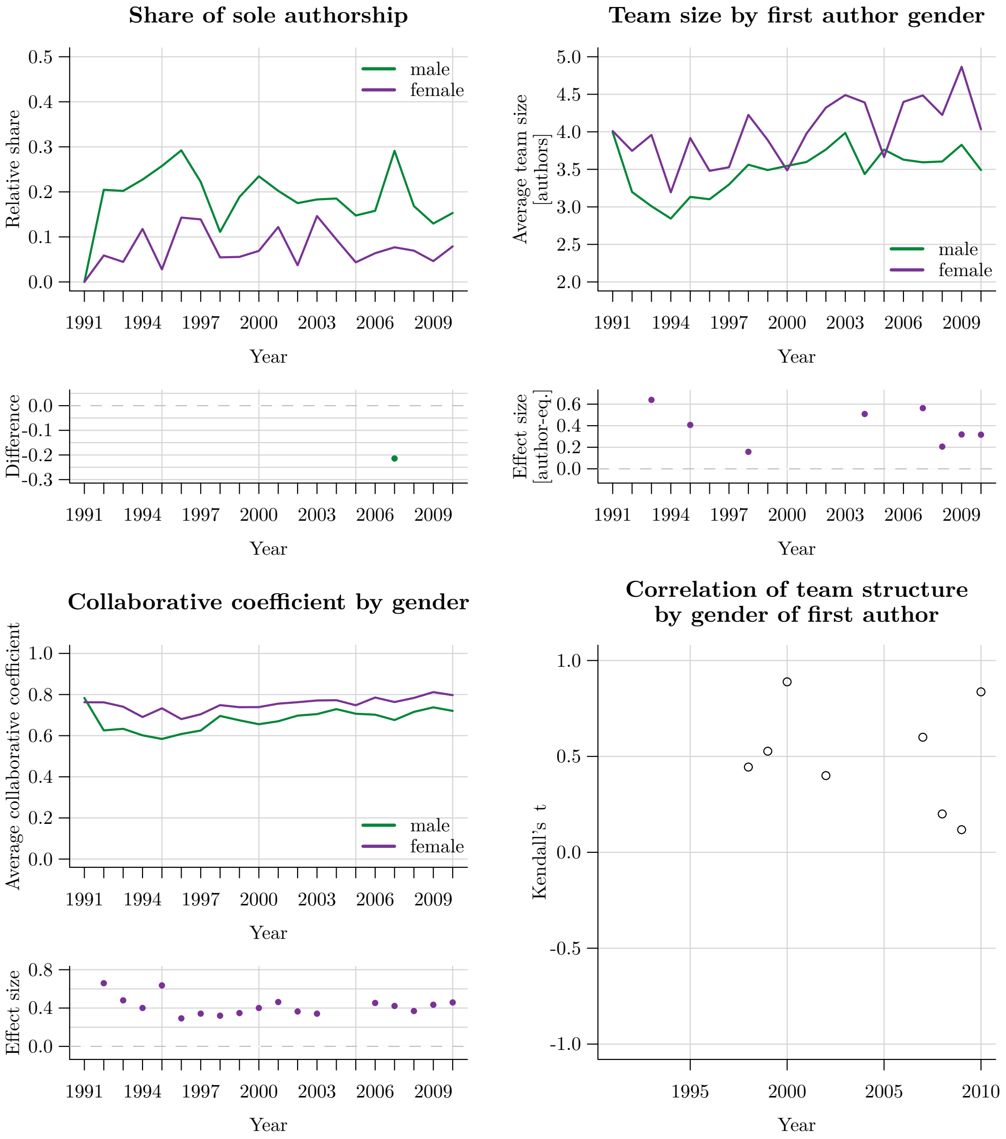


7 Latvia

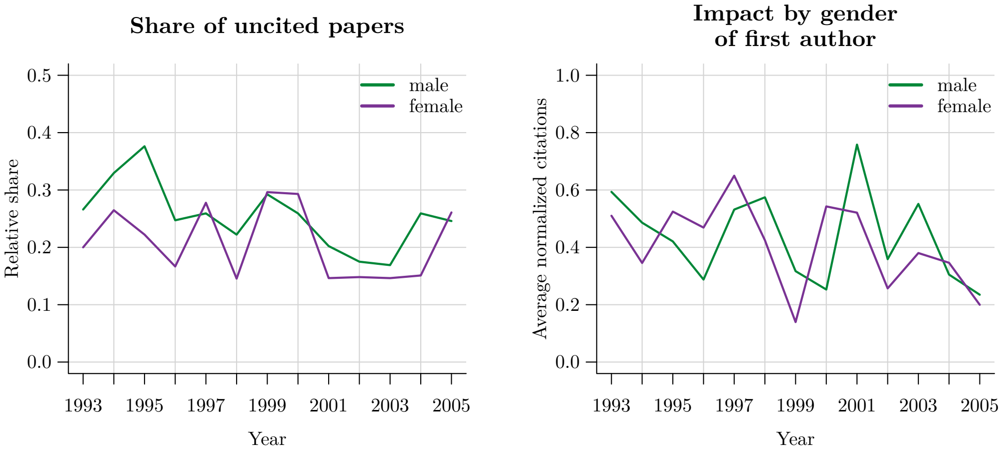
7.1 Productivity



7.2 Collaboration

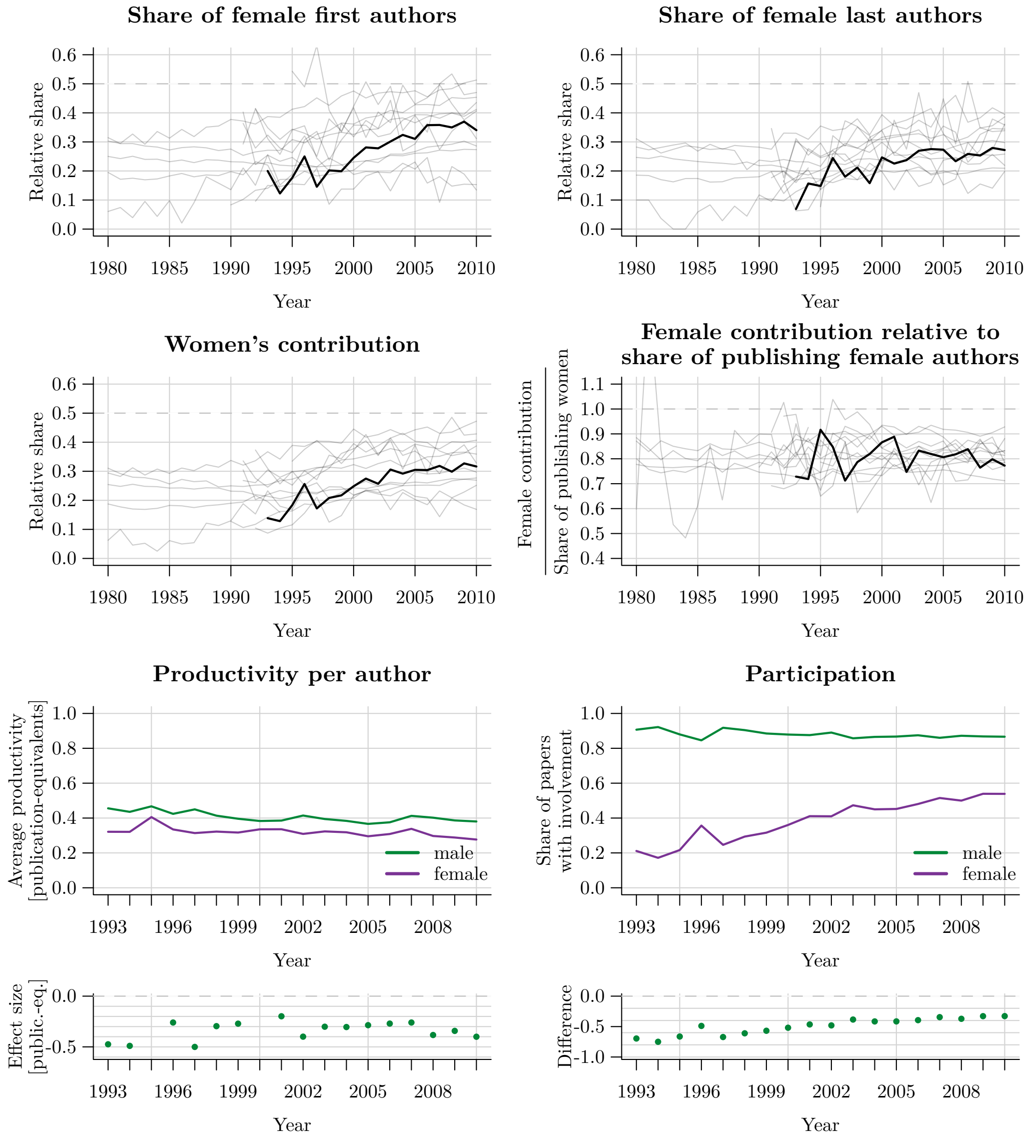


7.3 Citation analysis

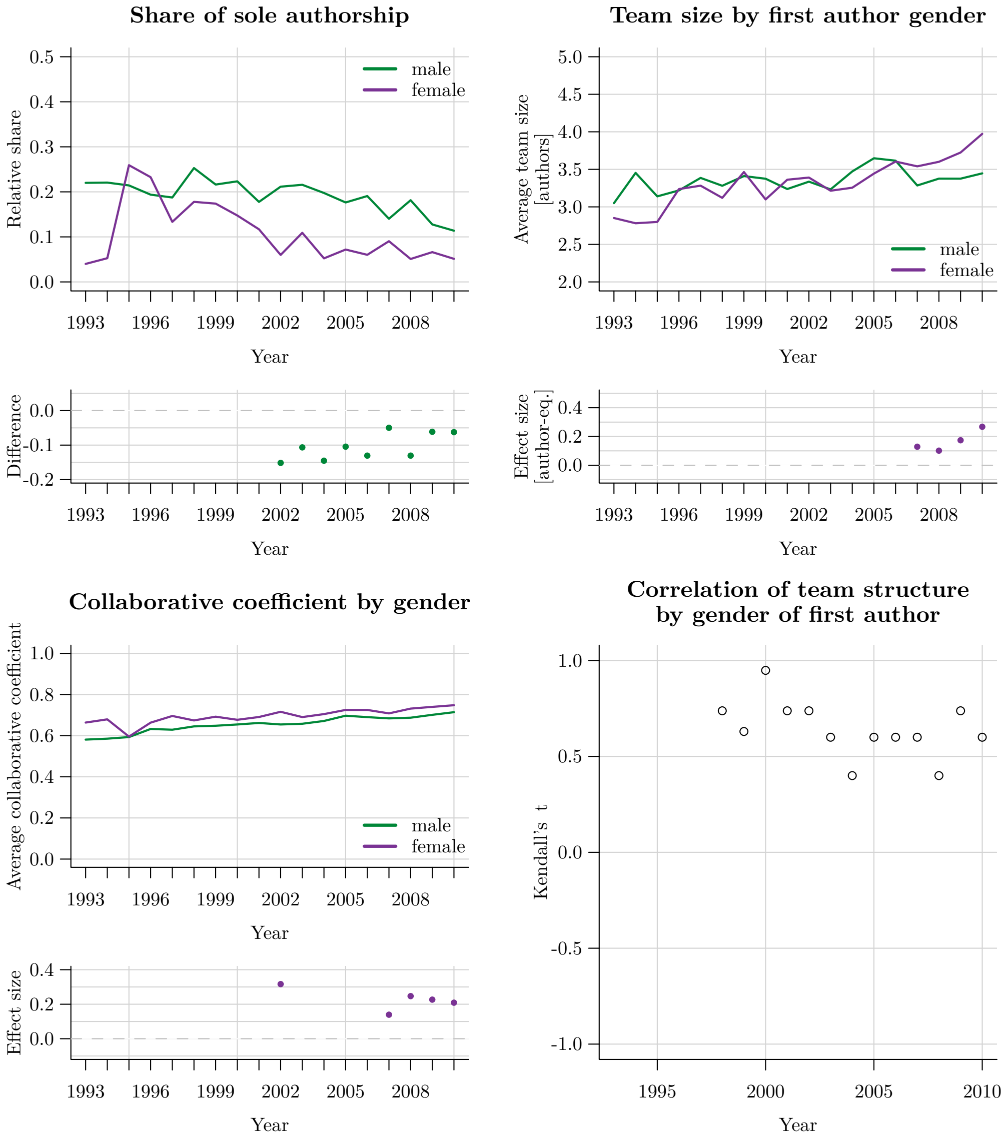


8 Lithuania

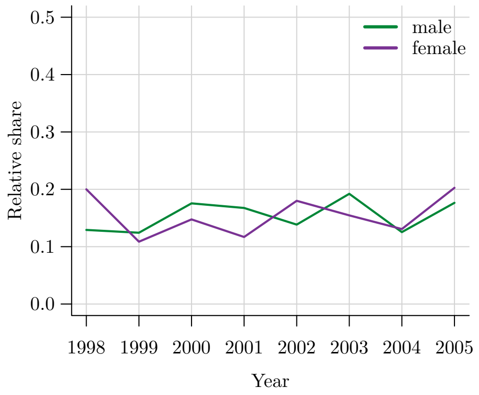
8.1 Productivity



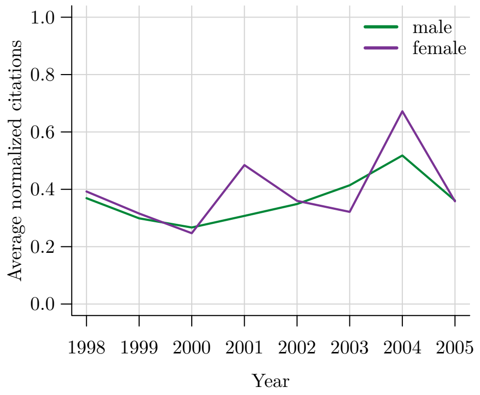
