Professor Marek Kwiek
Center for Public Policy Studies, Director
UNESCO Chair in Institutional Research and Higher Education Policy
University of Poznan, Poland
kwiekm@amu.edu.pl

Dr. Wojciech Roszka
Poznan University of Economics and Business, Poznan, Poland
wojciech.roszka@ue.poznan.pl

Gender-Based Homophily in Research: A Large-Scale Study of Man-Woman Collaboration

Highlights

- We examined male-female collaboration practices of all internationally productive (25,000) Polish university professors based on their 160,000 Scopus-indexed publications.

- We merged a national registry of 100,000 scientists (with full administrative and biographical data) with the Scopus publication database.

- We examined the propensity to conduct same-sex collaboration across male-dominated, female-dominated, and gender-balanced disciplines.

- Across all age-cohorts and all academic positions, the majority of male scientists collaborate solely with males, and the majority of female scientists, in contrast, do not collaborate with females.

- The gender homophily principle (publishing predominantly with scientists of the same sex) works powerfully for male scientists but does not seem to work for female scientists.

- Articles written in mixed-sex collaboration are, on average, published in more prestigious journals than are those written in same-sex collaboration.

- Logistic regression analysis shows that the propensity to conduct same-sex collaboration for males is more than three times that for females.

Abstract

This paper investigates the respective impacts of (1) biological age, (2) academic position, (3) academic discipline, (4) journal prestige, and (5) type of institution of employment on the propensity to conduct same-sex collaboration in research. The gender homophily principle was found to work for male scientists—but not for females. The majority of male scientists collaborate solely with males; most female scientists, in contrast, do not collaborate with females at all. The propensity for same-sex collaboration of males is three times that of females. Across all age cohorts, scientists of both genders tend to collaborate more with males. All-female collaboration is marginal, while all-male collaboration is pervasive. However, a year-by-year approach confirmed a downward trend in same-sex collaboration among males and an upward trend among females. Additionally, gender homophily in research-intensive institutions proved stronger for males than for females. Finally, we estimated odds ratios of high homophily in publishing and used linear regression to explain the variability of the same-sex collaboration ratio in publishing. A comprehensive, fully integrated, biographical, administrative, publication, and citation database was used, and our sample (N = 25,463) included all Polish professors employed in 85 research-involved universities, grouped into 27 disciplines, with all their Scopus-indexed 2009–2018 publications (158,743 articles).

Keywords

Collaboration; co-authorships; gender gap; sociology of science; homophily; scientific careers

1. Introduction

The research collaboration patterns of male and female scientists are contrasted in this paper through five lenses: biological age, academic position, academic discipline, gender-defined research collaboration type, journal prestige, and type of institution of employment. The individual scientist, rather than the individual article, is the unit of analysis. The key methodological step is the determination of what we term an "individual publication portfolio" (for the decade of 2009–2018) for every internationally productive Polish scientist (N = 25,463 university professors from 85 universities, grouped into 27 disciplines, along with their 164,908 international collaborators, who together authored 158,743 Scopus-indexed publications). Co-authorships are used for the operationalization of research collaboration, following standard bibliometric practice.

The individual publication portfolio reflects the distribution of gender-defined research collaboration types (same-sex collaboration and mixed-sex collaboration) for every individual scientist. Team formation in academia, understood as publishing with co-authors of varying numbers and different genders, is voluntary (McDowell & Smith, 1992): researchers team up when they think that they are better-off collaborating than publishing alone. The teams formed, or the articles published, are likely to reflect "individual tastes and perceptions of the returns to collaboration, as well as the costs of coordination" (Boschini & Sjögren, 2007, p. 327). Some male scientists collaborate predominantly with other males, some female scientists collaborate predominantly with other females, and still others prefer to publish in mixed-sex collaborations (or to author individually). We examine the propensity to conduct same-sex research at an individual level of every internationally productive Polish scientist and generalize the results from the individual level to the level of the national higher education system. The research question of this paper is as follows: What is the impact of the following independent variables on the propensity to conduct same-sex collaboration in research: (1) biological age, (2) academic position, (3) academic discipline, (4) journal prestige, and (5) type of institution of employment?

2. Literature Review

2.1. The Gender Context of Science

The two guiding themes of "research collaboration" and "women in science" combined in this research have been widely studied for about half a century. However, the role of female scientists in the academic enterprise has been far from static since the 1960s and 1970s (Huang, Gates, Sinatra, & Barabàsi, 2020). The reason is simple: the gender context of academic science has changed substantially (Halevi, 2019; Larivière, Ni, Gingras, Cronin, & Sugimoto, 2013), with more female scientists entering and remaining in the higher education sector every decade and increasingly occupying high academic positions (Madison & Fahlman, 2020) in ever greater proportions (Zippel, 2017) and in an increasing number of disciplines (Diezmann & Grieshaber, 2019). These changing numbers and changing proportions have together transformed the traditional context, in which female scientists had been regarded as newcomers to science. The gender productivity, citation, and promotion gaps have been changing over time, albeit slowly. Male and female scientists often pursued or were pushed onto somewhat different career tracks and were located in different academic structures, with "differential access to valuable resources" (Xie & Shauman, 2003, p. 193). Females, as new entrants into a male-dominated academic profession, did not have equal access to professional networks (McDowell, Singell, & Stater, 2006), but the academic world is changing. Specifically, in the Polish context, females constitute a substantial, highly productive, and highly internationalized part of the academic workforce, which is often the case in formerly communist European countries, which exhibit greater gender parity than the OECD average (Larivière et al., 2013, p. 212). Poland has a higher proportion of full professors than any country studied in Larivière et al. (2013) or in Diezmann and Grieshaber (2019), reaching 24% in 2017, even though there is a clear "the higher the fewer" pattern across all institutional types (see Kwiek 2020b on "internationalists" contrasted with "locals" in Polish academic science by gender).

Females' rising participation in academic science changes both the global and national contexts in which gender disparities in research collaboration can be analyzed, especially gender homophily in academic publishing. New bibliometric literatures applying the various gender-determination methods to authors and authorships (Halevi, 2019) bring new data-driven insights to the "research collaboration" and "women in science" fields. Gender disparities in science have been changing (Zippel, 2017; Diezmann & Grieshaber 2019), and literatures have become much less based on anecdotal and localized studies (as Larivière et al., 2013 note). For instance, Madison and Fahlman (2020) demonstrated, for the entire population of Swedish full professors, that no bias against females occurred in attaining the rank of full professor in relation to their publication metrics. Women are plugging into networks over time as the profession becomes more gender representative (as shown for academic economists by McDowell et al., 2006, p. 154). However, somewhat paradoxically, the increased participation of women in STEM disciplines over the past 50 years is reported to have been accompanied by an increase in gender differences regarding both productivity and impact (Huang et al., 2020, p. 8; Kwiek, 2016). The lower social capital of female scientists (van Emmerik, 2006) has been traditionally linked to gender-based homophily in research collaborations—the tendency for scientists to collaborate (and co-author) with individuals of the same gender. Anticipating the results of this paper: in Poland, males tend to collaborate with males—but females tend *not* to collaborate with females. Thus, gender homophily is high among Polish males and low among females, the latter constituting 41.5% of Polish university professors (of all ranks) in our sample.

Female scientists still occupy more junior positions with lower salaries, are more often in non-tenure track and teaching-only positions, receive less grant money, are promoted more slowly, are less likely to be listed as either first or last author on a paper, and are allocated fewer resources and less research funding from national research councils. Women also tend to be less involved in international collaboration; female collaborations are more domestically oriented than are the collaborations of males from the same country; and females have less-prestigious collaborations and fewer collaborations overall, as the past decade of research highlights (see Holman & Morandin, 2019; Halevi, 2019; Larivière et al., 2013; Larivière et al., 2011; Aksnes, Rørstad, Piro, & Sivertsen, 2011; Aksnes, Piro, & Rørstad, 2019; Huang et al., 2020; Maddi, Larivière, & Gingras, 2019; Fell & König, 2016; van den Besselaar & Sandström, 2016; Nielsen, 2016).

A recent cohort analysis of the effects of gender on the publication patterns in mathematics (Mihaljević-Brandt, Santamaria, & Tullney, 2016, pp. 8–13), one of the most heavily male-dominated academic fields, based on the scholarly output of 150,000 mathematicians, shows that women publish less at the beginning of their careers; they leave academia at a higher rate than men; and high-ranked mathematics journals publish fewer articles authored by women. Women may also suffer from "biased attention" to their work, even if their work is of comparable quality (Lerchenmueller, Hoisl, & Schmallenbach, 2019, p. 10). The authors' gender is also reported to affect the citations received (Potthof & Zimmermann, 2017): as the proportion of women per article increases, the citations tend to decrease (as Maddi et al., 2019, show for economics). The gender citation gap matters because citations are one of the chief metrics used in academia to evaluate a scholar's performance and influence and to distribute resources, including salary (Maliniak, Powers, & Walter, 2013, p. 895), with the citation measure being increasingly used as a "reward currency in science" upon which decisions on all major aspects of an academic career are often based (Ghiasi, Mongeon, Sugimoto, & Larivière, 2018, p. 1519). Female lead authors are reported to receive up to 29% fewer citations for work published in the most influential journals (as shown for publications from the PubMed database of 3,233 recipients of prestigious fellowships in life sciences in the U.S.: Lerchenmueller et al., 2019, p. 4).

Furthermore, gender-based homophily in citations exists in all disciplines, as a study of the citation data of 7 million articles published in 2008–2016 shows: the citer disproportionately cites references from authors who are of the same gender, male scientists disproportionately citing other male scientists, possibly leading to a "perpetual disparity" in citations in favor of men as men represent about 70% of all authorships (Ghiasi et al., 2018, p. 1520). Moreover, recent research based on a sample of CVs of U.S. economists reports that gender influences the attribution of credit for group work, that is, co-authorship matters differently for tenure for men and women, with women being less likely to receive tenure the more they co-author (Sarsons, Gërxhani, Reuben, & Schram, 2020). This differential attribution of credit contributes to the gender promotion gap (Fell & König, 2016; Abramo, D'Angelo, & Rosati, 2015). Furthermore, the gender citation gap persists: even though female scientists may publish more in journals with higher impact factors than their male peers, their work may receive lower recognition (fewer citations) from the scientific community (as Ghiasi, Larivière, & Sugimoto, 2015, have shown for female engineers, using a sample of 680,000 articles from 2008–2013, and Maliniak et al., 2013, for top journals in international relations).

2.2. Female Scientists and Competition

Of the various approaches to studying the "increasing and persistent" gender gap in science (Huang et al., 2020, p. 3), an approach centered on competition is especially relevant in the context of homophilous and heterophilous collaboration patterns. There have been ongoing discussions in experimental and personnel economics (often with laboratory-based evidence) about whether women are deterred by competition in some areas of science (and in some workplaces more generally; Flory, Leibbrandt, & List, 2015; Dargnies, 2012). The systematic shying away from competition could have implications not only for the gender distribution of females across academic disciplines and their sub-disciplines but also for team formation in research collaboration, prestige in academic publishing, and authorship composition. Laboratory experiments show that women may shy away from competition and men may embrace it, with gender implications for publishing in top academic journals, where competition is stiff and the risk of rejection high (Sonnert & Holton, 1996). Women are extremely underrepresented in top journals in some disciplines, such as mathematics (Mihaljević-Brandt et al., 2016, p. 19), and they can self-select into lower-ranked journals. Gender differences in the propensity to choose competitive environments (in our case, highly selective journals) are reported to be driven by gender differences in confidence and preferences for entering and performing in a competition (Niederle & Vesterlund, 2007, pp. 1098–1100). Gender differences in choices over competition may be driven partly by men preferring competitive to non-competitive settings and by a significantly stronger aversion to competitive workplaces among women compared to men (Flory et al., 2015). Not surprisingly, male scientists overcite (King et al., 2017; Maliniak et al., 2013), are better represented in top journals, and have higher visibility in science (Maddi et al., 2019).

Social norms or expectations of conventional behavior may also matter: there may be a common social practice, particularly in male-dominated disciplines of science, that "holds women up to more scrutiny than men" (Gupta, Poulsen, & Villeval, 2011, p. 16). Sonnert and Holton (1996, p. 69), in their study of gender disparities in career patterns of especially promising scientists, conclude that women might be seen as socialized to be less competitive "so that they choose their own niche rather than enter the fray with numerous competitors working on the same topic," often feeling they are "under the magnifying glass." Male scientists may be "more aggressive, combative and self-promoting in their pursuit of career success, and so they achieve higher visibility" (Sonnert & Holton, 1996, p. 67).

At the same time, in more firmly male-dominated disciplines (such as physics, astronomy, engineering, and computing, in the Polish case), female scientists may feel more intense performance pressure due to their high visibility among the overwhelming majority of male scientists and carrying the burden of representing women in these disciplines. They may have to work "twice as hard to prove their competence," with all their actions being public, as Kanter (1977, p. 973) suggested in her classic study of the role of male-female proportions in workplace settings. Being less competitively inclined in an increasingly competitive environment of global science may hurt female scientists, especially in their early careers, at an individual level of obtaining tenure, salary increases, and research funding (Van den Besselaar & Sandström, 2015; Sarsons et al., 2020; Kwiek, 2018a). In Polish academia, the list of gender-balanced disciplines goes beyond the social sciences and humanities (to include also business, economics, agricultural and biological sciences, medicine, chemistry and biochemistry, genetics, and psychology). Out of the 24 ASJC Scopus disciplines studied in this paper, female representation reaches at least 50% in 13 of them, which is a slight majority.

2.3. Gender Homophily in Research Collaboration Defined

The literature investigating gender homophily in academic publishing is based both on research on selected institutions (e.g., McDowell & Smith, 1992), selected disciplines (predominantly economics, as in Boschini & Sjörgen, 2007, or McDowell, Singell, & Stater, 2006), and large-scale bibliometric data (see Wang, Lee, West, Bergstrom, & Erosheva, 2019, who examined 252,413 papers with 807,588 authorships from the JSTOR corpus, or Ghiasi et al., 2015, who studied approximately one million Web of Science authorships in engineering).

