Cassatts in the Attic

Marlène Koffi & Matt Marx

July 2022

Abstract

We characterize the gender dynamics of the commercialization of science at scale. Analyzing more than 70 million scientific articles, we find approximately a 13% (and growing) gender gap in the commercialization of science. However, the gender dynamics are more subtle, as scientific teams with women as last authors suffer a higher penalty, even when controlling for latent commercial potential via "twin" scientific discoveries. What drives this effect? We find tentative support for supply-side factors including access to networks and representation in scientific fields. On the demand side, a natural experiment involving staggered open access to Federally-funded articles seems to spur commercialization generally but does not close the gender gap; worse, it widens it. Further, the use of boastful language, such as describing one's findings as a "breakthrough" appears to attract attention from commercializing firms, especially when the teams involve male last authors. Lastly, we find suggestive evidence that gender homophily might be a source of bias on the part of firms.

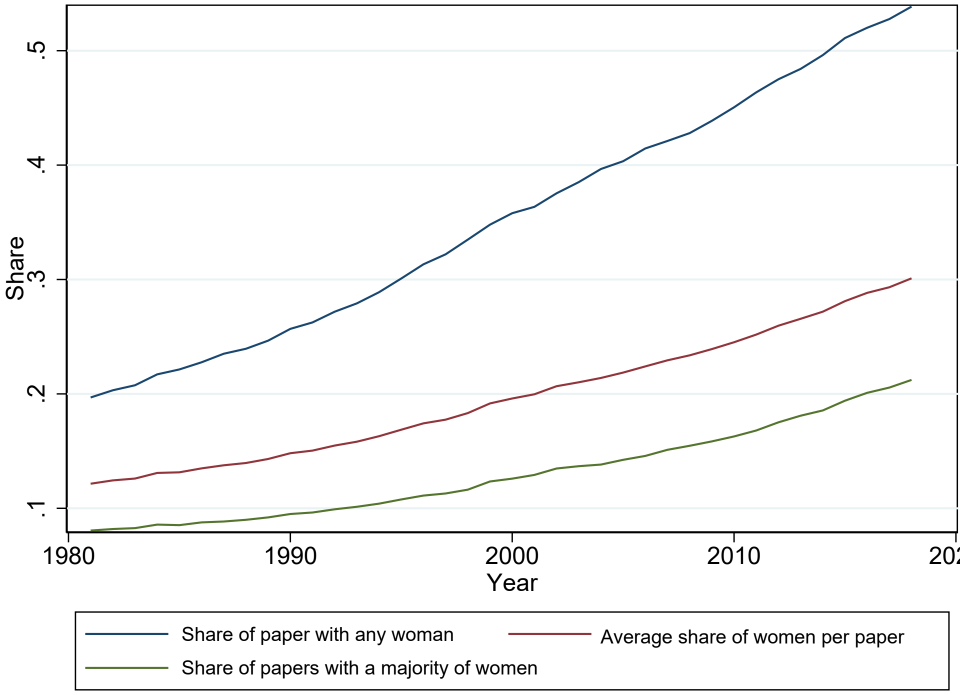
Keywords: gender, science, technology commercialization.

Koffi: Department of Economics, University of Toronto (marlene.koffi@utoronto.ca) and NBER. Marx: Cornell SC Johnson College of Business and NBER (mmarx@cornell.edu). We thank participants at the The African Econometric Society, NBER Productivity Seminar, the Cornell Innovation/Entrepreneurship/Technology Seminar, the Munich Summer Institute, the Wharton Technology Conference, the International Conference on Computational Social Science, the Asia Innovation and Entrepreneurship Conference, and The Banff Empirical Microeconomics Workshop, for insightful feedback. We also thank Joey Ryu for excellent research assistance.

1 Introduction

Among the more remarkable transformations in science during the past half-century has been the dramatic increase in women's participation. As shown in Figure 1, whereas in 1980 barely one in five published papers included a female author, now most papers include at least one woman on the authorship team. In fact, the percentage of papers in 2020 with a *majority* of female authors exceeds the percentage in 1980 with even a single woman. Broader gender diversity in scientific discovery can contribute to social welfare because the process of research and development itself is gendered (Koning, Samila, and Ferguson 2021). Indeed, prior work has showed that scientific teams combining male and female scientists can lead to discoveries that are more highly recognized within the academic community (Campbell et al. 2013; Yang et al. 2021).

Figure 1: Representation of women in authorship of scientific articles, 1980-2020



Notes: Data are counts of published articles captured by the Microsoft Academic Graph, limited to those with Digital Object Identifiers. Author gender is determined via algorithm as described in Section 2.2.

Even if published, however, the full impact of a scientific discovery on society requires that it be brought to market or *commercialized*. As one example, the breakthrough drug erythropoietin (EPO), which treats anemia by stimulating the production of red blood cells, was purified in a lab at the University of Chicago in 1971 and subsequently tested in rats, with a flurry of academic papers published by Goldwasser and collaborators in the mid-1970s: (Goldwasser, Kung, and Eliason 1974; Miyake, Kung, and Goldwasser 1977; Sherwood and Goldwasser 1979). But commercialization did not occur until the late 1980s, when the biotech startup Applied Molecular Genetics (now Amgen) patented the production of a recombinant formulation later marketed as EPOGEN.[1]

Rivette and Kline 2000 coined the term *Rembrandts in the Attic*[2] to describe discoveries like EPO, which remain uncommercialized for years, or perhaps permanently. Given that in 2019 alone, the U.S. federal government invested $43.9 billion in basic research, there is substantial societal interest in avoiding scientific discoveries becoming trapped in the ivory tower. More than 175 academic articles have investigated the process of commercializing scientific discoveries (see Rothaermel, Agung, and Jiang 2007 for a review), not to mention countless task forces. In 2022, the National Science Foundation launched its first new directorate in 32 years: TIP (Technology, Innovation and Partnerships), with the aim of "translation of research results to the market and society."[3]

The rise of women in the production of science, and the frequent failure of scientific discoveries to be commercialized, highlight and reinforce a concern raised frequently both in scholarly and policy circles: although women have made substantial strides in the production of science, they nonetheless appear to be less involved than men in the commercialization of those same scientific discoveries. The implications of this potential gender-gap in commercialization are myriad and worrisome, including missed opportunities for career advancement and wealth creation.

Although many programs have arisen to boost women's participation in commercialization, such as REACH for Commercialization[4] and STEM to Market,[5] we nonetheless lack large-scale data regarding the gender dynamics of commercialization. Most studies of commercialization have focused on patenting behavior and within a single field (most often, medicine or life sciences (Ding, Murray, and Stuart 2006; Fechner and Shapanka 2018; Thursby and Thursby 2005; Whittington and Smith-Doerr 2005). Alternatively, scholars have conducted surveys and interviews (Murray and Graham 2007) to characterize commercialization, which have tended to focus on choices by *individuals* whereas the role of teams in the production of science is ever-increasing. Noting that even the title "Rembrandts in the Attic" is gendered, the authors of the book having selected a male painter, our goal in this paper is to analyze the commercialization in the full canon of scientific inquiry.

We analyze commercialize among all published articles reported by the Microsoft Academic Graph (MAG) from 1800-2020. Automatically classifying the gender of the authors via forenames, and hand-coding gender for a substantial subsample, we explore the gender dynamics involved in the commercialization of these scientific discoveries. Commercialization is captured via a "patent-paper pair" (Ducor 2000) where (a) the patent cites the paper (b) the patent assignee is a firm (c) where the inventors on the patent overlap with the authors of the paper.

At first blush, it appears there is a "gender gap" of about 13% in commercialization for papers with

[1] https://news.uchicago.edu/story/eugene-goldwasser-biochemist-behind-blockbuster-anemia-drug-1922-2010.

[2] Rivette and Kline 2000 use the phrase to refer to firms' failure to capitalize on technologies they developed but did not bring to market. Although they did not directly consider universities, we believe the notion of missed commercialization opportunities is relevant to the academic sector as well.

[3] https://beta.nsf.gov/tip/latest

[4] https://ccts.osu.edu/content/reach-commercialization-inspiring-female-entrepreneurship

[5] https://www.awis.org/stem-to-market/

at least one woman on the scientific team, a gap that is increasing in the share of women on the team and over time. But the gender dynamics are more subtle, with the gap widening as women become the most prominent authors on a paper. We refer to these disproportionately uncommercialized discoveries as *Cassatts in the Attic* after Pittsburgh-born painter and printmaker Mary Cassatt.

We investigate possible mechanisms to explain these gender dynamics. First, we check that our results are not driven by the selection of less-commercializable projects by women. This is essential because, consistent with prior findings in specific fields (Yang et al. 2021), we find that in the full population of academic articles, those involving women are more highly cited by academic articles. Therefore, it might be that women focus on basic research as opposed to applied projects with greater commercial potential. We borrow the method of Bikard 2020, identifying "twin" scientific articles that report the same scientific discovery (scaling his method up to the entirety of MAG) in order to control for unobserved commercial potential. Doing so, we replicate the findings from the cross-section and show that the effect is strongest among the Natural Sciences and Engineering.

We proceed to consider both supply-side and demand-side mechanisms underlying these gender dynamics. Given that prior work has suggested that lack of access to networks, as well as under-representation of women in science, may explain the reluctance to engage in venturing "beyond the lab" (Murray and Graham 2007), we explore the possible role of such factors in commercialization. We find some (though limited) support for the notion that women are more likely to commercialize in scientific fields where a higher share of scientists are female. As well, we confirm that women generally have weaker commercial networks than men, and we also find that, if anything, men will tend to benefit more from the commercial networks they do have when it comes to commercialization.

Turning to demand-side factors, we leverage a natural experiment regarding Federal open-access mandates to ascertain whether the simple availability of science affects the gender dynamics of commercialization. Although we note that open access may drive commercialization generally, we see no gender-specific effects from the staggered introduction of these policies when considering all-male teams vs. teams that include at least one woman. Refining this analysis to include authors' positionality and more complete teams' gender dynamics, we observe a noticeable deepening of the gender gap as articles become more accessible. This is suggestive of a bias from the firm side.

Moreover, it does appear that even though articles are available, firms' attention is drawn to some and not others due to simple word choice. We replicate the Lerchenmueller, Sorenson, and Jena 2019 study showing that papers with women are less likely to use "boastful" language when publishing their scientific discoveries and that this is true even among twin articles that report the same scientific discovery. Although academics are not influenced by and do not differentially cite papers that tout themselves as a "breakthrough" or "unprecedented", this sort of word choice appears to attract the attention of firms. Moreover, teams with women, but with men as the last author, appear to benefit disproportionately from boasting when it comes to commercialization. Finally, the analysis concludes by showing suggestive evidence that homophily might be one of the sources of bias on the firms' side that could partly explain the gender gap in commercialization.

1.1 Related Literature

Our work is related to two broad literatures. First, nearly 200 articles characterizing the technology-transfer process whereby university scientific discoveries are licensed or commercialized have been written (see Rothaermel, Agung, and Jiang 2007 for an overview). Second, a smaller but growing literature observes gender bias in the production and recognition of science (Ding, Murray, and Stuart 2006; Fechner and Shapanka 2018; Murray and Graham 2007; Tartari and Salter 2015; Thursby and Thursby 2005; Whittington and Smith-Doerr 2005). Recent work has also branched out to consider the allocation of credit in the social sciences (Koffi 2021; Sarsons 2017).

One article deserves particular mention. Bikard and Fernandez-Mateo 2022 also use the twins methodology to examine whether academic papers have impact beyond collecting citations from other academic papers. There are a few key differences. First is the dependent variable, where they use the patent-to-paper citation datasets of Marx and Fuegi (2020, 2021) to measure impact, whereas we examine only those patents citing papers where the inventors overlap with the authors and the patent assignee was a firm (i.e., this a patent-paper pair that represents an act of commercialization). Thus their paper is more about cumulative innovation whereas we focus on commercialization. Second, whereas they focus on a subset of 295 academic publications and hand-code the gender of the authors in their sample, we explore the whole universe of academic research with 70,013,980 articles while relying both on forename-classification algorithms and hand-coded classification.

The remainder of the paper is organized as follows. Section 2 describes the data. Section 3 presents the empirical strategy and the results. Section 4 discusses some potential mechanisms. Finally, section 5 shows how robust the results are to alternative model specifications, controls, and restrictions on variables, and section 6 concludes.

2 Data and Variable Construction

We conduct our analysis using the Microsoft Academic Graph (MAG) (Sinha et al. 2015), which contains bibliometric metadata on nearly 190 million academic articles, conference proceedings, and working papers. Other sources of such metadata are available, such as the Clarivate Web of Science and Elsevier Scopus, we select MAG for three reasons. First, it is an open database, which permits replication and cumulative development of our results. (PubMed is also open but is limited to the life sciences whereas we seek to analyze the sciences more generally.) Second, its coverage has been found to outstrip proprietary databases, particularly with respect to conference proceedings (Hug and

Brändle 2017).[6] Third, and most importantly for our analysis, in MAG seems to do particularly well at capturing *given* names (hereafter, "forenames" of authors. As we rely on algorithmic determination and gender, forenames are critical to our analysis.

MAG captures metadata regarding the year of publication, the journal name, as well as the authors of the article and their affiliations. It also reports article-to-article citations. We use these fields to calculate our dependent variable *Citations from Scientific Articles*, which we limit to a fixed five-year window from the publication of the article. We also use these data to calculate the *Journal Impact Factor* (not included in MAG) as well as the number of citations for each author and institution on each paper; each of these is averaged to create a series of control variables: *Average citations per author*, *Average citations per institution*.

Finally, unlike the Clarivate Web of Science, MAG does not provide high-level categorizations of scientific fields. Instead, Microsoft automatically extracts more than 200,000 keywords from the abstracts and titles of the papers themselves. We mapped the MAG subjects to 6 OECD fields and 39 subfields.[7] Clarivate provides a public crosswalk between the OECD classifications and the 251 Web of Science scientific "categories", so we probabilistically mapped the MAG keywords to WoS categories. These probabilistic mappings are used as fixed effects in our cross-sectional models and for subsetting to examine field-specific findings.

Some of our analyses count citations to articles not from other articles but from patents. These are obtained from Marx and Fuegi 2020 for "front page" citations and Marx and Fuegi 2021 for "in-text" citations. Patent-to-article citations are also critical to identifying patent-paper pairs and therefore commercialization.

2.1 Identifying the commercialization of scientific discoveries

What do we mean by "commercializing science", and how do we operationalize this construct? Must commercialization include that a product was actually brought to market, or does the attempt alone suffice, even if it fails? Is the simple referencing or acknowledgement of science by a commercial effort enough? And how can such activity be captured at scale?

