Gender Identity, Peer Effects, and Research Focus: Evidence from PhD Dissertations *

Paul W. Dai † Hongyuan Xia ‡

This Version (Preliminary): November 24, 2023

Abstract

Using a universe record of PhD dissertations in the US, we find a substantial amount of missing female PhD and female-focus research from 1985 to 2015. Our decomposition exercise quantifies the driving forces of this phenomenon. Besides, we find a close connection between PhD gender identity and their research focus: females are more likely to do female-focus research (FFR), while males are less likely. We further causally identify the peer-effects of producing female-focus research by exploiting year-to-year variation of the share of females within a PhD cohort: a higher female ratio encourages females but discourages male PhDs doing FFR, which is in favor of a competition effect in addition to knowledge diffusion in FFR production. These effects last for at least 5 years after graduation. Further investigation indicates that the homophily and coauthor networks explain the differential effects on female and male PhDs.

*We are grateful for the encouragement and endless support from Chris Forman and Matt Marx. We thank the University of Chicago for data access from ProQuest and a dedicated technical support team from Cornell Center for Social Science (CCSS).

†paulwdai@mit.edu

‡hx276@cornell.edu

1 Introduction

Women face tremendous barriers to seeking equal opportunities, including attaining a profession in academia and conducting scientific research (Ding et al., 2006; Azoulay et al., 2007; Lerchenmueller and Sorenson, 2018). Despite the progress in recent decades on female empowerment and affirmative action (Goldin, 2014; Hyland et al., 2020), women are still considered as an underrepresented group in academia, due to discrimination (Wu, 2018), career-family trade-off (Kim and Moser, 2021), and lack of recognition (Koffi, 2021), etc. Meanwhile, this gender gap in academic participation has far-reaching consequences:

First, underrepresented groups, like female researchers, produce more novel ideas (Hofstra et al., 2020). In this way, this gender gap results in the *Missing Marie Curies* and represents a misallocation of talents, impeding long-run economic growth (Hsieh et al., 2019), especially in an era when ideas are getting harder to find (Bloom et al., 2020). Second, numerous research highlights the close connection between the identity of researchers and their research direction [1]. For example, if female researchers are more likely to do research that can further benefit other females, this gender gap reshapes the landscape of our knowledge and may have a negative implication on female welfare [2].

Despite the first order effect of gender gap on science, little attention has been paid on the consequences of the academic environment caused by the gender gap. A gender-diverse environment is an incubator to generate new ideas. Lowering the barrier for females doing scientific research affects the idea production in the PhD community through competition and knowledge spillover. In this paper, we provide *causal* evidence on how the diverse environment affects researchers' choice of research topics, in particular, the topics focusing on the female.

We try to address this question empirically by leveraging a universe record on PhD thesis in the United States over the past three decades. Our focal point is PhDs because they are among the most innovative group of people in the society (Akcigit et al., 2022), continuously pushing the knowledge frontier. Meanwhile, universities are very important sources of innovation, as suggested by Alon et al. (2022). We start our analysis by evaluating the phenomenon of *Missing Marie Curies*, motivated by the growing number of female PhD in the last 30 years. We follow up a similar exercise accounting for the missing number of female-focus research in this time window. We find the increasing number of female-focus research in the past 30 years is not only driven by the shrinking gender gap, but also by the female's inclination to do female-focus research. The decomposition exercise highlights the importance of second order factors, such as the academic environment, caused by the under-representation of females in academia. To causally identify the effect of a diverse academic environment on researchers' to inclination to do female-focus research, we narrow our horizon into one aspect of the academic environment, the *peer effects* within each PhD cohort. Our key question is to evaluate the impact of female peers on the researchers' probability of producing female-focus

[1] Koning et al. (2020) and Koning et al. (2021) find patents with all-female inventor teams are 35% more likely than all-male teams to focus on women's health. Nielsen et al. (2017) show a robust positive correlation between women's authorship and the likelihood of conducting gender- and sex-related research.

[2] This channel can be persistent, amplified by the homophily in scientific research that males are more likely to find coauthorship from males Kwiek and Roszka (2021).

research.