8.2 Collaboration



8.3 Citation analysis



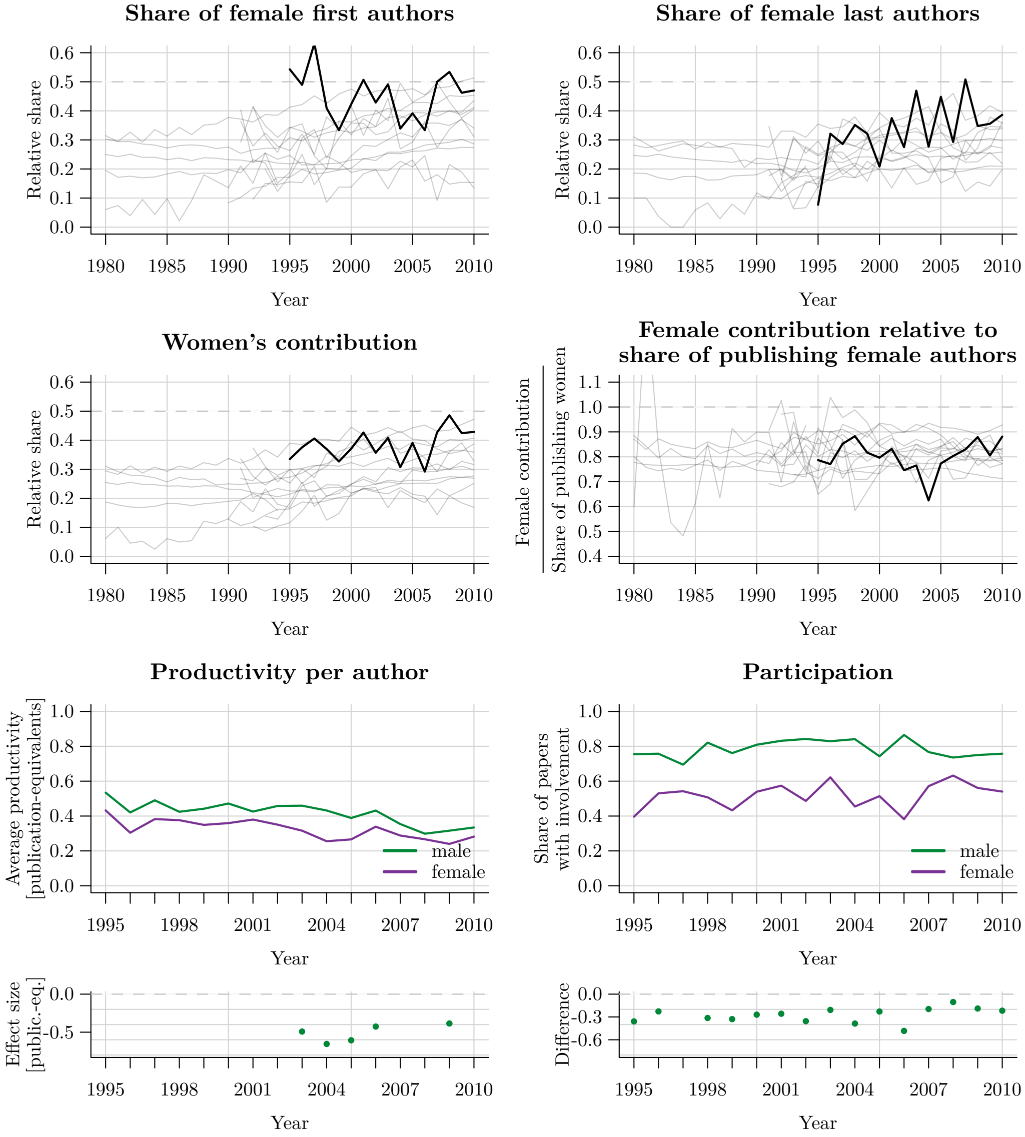
Share of uncited papers



Impact by gender of first author

9 Macedonia

9.1 Productivity



Share of female first authors

Share of female last authors

Women's contribution

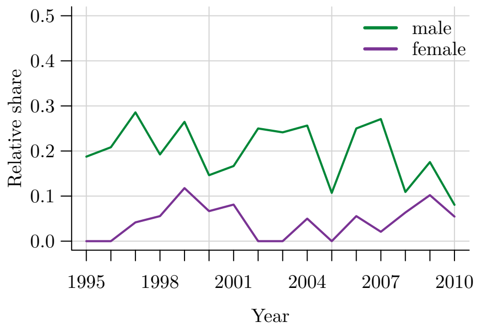
Female contribution relative to share of publishing female authors