One measure of commercialization could be licensing data from universities. Even if a product does not launch, a license might be written in anticipation of such an effort. The Association for University Technology Managers conducts surveys of university licensing activity, but these data are counts and cannot be mapped to the recipient of the license. Gathering licensing data from individual universities is another alternative and has been pursued by various authors (e.g., Lach and Schankerman 2008), but this approach has two limitations. First, such data are difficult to collect at scale as no central

[6] That said, one critique of MAG is that by collecting preprints and working papers from the web via the Bing web crawler, it includes many incomplete/duplicate drafts, inflating the count of papers. Therefore we restrict our analysis to entries in MAG that contain Digital Object Identifiers, reducing the number of academic articles from 190 million to approximately 95 million.

[7] These are defined at http://www.oecd.org/science/inno/38235147.pdf.

repository of licensing data exists and thus individual arrangements must be negotiated with each university. Second, even if each university's data could be collected, formal licenses may not represent the full scale of commercial activity.

An alternative is to use citations from patents to papers as a proxy for the original scientific discovery. Arguably, if a (commercial) patent references an academic article, it is relying upon the underlying science in a commercially-meaningful way. But citations can be made as background material or can be done strategically (Lampe 2012). Moreover, given the more than 30 million patent-to-paper citations, can all of these truly be said to be instances of commercializing science? It may be more useful to view citations from patents to articles as a *superset* of actual commercializations.

2.1.1 Capturing commercialization with patent-paper pairs

Our measure of commercialization depends on "patent-paper pairs" (PPPs). The concept of a PPP originates with Ducor 2000, who detected overlapping gene sequences between papers and patents as evidence of the same material being published and patented. Since Ducor, more than two dozen articles have employed the PPP method (see for example Murray 2002, Thompson, Ziedonis, and Mowery 2018, Ranaei et al. 2016, and Haeussler and Sauermann 2013). The intuition is that if we find a patent citing a paper *where the author(s) of the paper were also inventors on the patent*, we might reasonably assume that there was some act of commercialization involved.

Following Marx and Hsu 2021, we generalize their prior PPP algorithm to the entire MAG. As we are focused not only on genetics, we cannot rely on genetic-sequence overlap; instead, we use techniques common to other PPP studies including overlap of authors and inventors as well as temporal similarity. Given the finding of Haeussler and Sauermann 2013 that papers often have "extra" authors who do not appear on a patent, and also being mindful of the critique of Ranaei et al. 2016 that misspellings and commonality can lead to both "lumping" and "splitting" problems in matching names, we adopt a probabilistic approach by taking into account prior probabilities of author and inventor names in the respective corpora. For example, even if "John Smith" is an inventor on the patent and an author on the paper, we do not consider this a strong indicator of overlap unless there is also a match on the middle initial (stronger still if there is a match on the full middle name). The rarer the name, the stronger the match. (The full name-matching algorithm is detailed in Marx and Hsu 2021.) Moreover, we deviate from prior approaches in the interest of avoiding false-positive PPPs by requiring that the patent cite the paper.

In a second step, we restrict matches in terms of temporal proximity. Although some PPP authors have required the patent and paper in a pair to appear in the same year (Murray 2002), others have permitted a wider window. Our approach requires the paper's publication not to come after the filing of the patent, lest public disclosure have occurred previously, and moreover no earlier than five years before the patent is filed. (In robustness checks, we shrink this window further.) In some cases, we end up with multiple patents mapped to a single paper.

2.1.2 "Transitive" patent-paper pairs

Because we are interested in the commercialization of science by firms, the PPP is not necessarily our end-point dependent variable. Rather, we are interested in whether the original science is commercialized *by a firm* and therefore limit our scope to PPPs case where the patent assignee is a commercial entity. We recognize however that in some cases a firm may commercialize a scientific discovery where the university has already claimed patent protection, often, by issuing a license to the patent.

We cannot observe university-licensing data at scale, but we take a step toward including such arrangements by also capturing "transitive" patent-paper pairs. In this scenario a university patents a paper, and then a firm files a patent that cites that same PPP and with overlap between the inventors on the firm-assigned patent and the university-assigned patent it cites (which is paired with the original scientific article). These "transitive" paper pairs account for approximately 14 percent of our commercialization instances; removing them does not affect the results.

2.1.3 Characterization

The PPP method of detecting commercialization is not without limitations. PPPs might detect instances of sponsored research as commercialization (i.e., false positives), which we cannot algorithmically rule out. It might also create false negatives by failing to detect instances of commercialization where, for example, a company was granted a license to a university technology, and the ensuing corporate patent(s) cited the original academic article, but because the company had different staff work on the project not including the original authors, our algorithm would not pick up that PPP. Employing a similar algorithm to detect instances of startup commercialization, Marx and Hsu 2021 hand-checked a random sample of 40 such PPPs and found that 39 were correct. This suggests that the false-positive rate is low. We lack a "golden" dataset to characterize false negatives and therefore view our PPP method as a lower bound.[8] Table 1 shows the top 15 commercializers of scientific articles. Figure 3 shows commercialization rates over time. Figure 4 shows commercialization rates by OECD scientific category.

Table 1, Figure 3, and Figure 4 about here

2.2 Gender classification

The Microsoft Academic Graph does not contain gender records for individual authors. Given the volume of data, it is impractial to carry out a manual verification of every author's gender. Thus, we

[8] It is not immediately obvious to us why the algorithm should be biased toward or away from detecting PPPs involving men vs. women, though if it were the case that, say, patenting or publishing women had much more rare (common) forenames than men, it would be easier (harder) to detect author/inventor overlap.

determine the gender of the authors from inference based on the first name (a common practice in big data analysis). We rely on Genderize.io API, a corpus-based dictionary that uses large databases of names collected from the US Census, international dictionaries, and social media and calculates the probability that a specific first name is associated with the male or female gender. In other words, for a name, like “Anna”, it computes the fraction of all individuals with that name that are women or girls and men or boys.

Further, We use a restrictive procedure applying a threshold of 90% to assign gender. Although Hofstra et al. 2020 found that 70% maximized the level of accuracy between algorithmic-assigned gender and self-reported gender, when comparing automatic classification vs. hand-classification in section 2.2.1, we found substantial errors for confidence scores lower than 90% In addition, we do not assign a gender to authors without first names.

Then, we identify the gender composition of each team based on the gender composition of the authors. In doing so, we consider alternative definitions of the gender composition of a paper. We use either the share of female authors among the authors directly or binary variables such as *papers with any female author* to characterize articles with at least one woman and *papers with all-female authors* for articles with only women.

Finally, we discard papers where we cannot classify the gender of at least one-third of the authors. For example, if there are two authors and only one can be classified, the paper is discarded. If there are three authors, two must be classified. However, the results are not overly sensitive to those restrictions. In the Appendix we check that our main results are robust to instead imposing a threshold of one-half, three-quarters, or that all authors’ gender must be classifiable by algorithm at a confidence level at or above 90%.

2.2.1 Manual classification

To increase confidence in the results, we hand-code the gender of every author for a subsample of papers. As described in Section 3.3, our primary identification strategy is to control for unobserved commercial potential by analyzing “twin” discoveries that yield multiple papers reporting the same science. There are 24,682 with more than 120,000 authors, still a prohibitively large number to hand-code. However, our estimates are principally driven by the subset of 1,704 articles reporting twin discoveries where one article was commercialized and the other was not. (Indeed, our robustness checks using clogit instead of OLS necessarily omit twin discoveries where neither, or both, of the articles reporting the discoveries was commercialized.) We hand-coded 13,729 authors including all of those for the twin discoveries with variation in the outcome variable.

A team of research assistants (RAs) hand-checked the gender for all of these authors. RAs were instructed to find the author’s website, LinkedIn page, or other indivdually-identifying information. All but 527 of the authors were able to be located, most of which had confidence scores of 99% or

100% from the algorithm. For the remaining authors, an RA checked for people with the same name to identify gender by proxy. Of the 7049 names that were automatically classified at 99% or 100% confidence, only one was determined to be incorrect.

We present results with this subsample of author gender classified by hand, while also retaining the 90% of twins that are algorithmically classified. However, as shown in Table 20 the main results are robust to analyzing only the hand-classified articles.

Descriptive statistics for the 70,013,980 papers analyzed are are in Table 2.

Table 2 about here

3 Results

3.1 Empirical specification

Let us define by $COMM_{it}$ the commercialization status of a given scientific article i published at time t, to be the outcome of interest. $COMM_{it}$ is a dummy variable equal to 1 if the scientific article i has been commercialized and 0 otherwise.[9] $Gender_i$ refers to the gender structure of the authors of the scientific article i. X_{it} captures a set of control variables such as the number of authors, “prestige” of the authors, and their institutions as calculated by the average forward citations (in a five-year window) for each and the impact factor for the article’s journal. Finally, we include a battery of fixed effects in the model with $Time_{FE}$ standing for the publication year fixed effects and $Field_{FE}$ for the article’s field fixed effect where the field is defined by WoS field category. We define the error term by ϵ_{it}. Then, we use the following specification to analyze the gender gap in commercialization:

$$COMM_{it} = \alpha_0 + \alpha_1 Gender_i + \beta X_{it} + Time_{FE} + Field_{FE} + \epsilon_{it} \tag{1}$$

We estimate equation 3 using linear probability models with robust standard errors. The coefficient of interest is α_1 which captures the gender gap in commercialization.

3.2 Cross-sectional Results

Table 3 examines the gender gap in commercialization. In Panel A, we observe higher commercialization rates for all-male scientific teams vs. those with at least one female member (column 1). In fact, having at least one woman in a scientific team is associated with a reduction of the likelihood of commercialization by 19 percent (–0.000640/0.0033). The same is true when observing a continuous measure in column 2: a one-standard deviation increase in the number of women in a scientific team predicts a 12.90 percent drop in the likelihood of that article being commercialized.

[9] Because commercialization could occur in a different year than the year of publication of the articles, we play around with the time window between the publication and the commercialization in the robustness section.

Table 3 about here

Although Table 3 shows lower rates of commercialization among science conducted by women, this result could be mechanical if there is a quality difference that matches the different gender categories. To address these questions, in columns 3 and 4 of Table 3 we examine the quality of papers by counting forward citations from other academic articles in a fixed, five-year window. The same controls and fixed effects from Table 3 apply. Whether a team of scientists has at least one woman or more, their citation count increases.[10]

The contrasting findings in Table 3 are striking. While increasing the fraction of women in a scientific team increases the scientific community citation, the commercialization pattern suggests a decrease in the likelihood of commercializing for teams with women.

However, there are potential threats to identifying the gender gap in commercialization in the cross-sectional data. Perhaps the most immediate interpretation is that women conduct more theoretical science (and are therefore appreciated by the academic community) but which is less applied and therefore ignored by commercial firms due to "latent commercializability" (Marx and Hsu 2021). Ruling this out is difficult because it is hard to know ex-ante the commercial potential of a scientific discovery or have a good proxy of paper quality. In the next section, we attempt to account for latent commercializability and quality by comparing commercialization outcomes for pairs of "twin" scientific discoveries with varying percentages of female authors.

3.3 "Twin" scientific discoveries

We attempt to control for latent commercializability by adopting the Bikard 2020 method of identifying "twin" discoveries based on co-citation. Although co-discoveries are uncommon in the social sciences, they happen routinely in the hard sciences. Twin discoveries are identified as papers satisfying the following five conditions: (1) they are published no more than a year apart; (2) they have zero overlap in among authors; (3) they are cited at least 5 times; (4) they share at least 50% of forward citations; and (5) they are cited *adjacently* (i.e., within the same parenthesis) at least once. Bikard and Marx 2019 confirmed no false positives in a random sample of twin papers identified via the above algorithm.

We combine the public twins made available by Bikard 2020 with those generated by Marx and Hsu 2021, crosswalking the Marx/Hsu twins from the Web of Science to MAG.[11] This exercise resulted in a set of 24,682 twins in MAG. Descriptive statistics for this subsample are in Table 4.

Table 4 about here

The empirical specification follows that of twin studies in the epidemiological literature (Carlin

[10] Our findings are reminiscent of Yang et al. 2021 in Medicine, although we examine all fields of scientific inquiry.

[11] The majority of WoS-based twins from Marx and Hsu 2021 could be mapped directly via Digital Object Identifiers; the remainder were fuzzy-matched using author, title, year, journal, volume, and page(s).

et al. 2005), introducing a fixed effect for the twin discovery (TwinDiscoveryFE). Year and field fixed effects are omitted, given the nature of the twin-matching exercise.

$$COMM_{it} = \alpha_0 + \alpha_1 Gender_i + \beta X_{it} + TwinDiscoveryFE + \epsilon_{it} \tag{2}$$

Table 5 refines the cross-sectional commercialization estimates of Table 3 while controlling for latent commercializability and quality. The effect of doing so becomes apparent in column 1, where the estimated coefficient of having at least one female author on the paper is associated with approximately a 1.13 percentage point decrease in commercialization . In column 2, a one-standard-deviation in the percentage of women among the scientific authors is associated with a 0.5 percentage-point decrease in the likelihood of a discovery being commercialized. Given that the twin articles have on average 25 percent female authors with a standard deviation of .2623 and a baseline rate of 4% commercialization, moving from one-quarter female authors to one-half corresponds to approximately a 13% drop in the likelihood of being commercialized.[12]

Table 5 about here

The same pattern is true when considering a two-dummy model with mixed-gender teams (teams with at least one man and at least one woman) and all-female teams. First, relative to all-male teams, mixed-gender teams are about 30 percent less likely to commercialize. Interestingly, columns 3 and 4 of Table 5 suggest a non-monotonic relationship between the share of women in the scientific team and the likelihood of commercializing. Adding the square of the percentage of females in the scientific teams reveals a somewhat significant curvilinear pattern where the relation between commercialization and the share of women is negative up to a certain cutoff. Once we cross the cutoff, the coefficient flips to a positive and significant value. We follow (Haans, Pieters, and He 2016) in testing the truthfulness of this relationship.