Identifying the peer effects of females is challenging due to several reasons: First, peers are not randomly assigned, and those who care about female-related topics are more likely to group together, which yields selection bias. Second, the research topics and directions of PhD students are heavily influenced by their advisors. It is difficult to address this concern because, to our best knowledge, there is no existing database that contains information about the research focus and publication of both advisors and PhD students. We contribute to filling this gap in this paper.

To deal with the above challenges, our paper makes several attempts: First, we link the ProQuest data with Microsoft Academic Graph (MAG) data by using the name, affiliation, year, and research fields. Microsoft Academic Graph (MAG) contains information of more than 200 million research papers. This linked data not only allows us to measure the research focus of the advisor, but also enables us to get a comprehensive picture of both short-term and long-term effects of cohort gender composition. Second, By taking advantage of the PhD thesis database, ProQuest, we calculate how many female PhD students graduate in each department each year. Considering the female ratio is different and somewhat exogenous across different cohorts in the same department, we can identify the causal relationship between cohort gender composition and research focus. This empirical strategy, the exogenous cohort female ratio, is introduced by Hoxby (2000) and has been widely used in the labor literature (Bostwick and Weinberg, 2022; Mouganie and Wang, 2020). More specifically, it leverages the fact that there is uncertainty, both on the part of admissions and on the part of potential doctoral students, as to the gender composition of each incoming cohort. There is possibility that a doctoral program's admissions committee might target a specific gender mix and an incoming student might know the average gender mix of past cohorts in a program, neither party can fully anticipate the realized gender composition of an incoming cohort of students. Moreover, We use Monte Carlo simulations to demonstrate that the observed within-cohort variation in the female ratio is consistent with variation generated from a random process.

Utilizing the empirical strategy above, we find that (1). men in a cohort with no female peers are 1.72 percentage points less likely than their female peers to do female-focus research in their PhD thesis; (2). for each additional 10 percentage points of female students in a cohort, women are 0.61 percentage points more likely to do female-focus research in the dissertation; (3). the differentiation effect on men is -9.4 percentage points and is statistically significant, which indicates the effect of additional 10 percentage points of female students in a cohort for a male is 0.32 percentage points less likely to female-focus research in PhD thesis. These results are consistent in several robustness checks.

By linking the ProQuest PhD data to MAG data, we extend our analysis to a more comprehensive scope. We find that PhDs with more female peers are more likely to do female-focus research within at least 5 years after graduation. There are 0.17 percentage points increase in the likelihood of doing female focus research, 0.16 papers increase in the female-focus research and 0.05 percentage points increase in the ratio of female-focus research, for each 10 percentage points increase in cohort female ratio. Meanwhile, we find cohort gender composition has no effect on pre-PhD publication, which

further supports our empirical strategy. Besides, we find females' higher inclination to do female-focus research is mainly driven by female-dominant teams.

Further investigation indicates that the cohort gender composition may affect the choice of research topics through coauthor networks. Our results show that PhD students in a cohort with more female peers are more likely to collaborate with women and have more female coauthors. In addition to the increasing width of coauthorship with women, we also document the collaboration with female researchers is deeper, measured by the length of collaboration and the number of coauthored papers, for scientists who receive PhD education with more female peers.

Our study makes several contributions. First, this paper furthers our understanding of the relationship between diversity and innovation. A series of studies suggest that female scientists, inventors, and entrepreneurs are more likely to produce ideas, inventions, and companies that benefit women (Koning et al., 2020; Nielsen et al., 2017). Substantial research has also shown that gender and racial diversity of team members are linked to higher creativity and higher team performance (Hofstra et al., 2020). However, these studies focus mainly on the diversity of identity without causally identifying the effect of a diverse environment. Recent papers reveal that a toxic workplace environment may be contributing to female under-representation (Wu, 2020; Dupas et al., 2021). The research that is most closely related to our paper is Truffa and Wong (2022), which takes advantage of the universities' transition to coeducation to identify the impacts of the diversity of the academic environment on institution-level research focus. Our paper distinguishes itself from their research since we focus on another aspect of the diversity of the academic environment, the cohort gender composition, and implement detailed individual-level analysis instead of aggregate university-level analysis.