Most recent bibliometric studies on gender differences in research collaboration patterns suggest that men tend to co-author with men and women with women—leading to the research theme of "gender homophily" in science (Ghiasi et al., 2018; Potthoff & Zimmermann, 2017; Lerchenmueller et al., 2019; Kegen, 2013; Wang et al., 2019; Boschini & Sjögren, 2007). At the same time, however, collaboration in research, traditionally operationalized as co-authored publications, influences career progress. Excessive gender homophily among women, while supportive for early-career female researchers, may also harm their careers. This is especially relevant for particularly able female scientists publishing in high-impact journals (as Lerchenmueller et al., 2019, show with powerful empirical evidence). Women may place themselves at a disadvantage when collaborating disproportionately with other women because, for example, "women tend to be part of less resource-rich and influential networks or because women's work may receive less attention than men's, likely harming career progress" (Lerchenmueller et al., 2019, p. 3). This is not the case in Poland, though, as we shall demonstrate, since the Polish female scientists studied tend to avoid publishing exclusively with other female scientists at all levels of their careers and for all age cohorts.

The homophily principle maintains that "similarity breeds connection": consequently, personal networks are homogeneous with regard to many sociodemographic, behavioral, and intrapersonal characteristics. Homophily is known to "limit people's social worlds in a way that has powerful implications for the information they receive, the attitudes they form, and the interactions they experience" (McPherson, Smith-Lovin, & Cook, 2001, p. 415). Research collaboration in science and team formation or gender co-authorship patterns provide fertile ground to test the homophily principle. According to this principle, contact between similar people occurs at a higher rate than among dissimilar people; in other words, "birds of a feather flock together" (McPherson et al., 2001, p. 417). Thus, males should co-author with males in a disproportionate fashion, while females should co-author disproportionately with females, across countries, disciplines, and institutions.

Homophily, in general, (including the gender-based homophily examined in this research) is reported to simplify communication, enhance the predictability of behavior, entail reciprocity in collaboration, and increase trust between collaborating parties (McPherson et al., 2001, p. 435; Kegen, 2013, p. 63). While the behavior of collaborators might be more predictable and collaboration potentially less costly and less risky, gender homophily might also exclude women from powerful informal networks. Furthermore, embeddedness in academic social networks—especially informal networks—is crucial both for doing research and for achieving a career (Kegen, 2013, p. 65). "Networks matter. Producing high-quality work is not sufficient for research to gain the attention of the widest number of scholars or have the greatest impact" (Maliniak et al., 2013, p. 918).

If homophily means "the tendency of people to choose to interact with similar others" (McPherson et al., 2001, p. 435), then gender-based homophily in this research means Polish male scientists disproportionately co-authoring with other male scientists, and Polish female scientists co-authoring disproportionately with other female scientists. Recent research tends to indicate that female scientists exhibit stronger gender homophily than male scientists (Jadidid, Karimi, Lietz, & Wagner, 2018): females are reported to collaborate more often with females than males with males (Kegen, 2013;

Lerchenmueller et al., 2019; Ghiasi et al., 2018). Evidence from co-authorship patterns in economics indicates that team formation in academic publishing is not gender-neutral: rather, there is powerful gender sorting in team formation (Boschini & Sjögren, 2007). However, the practices of collaboration between males and females differ across disciplines (Maddi et al., 2019); the patterns of international research collaboration differ cross-nationally (Kwiek 2020a on 28 European countries) and between genders intra-nationally (Kwiek 2020b and Kwiek & Roszka 2020 on Poland).

Male-female collaboration practices in research will be tested in this paper against the homophily principle: does similarity breed connection between individuals (McPherson et al., 2001), and does it structure publishing ties? There are many types of homophily (age-based, race-based, education-based, wealth-based, etc.), but this paper explores a single dimension: gender-based homophily.

2.4. Hypotheses of This Research

Following a comprehensive literature review and based on prior in-depth knowledge of the Polish academic science system, we have formulated the following eight research hypotheses (with the results of our research, Table 1):

Table 1. Research hypotheses and results (summary).

Hypothesis	Result
Hypothesis 1. The propensity to conduct same-sex collaboration is higher for female than for male scientists	Not confirmed
Hypothesis 2. The prestige level of mixed-sex publications is higher than that of same-sex publications for both male and female scientists	Confirmed
Hypothesis 3. The propensity to conduct same-sex collaboration decreases with age for both male and female scientists	Confirmed for males only
Hypothesis 4. The propensity to conduct same-sex collaboration decreases with academic position for both male and female scientists	Confirmed for males only
Hypothesis 5. The propensity to conduct same-sex collaboration is higher in male-dominated academic disciplines	Confirmed
Hypothesis 6. The propensity to conduct same-sex collaboration is higher in research-intensive universities	Confirmed for males only
Hypothesis 7. In logistic regression analysis, individual-level independent variables are more influential in predicting whether a scientist is highly homophilous than are institutional-level independent variables	Confirmed
Hypothesis 8. In linear logistic regression analysis, the percentage of articles written in same-sex collaboration is influenced by both individual-level and institutional-level independent variables	Confirmed

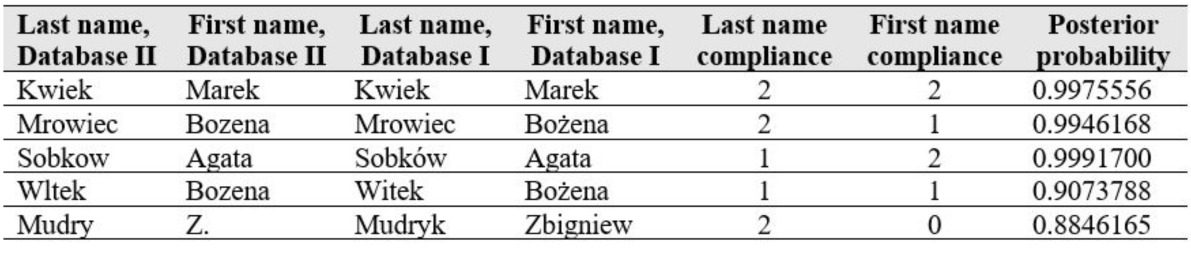
3. Data and Methods

3.1. Dataset

Two large databases were merged: Database I was an official national administrative and biographical register of all Polish scientists; Database II was the Scopus database, an official publication and citation source used for individual- and institutional-level evaluation in Poland. The two were merged to create “The Observatory of Polish Science,” which was maintained and periodically updated by the two authors. Database I (created by the OPI National Research Institute) comprised 99,535 scientists employed in the Polish science sector as of November 21, 2017. Only scientists with at least a doctoral degree (70,272) and employed in the higher education sector were selected for further analysis (54,448 or 54.70% of all scientists with at least a doctoral degree working at 85 universities of various types). The data used were both demographic (gender and date of birth) and professional (highest degree awarded; award date of Ph.D., habilitation, and full professorship; and institutional affiliation), with each scientist identified by a unique ID. Database II, the original Scopus publication and citation database, included 169,775 names from 85 institutions whose publications for the decade analyzed (2009–2018) were included in the database. Authors in Database II were defined by their institutional affiliations, Scopus documents, and individual Scopus IDs. We did not reconstruct full publishing careers (as in Huang et al., 2020) of Polish scientists but the last decade, when their Scopus publications increased markedly.

The key procedure was to appropriately identify authors with their different individual IDs in the two databases and to provide them with a new ID in the integrated “Observatory” database. Probabilistic methods of data integration were used (as defined in Fellegi & Sunter, 1969; Herzog et al., 2007; and Enamorado, Fifield, & Imai, 2019). Separately within each of the 85 universities, the first name and last name records of each record in Database I were compared with each of the records in Database II using the Jaro-Winkler string distance (with values from 0 to 1; see Jaro, 1989; Winkler, 1990). Pairs of strings with a distance greater than 0.94 were considered identical (signified by 2) (see Table 2), pairs with a distance greater than 0.88 but less than 0.94 were considered similar (signified by 1), while those with a distance less than 0.88 were considered disparate (signified by 0). Next, using an expectation maximization algorithm (Enamorado et al., 2019), the posterior probability that a given pair of records belongs to the same unit was estimated. If the probability was greater than 0.85, the pair was considered to be part of the same unit (Harron et al., 2017). The computation was made using the fastLink R package (version 0.6.0).

Table 2. An example of probabilistic integration output (identical, similar, and disparate pairs of strings).



Last name, Database II	First name, Database II	Last name, Database I	First name, Database I	Last name compliance	First name compliance	Posterior probability
Kwiek	Marek	Kwiek	Marek	2	2	0.9975556
Mrowiec	Bozena	Mrowiec	Bożena	2	1	0.9946168
Sobkow	Agata	Sobków	Agata	1	2	0.9991700
Wltek	Bozena	Witek	Bożena	1	1	0.9073788
Mudry	Z.	Mudryk	Zbigniew	2	0	0.8846165

By employing a probabilistic approach to the merging of the data sets, it was possible to estimate the uncertainty of the process and, thus, assess the quality of the new integrated database by calculating the percentage of records incorrectly classified as matches (false discovery rate, FDR) and the percentage of records incorrectly classified as non-matches (false negative rate, FNR). An integrated database obtained in accordance with the above procedures and used in our research finally included 37,081 records.[1] Database I contained biographical and professional career information on all authors affiliated with the 85 largest Polish universities in the 2009–2018 reference period. Database II contained metadata on 377,886 papers. From among the 377,886 papers in the original Database II, 230,007 were written by the authors included in Database I. Subsequently, only articles written in journals were selected for further analysis, with the number of papers in the database reducing to 158,743 articles. Approximately half of the Polish scientists from the higher education sector did not publish a paper indexed in the Scopus database in the reference period—which is in line with previous findings regarding the distribution of Polish publications—with the overwhelming majority of publications belonging to national publication outlets.

3.2. Methods

Every Polish scientist represented in our integrated database was ascribed to one of 334 ASJC disciplines at the four-digit level and one of 27 ASJC disciplines at the two-digit level (following Abramo, Aksnes, & D'Angelo, 2020, who defined in their study the dominant Web of Science subject category for each Italian and Norwegian professor). In the ASJC system used, a given paper can have one or multiple disciplinary classifications.[2] The dominant ASJC for each scientist was taken as the mode for each of them: the most frequently occurring value. In the case when no

[1] There were 38,750 records referring to 32,937 unique authors (more than one occurrence in Database II was found for 4,452 people or 13.51% of unique authors). With regard to quality, FDR was 0.21% and FNR was 39.91%. The high value of FNR is the result of duplicate instances in the database due to errors in the Scopus database sometimes assigning one author to different Scopus IDs. There were 9,931 records that referred to more than one person, where 3,679 (82.63%) occurred twice, 609 (13.68%) occurred three times, and 169 (3.68%) occurred four or more times. Therefore, for duplicated records, a clerical review was performed (as suggested in Herzog et al., 2007). Manual verification of duplicate records revealed that 1,207 records (12.15% in terms of duplicated records and 3.11% of all integrated records) were incorrectly assigned to the same person. These records were deleted from the integrated database, yielding N = 37,081 records.

[2] The ASJC discipline codes were described in the paper in the following manner: AGRI Agricultural and Biological Sciences; HUM Arts and Humanities; BIO Biochemistry, Genetics, and Molecular Biology; BUS Business, Management, and Accounting; CHEMENG Chemical Engineering; CHEM Chemistry; COMP Computer Science; DEC Decision Science; DENT Dentistry, EARTH Earth and Planetary Sciences; ECON Economics, Econometrics, and Finance; ENER Energy; ENG Engineering; ENVIR Environmental Science; IMMU Immunology and Microbiology; MATER Materials Science; MATH Mathematics; MED Medicine; NEURO Neuroscience; NURS Nursing; PHARM Pharmacology, Toxicology, and Pharmaceutics; PHYS Physics and Astronomy; PSYCH Psychology; SOC Social Sciences; VET Veterinary; DENT Dentistry; and HEALTH Health Professions. Non-STEM disciplines in our analysis include BUS, DENT, ECON, HEALTH, HUM, MED, PSYCH, SOC, and VET.

single mode occurred, the dominant ASJC was randomly selected. Consequently, we had Polish scientists defined by their gender and ASJC discipline, along with all their publications written solo and in male-only, female-only, and mixed collaborations. We also had a proportion of female scientists in every ASJC discipline. Furthermore, three disciplines were omitted from analysis as they did not meet an arbitrary minimum threshold of 50 scientists per discipline (GEN, NEURO, and NURS).

In the present research, in which the unit of analysis was an individual scientist, every scientist in our integrated database had solo or collaborative articles. Collaborative articles include same-sex and mixed-sex articles. Collaborative articles with authors included in our database are defined in terms of the gender of the authors. Of the Polish scientists included in the integrated database of 54,448 scientists, 100% had their gender defined. In contrast, there are Polish co-authors outside of our database (e.g., affiliated with the Polish Academy of Sciences) and international co-authors of publications with Polish co-authors whose gender is not defined.

Regarding international collaborators of Polish authors and their gender, we analyzed 158,743 articles with individual EIDs (Scopus individual publication IDs). There were 15,149 articles (9.54%) written solely by female scientists, 39,089 (24.62%) written solely by male scientists, 78,419 (49.40%) written in mixed female-male collaboration, and 18,109 (11.41%) solo-written articles. There were 7,979 articles (5.03%) for which only the gender of Polish co-authors was known.

For the purpose of determining the gender of the international co-authors, we used another dataset at our disposal: a dataset of 27.4 million articles published in the same period of 2009–2018 in the OECD area and indexed in Scopus. Our "OECD" dataset includes all metadata about all publications produced in the study period in 1,674 research-active institutions located in 40 OECD economies (the threshold we used was 3,000 Scopus-indexed articles published in the past 10 years). Specifically, we used a subset of our OECD dataset of authors (with 11,087,392 individual Scopus IDs). In the next step, we used the R package of GenderizeR to estimate the gender of the OECD authors from our OECD dataset (see Wais, 2016, on the various gender determination methods, including via the R package).

GenderizeR was previously used for gender prediction in Topaz and Sen (2016) for gender representation in editorial boards in 435 journals in mathematical sciences; Fell and König (2016) studied gender difference in co-authorships among 4,234 industrial-organizational psychologists; Huang et al. (2020) examined gender inequality in academic careers of 7.9 million Web of Science authors. Finally, Wang et al. (2019) also used the R package to study gender-based homophily in JSTORE publication data. For each first name provided, genderize.io returns a count of the number of times that name appears in the corpus and corresponding probabilities of gender (binary: male, female) based on frequency counts. In order to establish optimal values of gender prediction indicators, we can manipulate the threshold of probability and count values.