We find that while the coefficient of the squared of the percentage of women on the team is positive and significant, we reject the hypothesis test on the steepness at both ends of the range of this variable. At the lower range of 0, the linear combination of the coefficient is negative and significant. At the upper range, the value of the coefficient is positive but insignificant. The t-statistics, computed using the variance-covariance matrix of panel A Table 6, is 1.3196.[13] Moreover, we also split the sample using the suggested turning point, which in this case is estimated at 0.63. Panel B of Table 6 shows

[12] Although our main objective is to verify that the findings in Table 3 are not driven by unobserved heterogeneity, we also confirm using the twins method that the science published by women is in fact of higher quality. One might allege that, for some reasons, the results in our Table 3 or in (Yang et al. 2021) are due to women accruing artificially high numbers of citations from other academics for reasons having to do with favoritism. Disproving this is difficult in the cross-section, but when examining twin discoveries one can test whether men vs. women receive more citations for the same scientific discovery. Table A1 in the appendix confirm that there is no gender bias in citation patterns from academic articles.

[13] The corresponding formula in (Haans, Pieters, and He 2016) is given by the model $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \epsilon$. To test for steepness, let us call X_L the lower bound of X and X_H its upper bound. $\beta_1 + 2\beta_2 X_L$ should be negative and significant, while $\beta_1 + 2\beta_2 X_H$ should be positive and significant.

that below the cutoff, the coefficient is negative and significant. Above the cutoff, the coefficient is still negative but significant. Therefore, we do not find sufficient evidence of a curvilinear relationship. Rather, the weaker sample size as the share of women increases and the aggregation scheme in the linear regression estimation with the twins' fixed effects may produce what could resemble a U-shape relationship.[14]

Table 6 about here

Figure 5 explores this relationship in greater detail, using the twins specification. Each confidence interval plots the results of a regression resembling column 1 of Table 5 but with a dummy for teams at or above a given percentage of female authors. It suggests that, when controlling for latent commercializability, all-male teams remain considerably more likely to commercialize.

Figure 5 about here

The twins model shows that even for very similar academic papers, controlling for unobserved differences in quality and latent commercializability, papers with women experience a lower likelihood of commercializing. This result is at odds with the fact that those very same papers are also the ones with more academic recognition as measured by citations (column 3, Panel B of Table 3).

3.4 First and last authorship

We also explore the visibility of women as contributors to scientific articles. If firms are somehow biased against scientific discoveries published by women, we would expect that bias to be amplified when women are more easily visible as authors of the publication. Thus, we consider the first author and the last author.

The first author may in some cases be the lab manager but is also frequently the scientist who did the bulk of the work on the paper. In any case, the first author may be more often seen when others authors are "lost in the et al." (Simcoe and Waguespack 2011). We also include the last author because, for many labs, the last author is the Principal Investigator and/or would be easily seen at the end of the list. Accordingly, we label each paper in one of the following four categories: (a) the first and last authors are male; (b) the first author is female, and the last author is male; (c) the first author is male, and the last author is female; (d) the first author is female and the last author male. Approximately 15% of articles have a female last author, and 19% have a female first author and a male last author.

In panel A of Table 7, we assess whether firms are more vs. less likely to commercialize science where a woman is the paper's first and/or last author. Column 1 indicates that when the last author

[14] Moreover, in the robustness section where we use the conditional logit model, we do not find evidence of such a curvilinear pattern.

is female and the first author male, the paper is not statistically significantly penalized in terms of commercialization likelihood by comparison to papers where both the first and the last authors are male. However, when we have women being the first and the last author on the paper, the penalty is estimated at -2 percentage points. To avoid conflating the results with the presence of a female author, we add additional controls on the share of women on the paper. Columns (4) and (5) show the results controlling for the fraction of women among the other coauthors, excluding the first and the last author (to avoid multicollinearity). One can see an apparent effect of the positionality, where the commercialization penalty tends to be more significant when women appear to be the last author. This result is confirmed in panel B of Table 7 where categories (c) and (d) are combined.

Table 7 about here

Therefore, commercialization is less likely to happen when women are in more visible author positions, especially when they are the last author on the paper. This result is over and above the percentage of women in the scientific team. We will therefore include in the next sections the categorization depending on the gender of the last author in the mixed-gender teams.

3.5 Heterogeneous effect by scientific field

Table 8 shows where in the attic the “Cassatts” are, segmenting the scientific space into four top-level OECD categories: columns 1-2) Natural Sciences 3-4) Engineering 5-6) Medical 7-8) Social Sciences. Each pair of models shows first the cross-section, and then the subset of twin papers for that category.

Although in the cross-section (odd-numbered columns), a gender gap appears in every category, when applying the twins’ methodology (even-numbered columns), this is limited to Natural Sciences and Engineering. Medical Sciences display a negative sign, although insignificant. Introducing the percentage of women on the paper in Panel B and the three-model dummies highlighting the effect of positionality yield similar results. In other words, in the “attics” of natural sciences and engineering, there appear to be Cassatts. However, we may face a limiting power due to sample size constraints to correctly estimate the commercialization gap in those fields. Not all estimates in the twins’ sample are precise enough to achieve statistical significance.

Overall, these results reveal that all the fields are not guided by the same level of gender commercialization gap. In particular, the coefficient of social sciences could also testify of a non-systematic lesser likelihood to commercialize for women teams across all the fields (but note that the sample size for this field is meager in the twins’ sample).

Table 8 about here

3.6 Effect over time

As more women participate in science, we would hope that any bias against women would begin to dissipate, particularly with the myriad of examples of women doing high-quality research. Thus, we could expect a reduction in the commercialization gap as time passes.

In Table 9, we re-estimate the baseline model but interact the gender variable with a linear time trend. Column 1 shows an increasing trend of the gender gap over time. Scientific publications by women are still less likely to be commercialized. The interaction coefficient "Share of female time trend" is negative and significant. A similar pattern is observable when breaking down the timing into sub-periods. Column 3 attests that the gap observed in 2016-2021 is greater than any sub-periods considered. Interestingly, this period registers a number of publications by women higher than ever.

Nevertheless, examining either the absolute value of the coefficient in column 3 or the coefficients in column 2 also suggests a non-monotonic relation over time. This again could provide some evidence that the commercialization gap observed is due to other factors external to the share of women in science. If the latter is true, we should expect a reduction over time, especially in the recent decade.

Table 9 about here

4 Mechanisms

Although the twins regressions in Table 5 control for unobserved quality and show lower rates of commercialization among science conducted by gender-diverse teams, it is unclear what mechanisms drive these dynamics. We explore both supply- and demand-side factors.

On the supply side, we could observe a gender gap in commercialization if female scientists have different preferences, i.e., not interested in commercializing and patent less (Whittington and Smith-Doerr 2005). In this case, the PPP-based result will be a mechanical derivation from the argument of women patenting less. On the other hand, conditional on being interested in the patenting process, they may lack networks, connections, and support to commercialize (Murray and Graham 2007).

On the demand side, firms could be biased against female scientists and discriminate against women in searching for commercializable discoveries, as has been found with investors and female entrepreneurs (Brooks et al. 2014). Alternatively, it could be that the lack of self-promotion of their work attracts less attention from firms compared to similar male-authored works.

Table A8 about here

4.1 Supply side: Commercialization, representation, and networks

Several scholars have suggested that gender disparities in commercial activity by scientists may be attributable less to intention than to access. For example, Murray and Graham 2007 report that female scientists who wish to commercialize their work are excluded due to lack of access to networks. Similarly, Tartari and Salter 2015 suggest that women are more likely to engage in commercialization in fields with higher representation of women. In Tables 10-12 we explore whether we can find evidence for these mechanisms at scale.

For each of the 251 fields defined by the Clarivate Web of Science and crosswalked to MAG, we calculate the percentage of women publishing in that field each year. This variable is entered into Equation 2 as a covariate in column (1) of Table 10 and is then interacted with the presence of a female author in column (2) and in column (3) with the percentage of female authors. The estimated coefficient on the interaction in (3) is positive and statistically significant, consistent with the finding of Tartari and Salter 2015. However, such precision is not achieved when interacting with a dummy for any female authors in (2), or either mixed-gender with male last author or female last author or all-female teams in (4). We take this as only weakly suggestive evidence that higher gender representation in a scientific field promotes commercialization.

Table 10 about here

Next, we explore whether access to boundary-spanning networks may play a role in commercialization. For each author on each paper, we count the number of coauthors of that author who satisfy two conditions: 1) they are not coauthors on the focal paper 2) they are affiliated with commercial firms. For each author, this reflects the first-degree commercial reach of their coauthorship networks. This count is then averaged across all authors on a focal paper to reflect the commercial reach of the scientific team's network. Consistent with prior work (Murray and Graham 2007), in Table 11, papers with (more) female authors have weaker commercial reach as represented by the average number of coauthors of the paper authors. This is true whether measuring female representation as one-or-more (column 1), linearly (column 2), as mixed-gender with male last author or female last author vs. all-female teams (3).

Table 11 about here

We then enter this commercial-network-reach variable as a covariate in columns (2,4,6) of Table 12 and then interact it with the measures of female authors on the paper. Although the interaction with one-or-more female authors is imprecisely estimated in column (2), the interaction with the percentage of female authors in column (4) is negative and significant at the 5% level. Curiously, this suggests that gender commercialization gap appears even at higher degree of commercial-network-reach; implying that men may benefit more from networks of coauthors at firms when it comes to commercializing scientific discoveries. Similar results are obtained in Appendix tables A2 and A3 via alternative measures of commercial reach.

Table 12 about here

As with our analysis of representation, the result regarding commercial networks is not strongly robust. In column (3) we estimate the three-dummy model, with a somewhat imprecise coefficient for mixed-gender with male last author or female last author (though negative for all female, as in column (2)). Taken together with Table 10, we conclude that there is suggestive but not conclusive evidence for role of representation and network access in the commercialization process.

4.2 Demand side: accessibility of articles

The gender gap in commercialization could also be salient in an environment where access to information on scientific articles is not perfectly distributed and gender-specific.[15] In particular, if scientific articles with women are less visible than male-authored scientific articles, this could prevent companies from accessing the former's publications and therefore reduce their probability to commercialize relative to the latter. If such a hypothesis turns out to be accurate, then a shock that would increase awareness and access to scientific research should contribute to reducing the gender gap in commercialization.

To test this, we use a natural experiment provided by the open-access mandates for federally-funded research. In many fields in sciences, most articles and working papers are not freely available (Bjork, Roos, and Lauri 2009, Khabsa and Giles 2014, Ware and Mabe 2015). At the same time, one of the most common rationales behind the evolution of scientific discovery is to expand the frontier of knowledge by building upon previously available research. This channel could be more important in the commercialization of academic research as firms may need to explicitly collaborate with a researcher from an academic institution.

In 2008, the National Institutes of Health (NIH) had leveraged an initiative to make freely available the academic research they founded so that any funded article accepted for publication after April 7, 2008, must be archived in the open access PubMed Central (PMC) database within 12 months of publication.[16] In 2013, the White House Office of Science Technology Policy mandated agencies with an RD budget of $100M in order to develop plans to make the results of the federally funded research freely available. This gives rise to a staggered implementation of the "Public Access policy"(PAP) with, for example, the Department of Energy (DOE) implementing this policy in 2014 and the National Science Foundation (NSF) in 2016.

Our empirical model takes advantage of the gradual implementation of the PAP by constructing an event study where the event date is the starting year of the PAP for one agency. Therefore, an article in the database is considered to be "treated" in a given year if a federal agency financed this paper, and during that year, this agency has started to implement the PAP. In this setup, we are particularly

[15] Indeed, dissemination of academic research via social media, for example, have been shown to increase the visibility and the likelihood of citation (Eysenbach 2011 and Klar et al. 2020)

[16] Most of the paper in the literature of open access on academic citation finds a non-negative effect. In particular, Bryan and Ozcan 2021 shows that after 2008, NIH-funded research were 12 to 27% more likely to get cited, while Staudt 2020 finds a positive but more moderate effect.

interested in the triple difference that captures the effect on the commercialization of federally-funded publications written by women after the implementation of the PAP relative to those written by men. Therefore, assessing the effect of the PAP on narrowing the gender gap in commercialization. We further add the control variables similar to our baseline and include the journal, year, and field fixed effects. Our identifying assumption is that there are no shocks correlated with the introduction of the PAP that differentially affects scientific teams with men/women commercialization likelihood. To address concerns regarding heterogeneous treatment effects, we use a robust staggered difference and difference approach by Sun and Abraham 2021. Other procedures to solve this issue have been proposed by Goodman-Bacon 2021, Callaway and Sant'Anna 2021. Baker, Larcker, and Wang 2022 shows an interesting equivalence between those different procedures.

Figure 6 shows the result of this estimation (Table 13 refers to the corresponding numerical values). We use one year before the introduction of the PAP as the reference year. Panel A presents the double-difference for each gender type. Although the pre-trend for the sample difference and difference is not non-significant, impeding the interpretation of the simple difference, we clearly see that both genders are moving in an almost perfect one-to-one mapping. Therefore, although as is visible in Panel A, there seems to be a sharp jump in the commercialization of science following the advent of Open Access mandates, we do not see a material convergence of the gender gap. Panel B plots the result of the triple difference exercise. There is no statistically significant pre-trend and no effect after the PAP. This means that the PAP has not changed the gender gap in commercialization, as we would expect if a differential in access to information on a paper exists.

Table 13 and Figure 6 about here

Since the "any woman" variable could encompass different team structures, we further assess whether we observe a difference in the gender gap when considering other ways to define an "women's team." Figure 7 repeats this analysis but shifting the focus to papers with women in leading positions of responsibility. Although MAG does not contain information about the exact roles played by each other, it does contain the order of authors on the paper. We focus on the last author of the paper (as it happens to be the most important positionality factor as shown in the previous sections) and present also detailed evidence with the first author.

Figure 7 about here

Panel A compares papers with a women as last author (i.e., manager of the lab). Approximately two years following the introduction of open access, commercialization rates diverge sharply with papers that have a woman as the last author considerably less likely to be commercialized. The same is true in Panel B for papers where a woman is either the first or the last author. In Panel C, the variable of interest is whether a woman was both first and last author. This is somewhat less precisely estimated but still shows a divergence following open access mandates.

We conclude that, contrary to priors that increased information might help to close the gender gap, the introduction of open access mandates in fact exacerbated the gender gap in commercialization for scientific projects led by women. Thus, it seems like the "any female" teams that tend to benefit from the increased accessibility to information are those with women, but where men are in leading positions (first or last). By contrast, for teams where women are last and/or first authors, the open access policy causes a widening of the commercialization gap. This is suggestive of bias on the part of firms.

4.3 Demand side: firms' attention to word choice

Our results thus far suggest that firms pay less attention to scientific discoveries with mixed-gender teams where women are not in a strong majority. One possibility raised in the literature is that women use less boastful language when writing scientific articles Lerchenmueller, Sorenson, and Jena 2019 (hereafter, "LSJ"). If so, it might be that firms are more enticed by scientific articles by men if they are more prone to describe their research findings using words such as "breakthrough" or "unprecedented."