Second, our research contributes to the rising literature investigating the gender peer effects and under-representation of females in academic settings. A number of studies have shown that girls and boys benefit academically from an increase in the number of female peers (Hoxby, 2000; Lavy and Schlosser, 2011; Mouganie and Wang, 2020). One example is Bostwick and Weinberg (2022) focusing on the effects of peer gender composition in STEM doctoral programs on persistence and degree completion. Instead of concentrating on degree completion, our study investigates the short-term and long-term peer effects on the research focus of PhDs. Our results are consistent with existing literature that females benefit from having more female peers. However, we also find that males are less likely to be affected by the increasing number of female peers, and sometimes, more female PhDs may even crowd out male PhDs in doing female-focus research.

Third, this paper is also related to a large literature on homophily, which has been thought as a foundation in the social network. Sociology and economics literature has documented the "birds of a feather" phenomenon (McPherson et al., 2001; Agrawal et al., 2008) which reveals the tendency of individuals to associate, interact, and bond with others who possess similar characteristics. Research on the performance consequences of homophily for individuals has presented conflicting results (Ertug et al., 2022). One example is Freeman and Huang (2015) who finds researchers of similar ethnicity coauthor together more frequently, and they then associated this homophily with publica-

tion in lower-impact journals and with fewer citations. Our paper suggests for under-represented groups like the female, homophily helps female PhDs to perform better in doing female-focus research. Moreover, our results empirically identify that the environment factor, cohort gender ratio, may shape the preference and consequently promote homophily.

The remainder of this paper is organized as follows: Section 2 describes the data used in this paper and how the variables are constructed. Section 3 implements the decomposition exercise. Section 4 reviews the empirical strategy and reports estimation results. We implement the robustness check in section 5. Section 6 discusses the mechanism and section 7 concludes.

2 Data and Measurement

In this section, we describe the data and how we construct the main variables.

2.1 PhD data

For data on PhD students' thesis, research fields, graduation universities, and graduation year, we use ProQuest Dissertations & Theses Global Data [3], which is the official offsite dissertation repository for the U.S. Library of Congress. It includes records of nearly all US PhD theses metadata from 1985 to 2015, including student names, advisors, committee members, institutions, thesis titles, abstracts, research subjects, and degree date. These structural and semantic footprints enable us to scrutinize students' research focus at the very onset of their scholarly careers. In this paper, we analyze 771,011 dissertations from Ph.D. recipients between 1985 and 2015 in universities listed in the Association of American Universities.

2.2 Publications data

To further understand the long-term effects of cohort gender composition and the dynamics of the effects, we use the Microsoft Academic Graph (MAG) database (Sinha et al. (2015)), a large database with information on over 207 million papers published in journals and conferences. By using the information of PhD name, advisor name, affiliation, degree year, and research field from ProQuest, we link the publication information to our PhD data.

The data matching procedure between ProQuest and MAG is as below: First, we screen out all the authors in MAG with affiliations that are the universities listed in the Association of American Universities. Second, we match the last names and calculate the first name similarity within each affiliation, the key advantages of our approach are that (1). it is more flexible and allows the linkage of different forms of names, such as nicknames, abbreviations, and so on, compared with direct name matching. (2). it should be more accurate by implementing a stricter role on the last name

[3] https://www.proquest.com/

compared with computing similarity for all the names. Third, for the matched advisor to MAG author, we drop the observations with research fields are quite different or the first publication is more than 5 years after graduation, for the matched PhD student to MAG author, we drop observations with more than 5 publications before PhD, the first publication being published more than 5 years after graduation, or research fields are not quite similar. We use coauthors of each matched PhD advisor and PhD student to further supplement the matching data sets. We finally get 387,968 PhDs (about 50.3%)matched to MAG data, and 540,141 advisors (about 83.1%, since only 650,170 PhDs have advisor name information) matched to MAG data.

2.3 Main variables

In this section, we describe our identification of the gender of PhD students, assigning the field of research, classification of female-focus research, and how we group PhD into different cohorts.