Productivity per author

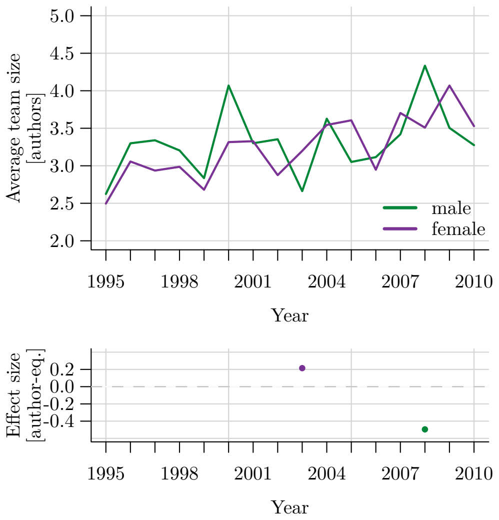
Participation

9.2 Collaboration and citation analysis

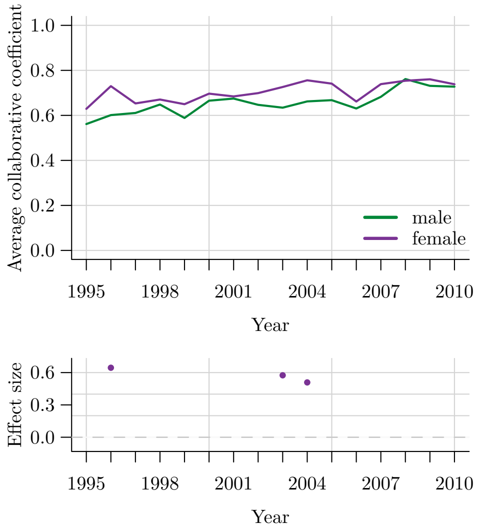
Share of sole authorship



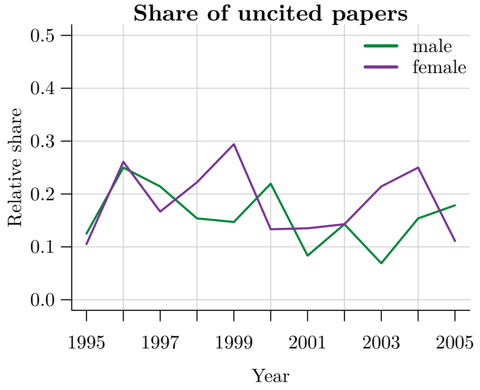
Team size by first author gender



Collaborative coefficient by gender

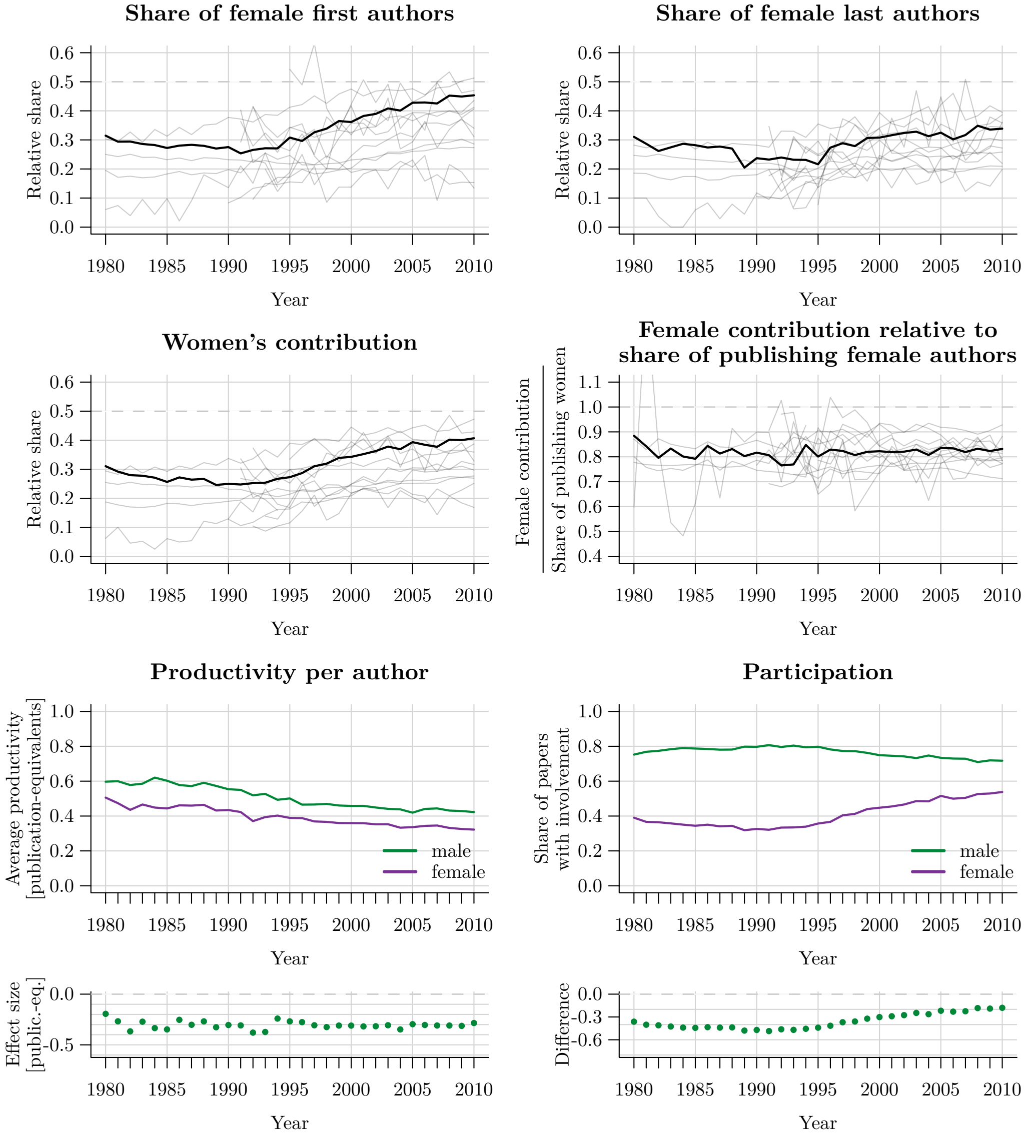


Share of uncited papers



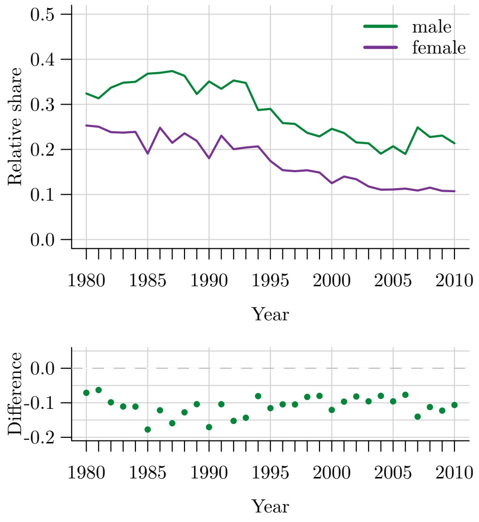
10 Poland

10.1 Productivity

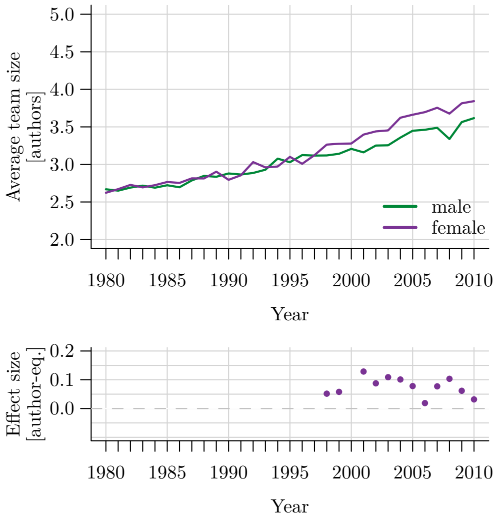


10.2 Collaboration

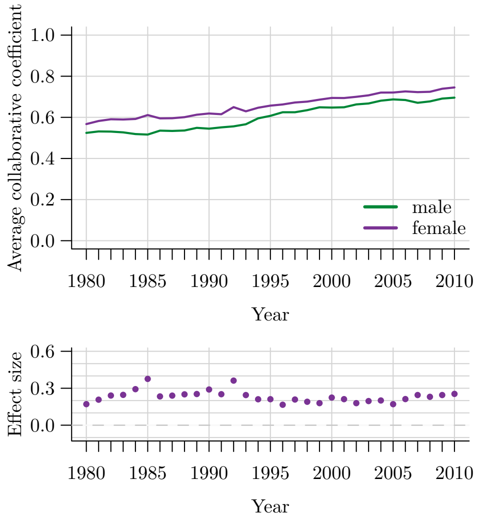
Share of sole authorship



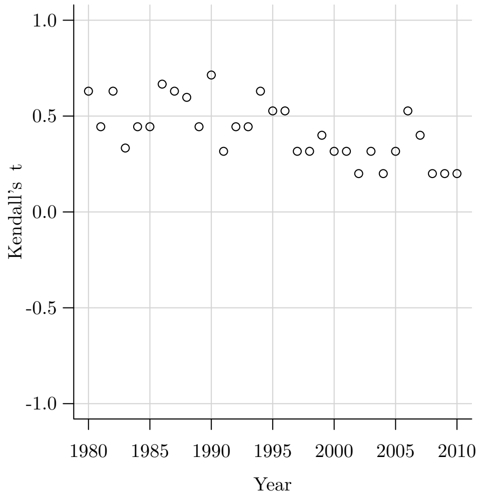
Team size by first author gender



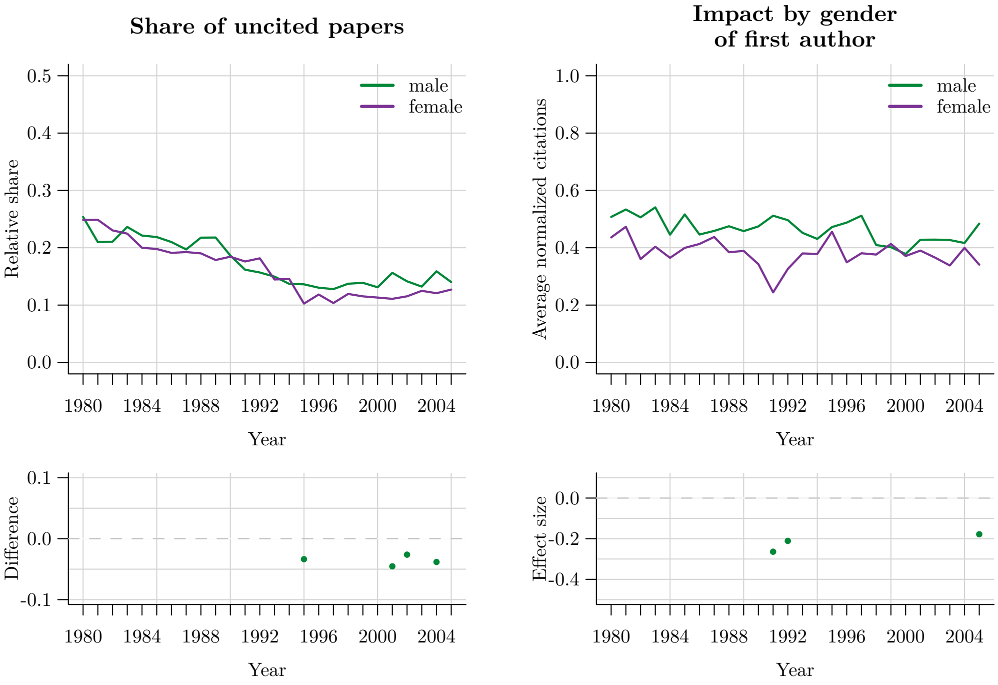
Collaborative coefficient by gender