We collected the titles and abstracts for all 24,682 twin articles and counted the number of boastful words used by LSJ, with one exception. The most commonly found boastful word in LSJ's analysis was "novel.". Worrying that "novel" is often be used not for boasting but rather to identify novelty as in *novel coronavirus*, we excluded such bigrams.

With these refinements, in Table 14 we confirm LSJ's finding that overall, papers with female authors tend to use boastful words less often. (Note that this replication of LSJ holds in our subsample of "twin" papers, controlling very closely for content of the article.) This holds whether looking at papers with at least one woman (column 1), measuring the share of women (column 2), or separating mixed-gender teams with male last authors from mixed-gender teams with female last authors from all-female teams (3).

Table 14 about here

Next, we add the count of boastful words found in the title or abstract of the twin articles to our analysis. In Table 15 we analyze whether boasting is associated with the likelihood of being cited by academic papers. Regardless of the means of measuring the share of women who are authors on a paper, we see no correlation between word choice and citation counts. In other words, academic scientists are not swayed by boasting.

Table 15 about here

Firms, however, appear to pay more attention to articles that describe themselves as "breakthroughs" and the like, even when controlling for content via twins. In column (1) of Table 16, articles with more boastful words are commercialized at a higher rate. Moreover, this effect is amplified for teams with at least one female author. We also note a commercialization premium for mixed-gender teams with male last authors who boast in column (6). We do not find sufficient evidence of a boasting effect for mixed-gender teams with female last author or all-female teams.[17]

We conclude that the less common use of boastful words by female scientists may explain in part the gender gap in commercialization, especially when those female scientists collaborate with a male author who is last author. Although, as shown in Section 4.2, the gender gap in commercialization is not explained by mere access to scientific articles, it does appear that firms are more drawn to published papers that do not simply report their findings but embellish the presentation. Indeed, comparing the magnitude of the estimated coefficients in most columns of Table 16 would suggest that the penalty for mixed-gender scientific teams with male last authors is nearly offset by include a word such as "breakthrough" in the title or abstract.

Table 16 about here

Overall, controlling for unobserved differences in quality and commercializability potential, the way one paper is sold matters. For example, this could explain why we observe a difference in the commercialization pattern between mixed-gender teams, especially those with males in last positions and all-male teams in the full sample.[18] Table 17 shows that those teams are indeed less likely to use boastful words in the universe of academic publications. However, when they do so (Table 18, column 4), their likelihood of commercialization increases by 0.2 percentage points, offsetting the gap of 0.1 percentage points. Unfortunately, we can see again in the cross-sectional analysis that this benefit does not extend to articles with women as last authors. Although we should not overthink the cross-sectional regression because of the possibility of selection, this apparent double-standard for the same behavior remains interesting.

Tables 17 and 18 about here

4.4 Commercialization and gender homophily

In this section we present evidence of gender homophily in the commercialization process. In an ideal experiment, we would randomly seed commercialization partners with heterogenous gender composi-

[17] As a placebo test of the boasting effect, in Appendix table A7 instead of PPPs involving firms we instead test PPPs formed via citation by university patents (with overlapping authors to the paper). This is, in a sense, a subset of our "transitive" PPPs but not requiring a firm to subsequently cite the university PPP. Here, although we see again a commercialization gap among papers with women, but we do not see any difference when women use more boastful language in their papers.

[18] Note: in capturing "boastful" for the full sample of papers in MAG, we did not count the number of such words but rather the occurrence of at least one "boastful" word. As in the twins subsample, we did not count papers with the word "novel" in the title or abstract where the word "coronavirus" also appeared.

tion and assess the likelihood of commercializing scientific articles of otherwise identical quality but heterogeous composition of the scientific teams.

The analytic approach used above cannot accomplish this because the setup is at the level of the academic paper. Here, we switch the level of analysis from the paper to the dyad of the paper and a patent that potentially form a patent-paper pair. As before, we account for the quality and nature of the paper with "twin" discoveries. To approximate random seeding of patents that could form a patent-paper pair, we adopt a case-control setup. We reduce our set of twin papers to those where one or the other indeed formed a pair with some patent. A dyad is formed both for the patent and the paper with which it is paired as well as the paper's twin, with which the paper was *not* paired but should be equally likely to have been paired.

The gender characteristics of the paper are as in the above models. For the patent, we calculate the percentage of male vs. female inventors on the patent using USPTO inventor-level classification (therefore this step is limited to USPTO-issued patents only). To avoid biasing this measure in favor of the paper that was actually cited, in calculating the gender balance we omit the inventor(s) who were matched to authors on the paper in the patent-paper pair. We then estimate the following equation:

$$\begin{aligned} COMM_{ijt} = \alpha_0 + \alpha_1 PaperGender_i + \alpha_2 PatentGender_j + \\ \alpha_3 PaperGender_i X PatentGender_j + \\ \beta X_{it} + TwinDiscoveryFE + Patent_j FE + \epsilon_{ijt} \end{aligned} \tag{3}$$

This equation deviates from Equation 2 with the inclusion of the gender composition of patent j as well as fixed effects for patent j.

Column (1) of Table 19 is a brief replication of our main (i.e., non-homophily) finding, using a dummy for any female authors on the patent. In column (2), the indicator for having at least one female author on the focal paper is interacted with the percentage of male inventors on the focal patent. (The base coefficient is not estimated due to the patent FE.) The negative and statistically-significant coefficient on 0nteraction of any-female-authors on the paper and percentage-of-male-inventors on the patent indicates that papers with a female author are less likely to be commercialized by a patent with more male inventors than its counterpart "twin" paper without any female authors.

Table 19 about here

This preference for same-gender findings in commercialization holds in columns 3 and 4 for a variety of gender composition measures of the paper, including percentage (3) and three-dummy model (4). In column (4), we find the same when interacting the gender composition of the patent with an indicator for whether the last author on the paper—often the lab manager—is female. The estimated interaction coefficient is likewise negative but not significant at conventional levels for papers where all authors are female (4). While this result tentatively points toward a homophilic pattern, it is worth acknowledging that teams' compositions for collaboration may be endogenous. In fact, it could rely on (1) the firms wanting to reduce communication costs by having a less diverse gendered group given the gender of the author with whom the collaboration will occur (Reagans and Zuckerman 2001); or (2) this could also be an implicit or explicit preference of the author with whom the collaboration may occur.

To put more arguments on the table for the possible homophily interpretation, we use a similar framework and consider the whole set of patents (not necessarily a patent-paper pair), citing one twin paper but not the other (columns 5-7 Table 19). In fact, the citation is a one-sided behavior invalidating the previous arguments. Moreover, patent citations are also a common way to measure the contribution of one academic paper to innovation. So, it could testify of the prevalence of such a homophilic pattern or not from the firm perspective in the use of academic research (not just their commercialization which involves collaboration with academic authors). We find qualitatively similar results pointing toward homophily from firms.

5 Robustness

We test the robustness of our findings in several ways. As described above, we hand-code gender for all authors including 1,704 'twin' articles reporting a simultaneous discovery where one article was commercialized but not the other. Columns (1-2) of Table 20, Panel A, reveals that the main results are robust to this restriction, even though 90% of the twin papers are omitted. This is true whether examining the any-female dummy (column 1), the percentage of women (column 2), or the separation into mixed-gender and all-female teams (column 3). In Appendix table A4 we moreover raise the required percentage of authors on a paper whose gender can be classified automatically (with a 90% confidence cutoff) from 33% to 50, 75, and 100%. Raising the cutoff reduces sample size but preserves the main result. Also, in Appendix table A10 we exclude "transitive" PPPs, retaining the main result.

Columns (4-6) of Table 20 Panel A follow Beck 2020 by re-estimating our main twins model using conditional logit, which omits any twin discovery where neither (or both) of the twin articles reporting the discovery is commercialized. Observations counts are closer to those of the hand-coded only models, with similar estimated coefficients. Appendix tables A5 and A6 establish the robustness of the results to including measures of authors' and institutions' prior experience with commercialization.

Table 20 about here

We defined a patent-paper pair as having a gap of no more than five years between the paper and the patent. In Panel B of Table 20 we shorten that maximum gap to two years. Examining both the percentage-female continuous measure and the dummy for teams at least one woman, we find in columns (1-2) a similar effect.

5.1 Biological twin definition

Finally, in Panel C we consider alternative "twin" formulations. Following Hill and Stein 2019, instead of relying on co-citation we construct pairs of papers according to unique biological sequence and structure. Our methodology is as follows. We begin with all proteins deposited to the Protein Data Bank (PDB) as of June 2022, which is a repository used in structural biology. PDB clusters proteins with similar underlying structural entities using its MMseqs2 sequence-clustering algorithm.[19] The MMseqs2 clustering algorithm can be implemented at varying levels of similarity; for example, Hill and Stein 2019 employ a 50% similarity match. Here, we employ a 100% similarity match.

Even so, a 100% similarity match in PDB can be misleading because this match can be, as Hill and Stein describe, for one of of many structural entities between proteins. We therefore employ the Uniprot database to ensure a unique sequence for proteins that share even identical portions of their substructure. Uniprot is complementary protein database; whereas PDB focuses on structure, Uniprot focuses on sequence. We submit the PDB identifiers from the previous step to Uniprot, which returns a Uniprot identifier for each PDB identifier. Because a unique protein substructure may be employed by multiple proteins with different sequences, Uniprot may map a single Uniprot identifier to multiple PDB identifiers. We therefore retain only the unique PDB-Uniprot identifier mappings to obtain a list of proteins that are unique in both structure and sequence.

Uniprot moreover returns a list of published articles (from PubMed) for each of its unique identifiers. For our list of proteins with unique structure and sequence we collect all PubMed identifiers (PMIDs). Some Uniprot identifiers are associated with many PMIDs because a single protein can be employed for a variety of studies. In Panel C of Table 20 we examine only proteins with a unique structure and sequence that are reflected in exactly two published articles. This exercise yields 5974 articles from 2987 proteins with unique structure and sequence. Because we define twins according to biological structure and sequence, we employ fixed effects for each PDB identifier. Column (2) of Table 20, Panel C suggests that among these "biological twins" there is a commercialization gap for scientific teams with more women as well as for mixed-gender teams where the last author is female (column 3). (The coefficient is imprecisely estimated with dummy variables for any women, in column (1)).

Table 20 about here

6 Conclusion

We provide the first large-scale characterization of the gender dynamics underlying the commercialization of science. Analyzing more than 70 million articles from the Microsoft Academic Graph, we find

[19] MMseqs2 is an upgraded version of the BLAST algorithm employed by Hill and Stein 2019.

that scientific teams with women suffer a commercialization penalty relative to all-male teams. These results are robust to controlling for latent commercializability using “twin” discoveries. We find some support for demand-side factors including the representation of women in a particular scientific field, though we do not find that women are aided by access to commercial networks. Neither does access to the articles themselves improve the gender dynamics of commercialization (though open access spurs commercialization more generally). On the contrary, we see the gap widening when women are in leading positions on the paper, suggesting that some bias might occur on the firm side. What does appear to matter (at least to firms) is word choice. However, even self promotion via boastful words does appear to benefit more to mixed gender teams with male last authors. We also find evidence of a two-sided gender effect where women are less likely to commercialize when the fraction of male inventors on the patent is high (this is also true for patent citations). Thus, this is likely to point to a homophilic behavior.

To summarize, our findings tend to point toward a bias from the firms’ side, not systematic across all fields, but with some homophilic features. The results of this analysis are keen in informing the public debate about the economic and welfare losses of such discriminatory behavior against scientific publications by women.

References

Baker, Andrew C, David F Larcker, and Charles CY Wang. 2022. "How much should we trust staggered difference-in-differences estimates?" *Journal of Financial Economics* 144 (2): 370–395.

Beck, Nathaniel. 2020. "Estimating grouped data models with a binary-dependent variable and fixed effects via a logit versus a linear probability model: The impact of dropped units." *Political Analysis* 28 (1): 139–145.

Bikard, Michaël. 2020. "Idea twins: Simultaneous discoveries as a research tool." *Strategic Management Journal* 41 (8): 1528–1543.

Bikard, Michaël, and Isabel Fernandez-Mateo. 2022. "Standing on the Shoulders of (Male) Giants: Gender Inequality and the Technological Impact of Scientific Ideas." *Available at SSRN 4059813.*

Bikard, Michaël, and Matt Marx. 2019. "Bridging academia and industry: How geographic hubs connect university science and corporate technology." *Management Science.*

Bjork, Bo-Christer, Annikki Roos, and Mari Lauri. 2009. "Scientific journal publishing: yearly volume and open access availability." *Information Research: An International Electronic Journal* 14 (1).

Brooks, Alison Wood, Laura Huang, Sarah Wood Kearney, and Fiona E Murray. 2014. "Investors prefer entrepreneurial ventures pitched by attractive men." *Proceedings of the National Academy of Sciences* 111 (12): 4427–4431.

Bryan, Kevin A, and Yasin Ozcan. 2021. "The impact of open access mandates on invention." *Review of Economics and Statistics* 103 (5): 954–967.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics* 225 (2): 200–230.

Campbell, Lesley G, Siya Mehtani, Mary E Dozier, and Janice Rinehart. 2013. "Gender-heterogeneous working groups produce higher quality science." *PloS one* 8 (10): e79147.

Carlin, John B, Lyle C Gurrin, Jonathan AC Sterne, Ruth Morley, and Terry Dwyer. 2005. "Regression models for twin studies: a critical review." *International Journal of Epidemiology* 34 (5): 1089–1099.

Ding, Waverly W, Fiona Murray, and Toby E Stuart. 2006. "Gender differences in patenting in the academic life sciences." *science* 313 (5787): 665–667.

Ducor, Philippe. 2000. "Coauthorship and coinventorship." *Science* 289 (5481): 873–875.

Eysenbach, Gunther. 2011. "Can tweets predict citations? Metrics of social impact based on Twitter and correlation with traditional metrics of scientific impact." *Journal of medical Internet research* 13 (4): e2012.

Fechner, Holly, and Matthew S Shapanka. 2018. "Closing diversity gaps in innovation: Gender, race, and income disparities in patenting and commercialization of inventions." *Technology & Innovation* 19 (4): 727–734.

Goldwasser, Eugene, Charles K-H Kung, and James Eliason. 1974. "On the mechanism of erythropoietin-induced differentiation: XIII. The role of sialic acid in erythropoietin action." *Journal of Biological Chemistry* 249 (13): 4202–4206.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics* 225 (2): 254–277.