Since the gender of PhD recipients is not provided in the ProQuest database, we classify the gender of the researcher by using PhD recipients' first names and Genderize.io [4], an API that has been employed by academia to identify gender and report its possibility (Topaz and Sen, 2016; Huang et al., 2020). For example, the first name "Paul" has a 0.99 probability to be a male, and the first name "Hongyuan" has a 0.86 probability to be a male. There are 769,712 PhDs with valid first names. In the process, 741,033 PhD recipients (about 96%) can be assigned to either female or male, and 650,300 PhD recipients (about 84%) can be assigned to one gender with a probability of no less than 0.9.

To assign the field of research to each PhD recipient, we use the class terms/ subject terms in the ProQuest database. There are 432 subject categories in total, falling into 21 large disciplines. For example, economics (subject number: 0501) and labor economics (subject number: 0510) are subjects within the social science disciplines. Since subject terms are very specific about the research topics, they may not reflect the actual program of the PhD recipients. To deal with this issue, we re-classify the subject terms to the first four digits of the CIP code, the Classification of Instructional Programs [5]. CIP provides a taxonomic scheme that supports the accurate tracking and reporting of fields of study and program completion activity. After classification, there are 182 fields. Based on that, we define a cohort as a group of Ph.D. recipients who graduate in the same field and same year within a university, for example, the PhD recipients who graduated from Cornell in the research field of economics in 2015 are considered in the same cohort.

Our main outcome variable is whether the thesis is female-focus research. We use a similar keyword approach adopted by Truffa and Wong (2022) to define whether a paper is female-focus or not. By using Datamuse API [6], a word-finding query search engine that is based on Google Books Ngrams data and other corpus-based datasets. We selected the top 20 most related words as "gender", "female", "women", and "sex", and we exclude male-related words because historically men are considered "standard" in research. The keyword list is provided in A.1. The female-focus research

[4] https://genderize.io/
[5] https://nces.ed.gov/ipeds/cipcode/browse.aspx?y=55
[6] https://www.datamuse.com/api/

is defined as one if at least one of the keywords, such as female, appears in either the title or the abstract. There are two advantages of this approach: (1) titles and abstracts are available for nearly all Ph.D. recipients; (2) this approach can be applied broadly to all fields instead of only one or two fields in some existing work (Koning et al., 2020). Under such definition, a research paper is considered female-focus if the research topic is about women or it highlights analysis pertaining to gender or women. There is 9 percent of dissertations can be defined as female-focus. Here we provide two examples of female-focus dissertations with the corresponding titles and abstracts in Appendix A.1.

(a) *Example 1: Welfare waivers and women's non-marital fertility decisions*

(b) *Example 2: Development of novel antiestrogens for the treatment of tamoxifen-resistant breast cancer*

For the PhD students and advisors who are matched to MAG data, they publish 14,841,538 and 85,568,244 papers, respectively, in total. we use the research fields assigned by MAG to define whether the paper is female-focus research. MAG classifies papers into 771,038 fields and we use the keyword approach mentioned above to define whether a paper is female-focus. There is about 1.8 percent of all papers are classified as female-focus, this ratio is much smaller than that using a keyword approach in titles and abstracts, we may take the fields defined by MAG as a stricter version of female-focus research. Based on that, we create three categories of variables: the first category is a dummy variable indicating whether the author does female focus research, the second category is the ratio of female-focus research, and the third category is the number of female-focus research papers. To make the data comparable across PhDs graduating in different years, we construct four variables in different time windows, that is, more than 5 years before graduation, within 5 years before graduation, within 5 years after graduation, more than 5 years but less than 10 years after graduation.