Using the R package, the gender of 7,640,123 authors (individual Scopus IDs) was estimated with a probability of greater than or equal to 0.85. With the data at our disposal, out of 11,087,392 authors, the genderizeR algorithm was unable to estimate the gender of 2,521,150 authors (22.74%), including a large number of authors from Japan and South Korea, with whom Poland collaborates only marginally. Out of 8,566,242 authors whose gender the algorithm estimated, in 926,119 (10.81%) of cases, gender was estimated with a probability lower than 0.85. In the next step, using individual Scopus IDs, the "Observatory" and the "OECD" datasets were merged to determine the gender of international collaborators of Polish authors. Out of 164,908 international collaborators, we were able to determine the gender of 83,702 (or 50,75%). Our reference database to estimate the gender of co-authors was restricted to 1,674 research-intensive OECD universities; consequently, we were not able to estimate the gender of collaborators from non-research-intensive universities in the OECD area or from non-OECD universities.

Having an individual scientist as the unit of analysis, we calculated the proportion of same-sex publications among collaborative articles within the individual publication portfolio of every Polish scientist in the sample. Thus, for all scientists, male and female, within their collaborative articles only, we determined the propensity to conduct same-sex collaboration (for male scientists collaborating only with male scientists, the propensity is 1). Analogously, the propensity of 0 is equivalent to conducting no same-sex collaboration—the scientist collaborates only with the other gender (i.e., there are only mixed-sex publications in the scientist's individual publication portfolio).

3.3. Sample

The structure of the sample (N = 25,463) is presented in Tables 3 and 4: approximately half of the scientists are in the 36–50 age bracket (51.5%), and over half of them are assistant professors (56.0%). Column percentages enable the analysis of the gender distribution of the Polish academic profession by age cohorts, academic positions, and disciplines, while row percentages enable the analysis of how male and female scientists are distributed according to a given age cohort, academic position, and discipline (Table 3). Table 4 shows age distributions for each academic position from a gender perspective. The three largest disciplines represented in the sample are agricultural and biological sciences, engineering, and medicine (AGR, ENG, and MED), representing over one-third of the scientists (37.8%).

Female participation in the academic profession decreases with age: while female scientists represent approximately half of all scientists aged 31–35, they represent only about a quarter of all scientists aged 61–65 years (49.8% and 26.7%, respectively). Female scientists are also clustered in lower academic positions: while females constitute about half of all assistant professors, they represent only about a quarter of full professors (48% and 24%, respectively, levels comparable to those in many other countries; for Sweden, see Madison & Fahlman, 2020, and for global overviews, see Halevi, 2019; Larivière et al., 2013; and Diezmann & Grieshaber, 2019). Polish assistant professors under 45 (our entire sample includes scientists with doctorates only) have an almost equal gender distribution. The older professors (aged 41–55 years) with a habilitation degree (a second, postdoctoral degree) are already dominated by male scientists (who represent approximately 60% of associate professors). In the case of full professors, the number of males is at least three times that of females (see Table 4) for every age cohort for both young full professors (aged 41–45) and the oldest ones (aged 61–65). All associate professors as defined in this paper hold doctoral degrees, all associate professors hold habilitations, and all full professors hold full professorships.

The age structure by gender of the sample is presented in Figure 1. Our sample contains only scientists who had at least a single publication in the Scopus database in the period 2009–2018 and, therefore, it includes all internationally productive Polish academic scientists (on research productivity of Polish scientists, see Kwiek 2018b). Additionally, our sample includes the international collaborators of Polish authors whose gender was determined using the algorithm described in the Data and Methods subsection (164,908 international co-authors). The differentiated proportions of female scientists can also be examined by academic discipline. Female scientists are severely underrepresented in computer science (COMP 16.5%), engineering (ENG 14.9%), physics and astronomy (PHYS 16.6%), and mathematics (MATHS 25.2%). In arts and humanities (HUM) and social sciences (SOC), the distribution of scientists by gender is practically equal (49.8%).

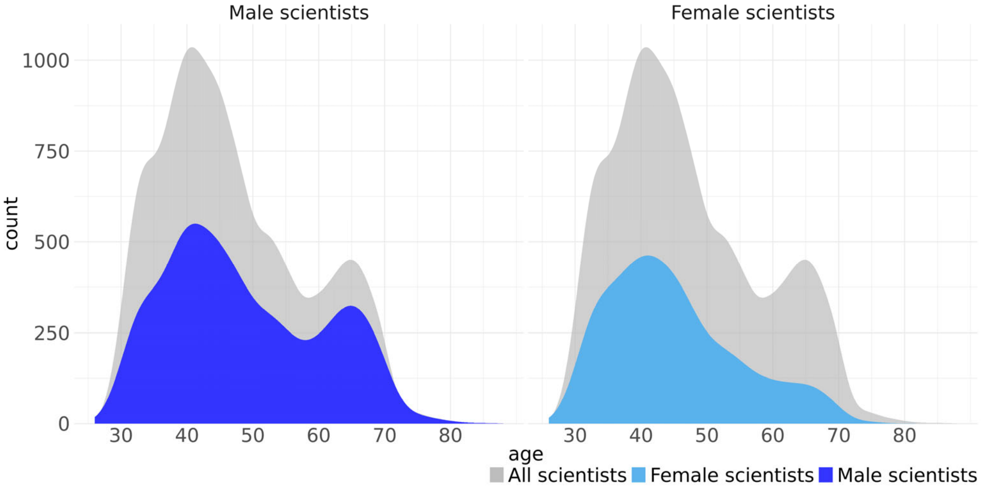


Figure 1. Age structure of the sample, all Polish internationally productive university professors (N = 25,463), by gender. All university professors in grey.

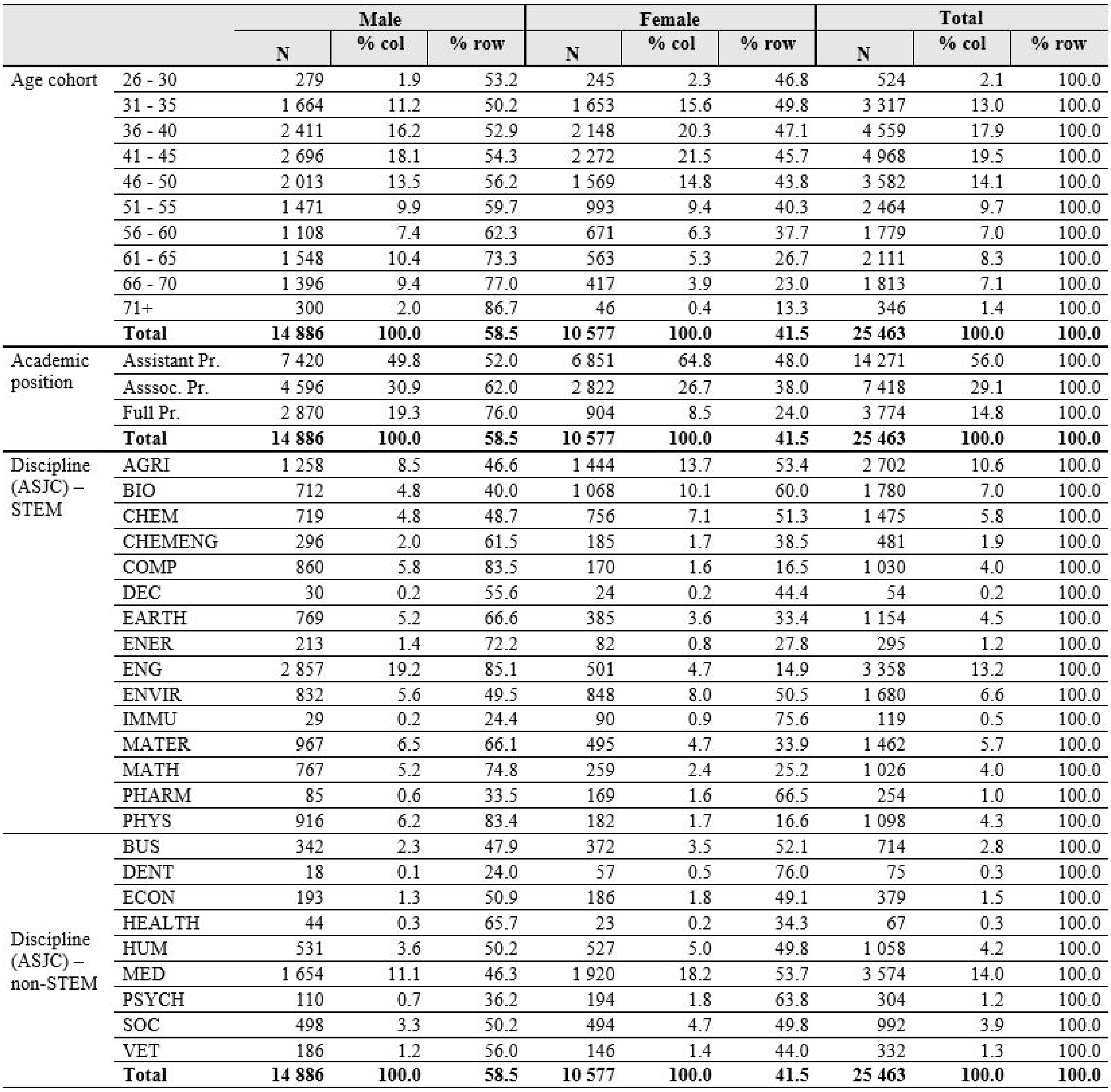
4. Results

4.1. Publication Prestige, Academic Disciplines, and Major Gender-Defined Research Collaboration Types

Hypothesis 1. The propensity to conduct same-sex collaboration is higher for female than for male scientists (not confirmed).
Hypothesis 2. The prestige level of mixed-sex publications is higher than that of same-sex publications for both male and female scientists (confirmed).

Gender homophily in publishing, or the propensity to conduct same-sex collaboration, falls within the range of 0 (no same-sex collaborative articles among collaborative articles in the individual publication portfolio) to 1 (exclusively same-sex collaborative articles among collaborative articles in the portfolio). As clearly seen in Table 5, the average propensity of males to be involved in same-sex collaboration is more than three times that of females (the median propensity for males is 0.500, compared with 0.153 for females). For the whole national sample, the median propensity is 0.333, meaning that at least 50% of authors conduct same-sex collaboration (males with males, females with females) at the 33.3% level. The Mann-Whitney's Z-test shows the gender difference to be significantly different at the significance level of 0.05. Thus, Hypothesis 1 is not confirmed.

Table 3. Structure of the sample, all Polish internationally productive university professors, by gender, age cohort, academic position, and discipline, presented with column and row percentages.

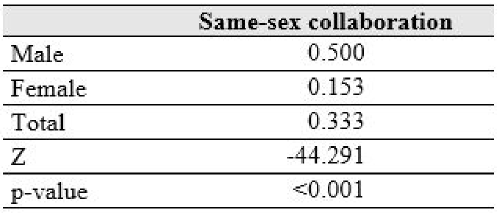


		Male			Female			Total		
		N	% col	% row	N	% col	% row	N	% col	% row
Age cohort	26 - 30	279	1.9	53.2	245	2.3	46.8	524	2.1	100.0
	31 - 35	1 664	11.2	50.2	1 653	15.6	49.8	3 317	13.0	100.0
	36 - 40	2 411	16.2	52.9	2 148	20.3	47.1	4 559	17.9	100.0
	41 - 45	2 696	18.1	54.3	2 272	21.5	45.7	4 968	19.5	100.0
	46 - 50	2 013	13.5	56.2	1 569	14.8	43.8	3 582	14.1	100.0
	51 - 55	1 471	9.9	59.7	993	9.4	40.3	2 464	9.7	100.0
	56 - 60	1 108	7.4	62.3	671	6.3	37.7	1 779	7.0	100.0
	61 - 65	1 548	10.4	73.3	563	5.3	26.7	2 111	8.3	100.0
	66 - 70	1 396	9.4	77.0	417	3.9	23.0	1 813	7.1	100.0
	71+	300	2.0	86.7	46	0.4	13.3	346	1.4	100.0
	Total	**14 886**	**100.0**	**58.5**	**10 577**	**100.0**	**41.5**	**25 463**	**100.0**	**100.0**
Academic position	Assistant Pr.	7 420	49.8	52.0	6 851	64.8	48.0	14 271	56.0	100.0
	Asssoc. Pr.	4 596	30.9	62.0	2 822	26.7	38.0	7 418	29.1	100.0
	Full Pr.	2 870	19.3	76.0	904	8.5	24.0	3 774	14.8	100.0
	Total	**14 886**	**100.0**	**58.5**	**10 577**	**100.0**	**41.5**	**25 463**	**100.0**	**100.0**
Discipline (ASJC) – STEM	AGRI	1 258	8.5	46.6	1 444	13.7	53.4	2 702	10.6	100.0
	BIO	712	4.8	40.0	1 068	10.1	60.0	1 780	7.0	100.0
	CHEM	719	4.8	48.7	756	7.1	51.3	1 475	5.8	100.0
	CHEMENG	296	2.0	61.5	185	1.7	38.5	481	1.9	100.0
	COMP	860	5.8	83.5	170	1.6	16.5	1 030	4.0	100.0
	DEC	30	0.2	55.6	24	0.2	44.4	54	0.2	100.0
	EARTH	769	5.2	66.6	385	3.6	33.4	1 154	4.5	100.0
	ENER	213	1.4	72.2	82	0.8	27.8	295	1.2	100.0
	ENG	2 857	19.2	85.1	501	4.7	14.9	3 358	13.2	100.0
	ENVIR	832	5.6	49.5	848	8.0	50.5	1 680	6.6	100.0
	IMMU	29	0.2	24.4	90	0.9	75.6	119	0.5	100.0
	MATER	967	6.5	66.1	495	4.7	33.9	1 462	5.7	100.0
	MATH	767	5.2	74.8	259	2.4	25.2	1 026	4.0	100.0
	PHARM	85	0.6	33.5	169	1.6	66.5	254	1.0	100.0
	PHYS	916	6.2	83.4	182	1.7	16.6	1 098	4.3	100.0
Discipline (ASJC) – non-STEM	BUS	342	2.3	47.9	372	3.5	52.1	714	2.8	100.0
	DENT	18	0.1	24.0	57	0.5	76.0	75	0.3	100.0
	ECON	193	1.3	50.9	186	1.8	49.1	379	1.5	100.0
	HEALTH	44	0.3	65.7	23	0.2	34.3	67	0.3	100.0
	HUM	531	3.6	50.2	527	5.0	49.8	1 058	4.2	100.0
	MED	1 654	11.1	46.3	1 920	18.2	53.7	3 574	14.0	100.0
	PSYCH	110	0.7	36.2	194	1.8	63.8	304	1.2	100.0
	SOC	498	3.3	50.2	494	4.7	49.8	992	3.9	100.0
	VET	186	1.2	56.0	146	1.4	44.0	332	1.3	100.0
	Total	**14 886**	**100.0**	**58.5**	**10 577**	**100.0**	**41.5**	**25 463**	**100.0**	**100.0**

Table 4. Structure of the sample, all Polish internationally productive university professors, by gender, age cohort, and academic position, presented with column and row percentages.