Haans, Richard FJ, Constant Pieters, and Zi-Lin He. 2016. "Thinking about U: Theorizing and testing U-and inverted U-shaped relationships in strategy research." *Strategic management journal* 37 (7): 1177–1195.

Haeussler, Carolin, and Henry Sauermann. 2013. "Credit where credit is due? The impact of project contributions and social factors on authorship and inventorship." *Research Policy* 42 (3): 688–703.

Hill, Ryan, and Carolyn Stein. 2019. "Scooped! Estimating rewards for priority in science." *Job Market Paper.*

Hofstra, Bas, Vivek V Kulkarni, Sebastian Munoz-Najar Galvez, Bryan He, Dan Jurafsky, and Daniel A McFarland. 2020. "The diversity–innovation paradox in science." *Proceedings of the National Academy of Sciences* 117 (17): 9284–9291.

Hug, Sven E, and Martin P Brändle. 2017. "The coverage of Microsoft Academic: Analyzing the publication output of a university." *Scientometrics* 113 (3): 1551–1571.

Khabsa, Madian, and C Lee Giles. 2014. “The number of scholarly documents on the public web.” *PloS one* 9 (5): e93949.

Klar, Samara, Yanna Krupnikov, John Barry Ryan, Kathleen Searles, and Yotam Shmargad. 2020. “Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work.” *PloS one* 15 (4): e0229446.

Koffi, Marlène. 2021. “Gendered citations at top economic journals.” In *AEA Papers and Proceedings,* 111:60–64.

Koning, Rembrand, Sampsa Samila, and John-Paul Ferguson. 2021. “Who do we invent for? Patents by women focus more on women’s health, but few women get to invent.” *Science* 372 (6548): 1345–1348.

Lach, Saul, and Mark Schankerman. 2008. “Incentives and invention in universities.” *The RAND Journal of Economics* 39 (2): 403–433.

Lampe, Ryan. 2012. “Strategic citation.” *Review of Economics and Statistics* 94 (1): 320–333.

Lerchenmueller, Marc J, Olav Sorenson, and Anupam B Jena. 2019. “Gender differences in how scientists present the importance of their research: observational study.” *bmj* 367.

Marx, Matt, and Aaron Fuegi. 2020. “Reliance on science: Worldwide front-page patent citations to scientific articles.” *Strategic Management Journal.*

———. 2021. *Reliance on science by inventors: Hybrid extraction of in-text patent-to-article citations.* Technical report. National Bureau of Economic Research.

Marx, Matt, and David H Hsu. 2021. “Revisiting the Entrepreneurial Commercialization of Academic Science: Evidence from “Twin” Discoveries.” *Management Science.*

Miyake, Takaji, Charles K Kung, and Eugene Goldwasser. 1977. “Purification of human erythropoietin.” *Journal of Biological Chemistry* 252 (15): 5558–5564.

Murray, Fiona. 2002. “Innovation as co-evolution of scientific and technological networks: exploring tissue engineering.” *Research Policy* 31 (8-9): 1389–1403.

Murray, Fiona, and Leigh Graham. 2007. “Buying science and selling science: Gender differences in the market for commercial science.” *Industrial and Corporate Change* 16 (4): 657–689.

Ranaei, Samira, Antti Knutas, Juho Salminen, and Arash Hajikhani. 2016. “Cloud-based Patent and Paper Analysis Tool for Comparative Analysis of Research.” In *CompSysTech,* 315–322.

Reagans, Ray, and Ezra W Zuckerman. 2001. “Networks, diversity, and productivity: The social capital of corporate R&D teams.” *Organization science* 12 (4): 502–517.

Rivette, Kevin G, and David Kline. 2000. *Rembrandts in the attic: Unlocking the hidden value of patents.* Harvard Business Press.

Rothaermel, Frank T, Shanti D Agung, and Lin Jiang. 2007. “University entrepreneurship: a taxonomy of the literature.” *Industrial and Corporate Change* 16 (4): 691–791.

Sarsons, Heather. 2017. “Recognition for group work: Gender differences in academia.” *American Economic Review* 107 (5): 141–45.

Sherwood, Judith B, and Eugene Goldwasser. 1979. “A radioimmunoassay for erythropoietin.”

Simcoe, Timothy S, and Dave M Waguespack. 2011. “Status, quality, and attention: What’s in a (missing) name?” *Management Science* 57 (2): 274–290.

Sinha, Arnab, Zhihong Shen, Yang Song, Hao Ma, Darrin Eide, Bo-June Hsu, and Kuansan Wang. 2015. “An overview of microsoft academic service (mas) and applications.” In *Proceedings of the 24th international conference on world wide web,* 243–246.

Staudt, Joseph. 2020. “Mandating access: assessing the NIH’s public access policy.” *Economic policy* 35 (102): 269–304.

Sun, Liyang, and Sarah Abraham. 2021. “Estimating dynamic treatment effects in event studies with heterogeneous treatment effects.” *Journal of Econometrics* 225 (2): 175–199.

Tartari, Valentina, and Ammon Salter. 2015. “The engagement gap:: Exploring gender differences in University–Industry collaboration activities.” *Research Policy* 44 (6): 1176–1191.

Thompson, Neil C, Arvids A Ziedonis, and David C Mowery. 2018. “University licensing and the flow of scientific knowledge.” *Research Policy* 47 (6): 1060–1069.

Thursby, Jerry G, and Marie C Thursby. 2005. “Gender patterns of research and licensing activity of science and engineering faculty.” *The Journal of Technology Transfer* 30 (4): 343–353.

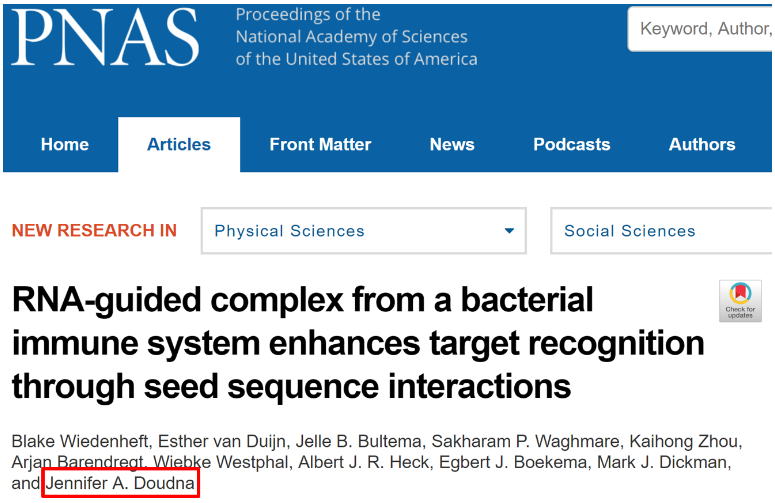
Ware, Mark, and Michael Mabe. 2015. “The STM report: An overview of scientific and scholarly journal publishing.”

Whittington, Kjersten Bunker, and Laurel Smith-Doerr. 2005. “Gender and commercial science: Women’s patenting in the life sciences.” *The Journal of Technology Transfer* 30 (4): 355–370.

Yang, Yang, Teresa Woodruff, Yuan Tian, Benjamin F Jones, and Brian Uzzi. 2021. “Gender Diverse Teams Produce More Innovative and Influential 2 Ideas in Medical Research 3.”

Figure 2: Patent-paper pair example

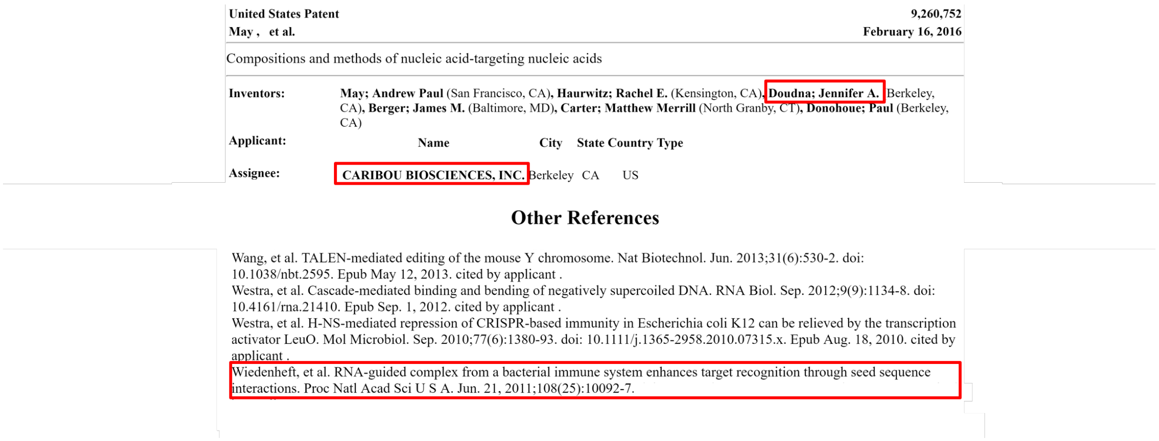
Panel A: PNAS Article in the Patent-Paper Pair



RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions

Blake Wiedenheft, Esther van Duijn, Jelle B. Bultema, Sakharam P. Waghmare, Kaihong Zhou, Arjan Barendregt, Wiebke Westphal, Albert J. R. Heck, Egbert J. Boekema, Mark J. Dickman, and Jennifer A. Doudna

Panel B: U.S. Patent 9,260,752 in the Patent-Paper Pair



United States Patent **9,260,752**
May , et al. **February 16, 2016**

Compositions and methods of nucleic acid-targeting nucleic acids

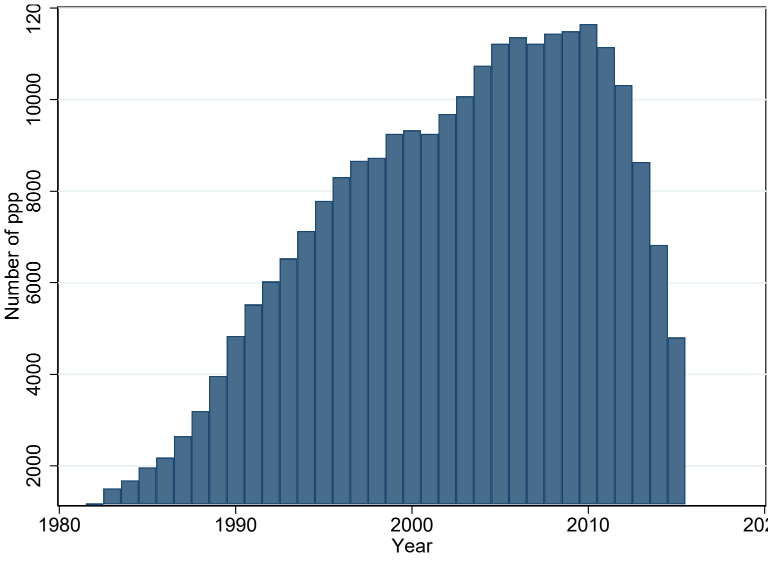
Inventors: **May; Andrew Paul** (San Francisco, CA), **Haurwitz; Rachel E.** (Kensington, CA), **Doudna; Jennifer A.** Berkeley, CA), **Berger; James M.** (Baltimore, MD), **Carter; Matthew Merrill** (North Granby, CT), **Donohoue; Paul** (Berkeley, CA)

	Name	City	State	Country	Type
Applicant:					
Assignee:	**CARIBOU BIOSCIENCES, INC.**	Berkeley	CA	US	

Other References

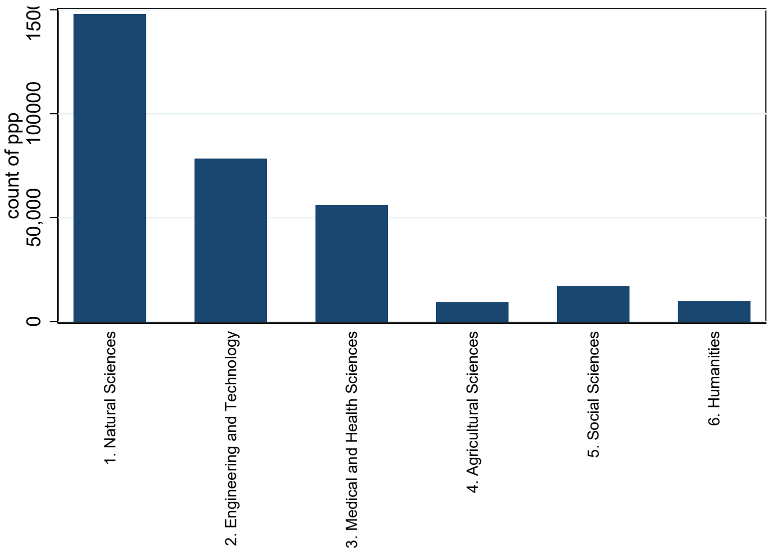
Wang, et al. TALEN-mediated editing of the mouse Y chromosome. Nat Biotechnol. Jun. 2013;31(6):530-2. doi: 10.1038/nbt.2595. Epub May 12, 2013. cited by applicant .
Westra, et al. Cascade-mediated binding and bending of negatively supercoiled DNA. RNA Biol. Sep. 2012;9(9):1134-8. doi: 10.4161/rna.21410. Epub Sep. 1, 2012. cited by applicant .
Westra, et al. H-NS-mediated repression of CRISPR-based immunity in Escherichia coli K12 can be relieved by the transcription activator LeuO. Mol Microbiol. Sep. 2010;77(6):1380-93. doi: 10.1111/j.1365-2958.2010.07315.x. Epub Aug. 18, 2010. cited by applicant .
Wiedenheft, et al. RNA-guided complex from a bacterial immune system enhances target recognition through seed sequence interactions. Proc Natl Acad Sci U S A. Jun. 21, 2011;108(25):10092-7.

Figure 3: Commercialization rates over time



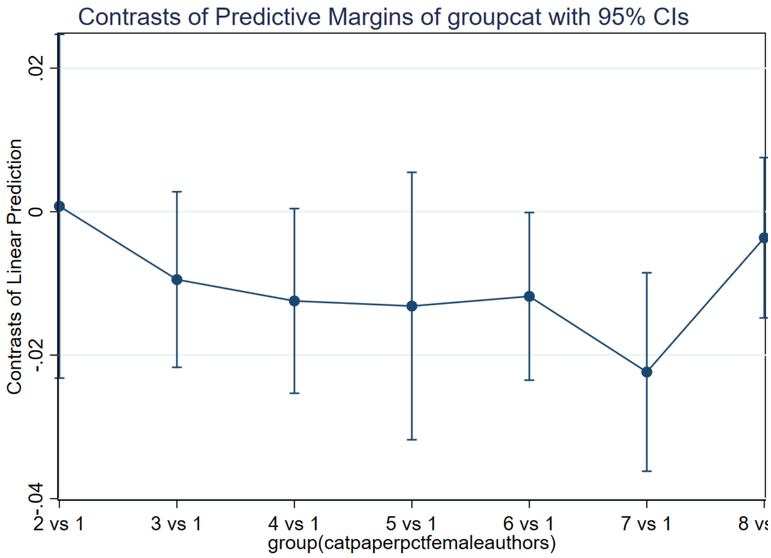
Notes: Count of commercialized papers per year as per our PPP methodology in Section 2.1.