Correlation of team structure by gender of first author

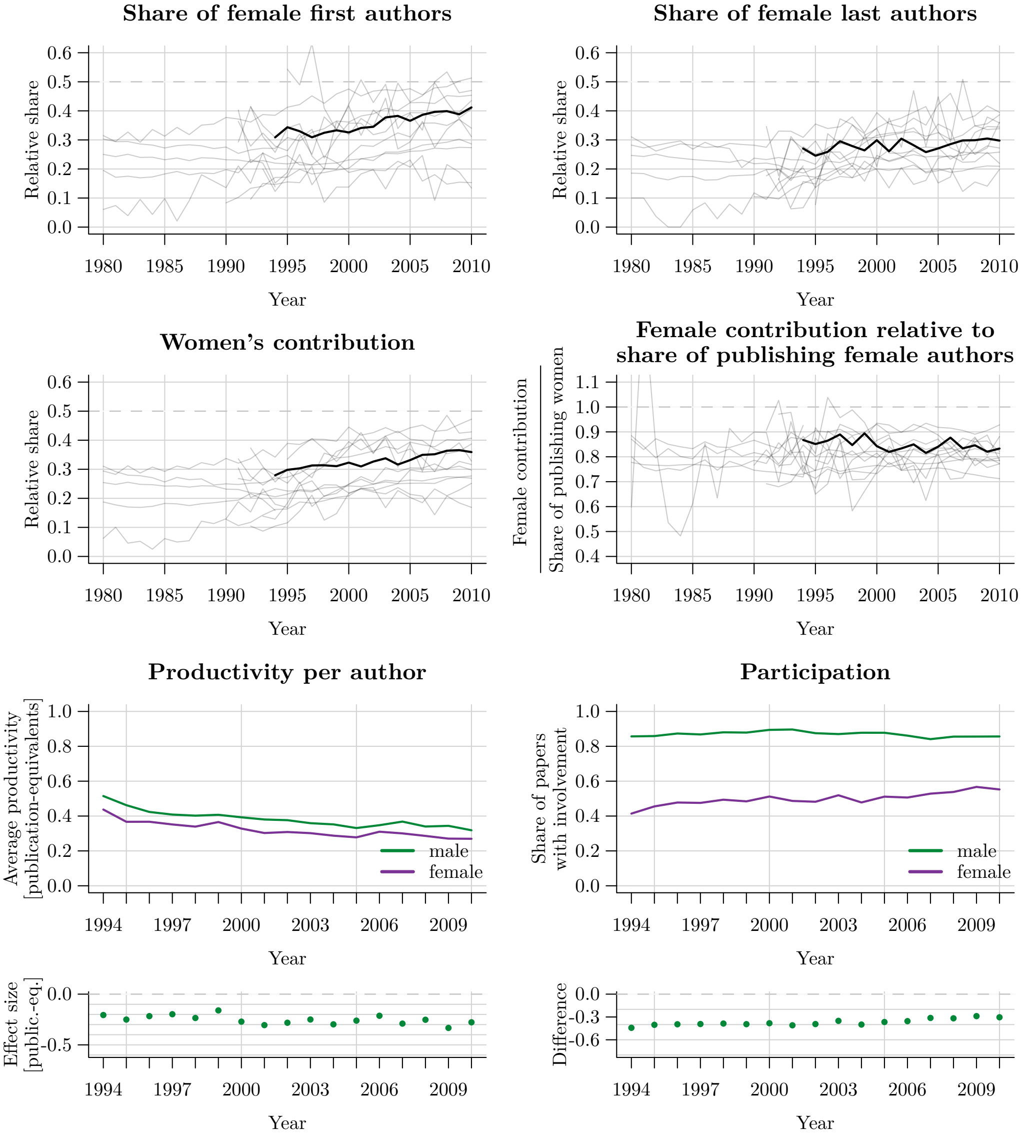


10.3 Citation analysis



11 Slovakia

11.1 Productivity



Share of female first authors

Share of female last authors

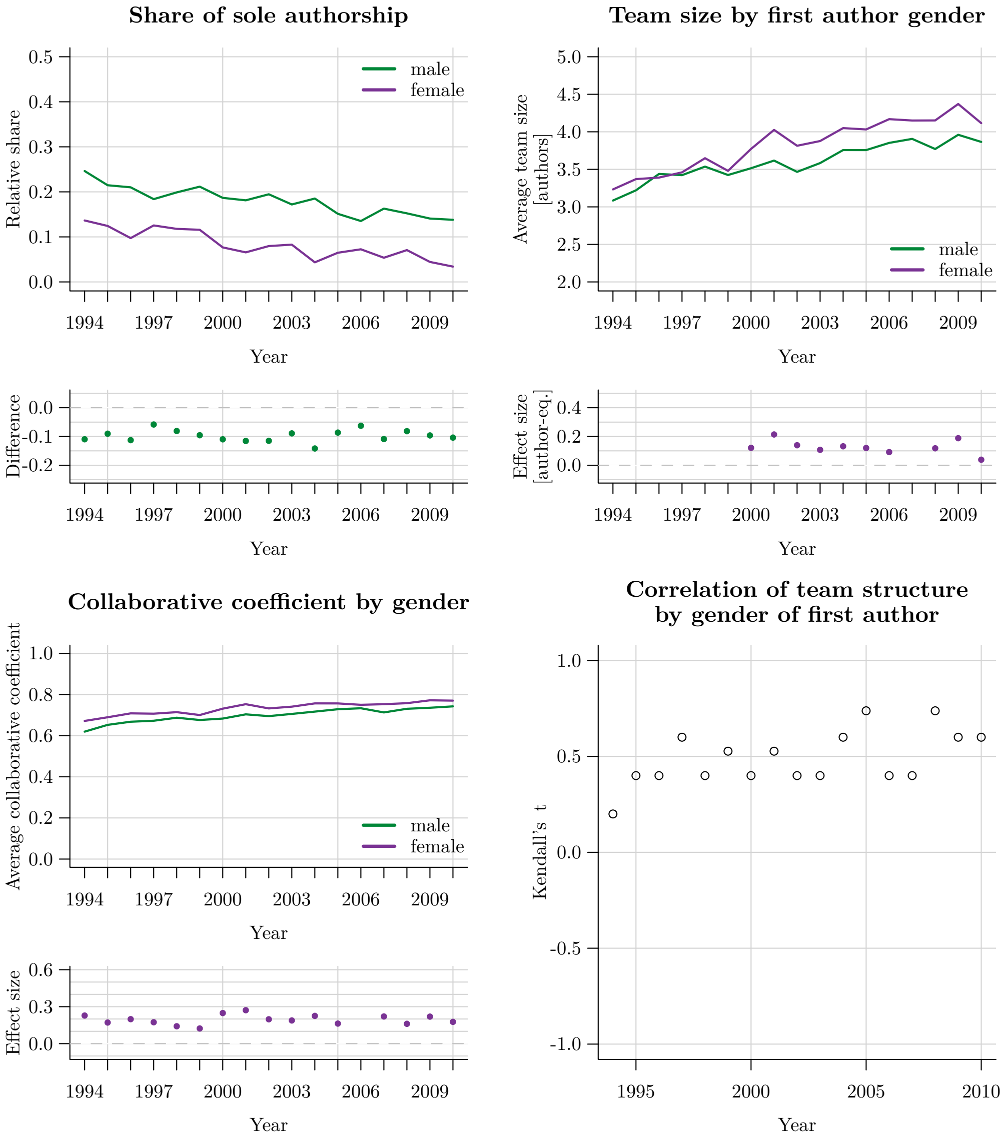
Women's contribution

Female contribution relative to share of publishing female authors

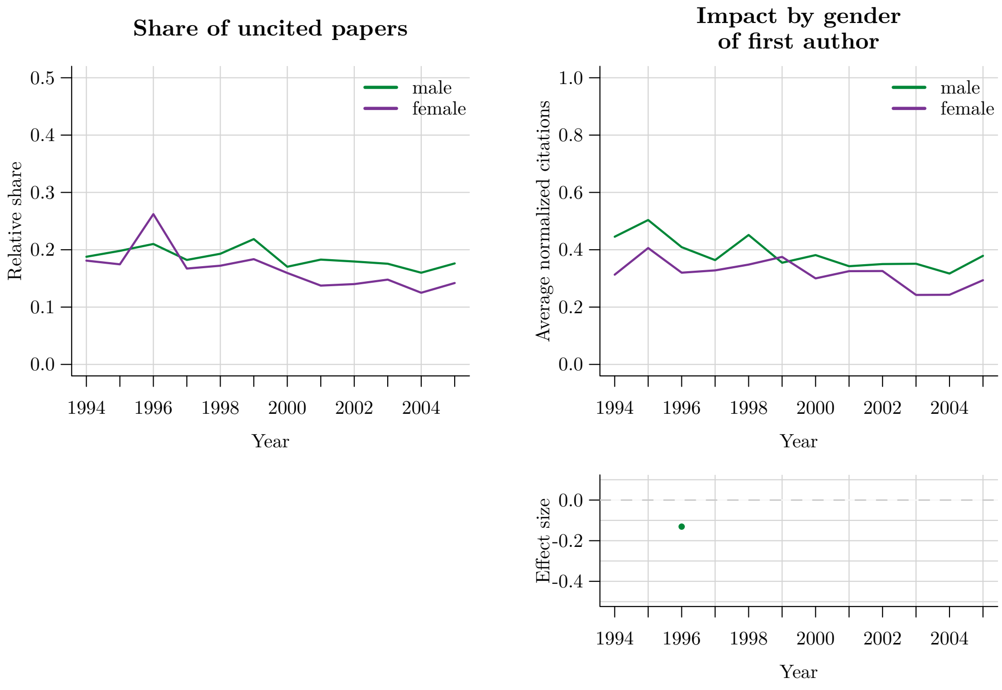
Productivity per author

Participation

11.2 Collaboration



11.3 Citation analysis

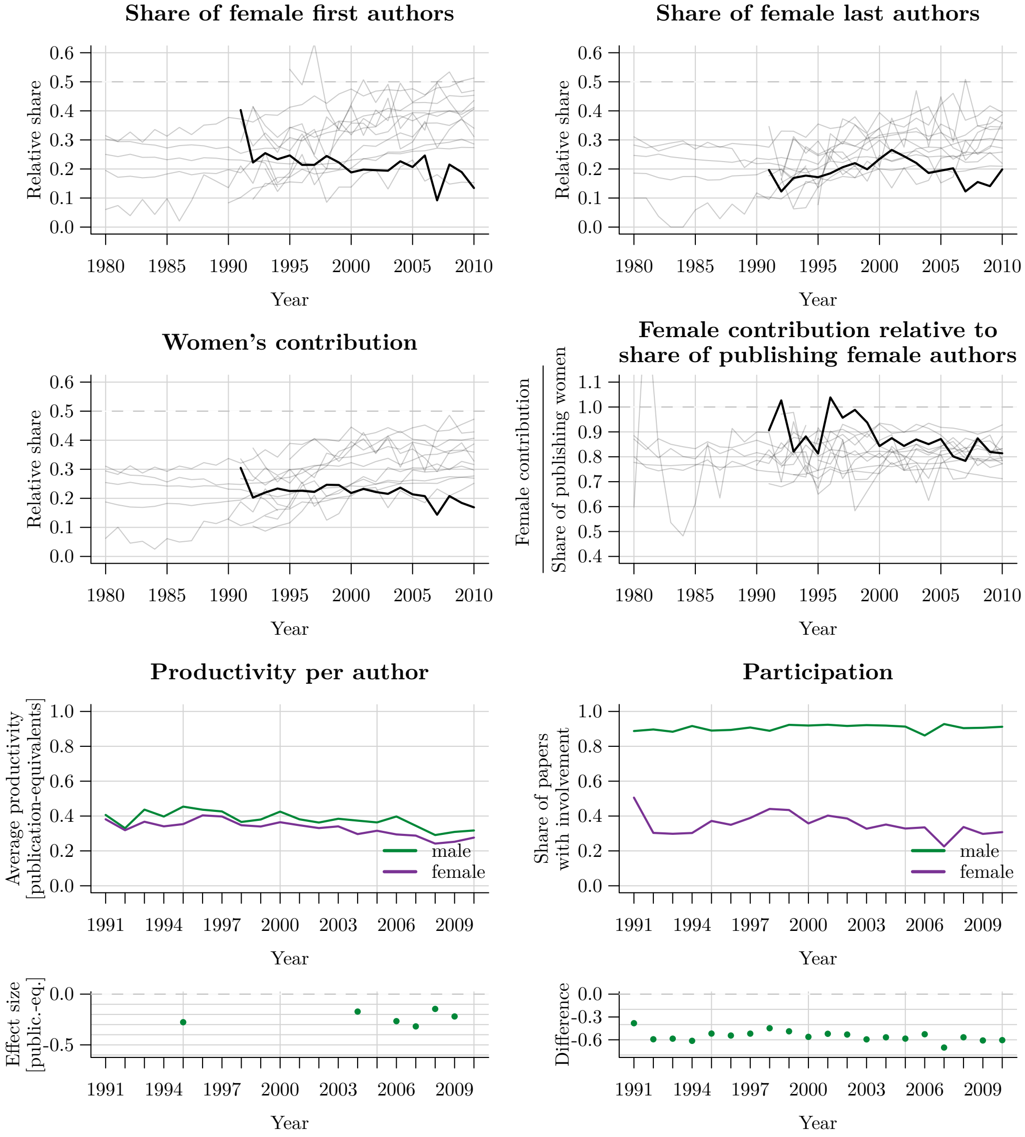


Share of uncited papers

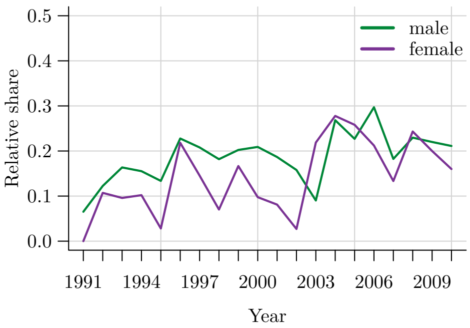
Impact by gender of first author

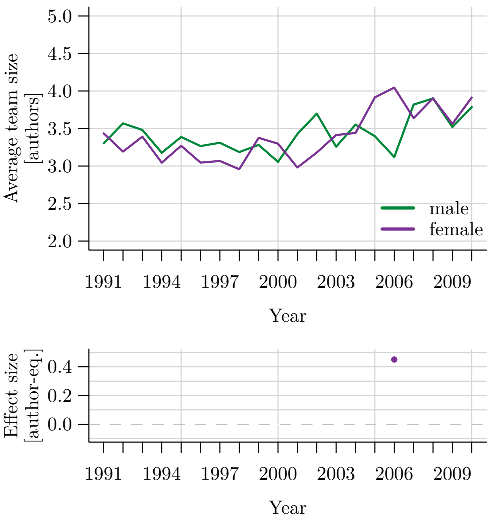
12 Uzbekistan

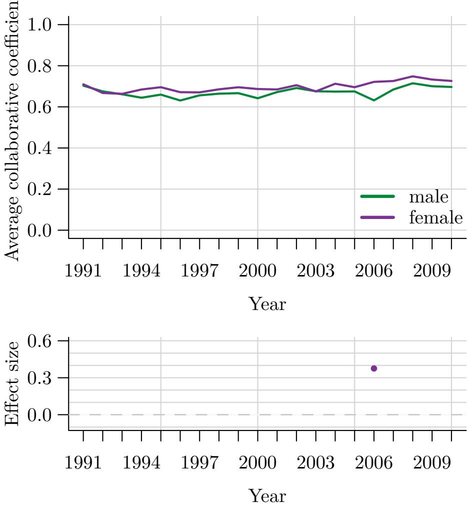