	Assistant Professor						Associate Professor						Full Professor						Total					
	Male			Female			Male			Female			Male			Female			Male			Female		
	N	% col	% row	N	% col	% row	N	% col	% row	N	% col	% row	N	% col	% row	N	% col	% row	N	% col	% row	N	% col	% row
26 - 30	280	3.8	53.2	246	3.6	46.8	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0	0	0.0	0.0	280	1.9	53.2	246	2.3	46.8
31 - 35	1621	21.8	49.8	1636	23.8	50.2	48	1.0	67.6	23	0.8	32.4	1	0.0	100.0	0	0.0	0.0	1670	11.2	50.2	1659	15.6	49.8
36 - 40	1956	26.3	50.2	1937	28.2	49.8	453	9.8	67.4	219	7.7	32.6	7	0.2	77.8	2	0.2	22.2	2416	16.2	52.8	2158	20.3	47.2
41 - 45	1652	22.2	51.0	1589	23.1	49.0	981	21.3	59.6	664	23.5	40.4	76	2.6	75.2	25	2.8	24.8	2709	18.1	54.3	2278	21.5	45.7
46 - 50	946	12.7	53.6	820	11.9	46.4	914	19.8	56.8	695	24.6	43.2	157	5.5	73.4	57	6.3	26.6	2017	13.5	56.2	1572	14.8	43.8
51 - 55	425	5.7	54.1	361	5.2	45.9	757	16.4	59.1	523	18.5	40.9	292	10.1	72.3	112	12.3	27.7	1474	9.9	59.7	996	9.4	40.3
56 - 60	223	3.0	53.3	195	2.8	46.7	513	11.1	60.4	336	11.9	39.6	374	13.0	72.2	144	15.9	27.8	1110	7.4	62.2	675	6.4	37.8
61 - 65	240	3.2	75.0	80	1.2	25.0	558	12.1	68.5	257	9.1	31.5	753	26.2	76.8	227	25.0	23.2	1551	10.4	73.3	564	5.3	26.7
66 - 70	88	1.2	87.1	13	0.2	12.9	338	7.3	76.3	105	3.7	23.7	974	33.9	76.3	302	33.3	23.7	1400	9.4	76.9	420	4.0	23.1
71+	13	0.2	100.0	0	0.0	0.0	45	1.0	84.9	8	0.3	15.1	243	8.4	86.5	38	4.2	13.5	301	2.0	86.7	46	0.4	13.3
Total	**7444**	**100.0**	**52.0**	**6877**	**100.0**	**48.0**	**4607**	**100.0**	**61.9**	**2830**	**100.0**	**38.1**	**2877**	**100.0**	**76.0**	**907**	**100.0**	**24.0**	**14928**	**100.0**	**58.4**	**10614**	**100.0**	**41.6**

Table 5. The median propensity to conduct same-sex collaboration by gender.



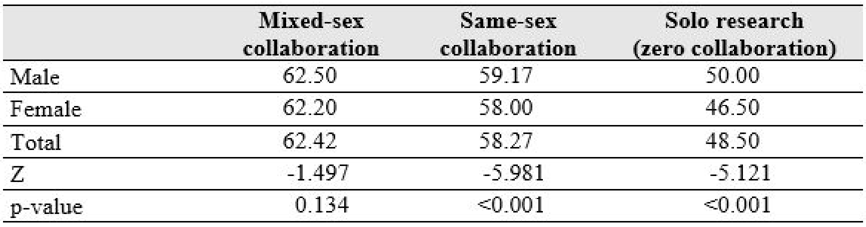
	Same-sex collaboration
Male	0.500
Female	0.153
Total	0.333
Z	-44.291
p-value	<0.001

Both the quantity and quality of output in academia are relatively easily measured (with all standard limitations) using the Scopus database: articles are published in journals of different ranks. The scientists in our sample have their own unique individual publication portfolio with publications, translatable into average individual prestige via Scopus citation metrics. The prestige of each article in this portfolio is derived from the prestige of the journal in which it was published and is defined by the percentile rank ascribed annually to each academic journal within its ASJC discipline. Top journals, including *Journal of Informetrics*, are usually located in the 95th percentile of journals within a discipline and above (the upper 5% of journals).

Importantly, the citation-based percentile ranking system used by Scopus is being systematically used in Poland, both in a new points-based research assessment exercise (expected in 2022) and in a complicated system of indicators used first to select (in 2019) and then to additionally finance (in 2020–2026) 10 research-intensive Polish universities. We used the measure of average prestige, which represents the median prestige value for all publications written by a given scientist in the study period of 2009–2018 for three categories of publications (same-sex, mixed-sex, and solo publications). For journals for which the Scopus database did not ascribe a percentile rank, we have ascribed the percentile rank of 0; Scopus ascribes percentiles to journals in the 25th to 99th percentile range, with the highest rank being the 99th percentile.

The median prestige level (in a range of 0–100) for publications written in same-sex and mixed-sex collaboration by gender does not differ much (Table 6): the median values for all-male publications and all-female publications by gender are almost identical (59.17 and 58.00, respectively). Also, the median value for mixed-sex collaborations does not differ significantly by gender. Both males and females, on average, regardless of the collaboration type, publish in journals with relatively low prestige. Articles written in mixed-sex collaboration are, on average, published in more prestigious journals than those written in same-sex collaboration, and in much more prestigious journals than solo articles (see the Total line in Table 6).

Table 6. The median prestige level distribution (by percentile from 0–99, with the 99th percentile being the highest) of publications by major gender collaboration type and gender.



	Mixed-sex collaboration	Same-sex collaboration	Solo research (zero collaboration)
Male	62.50	59.17	50.00
Female	62.20	58.00	46.50
Total	62.42	58.27	48.50
Z	-1.497	-5.981	-5.121
p-value	0.134	<0.001	<0.001

The distribution of the median journal prestige by discipline and collaboration type (mixed, same-sex, and solo publications, separately for males and females) shows both common patterns and substantial variations. Generally, for each ASJC discipline (Table 7), solo research is characterized by the lowest prestige level. BIO, CHEM, ENER, and PHARM belong to disciplines with the highest median prestige level, regardless of the collaboration type. Both mixed-sex and same-sex collaborations have higher average prestige levels than do solo articles.

The differences in prestige level by gender are as follows: for mixed-sex collaborations, they are marginal, but for same-sex collaboration, they are substantial (compare the same-sex collaboration columns for males and females in Table 6). Male-only collaborations have higher median prestige than do female-only collaborations, and this pattern is characteristic of a large number of disciplines. Males collaborating with males, on average, publish in substantially more prestigious journals than females collaborating with females do. Solo research by females exhibits lower median prestige levels than solo research by males in all except for nine disciplines (including BIO, CHEMENG, ENER, ENG, MATER, MED, and PHARM). Furthermore, solo research often sends clear signals of ability to employers, as opposed to mixed-sex research, for which female scientists may receive less credit than their male co-authors (Sarsons et al., 2020, p. 32; Fell & König, 2016). The median prestige level by ASJC discipline and gender is also shown graphically in the boxplots in Figure 2 to go beyond the median values and to highlight intra-disciplinary cross-gender variability, with three separate panels for the three gender-defined collaboration types. Thus, Hypothesis 2 is confirmed.

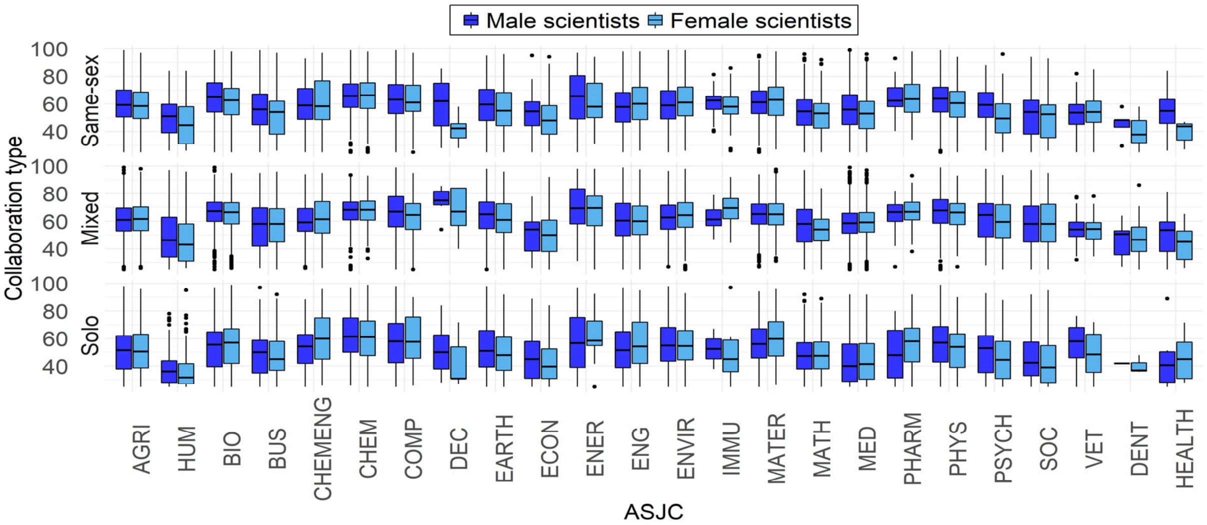


Figure 2. The prestige level distribution of publications (by Scopus percentile rank from 0–99, with the 99th percentile being the highest in prestige) by major collaboration type, gender, and discipline.

Table 7. The median prestige level for publications (by Scopus percentile ranks from 0–99, with the 99th percentile being the highest in prestige) by major collaboration type, gender and discipline (conditional formatting used to highlight differences).

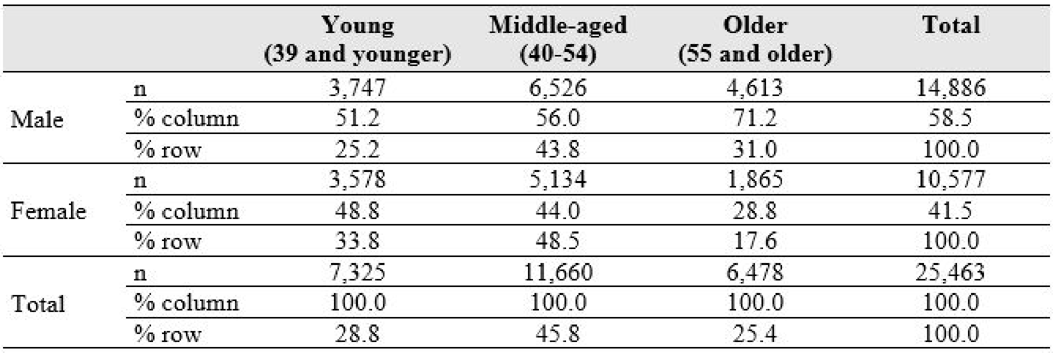
	Male			Female			Total			Mixed		Same sex		Solo	
	Mixed	Same sex	Solo	Mixed	Same sex	Solo	Mixed	Same sex	Solo	Z	p-value	Z	p-value	Z	p-value
AGRI	61	59	52	62	59	51	61	59	51	-1.446	0.148	-1.481	0.139	-0.181	0.857
BIO	67	65	56	67	63	57	67	64	56	-1.472	0.141	-3.439	0.001	-0.724	0.469
CHEM	68	66	61	68	66	61	68	66	61	-0.161	0.872	-0.381	0.703	-0.464	0.643
CHEMENG	59	59	54	62	58	60	60	59	58	-1.355	0.175	-0.424	0.671	-1.799	0.072
COMP	67	63	58	65	61	58	67	63	58	-1.831	0.067	-0.113	0.910	-0.381	0.703
DEC	75	62	50	67	42	31	74	46	45	-0.901	0.367	-2.018	0.044	-1.832	0.067
EARTH	65	60	51	61	55	48	63	58	50	-1.846	0.065	-2.727	0.006	-1.943	0.052
ENER	69	66	57	70	58	59	69	65	58	-1.084	0.278	-0.713	0.476	-0.662	0.508
ENG	61	58	52	60	60	54	60	58	52	-0.534	0.594	-2.067	0.039	-2.867	0.004
ENVIR	63	59	55	64	61	55	64	60	55	-2.057	0.040	-2.071	0.038	-0.033	0.974
HEALTH	54	55	41	45	44	45	50	48	45	-1.517	0.129	-2.668	0.008	-0.365	0.715
HUM	46	51	36	43	45	32	45	48	33	-0.992	0.321	-1.796	0.072	-2.157	0.031
IMMU	61	62	53	70	58	45	66	58	45	-2.695	0.007	-0.728	0.467	-0.471	0.637
MATER	65	61	56	65	63	60	65	62	58	-0.278	0.781	-1.293	0.196	-1.799	0.072
MATH	58	55	47	54	53	48	56	54	47	-2.446	0.014	-1.411	0.158	-0.103	0.918
PHARM	67	62	48	67	63	58	67	63	57	-0.702	0.483	-0.053	0.957	-0.298	0.765
PHYS	68	64	57	66	61	54	67	63	57	-1.461	0.144	-1.580	0.114	-1.529	0.126
BUS	58	56	50	58	54	45	58	55	47	-0.312	0.755	-2.198	0.028	-1.150	0.250
DENT	51	48	42	47	38	37	48	40	40	-0.037	0.970	-1.030	0.303	-0.447	0.655
ECON	54	55	45	50	48	40	53	52	42	-0.498	0.618	-1.337	0.181	-1.863	0.062
HEALTH	54	55	41	45	44	45	50	48	45	-1.517	0.129	-2.668	0.008	-0.365	0.715
HUM	46	51	36	43	45	32	45	48	33	-0.992	0.321	-1.796	0.072	-2.157	0.031
MED	59	56	40	59	53	42	59	54	41	-1.392	0.164	-4.630	0.000	-0.383	0.702
PSYCH	64	59	53	60	49	45	61	54	48	-0.679	0.497	-3.151	0.002	-0.985	0.325
SOC	58	54	43	58	53	39	58	53	41	-0.304	0.761	-1.320	0.187	-1.876	0.061
VET	54	54	58	54	54	49	54	54	52	-0.558	0.577	-1.034	0.301	-0.906	0.365
Total	63	59	50	62	58	47	62	58	49	-1.446	0.148	-1.481	0.139	-0.181	0.857

4.2. The Propensity to Conduct Same-sex Collaboration by Age and Academic Position

Hypothesis 3. The propensity to conduct same-sex collaboration decreases with age for both male and female scientists (confirmed for males but not for females).
Hypothesis 4. The propensity to conduct same-sex collaboration decreases with academic position for both male and female scientists (confirmed for males but not for females).