Figure 4: Commercialization rates by OECD category



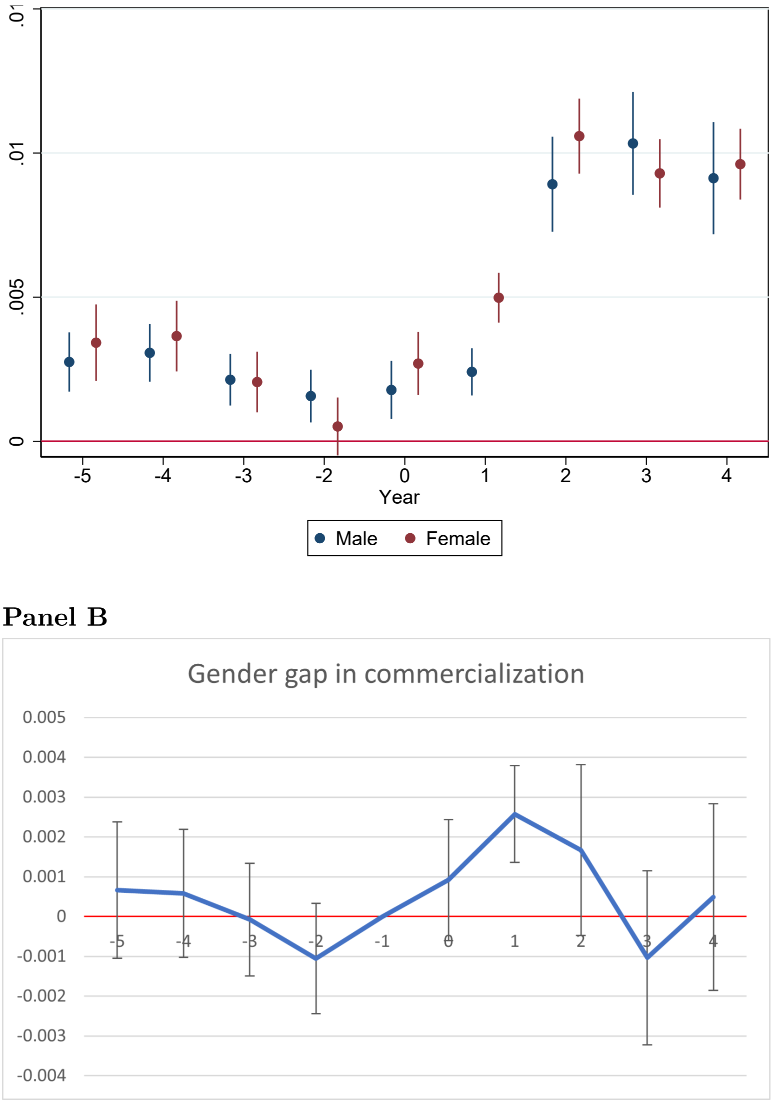
Notes: Count of commercialized papers as per our PPP methodology in Section 2.1, segmented by OECD top-level categories defined at http://www.oecd.org/science/inno/38235147.pdf.

Figure 5: Sensitivity of gender gap to team composition



Notes: Figure 5 plots estimated coefficients and 95% confidence intervals from a series of unreported estimations of Equation 2 where the indicator for gender is a progressively more restrictive threshold for the percentage of women, in 5-percent increments. The leftmost coefficient plotted is for papers with zero percent female authors, and the rightmost coefficient is for papers with all female authors. The coefficients in between correspond to papers with *at least* the depicted percentage of women.

Figure 6: Impact of Open Access mandates on commercialization



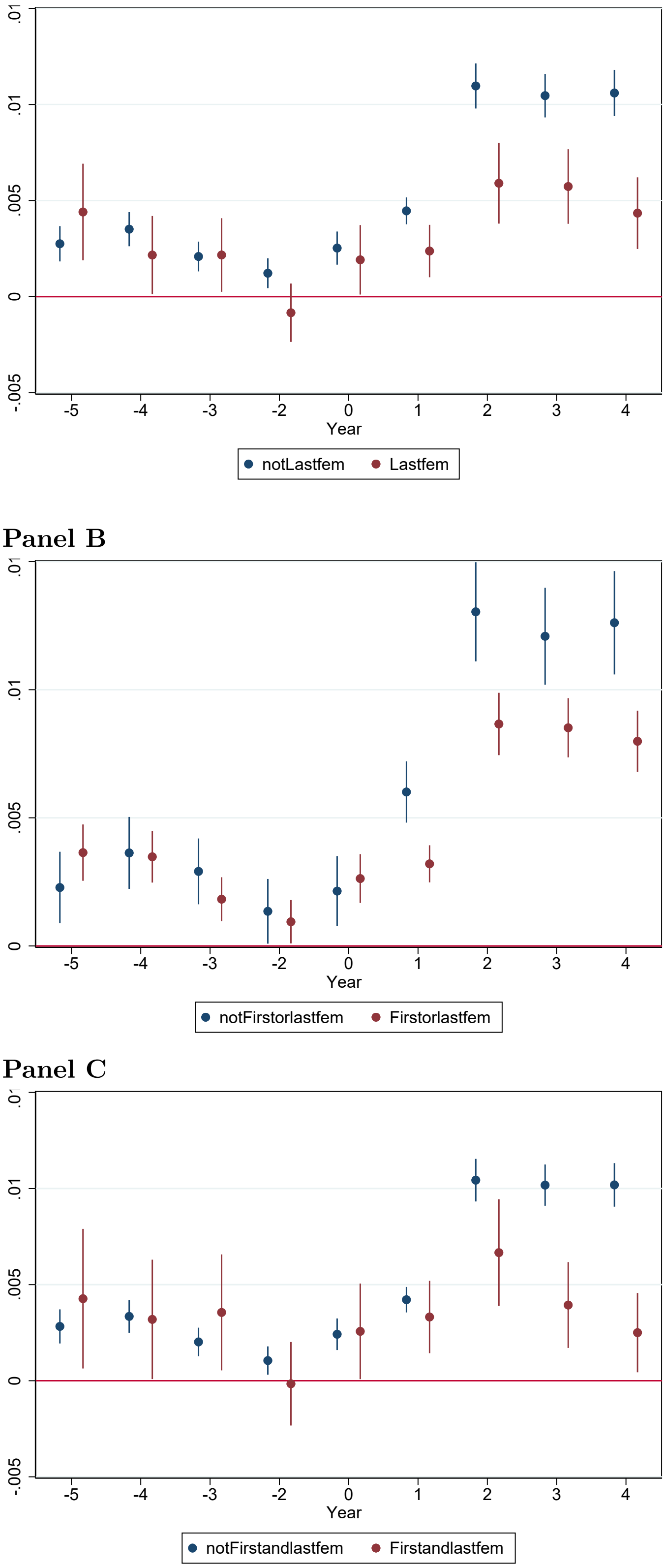
Panel A

Panel B

Notes: Figure 6 shows the estimation results of the staggered triple difference to assess the effect of Open Access on the gender gap in commercialization. Panel A shows the staggered difference-in-difference separately for male-authored and female-authored papers (papers with any woman). Panel B plots the triple difference results, therefore assessing the effect of the Open Access policy on the gender gap. The unit of observation is the academic article. The dependent variable is the commercialization measured by the patent-paper-pair whose assignee is a firm. All estimates include controls for the number of authors, the authors' average prominence and institutions, the fields, and years dummies. The coefficient for event time −1 is omitted to normalize the gender commercialization gap to zero in the year prior to the policy.

Figure 7: Impact of Open Access on senior women

Panel A



Panel B

Panel C

Notes: Figure 7 shows the estimation results of the staggered triple difference to assess the effect of Open Access on the gender gap in commercialization. Panel A shows the staggered difference-in-difference separately for papers with a woman as last author vs not. Panel B shows the staggered difference-in-difference separately for papers with a woman as either first or last author vs not. Panel C shows the staggered difference-in-difference separately for papers with a woman as both first and last author vs not. The unit of observation is the academic article. The dependent variable is the commercialization measured by the patent-paper-pair whose assignee is a firm. All estimates include controls for the number of authors, the authors' average prominence and institutions, the fields, and years dummies. The coefficient for event time −1 is omitted to normalize the gender commercialization gap to zero in the year prior to the policy.

Table 1: Top 15 Commercializers

1. International Business Machines Corporation
2. Semiconductor Energy Laboratory Co., Ltd
3. Microsoft Corporation
4. Ignis Innovation Inc.
5. Genentech, Inc.
6. Immunomedics, Inc.
7. Intel Corporation
8. Abbot Diabetes Care Inc.
9. Yeda Research and Development Co. Ltd.
10. Bristol-Myers Squibb Company
11. Immunex Corporation
12. Google Inc.
13. Hewlett-Packard
14. Lucent Technologies Inc.
15. Schlumberger Technology

Table 2: Descriptive statistics for 70,013,980 articles

Variable	Obs	Mean	Std. Dev.	Min	Max
Commercialization	70013980	.0033	.0575	0	1
Cooperative Commercialization with Existing Firms	70013980	.0032	.0568	0	1
Self-Commercialization via Startups	70013980	.0001	.0089	0	1
Ln scientific citations (5-years, forward)	70013980	1.0962	1.239	0	11.4892
Ln patent citations (forward, front-page)	70013980	.0633	.336	0	9.2406
Ln patent citations (forward, in-text)	70013980	.0317	.2357	0	9.9398
Ln authors	70013980	1.3103	.5362	.6931	4.5951
Ln average citations per author	70013980	4.6333	2.9354	0	12.5171
Ln average citations per institution	70013980	9.0133	9.6503	0	20.9692
Ln Journal Impact Factor	70013980	.4833	.6099	0	6.7052
At least one female	70013980	.4147	.4927	0	1
\% female	70013980	.2446	.349	0	1
Gender entropy	70013980	.2615	.4116	0	1
Mixed-gender teams	70013980	.2931	.4552	0	1
All female	70013980	.1215	.3268	0	1
Pct female in field-year	60521710	.2332	.1065	0	1
Ln avg num coauthors at firms	70013980	.822	1.0238	0	6.6425

.

Table 3: Female scientists: commercialization vs. academic citation (cross sectional)

Panel A

	Commercialization		Academic Citations	
	(1)	(2)	(3)	(4)
At least one female	-0.000640***		0.0359***	
	(0.0000175)		(0.000265)	
% female		-0.00122***		0.0408***
		(0.0000164)		(0.000337)
Observations	60521696	60521696	60521696	60521696
field/ year FE	y	y	y	y

Panel B

	Commercialization		Academic Citations	
	(1)	(2)	(3)	(4)
% female	-0.00505***		0.107***	
	(0.0000995)		(0.00143)	
% female × % female	0.00424***		-0.0727***	
	(0.0000986)		(0.00150)	
Mixed-gender teams		-0.000908***		0.0341***
		(0.0000243)		(0.000332)
All female		-0.000179***		0.0389***
		(0.0000129)		(0.000358)
Observations	60521696	60521696	60521696	60521696
field/ year FE	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 4: Descriptive statistics for “twin” articles

Variable	Obs	Mean	Std. Dev.	Min	Max
Commercialization	24682	.0437	.2044	0	1
Cooperative Commercialization with Existing Firms	24682	.0424	.2015	0	1
Self-Commercialization via Startups	24682	.0013	.0354	0	1
Ln scientific citations (5-years, forward)	24682	3.7651	1.1972	0	8.8364
Ln patent citations (forward, front-page)	24682	.6814	1.1218	0	7.1115
Ln patent citations (forward, in-text)	24682	.5797	1.0557	0	9.2014
Ln authors	24682	1.8329	.529	.6931	4.5433
Ln average citations per author	24682	7.5813	1.363	0	11.4417
Ln average citations per institution	24682	6.6875	8.026	0	20.9692
Ln Journal Impact Factor	24682	1.6913	.687	0	4.2995
At least one female	24682	.6423	.4793	0	1
\% female	24682	.2612	.2611	0	1
Gender entropy	24682	.5242	.4337	0	1
Mixed-gender teams	24682	.6103	.4877	0	1
All female	24682	.032	.176	0	1
Pct female in field-year	24665	.2514	.074	.0098	.5783
Ln avg num coauthors at firms	24682	1.6545	.9481	0	5.5858
Count of boastful words	24682	.1855	.4737	0	6

“female” refers to female authors on the scientific article.