3 The Missing Marie Curies and Female-focus Research

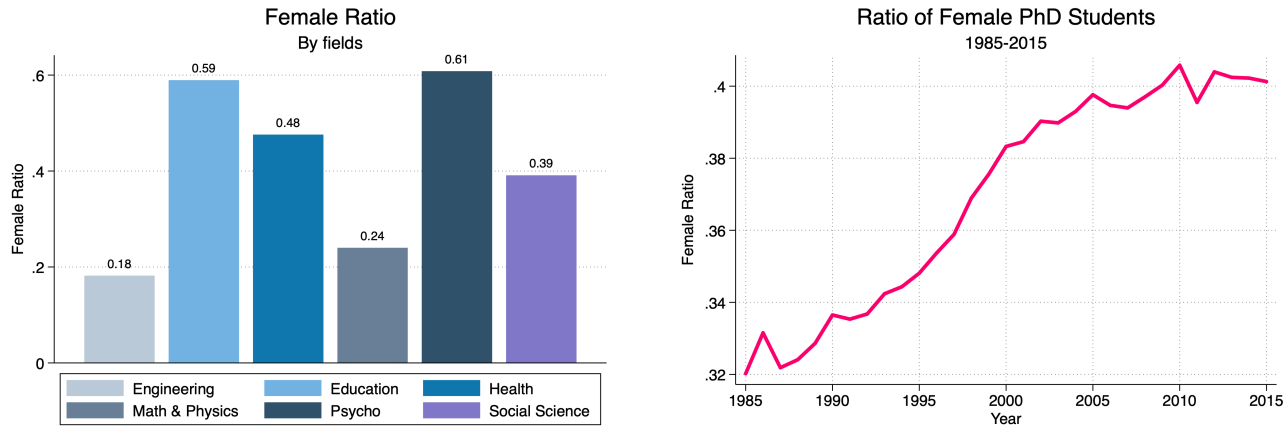
In this section, we document the fact of missing Marie Curies and female-focus research in the past three decades and investigate the reason behind using a simple decomposition exercise. We define "missing Marie Curies" (or "missing female-focus research") as the difference between the number of female PhDs (or number of female-focus research papers) in the actual year and that in 2015.

3.1 Summary Statistics

In this section, we start by presenting a brief overview of female PhDs and female-focus research.

Figure 1 demonstrates the female ratio in PhD fields and over time. The female ratio is calculated as the share of female PhD for a specific field or year. If there were no different barriers for females to choose PhD as their male counterparts, we would expect that the female ratio is roughly the same as the share of females in the total population. In the left panel, we divide the fields into the following six categories: engineering, education, health and biology, math and physics, psychology,

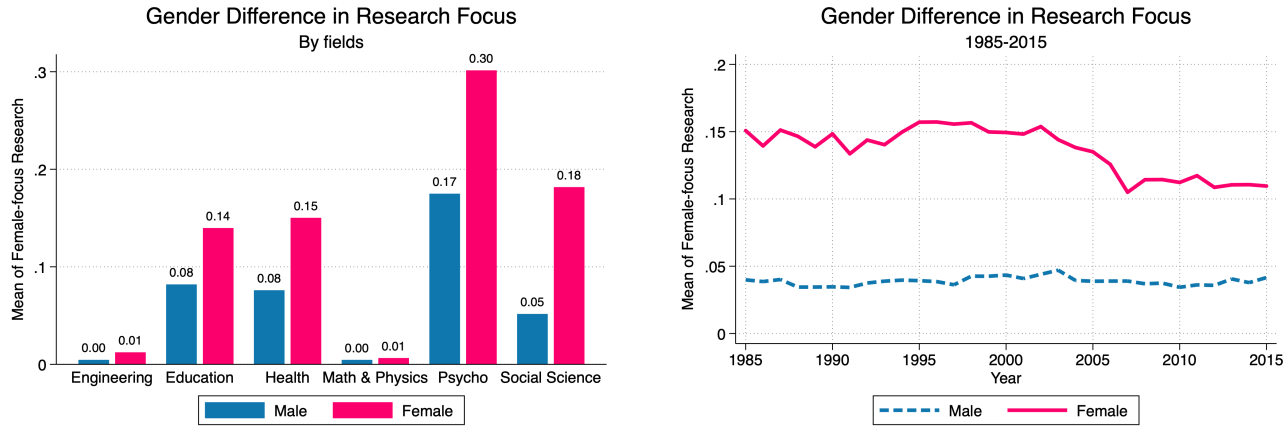
Figure 1: Ph.D. Female Ratio By Fields and Over Time