For the purposes of examining the propensity to conduct same-sex collaboration by age cohort, we divided our sample into the three categories: young scientists (aged 39 and younger), middle-aged scientists (aged 40–54) and older scientists (aged 55 and older), of which middle-aged scientists are the largest cohort (45.79%) (Table 8). The proportion of males and females is almost equal among young scientists—but females are less than 30% of older scientists (see % column).

Table 8. Distribution of the sample of Polish scientists by age cohort and gender.



		Young (39 and younger)	Middle-aged (40-54)	Older (55 and older)	Total
Male	n	3,747	6,526	4,613	14,886
	% column	51.2	56.0	71.2	58.5
	% row	25.2	43.8	31.0	100.0
Female	n	3,578	5,134	1,865	10,577
	% column	48.8	44.0	28.8	41.5
	% row	33.8	48.5	17.6	100.0
Total	n	7,325	11,660	6,478	25,463
	% column	100.0	100.0	100.0	100.0
	% row	28.8	45.8	25.4	100.0

Table 9 shows the distribution of the median value of the propensity to conduct same-sex collaboration by gender and age cohort. The median propensity by males slightly decreases with age. In contrast, the same median propensity for females substantially increases with age. While the propensity for females triples with age, it is still very low compared with that of males.

The difference in collaboration patterns for young scientists is striking: while half of young male scientists write at least 54% of their papers in collaboration with males, the same indicator for females is nine times lower (6.3%). Young males tend to collaborate with males—and young females tend *not* to collaborate with females. While 50% of young female scientists are characterized by the propensity to conduct same-sex collaboration at the level of 0.06, in the case of older females, the median propensity quadruples to 0.24: older females still tend to collaborate with males. For all age cohorts (see the Total line in Table 9), the difference in Polish science by gender is startling: while the median propensity to conduct same-sex collaboration for males is 0.5, the median for females is more than three times lower (0.15).

Table 9. The median propensity to conduct same-sex collaboration by age cohort and gender.

	Male	Female	Total	Z	p-value
Young (39 and younger)	0.5396	0.0625	0.2727	-29.676	<0.001
Middle-aged (40–54)	0.5000	0.1818	0.3333	-28.163	<0.001
Older (55 and older)	0.4762	0.2353	0.3750	-15.696	<0.001
Total	0.5000	0.1538	0.3333	-44.291	<0.001

What is clear in the two panels in Figure 3 is the predominance of extreme values (0 for no same-sex collaboration and 1 for exclusively same-sex collaboration) in individual publication portfolios. The total number of extreme values (0 and 1) is similar for both genders. The vast majority of collaborations are mixed-sex collaborations. The majority of collaborating male scientists (left panel, right peak) collaborated solely with males in the decade studied; the majority of collaborating female scientists (right panel, left peak), in contrast, *never* collaborated with females in the same period.

The distribution of the propensity to conduct same-sex collaboration for females is the mirror image of that for males. Apart from the two extreme values of 1 and 0, the distribution of the propensity in question for males is basically uniform. For females, a gradual decline in the propensity is clearly observed. Comparing the extremes, there are more females without same-sex collaboration than males for the same collaboration type; there are about three times more males who collaborate only with males compared with females who collaborate only with females.

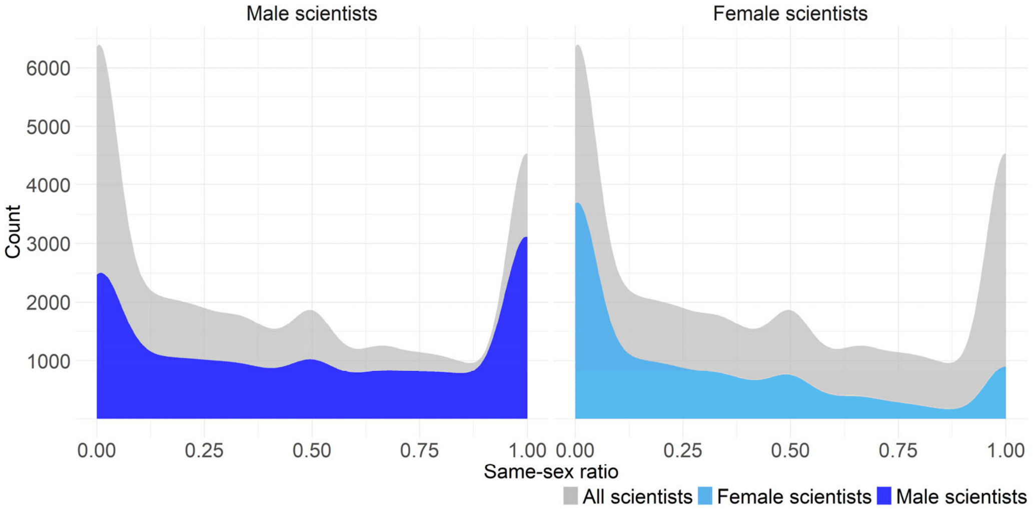
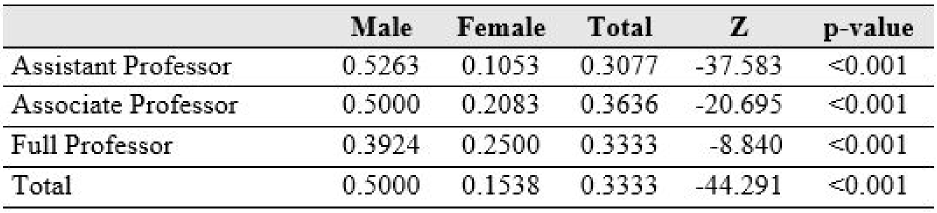


Figure 3. The distribution of the propensity to conduct same-sex collaboration by gender. The gray area is the overall distribution for both genders.

When we examine academic positions, in a similar vein, the propensity to conduct same-sex collaboration by males decreases with the highest academic position reached (Table 10). In contrast, the same propensity for females increases with academic positions, although its level is still very low for all females. While the median propensity level for females increases two and a half times when we move up the academic ladder, it is still much lower compared with that of males. While 50% of female assistant professors are characterized by the propensity to conduct same-sex collaboration at the level of 0.105, for female full professors, the propensity increases to 0.250.

Table 10. The median propensity to conduct same-sex collaboration by academic position and gender.



	Male	Female	Total	Z	p-value
Assistant Professor	0.5263	0.1053	0.3077	-37.583	<0.001
Associate Professor	0.5000	0.2083	0.3636	-20.695	<0.001
Full Professor	0.3924	0.2500	0.3333	-8.840	<0.001
Total	0.5000	0.1538	0.3333	-44.291	<0.001

The gender difference in collaboration patterns can be studied in more detail using boxplots and violin plots combined. The gender difference by age cohort (Figure 4) closely resembles the gender difference by academic position (Figure 5). Female scientists consistently, across the three age cohorts and across the three academic positions, tend *not* to collaborate with other females (compare the shapes for Propensity 1, i.e., females collaborating only with females, across the age cohorts and academic positions for females). Note that the median shown in boxplots is much lower for each cohort for females than for males, and it increases for females with age; it is also much lower for female assistant and associate professors and lower for female full professors.

Inverse proportionality in collaboration patterns between males and females is visible for each age cohort and each academic position. In terms of within-sex variation, male scientists are more differentiated than female scientists (compare the height of the boxes in the two columns) for each age cohort and each academic position studied. The vast majority of females, and especially young females and female assistant professors, tend not to collaborate with other females. As can be seen from Figures 4 and 5, generally, conclusions from a study of age cohorts resemble conclusions from a study of academic positions. In the specific Polish case, age and academic positions are strongly correlated as the principle of "up or out" has not been operative in the system for at least three decades. (Academic age stratification is one of the six major dimensions of social stratification in global science: see my recent monograph on academic performance stratification, academic salary stratification, academic power stratification, international research stratification, academic role stratification, and academic age stratification, Kwiek, 2019).

While above, we have studied three broad age cohorts, below, we focus on biological age as a numerical variable. The year-by-year approach illustrated by regression lines in Figure 6 generally confirms the two opposite trends for both genders, at least until the age of 60 for males and for all ages for females. Interestingly, the generally downward trend in the propensity to conduct same-sex collaboration for male scientists is reversed for those aged 60 and above: the propensity for the oldest males increases. In contrast, for female scientists, the damped growth characteristic of all ages until about 60 turns into exponential growth for the oldest female scientists (a cut-off point of 70 used, a retirement age for full professors). The dots in Figure 6 represent the median value of the propensity to conduct same-sex collaboration for each year of age. Relatively high variation of median values for very young male scientists and no variation for very young female scientists (see the respective dots in both panels) is caused by the low numbers of scientists in these age cohorts. Thus, Hypotheses 3 and 4 are confirmed for males but not for females.

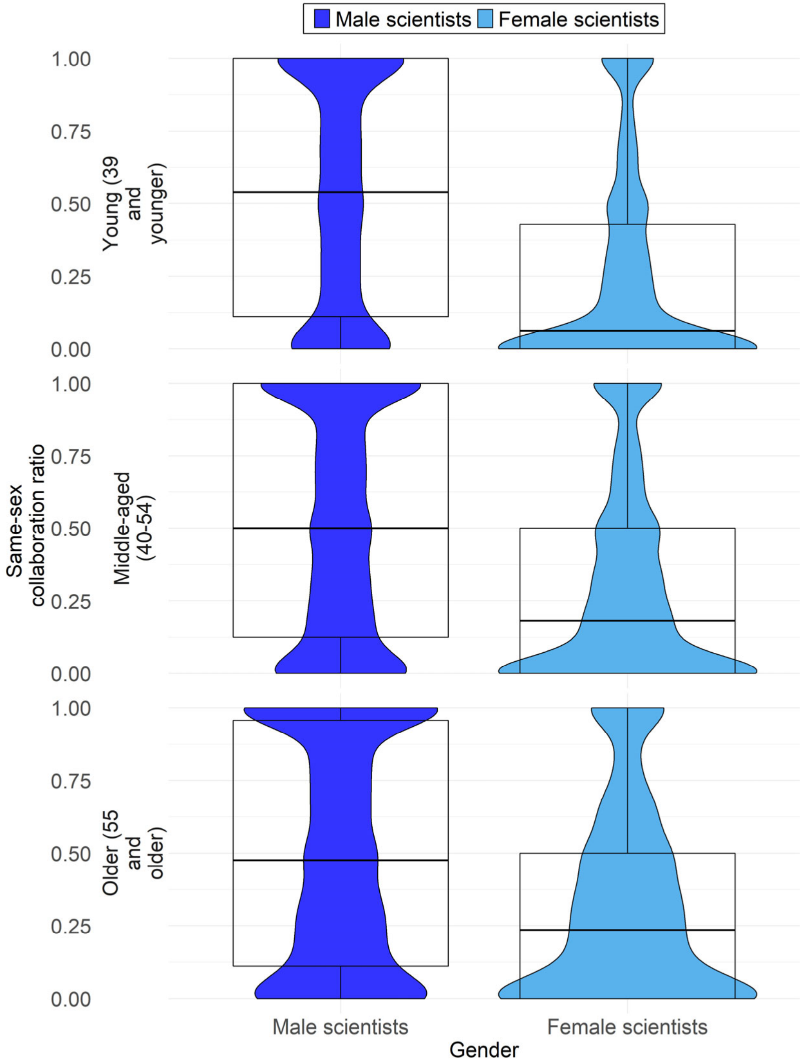


Figure 4. The propensity to conduct same-sex collaboration: distribution by age cohorts and gender (boxplots and violin plots combined).

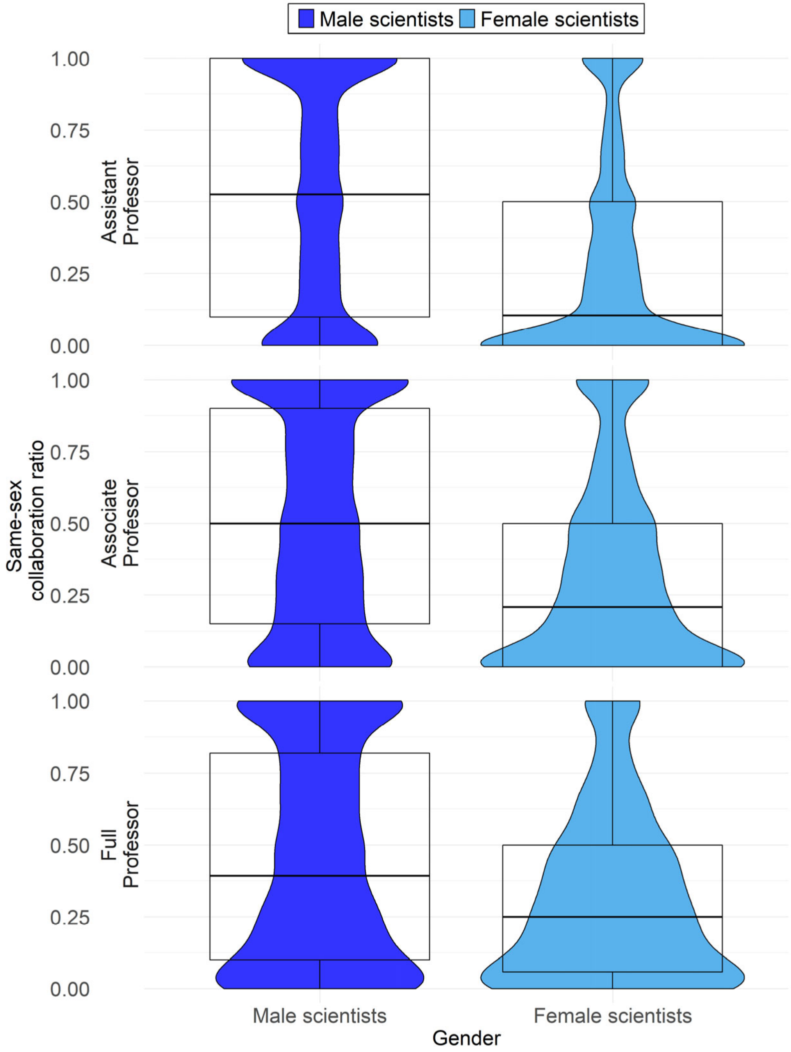


Figure 5. The propensity to conduct same-sex collaboration: distribution by academic position and gender (boxplots and violin plots combined).