Table 5: Female scientists: commercialization (twins)

	Commercialization			
	(1)	(2)	(3)	(4)
At least one female	-0.0113**			
	(0.00420)			
% female		-0.0212**	-0.0533**	
		(0.00651)	(0.0185)	
% female × % female			0.0418*	
			(0.0198)	
Mixed-gender teams				-0.0121**
				(0.00436)
All female				-0.00214
				(0.00569)
Observations	22698	22698	22698	22698
Twin FE	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 6: Testing curvilinear relationship

Panel A: Variance-covariance matrix

	% female	% female × % female
% female	.0003413	-.0003451
% female × % female	-.0003451	.0003916

Panel B: Split by suggested cutoff

	(1)	(2)	(3)
% female	-0.0304**	-0.0240	-0.0212**
	(0.0117)	(0.0246)	(0.00651)
Observations	18812	326	22698
twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 7: Positionality (twins)

Panel A: First vs. last author

	(1)	(2)	(3)	(4)	(5)
Female first author and male last author	-0.00907	-0.00216	-0.000508	-0.00459	-0.00497
	(0.00796)	(0.0104)	(0.0106)	(0.00885)	(0.00860)
Male first author and female last author	-0.0256*	-0.0185	-0.0163	-0.0204+	-0.0213+
	(0.0115)	(0.0131)	(0.0135)	(0.0123)	(0.0119)
Female first author and female last author	-0.0275*	-0.0132	-0.0161	-0.0243+	-0.0205+
	(0.0111)	(0.0173)	(0.0176)	(0.0133)	(0.0123)
% female		-0.0264	-0.0289		
		(0.0212)	(0.0221)		
% female authors (not first or last)				-0.0207	
				(0.0149)	
% female authors (not first or last, missing=0)					-0.0246+
					(0.0141)
Observations	12786	12786	12020	12020	12786
twin FE	y	y	y	y	y

Panel B: senior female author

	(1)	(2)	(3)	(4)
Female last author=1	-0.000721***	-0.0199**	-0.000162***	-0.0173*
	(0.0000174)	(0.00696)	(0.0000273)	(0.00722)
% female authors (not last)			-0.00104***	-0.0158+
			(0.0000285)	(0.00856)
Observations	51807800	17426	51807800	17426
twin FE	y	y	y	y

Omitted category in Panel A is articles where both the first and last author are male. Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 8: Female scientists and commercialization by field

Panel A

	Natural Sciences		Engineering		Medical		Social Sciences	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
At least one female	-0.000750***	-0.0164*	-0.000971***	-0.0692*	-0.000443***	-0.000339	-0.000290***	0.0126
	(0.0000359)	(0.00759)	(0.0000582)	(0.0325)	(0.0000293)	(0.00681)	(0.0000227)	(0.0142)
Observations	21029158	8736	9006116	1014	15090537	4188	10798779	878
year/field FE	n	y	n	y	n	y	n	y
twin FE	y	n	y	n	y	n	y	n

Panel B

	Natural Sciences		Engineering		Medical		Social Sciences	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
% female	-0.00201***	-0.0319**	-0.00202***	-0.0609	-0.000903***	-0.0144	-0.000418***	0.00617
	(0.0000399)	(0.0120)	(0.0000650)	(0.0583)	(0.0000255)	(0.0110)	(0.0000171)	(0.0216)
Observations	21029158	8736	9006116	1014	15090537	4188	10798779	878
year/field FE	n	y	n	y	n	y	n	y
twin FE	y	n	y	n	y	n	y	n

Panel C

	Natural Sciences		Engineering		Medical		Social Sciences	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Mixed-gender teams	0.000228***	-0.0143	-0.000165+	-0.0803*	-0.0000879+	0.00283	-0.000138**	0.0160
with male last author	(0.0000553)	(0.00908)	(0.0000917)	(0.0372)	(0.0000506)	(0.00851)	(0.0000488)	(0.0178)
Mixed-gender teams	-0.00209***	-0.0308*	-0.00238***	-0.0795	-0.00134***	-0.0169	-0.000931***	0.00871
with female last author	(0.0000624)	(0.0132)	(0.000106)	(0.0580)	(0.0000541)	(0.0124)	(0.0000446)	(0.0227)
All Female	-0.000714***	-0.0107	-0.000597***	-0.0437	-0.000159***	0.0134*	-0.000101***	0.0127
	(0.0000334)	(0.0138)	(0.0000563)	(0.0948)	(0.0000194)	(0.00610)	(0.0000137)	(0.0119)
Observations	20086857	7804	8632827	870	14453855	3620	10493689	778
year/field FE	n	y	n	y	n	y	n	y
twin FE	y	n	y	n	y	n	y	n

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 9: Female scientists: commercialization over time

	Commercialization		
	(1)	(2)	(3)
% female × year	-0.0000686*** (0.000000592)		
year19611965=1 × % female		0.000868*** (0.0000325)	-0.000824*** (0.0000259)
year19661970=1 × % female		0.000966*** (0.0000441)	-0.000380*** (0.0000412)
year19711975=1 × % female		0.000522*** (0.0000630)	-0.000959*** (0.0000607)
year19761980=1 × % female		0.000707*** (0.0000867)	-0.000977*** (0.0000844)
year19811985=1 × % female		0.000520*** (0.0000960)	-0.00116*** (0.0000939)
year19861990=1 × % female		0.000398*** (0.000108)	-0.00128*** (0.000106)
year19911995=1 × % female		0.000420*** (0.000116)	-0.00126*** (0.000114)
year19962000=1 × % female		-0.000734*** (0.0000990)	-0.00242*** (0.0000969)
year20012005=1 × % female		-0.00139*** (0.0000770)	-0.00308*** (0.0000744)
year20062010=1 × % female		-0.000984*** (0.0000630)	-0.00267*** (0.0000599)
year20112015=1 × % female		-0.0000721+ (0.0000429)	-0.00175*** (0.0000381)
year20162021=1 × % female		0.00250*** (0.0000265)	
year1960pre=1 × % female			-0.000508*** (0.0000170)
Observations	60521697	60521697	60521697
Twin FE	y	y	y

Notes: Year dummies and percent-female base variables not shown. All models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 10: Female scientists: commercialization and representation in scientific fields (“twins”)

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0115**	-0.0238+				
	(0.00422)	(0.0142)				
Pct female in field-year	0.0446	0.0113	0.0477	-0.00246	0.0558	0.0173
	(0.0329)	(0.0453)	(0.0329)	(0.0410)	(0.0383)	(0.0507)
At least one female × Pct female in field-year		0.0495				
		(0.0490)				
% female			-0.0217***	-0.0654**		
			(0.00654)	(0.0225)		
% female × Pct female in field-year				0.165*		
				(0.0719)		
Mixed-gender teams with male last author					-0.00948+	-0.0159
					(0.00518)	(0.0186)
Mixed-gender teams with female last author					-0.0260***	-0.0629*
					(0.00712)	(0.0254)
All Female					-0.00250	-0.0252
					(0.00621)	(0.0223)
Mixed-gender teams with male last author × Pct female in field-year						0.0283
						(0.0647)
Mixed-gender teams with female last author × Pct female in field-year						0.140+
						(0.0849)
All Female × Pct female in field-year						0.0862
						(0.0689)
Observations	22670	22670	22670	22670	19764	19764
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor . + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 11: Female scientists and commercial networks (twins)

	Avg. num coauthors at firm (ln)		
	(1)	(2)	(3)
At least one female	-0.0564***		
	(0.0141)		
% female		-0.132***	
		(0.0255)	
Mixed-gender teams with male last author			-0.0318*
			(0.0156)
Mixed-gender teams with female last author			-0.0797***
			(0.0206)
All Female			-0.0914*
			(0.0426)
Observations	22698	22698	19790
twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 12: Commercialization and Networks (twins)

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0112**	-0.0151**				
	(0.00421)	(0.00567)				
Ln avg num coauthors at firms	0.00236	0.000868	0.00222	0.00545+	0.00238	0.000388
	(0.00273)	(0.00333)	(0.00273)	(0.00312)	(0.00311)	(0.00365)
At least one female × Ln avg num coauthors at firms		0.00252				
		(0.00362)				
% female			-0.0209**	-0.00293		
			(0.00652)	(0.00786)		
% female × Ln avg num coauthors at firms				-0.0130*		
				(0.00601)		
Mixed-gender teams with male last author					-0.00923+	-0.0171*
					(0.00517)	(0.00756)
Mixed-gender teams with female last author					-0.0255***	-0.0291*
					(0.00711)	(0.0115)
All Female					-0.00165	0.00376
					(0.00615)	(0.00741)
Mixed-gender teams with male last author × Ln avg num coauthors at firms						0.00488
						(0.00457)
Mixed-gender teams with female last author × Ln avg num coauthors at firms						0.00248
						(0.00736)
All Female × Ln avg num coauthors at firms						-0.00682
						(0.00638)
Observations	22698	22698	22698	22698	19790	19790
Twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 13: Accessibility of articles

	Any Female	Male	Triple difference
	(1)	(2)	(3)
-5	0.00341939***	0.00275115***	0.00066824
	(0.00067777)	(0.00052428)	(0.000856824)
-4	0.00364887***	0.00306689***	0.00058198
	(0.0006253)	(0.00050915)	(0.000806386)
-3	0.00205633***	0.00213465***	-7.832E-05
	(0.00053755)	(0.00045627)	(0.000705083)
-2	0.00051376***	0.00156925	-0.00105549
	(0.00051271)	(0.0004672)	(0.000693648)
-1	0	0	0
	(0)	(0)	(0)
0	0.00269603***	0.00178032***	0.00091571
	(0.00055824)	(0.00051509)	(0.000759572)
1	0.00498088***	0.002408***	0.00257288***
	(0.00044047)	(0.00041735)	(0.000606791)
2	0.01058647***	0.00891711***	0.00166936
	(0.00066384)	(0.00084129)	(0.001071659)
3	0.00929429***	0.01033277***	-0.00103848
	(0.00060507)	(0.00091059)	(0.00109329)
4	0.00961399***	0.0091269***	0.00048709
	(0.0006259)	(0.00099224)	(0.001173154)
Observations	73690308	73690308	73690308

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 14: Commercialization, gender, and attention

	Count of “boastful” words		
	(1)	(2)	(3)
At least one female	-0.0233*		
	(0.0100)		
% female		-0.0392*	
		(0.0171)	
Mixed-gender teams with male last author			-0.0261*
			(0.0112)
Mixed-gender teams with female last author			-0.0407**
			(0.0153)
All Female			0.0242
			(0.0243)
Observations	22698	22698	19790
Twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 15: Academic citations and the use of “boastful” words (twins)

	Academic Citations					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female=1	0.00566	0.00617				
	(0.0134)	(0.0145)				
Count of boastful words	0.00662	0.00849	0.00694	-0.000488	-0.00325	0.00449
	(0.0138)	(0.0229)	(0.0137)	(0.0190)	(0.0142)	(0.0240)
At least one female=1 × Count of boastful words		-0.00274				
		(0.0261)				
% female			0.0377	0.0325		
			(0.0249)	(0.0269)		
% female × Count of boastful words				0.0286		
				(0.0484)		
Mixed-gender teams with male last author					0.00715	0.0116
					(0.0146)	(0.0156)
Mixed-gender teams with female last author					0.0198	0.0182
					(0.0208)	(0.0224)
All Female					0.0260	0.00561
					(0.0403)	(0.0443)
Mixed-gender teams with male last author × Count of boastful words						-0.0233
						(0.0284)
Mixed-gender teams with female last author × Count of boastful words						0.00561
						(0.0398)
All Female × Count of boastful words						0.108
						(0.0699)
Constant	1.857***	1.857***	1.845***	1.846***	1.840***	1.840***
	(0.0625)	(0.0627)	(0.0635)	(0.0635)	(0.0681)	(0.0682)
Observations	22698	22698	22698	22698	19790	19790
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 16: Commercialization and the use of “boastful” words (twins)

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female=1	-0.0111**	-0.0160***				
	(0.00420)	(0.00431)				
Count of boastful words	0.00976*	-0.00840	0.00974*	0.00385	0.0113*	-0.00978
	(0.00462)	(0.00707)	(0.00462)	(0.00603)	(0.00537)	(0.00749)
At least one female=1 × Count of boastful words		0.0267**				
		(0.00930)				
% female			-0.0208**	-0.0250***		
			(0.00650)	(0.00673)		
% female × Count of boastful words				0.0227		
				(0.0152)		
Mixed-gender teams with male last author					-0.00901+	-0.0158**
					(0.00516)	(0.00530)
Mixed-gender teams with female last author					-0.0253***	-0.0287***
					(0.00708)	(0.00742)
All Female					-0.00214	-0.00480
					(0.00616)	(0.00663)
Mixed-gender teams with male last author × Count of boastful words						0.0363**
						(0.0113)
Mixed-gender teams with female last author × Count of boastful words						0.0206
						(0.0147)
All Female × Count of boastful words						0.0155
						(0.0107)
Constant	-0.0705***	-0.0673***	-0.0652***	-0.0642***	-0.0841***	-0.0807***
	(0.0147)	(0.0147)	(0.0148)	(0.0148)	(0.0173)	(0.0173)
Observations	22698	22698	22698	22698	19790	19790
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 17: Commercialization, gender, and attention - full sample

	Count of “boastful” words		
	(1)	(2)	(3)
At least one female	-0.00493*** (0.0000830)		
% female		-0.00754*** (0.000102)	
Mixed-gender teams with male last author			-0.00419*** (0.000126)
Mixed-gender teams with female last author			-0.00793*** (0.000154)
All Female			-0.00461*** (0.000106)
Observations	60521696	60521696	58083572
year/field FE	y	y	y

Omitted category in Panel A is articles where both the first and last author are male. Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 18: Commercialization and the use of “boastful” words (full)

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female=1	-0.000620*** (0.0000175)	-0.000719*** (0.0000174)				
paperhasnovelwords=1	0.00387*** (0.0000421)	0.00337*** (0.0000546)	0.00387*** (0.0000421)	0.00452*** (0.0000523)	0.00384*** (0.0000432)	0.00334*** (0.0000546)
At least one female=1 × paperhasnovelwords=1		0.00107*** (0.0000844)				
% female			-0.00119*** (0.0000164)	-0.000984*** (0.0000160)		
paperhasnovelwords=1 × % female				-0.00278*** (0.0000987)		
Mixed-gender teams with male last author					-0.0000998*** (0.0000300)	-0.000485*** (0.0000302)
Mixed-gender teams with female last author					-0.00187*** (0.0000320)	-0.00185*** (0.0000316)
All Female					-0.000214*** (0.0000129)	-0.0000141 (0.0000126)
Mixed-gender teams with male last author × paperhasnovelwords=1						0.00329*** (0.000121)
Mixed-gender teams with female last author × paperhasnovelwords=1						-0.000113 (0.000137)
All Female × paperhasnovelwords=1						-0.00332*** (0.0000835)
Constant	-0.00539*** (0.0000293)	-0.00534*** (0.0000291)	-0.00512*** (0.0000295)	-0.00517*** (0.0000295)	-0.00545*** (0.0000318)	-0.00541*** (0.0000317)
Observations	60521696	60521696	60521696	60521696	58083572	58083572
year/field FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 19: Commercialization and gender homophily (Paper-patent twin dyads)

	Patent-paper pair dyads				Patent-to-paper citation dyads		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Any female authors	-0.113*** (0.0243)						
Any female authors=1		0.212* (0.0912)			0.0133 (0.00930)		
Any female authors=1 × Pct male inventors		-0.372*** (0.101)			-0.169*** (0.0101)		
Pct female authors			0.849*** (0.154)			0.301*** (0.0195)	
Pct female authors × Pct male inventors			-1.473*** (0.181)			-0.358*** (0.0222)	
Mixed-gender teams with male last author				0.144 (0.0985)			-0.0271** (0.00982)
Mixed-gender teams with female last author				0.580*** (0.121)			0.120*** (0.0159)
All Female				-0.140 (0.374)			0.241*** (0.0526)
Mixed-gender teams with male last author × Pct male inventors				-0.265* (0.108)			-0.181*** (0.0105)
Mixed-gender teams with female last author × Pct male inventors				-1.011*** (0.141)			-0.0319+ (0.0180)
All Female × Pct male inventors				-0.315 (0.378)			-0.306*** (0.0614)
Constant	-1.574*** (0.0856)	-1.574*** (0.0854)	-1.441*** (0.0871)	-1.488*** (0.0881)	-1.454*** (0.103)	-1.529*** (0.103)	-1.267*** (0.104)
Observations	5764	5764	5764	5636	283605	283605	269638
twin-paper FE	y	y	y	y	y	y	y
citing patent FE	y	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table 20: Robustness