12.1 Productivity

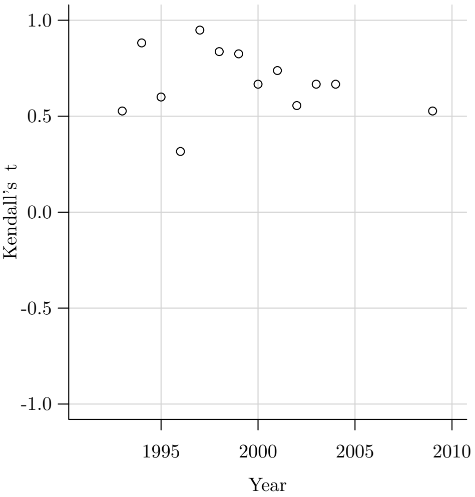


12.2 Collaboration

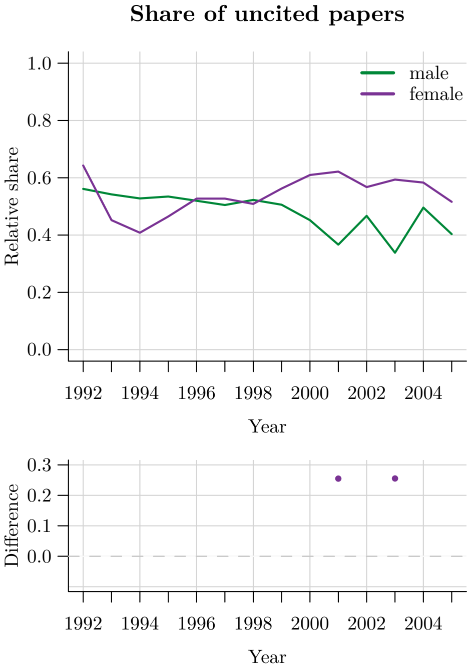

Share of sole authorship


Team size by first author gender


Collaborative coefficient by gender

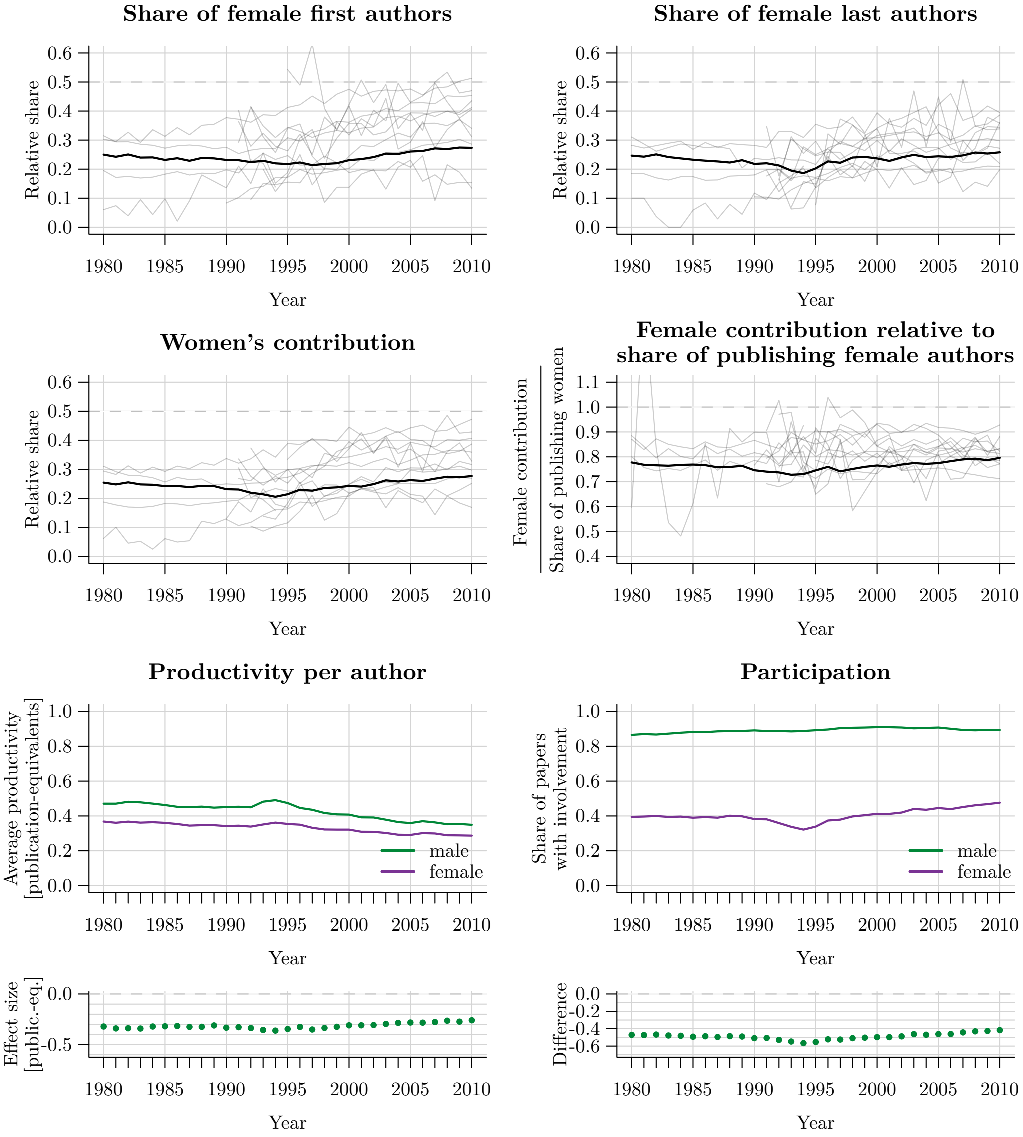

Correlation of team structure by gender of first author

12.3 Citation analysis



13 Soviet Russia and Russia

13.1 Productivity



Share of female first authors

Share of female last authors

Women's contribution

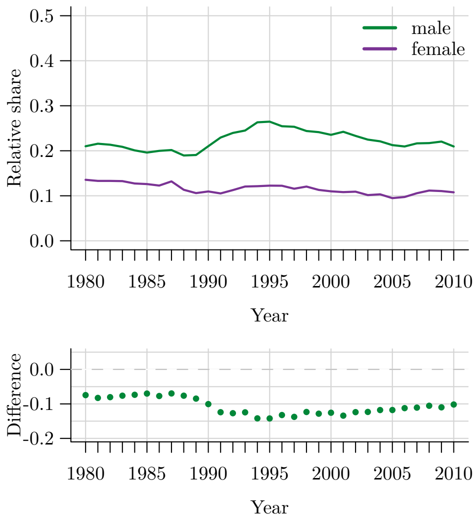
Female contribution relative to share of publishing female authors

Productivity per author

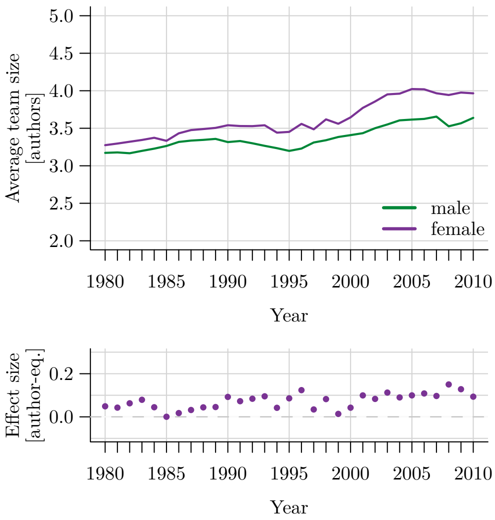
Participation

13.2 Collaboration

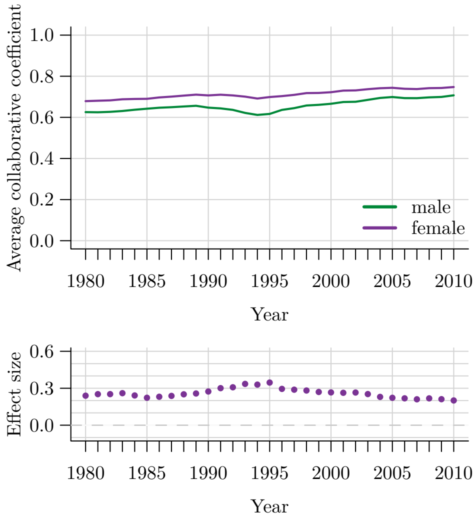