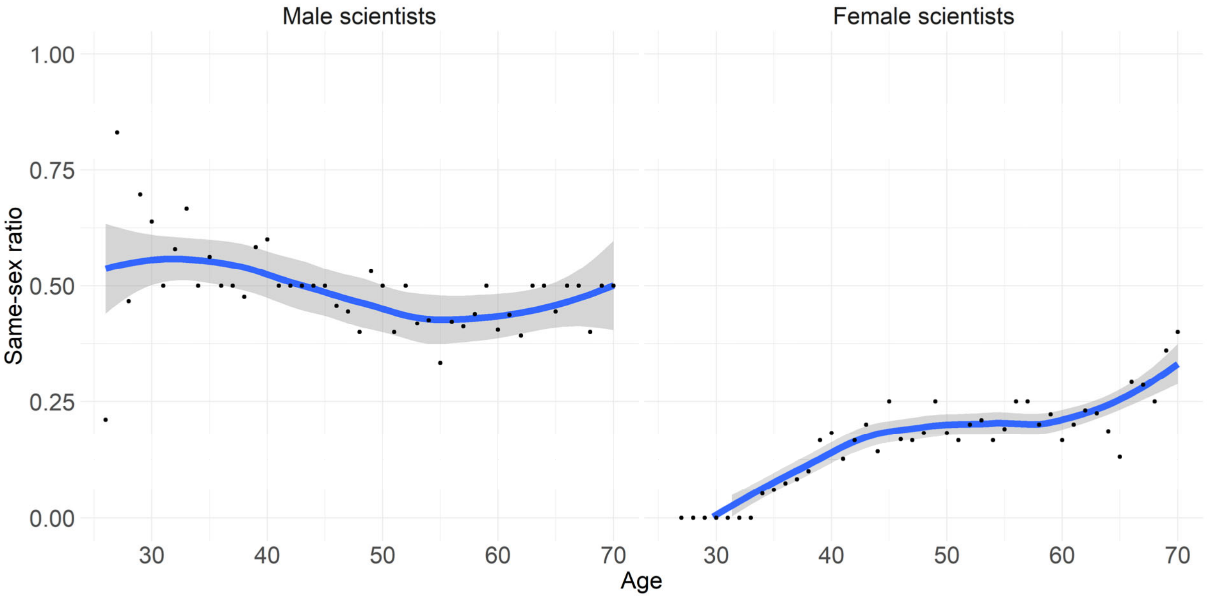


Figure 6. The propensity to conduct same-sex collaboration by gender and age. The regression line was estimated using the method of local polynomial regression fitting. The gray area represents 95% confidence intervals. Each year of age is represented by a single dot (a cut-off point of 70 used). Dots represent median values.

4.3. The Propensity to Conduct Same-Sex Collaboration by Academic Discipline

Hypothesis 5. The propensity to conduct same-sex collaboration is higher in male-dominated academic disciplines (confirmed).

The propensity to conduct same-sex collaboration differs vastly by discipline. Previous research shows that as the fraction of female researchers in a discipline increases, women increasingly tend to publish with other women; also, the male propensity to co-author with women is higher in disciplines with more women (Boschini & Sjögren, 2007, p. 339). A good way to visualize gender differences in the median propensity to conduct same-sex collaboration is through a heat map (the color palette in Table 11 changes from deep red for low values to deep green for high values). In the case of COMP, ENG, and MATH, with the high overrepresentation of male scientists, the propensity for males is extremely high (and the median values reach the level of 1 or almost 1). That is to say, at least half of male scientists in these disciplines collaborate only with males. In COMP, ENER, ENG, HEALTH, PHYS, and VET, at least half of females do not collaborate with females at all (and the median values reach the level of 0 or almost 0). In contrast, in disciplines such as PHARM, PSYCH, and SOC, the median value for females is significantly higher than for males. The median level by ASJC discipline is also shown graphically in boxplots in Figure 7. Thus, Hypothesis 5 is confirmed.

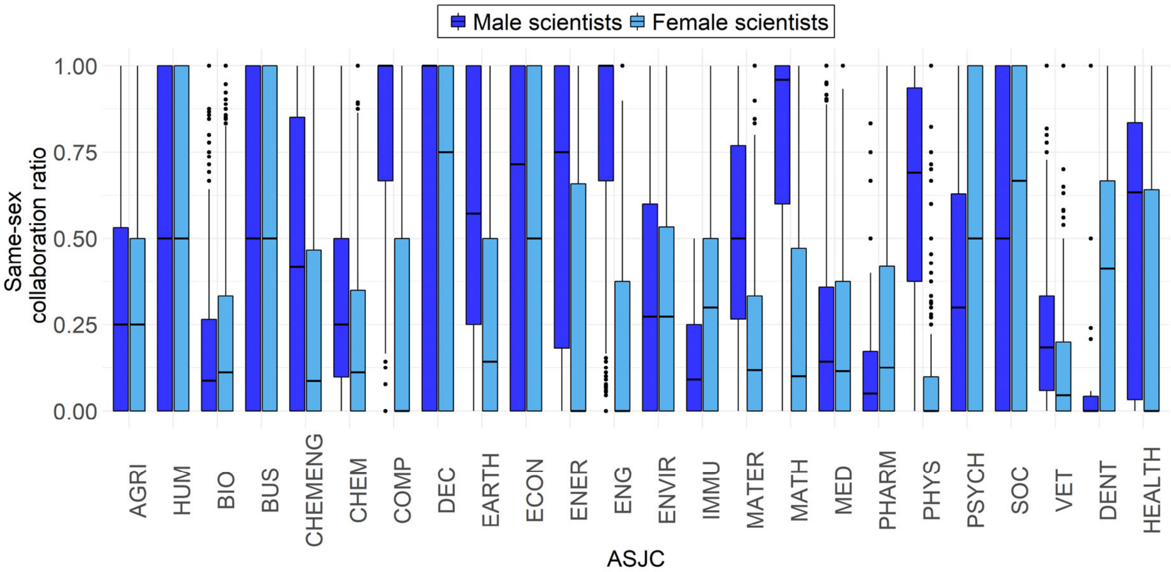


Figure 7. The propensity to conduct same-sex collaboration: distribution by discipline and gender.

Table 11. The median propensity to conduct same-sex collaboration by discipline and gender.

	Male	Female	Total	Z	p-value
AGRI	0.2500	0.2500	0.2500	-0.743	0.457
BIO	0.0870	0.1111	0.1000	-1.384	0.166
CHEM	0.2500	0.1111	0.1818	-8.753	<0.001
CHEMENG	0.4167	0.0861	0.2566	-5.403	<0.001
COMP	1.0000	0.0000	0.8750	-13.542	<0.001
DEC	1.0000	0.7500	1.0000	-0.518	0.604
EARTH	0.5714	0.1429	0.5000	-10.671	<0.001
ENER	0.7500	0.0000	0.6000	-4.570	<0.001
ENG	1.0000	0.0000	0.8889	-26.850	<0.001
ENVIR	0.2727	0.2727	0.2727	-1.310	0.190
IMMU	0.0909	0.3000	0.2500	-3.033	0.002
MATER	0.5000	0.1176	0.3750	-17.456	<0.001
MATH	0.9600	0.1000	0.7692	-16.172	<0.001
PHARM	0.0500	0.1250	0.0984	-3.319	0.001
PHYS	0.6903	0.0000	0.6000	-16.861	<0.001
BUS	0.5000	0.5000	0.5000	-0.490	0.624
DENT	0.0000	0.4118	0.2353	-3.270	0.001
ECON	0.7143	0.5000	0.6667	-1.456	0.145
HEALTH	0.6333	0.0000	0.5000	-2.146	0.032
HUM	0.5000	0.5000	0.5000	-0.150	0.880
MED	0.1429	0.1148	0.1250	-2.113	0.035
PSYCH	0.3000	0.5000	0.4710	-2.684	0.007
SOC	0.5000	0.6667	0.5000	-2.577	0.010
VET	0.1842	0.0455	0.1206	-5.178	<0.001
Total	0.5000	0.1538	0.3333	-44.291	<0.001

4.4. The Propensity to Conduct Same-Sex Collaboration by Institutional Type

Hypothesis 6. The propensity to conduct same-sex collaboration is higher in research-intensive universities (confirmed for males but not for females).

Previous literature indicates differences in gender homophily in research collaboration not only by discipline but also by institution. Therefore, finally, we will test whether the propensity to conduct same-sex collaboration also differs by institutional type: we contrast the 10 research-intensive institutions with 75 other institutions in the national system. The 10 institutions are the IDUB (or "Excellence Initiative–Research University") institutions, which were selected for additional research funding for the 2020–2026 period. The IDUB institutions include both top Polish universities and polytechnic institutes (similar results were achieved for top 10 and top 20 institutions in terms of publication numbers in the Scopus 95th-99th journal percentiles). For male scientists employed in the IDUB institutions, the propensity is high: the proportion of articles published only with males by the upper 50% of male scientists is at least 60% and is larger than the overall propensity for males in the system (see the Total line in Table 12: 50%). For female scientists, in contrast, the same proportion in the IDUB institutions is more than four times lower and is even lower than the overall propensity for females in the system. In other words, we reach the somewhat surprising conclusion that for males, the proportion of all-male collaboration in individual publication portfolios is higher in research-intensive institutions than the already high proportion for all institutions—while for females, the proportion of all-female collaboration is lower in research-intensive institutions than the already low proportion for all institutions.

In the Polish academic science system as a whole, the propensity to conduct same-sex collaboration for males is more than three times that for females (a finding which is confirmed by logistic regression in Section 4.6 below). Figure 8 shows the gender difference in the median propensity to conduct same-sex collaboration by institutional type and gender in more detail using boxplots and violin plots combined. The distribution of the median propensity for females is basically the same in both institutional types, and the within-sex variation is much higher for males than for females, as indicated by the height of the boxplots. The difference between the median values for males and females is much larger in the case of research-intensive institutions; the median value for males is much higher in these institutions, as it is for females.

This effectively means that in research-intensive institutions (see the top IDUB panel in Figure 8), males as well as females are more likely to collaborate with males. Gender homophily is thus stronger for males and weaker for females in research-intensive institutions. In other institutions (see the bottom panel), the number of males collaborating exclusively with males and the number of males collaborating exclusively with females are equal; the number of females collaborating exclusively with males and the number collaborating exclusively with females are similar in both institutional types (see the large base on which the two right columns rest for female scientists in both panels). Thus, Hypothesis 6 is confirmed for males but not confirmed for females.

Table 12. The median of the propensity to conduct same-sex collaboration by institutional type and gender.

Institutional type	Male	Female	Total	Z	p-value
Research-intensive (IDUB)	0.6000	0.1348	0.4138	-30.717	<0.001
Rest	0.4444	0.1667	0.2857	-31.992	<0.001
Total	0.5000	0.1538	0.3333	-44.291	<0.001

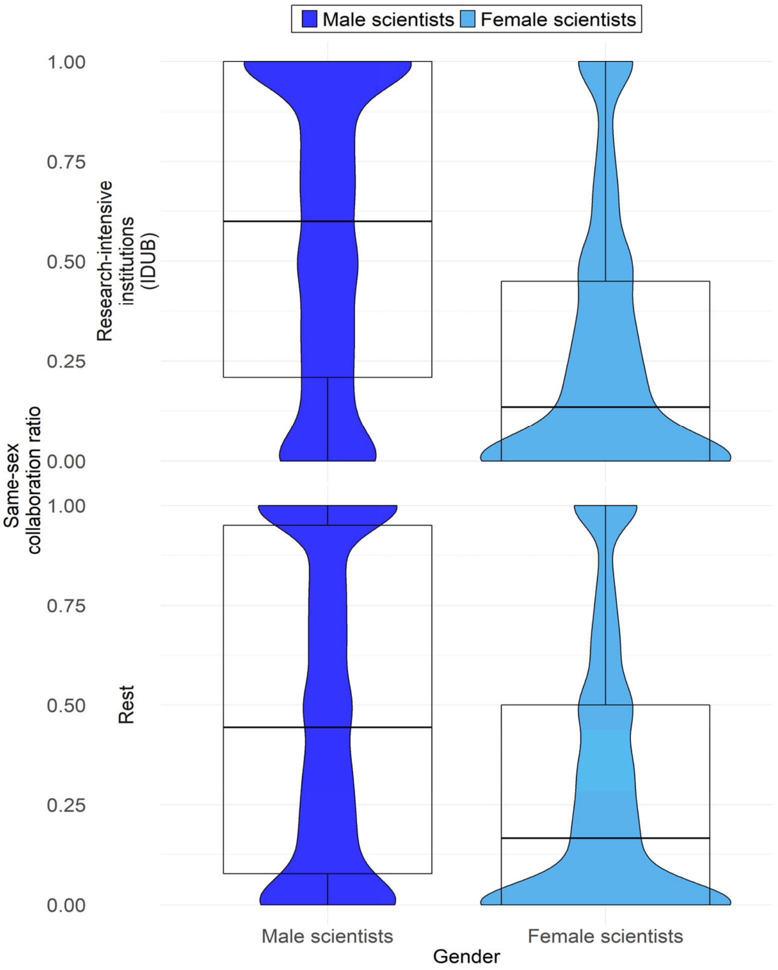


Figure 8. The propensity to conduct same-sex collaboration: distribution by institutional type and gender (boxplots and violin plots combined).

4.5. A Modeling Approach: Logistic Regression Analysis

Hypothesis 7. In logistic regression analysis, individual-level independent variables are more influential in predicting whether a scientist is highly homophilous than are institutional-level independent variables (confirmed).