Panel A: by-hand gender classification and conditional logit

	Commercialization					
	By-hand gender coding			conditional logit		
	(1)	(2)	(3)	(4)	(5)	(6)
main						
At least one female	-0.111*			-0.307*		
	(0.0497)			(0.135)		
% female		-0.294**			-0.796**	
		(0.0908)			(0.255)	
Mixed-gender teams with male last author			-0.0683			-0.211
			(0.0521)			(0.140)
Mixed-gender teams with female last author			-0.229***			-0.596**
			(0.0683)			(0.191)
All Female			-0.419*			-2.027*
			(0.173)			(0.997)
Constant	-1.212***	-1.146***	-1.177***			
	(0.202)	(0.204)	(0.202)			
Observations	1982	1982	1982	1704	1704	1698
twin FE	y	y	y	y	y	y

Panel B: 2-year commercialization window

	Commercialization		
	(1)	(2)	(3)
At least one female	-0.00648+		
	(0.00347)		
% female		-0.0114*	
		(0.00516)	
Mixed-gender teams with male last author			-0.00703
			(0.00430)
Mixed-gender teams with female last author			-0.0132*
			(0.00574)
All Female			0.00232
			(0.00447)
Constant	-0.0488***	-0.0460***	-0.0572***
	(0.0117)	(0.0119)	(0.0138)
Observations	22724	22724	19816
twin FE	y	y	y

Panel C: Alternative twins

	Commercialization		
	(1)	(2)	(3)
At least one female	-0.00972		
	(0.00891)		
% female		-0.0354*	
		(0.0171)	
Mixed-gender teams with male last author			-0.00311
			(0.0101)
Mixed-gender teams with female last author			-0.0265*
			(0.0132)
All Female			-0.0380
			(0.0283)
Constant	-0.00563	0.00501	0.00766
	(0.0283)	(0.0286)	(0.0306)
Observations	5917	5917	5091
twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A1: Female scientists: Citations from academic articles (twins)

	Citation		
	(1)	(2)	(3)
At least one female	0.00561 (0.0134)		
% female		0.0380 (0.0249)	
Mixed-gender teams with male last author			0.00748 (0.0145)
Mixed-gender teams with female last author			0.0198 (0.0208)
All Female			0.0254 (0.0403)
Observations	22724	22724	19816
Twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A2: Networks (twins): At least one of the authors is working at a firm

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.00988* (0.00434)	-0.00909* (0.00434)				
anyaffilfirm	0.0491*** (0.0142)	0.0685** (0.0249)	0.0486*** (0.0142)	0.0863*** (0.0209)	0.0601*** (0.0164)	0.0725** (0.0264)
At least one female × anyaffilfirm		-0.0242 (0.0286)				
% female			-0.0192** (0.00673)	-0.0153* (0.00672)		
% female × anyaffilfirm				-0.139* (0.0552)		
Mixed-gender teams with male last author					-0.00731 (0.00530)	-0.00765 (0.00530)
Mixed-gender teams with female last author					-0.0250*** (0.00730)	-0.0209** (0.00733)
All Female					-0.00149 (0.00649)	-0.000201 (0.00655)
Mixed-gender teams with male last author × anyaffilfirm						0.00157 (0.0330)
Mixed-gender teams with female last author × anyaffilfirm						-0.0901+ (0.0462)
All Female × anyaffilfirm						-0.0757* (0.0308)
Observations	21518	21518	21518	21518	18736	18736
Twin FE	y	y	y	y	y	y

Table A3: Networks (twins): Total number of coauthors working at a firm

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0106* (0.00420)	-0.0167** (0.00565)				
Num coauthors at firms	0.00684** (0.00208)	0.00506+ (0.00261)	0.00667** (0.00209)	0.00883*** (0.00239)	0.00780** (0.00239)	0.00519+ (0.00289)
At least one female × Num coauthors at firms		0.00269 (0.00248)				
% female			-0.0196** (0.00650)	-0.00258 (0.00799)		
% female × Num coauthors at firms				-0.00852* (0.00429)		
Mixed-gender teams with male last author					-0.00883+ (0.00516)	-0.0182* (0.00744)
Mixed-gender teams with female last author					-0.0247*** (0.00710)	-0.0395*** (0.0116)
All Female					-0.00144 (0.00614)	0.00493 (0.00754)
Mixed-gender teams with male last author × Num coauthors at firms						0.00415 (0.00304)
Mixed-gender teams with female last author × Num coauthors at firms						0.00646 (0.00485)
All Female × Num coauthors at firms						-0.00727 (0.00462)
Observations	22698	22698	22698	22698	19790	19790
Twin FE	y	y	y	y	y	y

Table A4: Robustness: Alternative percentage of known authors

Panel A: Percentage of 50%

	Commercialization		
	(1)	(2)	(3)
At least one female=1	-0.0128**		
	(0.00481)		
% female		-0.0245**	
		(0.00749)	
Mixed-gender teams with male last author			-0.0104+
			(0.00573)
Mixed-gender teams with female last author			-0.0277***
			(0.00769)
All Female			-0.00276
			(0.00734)
Constant	-0.0797***	-0.0738***	-0.0921***
	(0.0165)	(0.0166)	(0.0190)
Observations	20376	20376	18108
twin FE	y	y	y

Percentage of 75%

	Commercialization		
	(1)	(2)	(3)
At least one female=1	-0.0188*		
	(0.00763)		
% female		-0.0391**	
		(0.0124)	
Mixed-gender teams with male last author			-0.0150+
			(0.00848)
Mixed-gender teams with female last author			-0.0379***
			(0.0110)
All Female			-0.00554
			(0.0135)
Constant	-0.125***	-0.116***	-0.136***
	(0.0253)	(0.0256)	(0.0277)
Observations	13280	13280	12440
twin FE	y	y	y

Percentage of 100%

	Commercialization		
	(1)	(2)	(3)
At least one female=1	-0.0391**		
	(0.0143)		
% female		-0.0813***	
		(0.0234)	
Mixed-gender teams with male last author			-0.0294+
			(0.0156)
Mixed-gender teams with female last author			-0.0797***
			(0.0211)
All Female			-0.0141
			(0.0222)
Constant	-0.283***	-0.266***	-0.278***
	(0.0510)	(0.0514)	(0.0517)
Observations	6474	6474	6474
twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A5: Robustness: Authors' prior commercialization experience

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0105*					
	(0.00419)					
Auth. prev. comm.?=1	0.0515***	0.0511***	0.0592***	0.0491***	0.0583***	0.0528***
	(0.00414)	(0.00414)	(0.00475)	(0.00654)	(0.00556)	(0.00703)
% female		-0.0165*			-0.00588	
		(0.00650)			(0.00476)	
Mixed-gender teams with male last author			-0.00945+			-0.0179***
			(0.00515)			(0.00404)
Mixed-gender teams with female last author			-0.0226**			-0.0202***
			(0.00707)			(0.00579)
All Female			0.00187			0.00494
			(0.00608)			(0.00597)
At least one female=1				-0.0121***		
				(0.00314)		
At least one female=1 × Auth. prev. comm.?=1				0.00358		
				(0.00769)		
Auth. prev. comm.?=1 × % female					-0.0274+	
					(0.0142)	
Mixed-gender teams with male last author × Auth. prev. comm.?=1						0.0159+
						(0.00900)
Mixed-gender teams with female last author × Auth. prev. comm.?=1						-0.00339
						(0.0137)
All Female × Auth. prev. comm.?=1						-0.0289+
						(0.0159)
Constant	-0.0187	-0.0153	-0.0264	-0.0176	-0.0190	-0.0248
	(0.0154)	(0.0154)	(0.0181)	(0.0154)	(0.0155)	(0.0181)
Observations	22724	22724	19816	22724	22724	19816
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A6: Robustness: Institutions' historic experience with commercialization

	Commercialization					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0107** (0.00413)					
Inst. prev. comm.?=1	0.104*** (0.00596)	0.104*** (0.00596)	0.118*** (0.00674)	0.0825*** (0.00972)	0.109*** (0.00813)	0.0876*** (0.0103)
% female		-0.0171** (0.00643)			-0.0132* (0.00588)	
Mixed-gender teams with male last author			-0.00906+ (0.00507)			-0.0211*** (0.00489)
Mixed-gender teams with female last author			-0.0250*** (0.00695)			-0.0271*** (0.00653)
All Female			-0.000521 (0.00640)			0.00619 (0.00625)
At least one female=1				-0.0168*** (0.00391)		
At least one female=1 × Inst. prev. comm.?=1				0.0304* (0.0118)		
Inst. prev. comm.?=1 × % female					-0.0215 (0.0215)	
Mixed-gender teams with male last author × Inst. prev. comm.?=1						0.0559*** (0.0137)
Mixed-gender teams with female last author × Inst. prev. comm.?=1						0.0178 (0.0205)
All Female × Inst. prev. comm.?=1						-0.0721** (0.0225)
Constant	-0.0378** (0.0145)	-0.0339* (0.0146)	-0.0471** (0.0171)	-0.0339* (0.0146)	-0.0353* (0.0147)	-0.0447** (0.0171)
Observations	22724	22724	19816	22724	22724	19816
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A7: Counterfactual: “boasting” and patent-paper pairs from University patents only

	University patents					
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female=1	-0.00948*	-0.00765+				
	(0.00434)	(0.00454)				
paperhasnovelwords=1	-0.00139	0.00676	-0.00144	0.00129	-0.00185	0.00173
	(0.00603)	(0.00919)	(0.00603)	(0.00820)	(0.00657)	(0.00964)
At least one female=1 × paperhasnovelwords=1		-0.0122				
		(0.0114)				
% female			-0.0187**	-0.0172*		
			(0.00675)	(0.00711)		
paperhasnovelwords=1 × % female				-0.0104		
				(0.0190)		
Mixed-gender teams with male last author					-0.00881+	-0.00768
					(0.00514)	(0.00538)
Mixed-gender teams with female last author					-0.0154*	-0.0156*
					(0.00652)	(0.00686)
All Female					-0.0105	-0.0102
					(0.00873)	(0.00913)
Mixed-gender teams with male last author × paperhasnovelwords=1						-0.00735
						(0.0134)
Mixed-gender teams with female last author × paperhasnovelwords=1						0.000741
						(0.0175)
All Female × paperhasnovelwords=1						-0.00228
						(0.0253)
Constant	-0.0451**	-0.0463***	-0.0403**	-0.0407**	-0.0580***	-0.0585***
	(0.0139)	(0.0139)	(0.0140)	(0.0140)	(0.0160)	(0.0160)
Observations	22698	22698	22698	22698	19790	19790
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A8: Female scientists and commercialization “mode”

Panel A: OLS

	Cooperative Commercialization	with Existing Firms		*Self*-Commercialization	via Startups	
	(1)	(2)	(3)	(4)	(5)	(6)
At least one female	-0.0111**			-0.00134		
	(0.00420)			(0.00218)		
% female		-0.0208**			-0.00296	
		(0.00653)			(0.00295)	
Mixed-gender teams with male last author			-0.00866+			-0.00125
			(0.00513)			(0.00262)
Mixed-gender teams with female last author			-0.0281***			0.00312
			(0.00699)			(0.00290)
All Female			-0.00364			0.00240
			(0.00613)			(0.00148)
Observations	22938	22938	19995	22938	22938	19995
twin FE	y	y	y	y	y	y

Panel B: Conditional Logit

	Cooperative Commercialization	with Existing Firms		*Self*-Commercialization	via Startups	
	(1)	(2)	(3)	(4)	(5)	(6)
main						
At least one female	-0.299*			-0.156		
	(0.133)			(0.269)		
% female		-0.769**			-0.523	
		(0.251)			(0.583)	
Mixed-gender teams with male last author			-0.197			-0.0511
			(0.139)			(0.278)
Mixed-gender teams with female last author			-0.680***			0.535
			(0.192)			(0.475)
All Female			-2.085*			0
			(0.988)			(.)
Observations	1718	1718	1692	454	454	393
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors’ institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A9: PPP and citation level (full)

	(1)	(2)	(3)	(4)	(5)	(6)
Mixed-gender teams with male last author	-0.0000883***	-0.000169***	-0.000451***	-0.000601***	-0.000562***	-0.00116***
	(0.0000208)	(0.0000495)	(0.0000474)	(0.0000713)	(0.0000879)	(0.000134)
Mixed-gender teams with female last author	-0.000259***	-0.000492***	-0.000853***	-0.00138***	-0.00193***	-0.00536***
	(0.0000205)	(0.0000503)	(0.0000505)	(0.0000779)	(0.0000980)	(0.000154)
All Female	-0.0000757***	-0.000285***	-0.000571***	-0.00116***	-0.00154***	-0.00197***
	(0.00000563)	(0.0000219)	(0.0000296)	(0.0000537)	(0.0000780)	(0.000134)
Constant	-0.000225***	-0.000375***	-0.000737***	-0.00122***	-0.00241***	-0.0262***
	(0.0000166)	(0.0000543)	(0.0000638)	(0.000117)	(0.000168)	(0.000329)
Observations	22694988	6289887	9579516	6378900	6470933	6669313
Average number of citations (logarithm)	0	.6931	1.3177	1.9765	2.5680	3.5831
Percentile of citation distribution	<50 pct	50-60 pct	60-70 pct	70-80 pct	80-90 pct	>90 pct
twin FE	y	y	y	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor. + $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

Table A10: Exclude transitive PPP

	Citation		
	(1)	(2)	(3)
At least one female	-0.0112** (0.00403)		
% female		-0.0182** (0.00629)	
Mixed-gender teams with male last author			-0.00974+ (0.00497)
Mixed-gender teams with female last author			-0.0241*** (0.00682)
All Female			-0.000681 (0.00584)
Observations	22724	22724	19816
Twin FE	y	y	y

Notes: all models include controls for number of authors, prestige of authors, prestige of authors' institutions, and journal impact factor * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.