Share of sole authorship



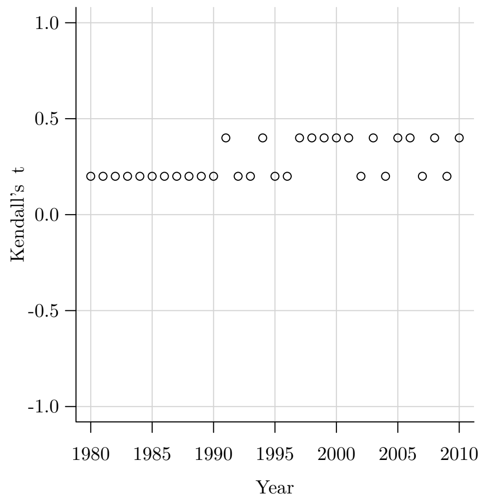
Team size by first author gender



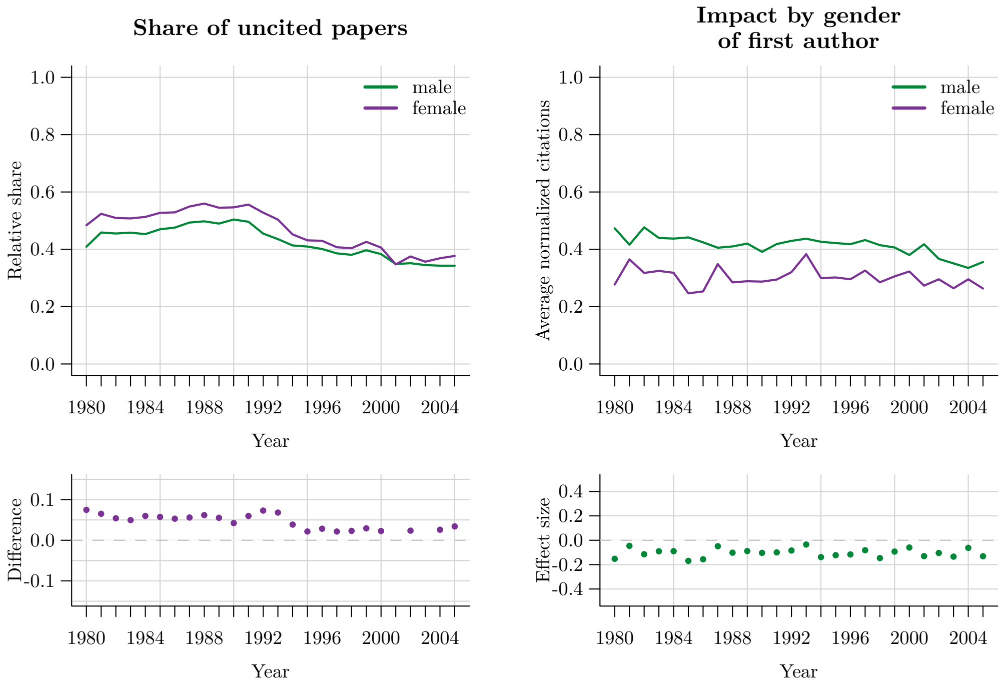
Collaborative coefficient by gender



Correlation of team structure by gender of first author



13.3 Citation analysis



Literaturverzeichnis

Isola Ajiferuke, Q. Burell, and Jean Tague. Collaborative coefficient: A single measure of the degree of collaboration in research. *Scientometrics*, 14(5-6):421–433, 1988.

Dag W. Aksnes. When different persons have an identical author name. How frequent are homonyms? *Journal of the American Society for Information Science and Technology*, 59(5):838–841, 2008.

Dag W. Aksnes, Kristoffer Rorstad, Fredrik Piro, and Gunnar Sivertsen. Are female researchers less cited? A large-scale study of Norwegian scientists. *Journal of the American Society for Information Science and Technology*, 62(4):628–636, 2011. ISSN 1532-2890. doi: 10.1002/asi.21486.

Jerome T. Bentley and Rebecca Adamson. Gender differences in the careers of academic scientists and engineers: A literature review. Special report, National Science Foundation, 2003.

Jacob Clark Blickenstaff. Women and science careers: leaky pipeline or gender filter? *Gender and Education*, 17(4):369–386, 2005.

Lutz Bornmann, Rüdiger Mutz, and Hans-Dieter Daniel. Gender differences in grant peer review: A meta-analysis. *Journal of Informetrics*, 1(3):226–238, 2007. URL `http://arxiv.org/abs/math/0701537`.