Hypothesis 8. In linear logistic regression analysis, the percentage of articles written in same-sex collaboration is influenced by both individual-level and institutional-level independent variables (confirmed).

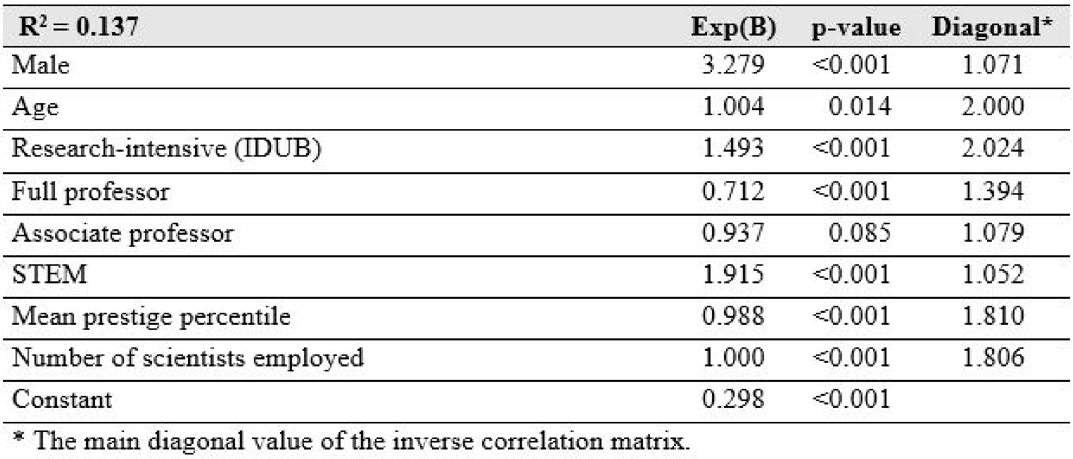
Finally, we estimate two regression models. In the first model, we estimate odds ratios of being highly homophilous in journal publishing, i.e., publishing predominantly with scientists of the same sex. We calculated the propensity for being homophilous as a percentage of same-sex collaboration articles in all the published collaborative articles of all the scientists' individual publication portfolios. We defined a highly homophilous scientist as one who conducts same-sex collaboration in more than 50% of their Scopus-indexed articles. Using a logistic regression approach, we estimated the probability of being highly homophilous using two types of independent variables.

The first type of independent variables refer to individual characteristics (demographic, biographical, and bibliometric ones): gender, biological age, mean individual publication prestige level, academic position, and the dominating Scopus-defined discipline. The second type of independent variables refer to two major institutional characteristics: a binary variable of an IDUB institution (being employed full-time in one of the 10 research-intensive institutions) or the rest, and the number of scientists employed in the author's institution (in full-time equivalents in 2018). The nonexistence of collinearity of the independent variables was confirmed through an analysis of the main diagonal of the inverse correlation matrix (see Table 13).

The distribution of residuals in our dataset was not normal (i.e., the K-S normality test statistic is equal to 0.24, with a p-value less than 0.001). The normality of a residuals distribution allows the performing of statistical inference on the model properties, as all statistical significance tests assume the normality of distribution. Still, a large sample size seems to limit the negative influence of not following model assumptions. A further step in the analysis of the residuals distribution indicates the lack of influential observations (as the range of standardized residuals does not exceed ± 3 standard deviations). A relatively good quality of the model is confirmed by Cook's distance analysis: these distances (see Table 14) are very low and very similar (in the range of 0–0.0027). Data points with large residuals (outliers) and/or a high Cook's distance may distort the outcome and accuracy of a regression. Cook's distance measures the effect of deleting a given observation. Points with a large Cook's distance are considered to merit closer examination in the analysis. This similarity confirms conclusions from the analysis of the residuals distribution that influential observations do not exist. Consequently, the conclusions drawn from our model are valid.

The model shows that male scientists' propensity for being highly homophilous in collaboration is, on average, more than three times that for females (all other predictors being equal; see Exp(B) = 3.279 in Table 12). When we consider academic positions, being a full professor, on average, decreases the odds by about 28.8%, with assistant professor being a reference category in the model. Associate professors do not differ significantly from assistant professors in their probability of being highly homophilous. Publishing in 15 STEM disciplines (as defined in the Methods section) substantially increases the odds (which are almost doubled at 91.5%). In the model, we have three quantitative variables (age, mean individual publication prestige percentile, and the number of scientists employed in an institution). Age increases the probability of being highly homophilous (on average, each year of age increases the chances of being highly homophilous by 0.4%, all other variables being constant). When the mean individual publication prestige percentile increases by one percentile, the probability of being highly homophilous decreases, on average, by 1.2% (a decrease of 10 percentile points decreases the odds, on average, by 12%). Among the institutional predictors, being employed in research-intensive institutions (in the 10 IDUB institutions) increases the same probability, on average, by half. As the number of scientists employed in an institution increases, a decrease in these odds occurs but is negligible (an increase of 1,000 scientists decreases the odds, on average, by about 1.16%). So, as expected from the literature, the propensity to conduct same-sex collaboration is generally lower in larger institutions.

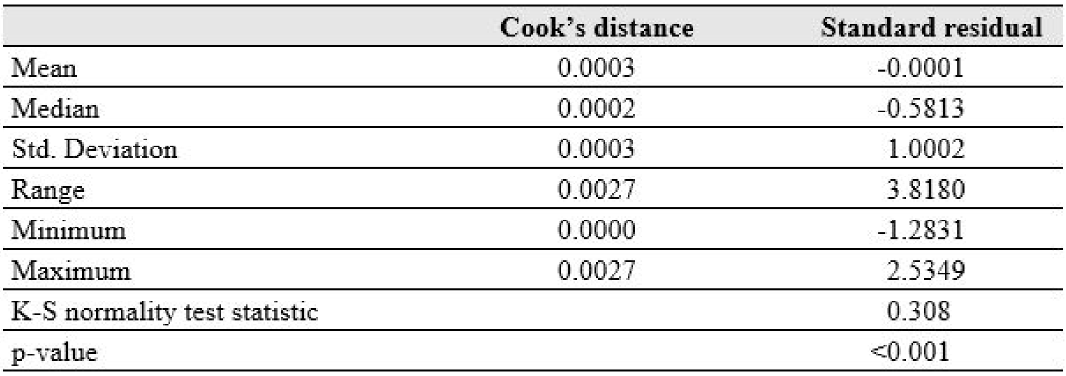
Table 12. Logistic regression model statistics, dependent variable: being a highly homophilous author in academic publishing (i.e., having the same-sex ratio higher than 50%).



$R^2 = 0.137$	Exp(B)	p-value	Diagonal*
Male	3.279	<0.001	1.071
Age	1.004	0.014	2.000
Research-intensive (IDUB)	1.493	<0.001	2.024
Full professor	0.712	<0.001	1.394
Associate professor	0.937	0.085	1.079
STEM	1.915	<0.001	1.052
Mean prestige percentile	0.988	<0.001	1.810
Number of scientists employed	1.000	<0.001	1.806
Constant	0.298	<0.001	

* The main diagonal value of the inverse correlation matrix.

Table 13. Residuals statistics of the logistic regression model.



	Cook's distance	Standard residual
Mean	0.0003	-0.0001
Median	0.0002	-0.5813
Std. Deviation	0.0003	1.0002
Range	0.0027	3.8180
Minimum	0.0000	-1.2831
Maximum	0.0027	2.5349
K-S normality test statistic		0.308
p-value		<0.001

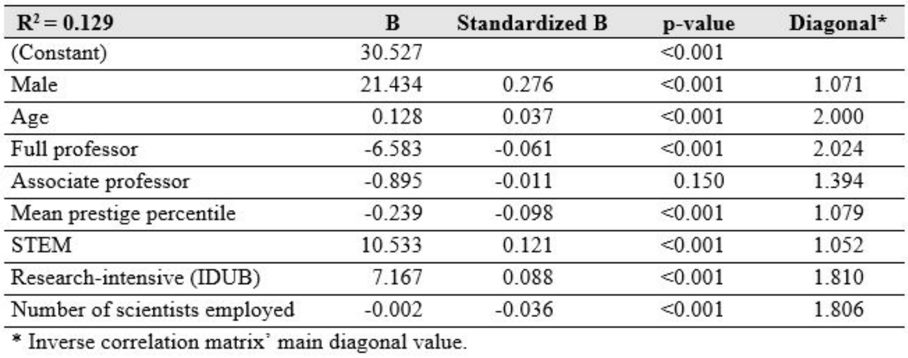
In the second model, we used linear regression to explain the variability of the same-sex ratio in research collaboration (for the sake of convenience, we multiplied the same-sex ratio by 100). The distribution of residuals was also not normal (i.e., the K-S normality test statistic is equal to 0.095, with a p-value of less than 0.001). Further analysis of the residuals distribution indicated the lack of influential observations as the range of standardized residuals does not exceed ± 3 standard deviations (see Table

15). A relatively good quality of the model is confirmed by Cook's distance analysis: these distances are very low and very similar (in the range of 0–0.0027). This similarity confirms conclusions from the analysis of the residuals distribution that influential observations do not exist in terms of the indicators used.

Being a male scientist increases the percentage of the same-sex ratio by 21.4 p.p., on average. Moreover, age significantly influences the level of the same-sex ratio. Being a full professor decreases the same-sex ratio by almost 7 p.p. on average (with assistant professor being a reference category). Being an associate professor does not significantly change the ratio. An increase of the mean individual publication prestige percentile by one percentile point decreases the ratio by 0.239 p.p., while publishing in STEM disciplines increases the ratio by 10.5 p.p. on average. Furthermore, being employed in an IDUB institution increases the ratio by 7.2 p.p. on average. Finally, as the number of scientists employed in an institution increases by 1, the ratio decreases by 0.002 p.p. on average.

A standardized beta coefficient compares the strength of the effect of each individual independent variable on the dependent variable. In other words, standardized beta coefficients are the coefficients that you would get if the variables in the regression were all converted to z-scores before running the analysis. As can be clearly seen in Table 14, it is gender that influences the ratio most strongly—but the power of this influence is still relatively low (0.28, in the range of ±1). There are three other major influential factors: STEM (0.12), mean individual publication prestige percentile (0.10), and IDUB (0.09).

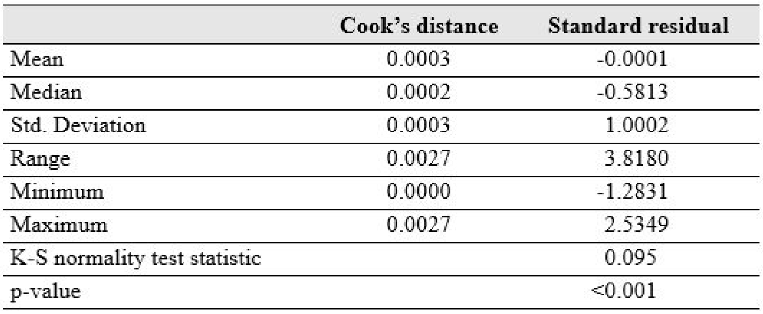
Table 14. Linear regression model statistics, dependent variable: Percentage of articles written in same-sex collaboration.



$R^2 = 0.129$	B	Standardized B	p-value	Diagonal*
(Constant)	30.527		<0.001	
Male	21.434	0.276	<0.001	1.071
Age	0.128	0.037	<0.001	2.000
Full professor	-6.583	-0.061	<0.001	2.024
Associate professor	-0.895	-0.011	0.150	1.394
Mean prestige percentile	-0.239	-0.098	<0.001	1.079
STEM	10.533	0.121	<0.001	1.052
Research-intensive (IDUB)	7.167	0.088	<0.001	1.810
Number of scientists employed	-0.002	-0.036	<0.001	1.806

* Inverse correlation matrix' main diagonal value.

Table 15. Residuals statistics of the linear regression model.



	Cook's distance	Standard residual
Mean	0.0003	-0.0001
Median	0.0002	-0.5813
Std. Deviation	0.0003	1.0002
Range	0.0027	3.8180
Minimum	0.0000	-1.2831
Maximum	0.0027	2.5349
K-S normality test statistic		0.095
p-value		<0.001

5. Summary of Findings, Discussion, and Conclusions

Our research differs from previous studies in several respects. First, we examined every internationally publishing Polish male and female scientist and the entirety of internationally visible (Scopus-indexed) Polish academic knowledge production for a decade (2009–2018). Second, owing to the characteristics of the database used, we had 100% gender determination for all scientists in the system (rather than probability thresholds in gender determination). Third, we defined what we termed the "individual publication portfolio" for every Polish scientist to examine the propensity to conduct same-sex collaboration at the level of the individual scientist. Fourth, our unit of analysis was the gender-defined individual scientist, rather than the individual publication, with its specific distribution of male/female authorships.

Finally, and most importantly, we used a comprehensive, fully integrated biographical, administrative, publication, and citation database (the "Observatory of Polish Science" database, which we constructed by merging the national registry of all 99,535 Polish scientists with the Scopus dataset comprised of all their publications in 2009–2018). Our sample (N = 25,463) included all the university professors holding at least a doctoral degree and employed in 85 research-involved universities, grouped into 27 disciplines with all their Scopus-indexed publications (158,743 articles).

While most previous literature highlights that women are much more likely to have a female than a male co-author (in three top economic journals, Boschini & Sjögren, 2007, p. 338; in life sciences, Holman & Morandin, 2019; and in industrial-organizational psychologists, Fell & König, 2016), or a female rather than a male collaborator in research projects (Lerchenmueller et al., 2019), leading to excessive gender homophily in female publishing, our findings, which are based on a large national sample, do not support this gender disparity in collaboration patterns.

Having a biographical, administrative, publication, and citation database at our disposal, we were able to examine the propensity to engage in same-sex collaboration across several dimensions, something which was previously usually either studied separately or studied based on small datasets. This research goes beyond traditional bibliometric studies of gender-based homophily in research collaboration by combining the following:

(1) The biographical and administrative data routinely inaccessible to large-scale studies: the biological age of all scientists (rather than a proxy of first publication), the three stages of their academic careers (assistant, associate and full professors, defined by the three major academic degrees used in the Polish science system, doctorate, habilitation, and full professorship, and

(2) the data routinely accessible in bibliometric studies, such as journal prestige, academic disciplines, and institutional type (operationalized as the journal percentile rank in Scopus, the dominant ASJC disciplines ascribed to each scientist, and a clear-

cut distinction between research-intensive and all other higher education institutions, respectively).