Angel Borrego, Maite Barrios, Anna Villarroya, and Candela Olle. Scientific output and impact of postdoctoral scientists: a gender perspective. *Scientometrics*, 83(1):93–101, 2010.

Anne Boschini and Anna Sjögren. Is team formation gender neutral? Evidence from coauthorship patterns. *Journal of Labor Economics*, 25(2):325–365, April 2007. URL `http://www.ifn.se/Wfiles/wp/WP658.pdf`.

Andrea Brendler and Silvio Brendler, editors. *Europäische Personennamensysteme. Ein Handbuch von Abasisch bis Zentralladinisch.* Lehr- und Handbücher zur Onomastik. Baar-Verlag, 2007.

Jonathan R. Cole and Harriet Zuckerman. The productivity puzzle: persistence and changes in patterns of publication of men and women scientists. *Advances in Motivation and Achievements*, 2:217–256, 1984.

Carolyn A. Copenheaver, Kyrille Goldbeck, and Paolo Cherubini. Lack of gender bias in citation rates of publications by dendrochronologists: What is unique about this discipline? *Tree-Ring Research*, 2010.

Laurel L. Cornell. Duplication of japanese names: a problem in citations and bibliographies. *Journal of the American Society for Information Science*, 33(2):102–104, 1982. ISSN 1097-4571.

Sally Jo Cunningham and S. M. Dillon. Authorship patterns in information systems. *Scientometrics*, 39(1):19–27, 1997.

Elisabeth Davenport and Herbert Snyder. Who cites women? Whom do women cite? An exploration of gender and scholarly citation in sociology. *Journal of Documentation*, 51(4):404–410, 1993.

Marianne A. Ferber. Citations: Are they an objective measure of scholarly merit? *Signs*, 11(2):381–389, 1986.

Marianne A. Ferber. Citations and networking. *Gender and Society*, 2(1):82–89, March 1988.

Mary Frank Fox. Gender, family characteristics, and publication productivity among scientists. *Social Studies of Science*, 35(1):131–150, 2005.

Rainer Frietsch, Inna Haller, Melanie Funken-Vrohlings, and Hariolf Grupp. Gender-specific patterns in patenting and publishing. *Research Policy*, 38:590–599, 2009.

Torsten Hothorn and Kurt Hornik. *exactRankTests: Exact Distributions for Rank and Permutation Tests*, 2011. URL `http://CRAN.R-project.org/package=exactRankTests`. R package version 0.8-22.

Scott R. Hutson. Gendered citation practices in American Antiquity and other archaeology journals. *American Antiquity*, 67(2):331–342, April 2002.

Malin Håkanson. The impact of gender on citations: An analysis of College & Research Libraries, Journal of Academic Librarianship, and Library Quarterly. *College & Research Libraries*, 66:312–322, 2005.

Anna Ledin, Lutz Bornmann, Frank Gannon, and Gerlind Wallon. A persistent problem. traditional gender roles hold back female scientists. *EMBO Reprts*, 8(11):982–987, 2007. URL `http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247380/`.

Grant Lewison. The quantity and quality of female researchers: A bibliometric study of Iceland. *Scientometrics*, 52(1):29–43, September 2001.

Grant Lewison and Valentina Markusova. Female researchers in Russia: have they become more visible? *Scientometrics*, 2011.

J. Scott Long. The origins of sex differences science. *Social Forces*, 68(4):1297–1315, 1990.

J. Scott Long. Measures of sex differences in scientific productivity. *Social Forces*, 71 (1):159–178, 1992.

Catherine Lutz. The erasure of women's writing in sociocultural anthropology. *American Ethnologist*, 17(4):611–258, November 1990.

Terttu Luukkonen-Gronow and Veronica Stolte-Heiskanen. Myths and realities of role incompatibility of women scientists. *Acta Sociologica*, 26(3-4):267–280, 1983. doi: 10.1177/000169938302600304.

Herbert W. Marsh, Lutz Bornmann, Rüdiger Mutz, Hans-Dieter Daniel, and Alison O'Mara. Gender effects in the peer reviews of grant proposals: A comprehensive meta-analysis comparing traditional and multilevel approaches. *Review of Educational Research*, 79(3):1290–1326, September 2009.

Werner Marx. Special features of historical papers from the viewpoint of bibliometrics. *Journal of the American Society for Information Science and Technology*, 62(3):433–439, 2011.

John McDowell and Janet Kiholm Smith. The effect of gender-sorting on propensity to coauthor: implications for academic promotion. *Economic Inquiry*, 30:68–82, 1992.

Bonnie McElhinny, Marijke Hols, Jeff Holtzkener, Susanne Unger, and Claire Hicks. Gender, publication and citation in sociolinguistics and linguistic anthropology: The construction of a scholarly canon. *Language in Society*, 32:299–328, 2003.

Jörg Michael. 40000 Namen, Anredebestimmung anhand des Vornamens. *c't. Magazin für Computer-Technik*, 24(17):182–183, 2007.

Fulvio Naldi and Ilaria Vannini Parenti. Scientific and technological performance by gender. A feasibility study on patent and bibliometric indicators. Vol. I: Statistical analysis. Technical report, Consiglio Nazionale delle Ricerche, 2002a.

Fulvio Naldi and Ilaria Vannini Parenti. Scientific and technological performance by gender. A feasibility study on patent and bibliometric indicators. Vol. II: Methodological report. Technical report, Consiglio Nazionale delle Ricerche, 2002b.

Lorraine J. Pellack and Lori Osmus Kappmeyer. The ripple effect of women's name changes in indexing, citation, and authority control. *Journal of the American Society for Information Science and Technology*, 62(3):440–448, 2011.

Katarina Prpić. Znanstvena produktivnost istraživača između minimalizma i maksimalizma [Scientific productivity between minimalism and maximalism]. In Katarina Prpić and B. Golub, editors, *Znanstvena produktivnosti potencijalni egzodus istraživača Hrvatske/Scientific productivity and the potential exodus of Croatian researchers*, pages 1–61. Institut za društvena istraživanja Sveučilišta u Zagrebu, 1990.

Katarina Prpić. Gender and productivity differentials in science. *Scientometrics*, 55(1):27–58, September 2002. URL `http://www.akademiai.com/content/u46530v572013k24/BodyRef/PDF/11192_2004_Article_400531.pdf`.

Hanna-Mari Puuska. Effects of scholar's gender and professional position on publishing productivity in different publication types. Analysis of a Finnish university. *Scientometrics*, 82:419–437, 2009.

R Development Core Team. *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, 2011. URL `http://www.R-project.org/`. ISBN 3-900051-07-0.

Linsay Reece-Evans. Gender and citation in two LIS e-journals: A bibliometric analysis of LIBRES and Information Research. *LIBRES*, 20(1), 2010. URL `https://libres.curtin.edu.au/libres20n1/Reece-Evans_Ref.pdf`.

Lothar Sachs and Jürgen Hedderich. *Angewandte Statistik. Methodensammlung mit R.* Springer, Berlin, 12., vollst. neu bearb. Aufl. edition, 2006. ISBN 3-540-32160-8.

A. Schubert and T. Braun. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, 9(5-6):281–291, 1986.

Gene D. Sprouse. Editorial: Which wei wang? *Physical Review Letters*, 99(23):230001, 2007.

Steven Stack. Gender, children and research productivity. *Research in Higher Education*, 45(8):891–920, 2004.

Celia Sánchez Peñas and Peter Willett. Gender differences in publication and citation counts in librarianship and information science research. *Journal of Information Science*, 32(5):480–485, 2006.

Kathryn B. Ward, Julie Gast, and Linda Grant. Visibility and dissemination of women's and men's sociological scholarship. *Social Problems*, 39(3):291–298, August 1992.

Berenika M. Webster. Polish women in science: a bibliometric analysis of Polish science and its publications, 1980–1999. *Research Evaluation*, 10(3):185–194, December 2001.

Wikipedia. Family name — Wikipedia, The Free Encyclopedia, 2011. URL `http://en.wikipedia.org/w/index.php?title=Family_name&oldid=455969282`. [Online; accessed 18-October-2011].

ZhaoRan Xu and Dan H. Nicolson. Don't abbreviate Chinese names. *Taxon*, 41(3):499–504, 1992.

Olaf Zawacki-Richter and Christine von Prümmer. Gender and collaboration patterns in distance education research. *Open Learning*, 25(2):95–114, June 2010.