Previous research tended to be restricted either (1) by focusing on selected institutions (Kegen, 2013) or selected disciplines (McDowell & Smith, 1992; Lerchenmueller et al., 2019; Fell & König, 2016; Maddi et al., 2019), sometimes with disciplines represented by their top journals (Potthoff & Zimmermann, 2017; Boschini & Sjörgen, 2007), or (2) by being large in scale but focused solely on bibliometric data (Huang et al., 2020; Wang et al., 2019; Ghiasi et al., 2015; Larivière et al., 2013; Ghiasi et al., 2018). This research, in contrast, reveals the opportunities that large-scale, comprehensive national databases may provide (such databases are currently available for Norway and Italy; see Abramo, Aksnes, & D'Angelo, 2020). Although our "Observatory" database is not an example of the Current Research Information System (CRIS) as a data source as recently defined by Sivertsen (2019), new Polish databases (such as POLON, a national registry of all higher education institutions; and PBN, a national registry of all publications and all scientists) are moving in the CRIS direction.

Our results show that in the Polish academic science system as a whole, the propensity to conduct same-sex collaboration for males is more than three times that for females (a finding which is confirmed by logistic regression analysis). The propensity of females to collaborate with females and of males to collaborate with males (or gender homophily in publishing patterns) showed clear patterns in accordance with biological age and academic seniority: across all age-cohorts, female scientists tend to collaborate with male scientists, and male scientists also tend to collaborate with male scientists. All-female collaboration, often discussed in literature (Boschini & Sjörgen, 2007; McDowell & Smith, 1992; McDowell, Singell, & Stater, 2006), is marginal, and all-male collaboration is pervasive. The gender patterns in publishing are stable across age cohorts and across academic positions. Both males and females, on average, regardless of the gender-defined collaboration type (same-sex, mixed-sex, or solo publications), publish in journals with prestige that is relatively low. Articles written in mixed-sex collaboration are, on average, published in more prestigious journals than are those written in same-sex collaboration (which is consistent with previous literature; see Campbell, Mehtani, Dozier, & Rinehart, 2013).

Our research shows that the gender differences in the collaboration patterns of young scientists (with equal participation of males and females) are striking: while young males tend to collaborate with other males, young females tend *not* to collaborate with other females. While half of young male scientists write at least 54% of their papers in collaboration with males, the same indicator for females collaborating with females is nine times lower (6.3%). For all age cohorts, the difference in collaboration patterns by gender is startling: while the median propensity to conduct same-sex collaboration for males is 0.5, the median for females is more than three times lower (0.15). Consequently, and interestingly, in the context of previous research, the gender homophily principle in the Polish empirical context works powerfully for male scientists but does not seem to work for female scientists. The majority of male scientists collaborate solely with males, and the majority of female scientists, in contrast, do not collaborate with females at all. The distribution of the propensity to conduct same-sex collaboration for females is the mirror image of the one for males. The propensity to conduct same-sex collaboration for males decreases with the highest academic position reached. In contrast, the same propensity for females increases with the highest academic positions.

The gender difference in the propensity to conduct same-sex collaboration by age cohort closely resembles the gender difference by academic position. Female scientists consistently, across the three age cohorts and across the three academic positions, tend not to collaborate with other females. Inverse proportionality in collaboration between the two genders is characteristic of each age cohort and each academic position. The vast majority of females, and especially young females (and female assistant professors), tend *not* to collaborate with other females. The year-by-year approach we used generally confirms the two opposite trends for both genders: the downward trend in the propensity to conduct same-sex collaboration for male scientists stands in contrast to the upward trend for female scientists.

We have examined the propensity to conduct same-sex collaboration across all disciplines. Differently than in most previous studies, we compared (1) male-dominated disciplines (where the participation of female scientists is lower than 20%) with (2) gender-balanced disciplines (about 50%); and with (3) female-dominated disciplines (in the 60-75% range). Our research supports the finding from previous research that as the fraction of female researchers in a discipline increases, women increasingly tend to write with other women (Boschini & Sjögren, 2007). In the case of male-dominated computer science, engineering, and mathematics, the propensity to conduct same-sex collaboration for males is prodigious: at least half of male scientists in these disciplines collaborate exclusively with males. In computer sciences, engineering, health professions, and physics and astronomy, at least half of females do not collaborate with females at all. In contrast, in several gender-balanced and female-dominated disciplines (e.g., social sciences and psychology), the median value of same-sex collaboration for females is significantly higher than for males.

The propensity to conduct same-sex collaboration also differs by institutional type. We contrasted 10 research-intensive institutions with 75 other institutions. The somewhat surprising conclusion is that for males, the proportion of all-male collaboration in individual publication portfolios is higher in research-intensive institutions than the already high proportion for all institutions—while for females, the proportion of all-female collaboration is lower in research-intensive institutions than the already low proportion for all institutions. Males in research-intensive institutions are more likely to collaborate with males, and females are also more likely to collaborate with males. Gender homophily in research-intensive institutions is thus stronger for males and weaker for females than in the rest of the higher education system, which might suggest that a stronger institutional research focus generally induces collaboration with male scientists.

Finally, we estimated two regression models. We estimated odds ratios of being highly homophilous in publishing (defined as publishing predominantly with scientists of the same sex). The model showed that male scientists' odds ratios are, on average, more than three times higher than those of females. Both males and females tend to collaborate with males. Age significantly increases the level of the same-sex collaboration ratio, publishing in STEM disciplines moderately increases it, and full professorship moderately decreases it. However, it is gender that influences the ratio most strongly. The two other major influential factors are the mean individual publication prestige percentile and working in a research-intensive university.

Male-female collaboration practices in research were tested against the homophily principle: our findings indicate that similarity indeed breeds connection between individual scientists and structures academic publishing ties. This is much more true, however, in the Polish case, for male rather than female scientists. Gender-based homophily has substantial implications for academic careers, with the citation measure being increasingly used as a "reward currency in science", often underlying decisions on all major aspects of an academic career (Ghiasi et al., 2018, p. 1519). While forming collaborative research teams—perhaps more intuitively than as a result of solid individual publishing strategies—Polish female scientists tend not to publish with other females and prefer male co-authors. This, in time, may contribute to the reduction of the gender productivity, citation, and promotion gaps in Polish science.

Acknowledgements

The authors gratefully acknowledge the support of the Ministry of Science and Higher Education through its Dialogue grant 0022/DLG/2019/10 (RESEARCH UNIVERSITIES).

References

Abramo, G., D'Angelo, C. A., & Rosati, F. (2015). Selection committees for academic recruitment: Does gender matter? *Res. Eval.*, *24*(4), 392–404.

Abramo, G., Aksnes, D. W., & D'Angelo, C. A. (2020). Comparison of research productivity of Italian and Norwegian professors and universities. *J. Informetr.*, *14*(2), 101023.

Aksnes, D. W., Rørstad, K., Piro, F. N., & Sivertsen, G. (2011). Are female researchers less cited? A large scale study of Norwegian researchers. *J. Am. Soc. Inf. Sci. Tech., 62*(4), 628–636.

Aksnes, D. W., Piro, F. N., & Rørstad, K. (2019) Gender gaps in international research collaboration: A bibliometric approach. *Scientometrics*. *120*, 747–774.

Boschini, A., & Sjögren, A. (2007) Is team formation gender neutral? Evidence from coauthorship patterns. *J. Labor Econ., 25*(2), 325–365.

Campbell, L. G., Mehtani, S., Dozier, M. E., & Rinehart, J. (2013). Gender-heterogeneous working groups produce higher quality science. *PLOS ONE*, *8*(10), e79147.

Diezmann, C., & Grieshaber, S. (2019). *Women professors. Who makes it and how?* Singapore: Springer Nature.

Enamorado, T., Fifield, B., & Imai, K. (2019). Using a probabilistic model to assist merging of large-scale administrative records. *Am. Political Sci. Rev., 113*(2), 353–371.

Fell, C. B., & König, C. J. (2016). Is there a gender difference in scientific collaboration? A scientometric examination of co-authorships among industrial-organizational psychologists. *Scientometrics, 108*(1), 113–141.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. *J. Am. Stat. Assoc., 64*, 1183–1210.

Ghiasi, G, Mongeon, P., Sugimoto, C., & Larivière, V. (2018). Gender homophily in citations. In *3rd International Conference on Science and Technology Indicators* (STI 2018) (pp. 1519–1525).

Ghiasi, G., Larivière, V., & Sugimoto, C. R. (2015). On the compliance of women engineers with a gendered scientific system. *PLOS ONE*, *10*(12).

Halevi, G. (2019). Bibliometric studies on gender disparities in science. In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp 563–580). Cham: Springer.

Herzog T. N., Scheuren F. J., & Winkler W. E. (2007) *Data quality and record linkage techniques*. Dordrecht: Springer.

Huang, J., Gates, A. J., Sinatra, R., & Barabási, A.-L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proc. Natl. Acad. Sci. U.S.A.*, *117*(9), 4609–4616.

Jadidi, M., Karimi, F., Lietz, H., & Wagner, C. (2018) Gender disparities in science? Dropout, productivity, collaborations, and success of male and female computer scientists. *Adv. Complex Syst., 21*(3–4), 1750011.

Jaro, M. A. (1989) Advances in record linkage methodology as applied to the 1985 census of Tampa Florida. *J. Am. Stat. Assoc.*, *84*(406), 414–420.

King, M. M., Bergstrom, C. T., Correll, S. J., Jacquet, J., & West, J. D. (2017). Men set their own cites high: Gender and self-citation across fields and over time. *Socius*, *3*.

Kwiek, M. (2016). The European research elite: A cross-national study of highly productive academics across 11 European systems. *Higher Education*, *71*(3), 379-397.

Kwiek, M. (2018a). Academic top earners. Research productivity, prestige generation and salary patterns in European universities. *Science and Public Policy*. *45*(1). February 2018. 1–13.

Kwiek, M. (2018b). High Research Productivity in Vertically Undifferentiated Higher Education Systems: Who Are the Top Performers?. *Scientometrics. 115*(1). 415–462.

Kwiek, M. (2019). *Changing European Academics. A Comparative Study of Social Stratification, Work Patterns and Research Productivity.* London and New York: Routledge.

Kwiek, M. (2020a). What Large-Scale Publication and Citation Data Tell Us About International Research Collaboration in Europe: Changing National Patterns in Global Contexts. *Studies in Higher Education.* Vol. 45. On-line first April 10, 2020. 1-21.

Kwiek, M. (2020b). Internationalists and Locals: International Research Collaboration in a Resource-Poor System. *Scientometrics*. Vol. 125. On-line first April 28, 2020

Kwiek, M., Roszka, W. (2020). Gender Disparities in International Research Collaboration: A Large-Scale Bibliometric Study of 25,000 University Professors (under reviews in *Journal of Economic Surveys*).

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

Larivière V., & Gingras Y. (2010). The impact factor's Matthew effect. A natural experiment in bibliometrics. *J. Am. Soc. Inf. Sci. Tech.*, *61*(2), 424–427.

Maddi, A., Larivière, V., & Gingras, Y. (2019). Man-woman collaboration behaviors and scientific visibility: Does gender affect the academic impact in economics and management? *Proceedings of the 17th International Conference on Scientometrics & Informetrics* (pp. 1687–1697). September 2–5, 2019.

Madison, G., & Fahlman, P. (2020). Sex differences in the number of scientific publications and citations when attaining the rank of professor in Sweden. *Stud. High. Educ.,* 1–22. doi:10.1080/03075079.2020.1723533.

McDowell, J. M., Larry, D., Singell, Jr., & Stater, M. (2006) Two to tango? Gender differences in the decisions to publish and coauthor. *Econ. Inq., 44*(1), 153–168.

McPherson, M., Smith-Lovin, L., and Cool, J. M. (2001). Birds of a feather: Homophily in social networks. *Annu. Rev. Sociol.,* 27, 415–444.

Mihaljević-Brandt, H., Santamaría, L., & Tullney, M. (2016). The effect of gender in the publication patterns in mathematics. *PLOS ONE*, *11*(10), e0165367.

Nielsen, M. W. (2016) Gender inequality and research performance: Moving beyond individual-meritocratic explanations of academic advancement. *Stud. High. Educ., 41*(11), 2044–2060.

Potthoff, M., & Zimmermann, F. (2017). Is there a gender-based fragmentation of communication science? An investigation of the reasons for the apparent gender homophily in citations. *Scientometrics*, *112*(2), 1047–1063.

Sarsons, H., Gërxhani, K., Reuben, E., and Schram, A. (2020). Gender differences in recognition for group work. Forthcoming in the *J. Political Econ.*

Sivertsen, G. (2019). Developing Current Research Information Systems (CRIS) as Data Sources for Studies of Research In W. Glänzel, H. F. Moed, U. Schmoch, & M. Thelwall (Eds.), *Springer handbook of science and technology indicators* (pp 667–683). Cham: Springer.

Topaz, C. M., & Sen, S. (2016). Gender representation on journal editorial boards in the mathematical sciences. *PLOS ONE*, *11*(8), e0161357.

Van Emmerik, I. H. (2006). Gender differences in the creation of different types of social capital: A multilevel study, *Soc. Netw.*, 28(1), 24–37.

Van den Besselaar, P., & Sandström, U. (2016). Gender differences in research performance and its impact on careers: A longitudinal case study. *Scientometrics, 106*(1), 143–162.

Van den Besselaar, P., & Sandström, U. (2015). Early career grants, performance, and careers: A study on predictive validity of grant decisions. *J. Informetr., 9*(4), 826–838.

Wang, Y. S., Lee, C. J., West, J. D., Bergstrom, C. T., & Erosheva, E. A. (2019). Gender-based homophily in collaborations across a heterogeneous scholarly landscape. *ArXiv:1909.01284 [Stat]*. http://arxiv.org/abs/1909.01284

Winkler W. (1990) String comparator metrics and enhanced decision rules in the Fellegi-Sunter model of record linkage. *Proceedings of the Section on Survey Research Methods*, *American Statistical Association* (pp. 354–359).

Xie, Y., & Shauman, K. A. (2003). *Women in science. Career processes and outcomes.* Cambridge, MA: Harvard University Press.

Zippel, K. (2017). *Women in global science.* Stanford: Stanford University Press.