Note: This figure shows the female ratio, measured by the ratio of the total number of female PhDs to the total number of PhDs, by fields and over time in the past three decades. We categories fields into 6 subjects: Engineering, Education, Health & Biology, Math & Physics, Psychology, and Social Science.

and social science. There is huge heterogeneity in the female ratio across fields. In particular, there are significantly more females in education and psychology. About 48% of PhD recipients are female in education and significantly fewer females in social science (39%), math and physics (24%), and engineering (18%). The right panel depicts a rising share of females in the last three decades, from 32% in 1985 to 40% in 2015.

Figure 2: Gender Difference in Research Focus By Fields and Over Time

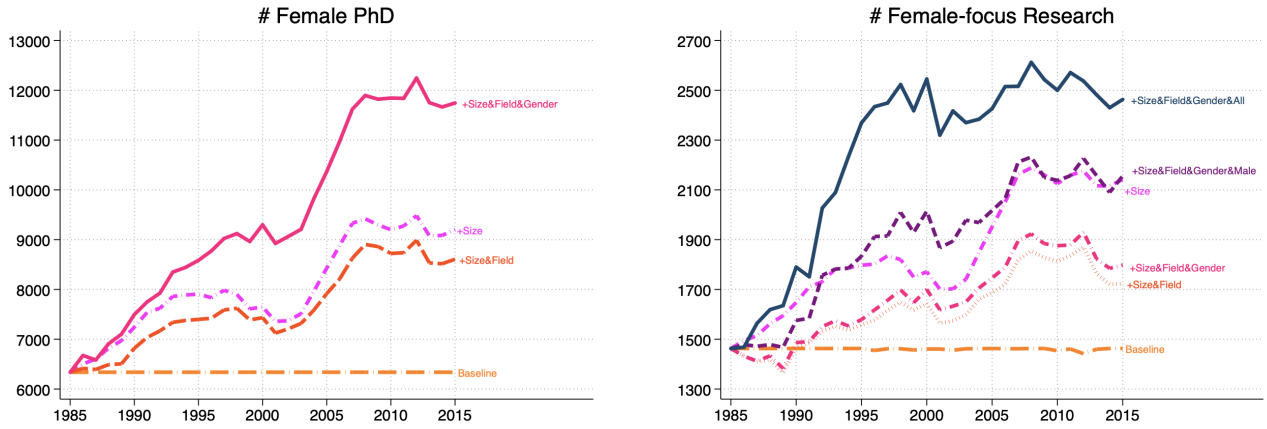


Note: This figure shows the gender difference in research-focus, measured by the proportion of female-focus research by genders, by fields and over time.

Next, we document the gender difference in research focus by fields and over time in Figure 2. More specifically, we ask the proportion of female-focus research by the genders of PhD recipients. There are several observations from Figure 2: first of all, female-focus research concentrates on psychology, social science, health & biology, and education. There is almost no gendered research in engineering and math and physics, which meets our common sense; secondly, there are huge gender differences across fields. For example, 30% and 18% of female PhDs write female-focus dissertations in psychology and social science, respectively, while such ratios are 17% and 5% for their

male counterparts.

Figure 3: Decomposition of the Rise of Female PhD and Female-focus Research



Note: This figure presents a decomposition of the growing number of female PhDs and female-focus research, following the accounting identity in equation 1 and 2. Panel A shows the result for the total number of female PhD. The baseline shows the level in the starting year 1985. Keeping all other effects at the 1985 level, we gradually add the size effect, field-composition effect, and gender composition effect. Panel B shows the result for the total number of female-focus research. We add the size effect, field-composition effect, gender composition effect, male-inclination effect, and female-inclination effect once at a time to compute corresponding counterfactual total number of female PhD or Female-focus research.

Lastly, this gender difference in research focus is quite persistent over time, as indicated by the right panel of Figure 2. Roughly 4% of males conduct research focusing on the female annually from 1985 to 2015. Such ratio is about 15% from before 2000, gradually declines to 10% in 2007 and stays at the same level afterward. [7]

3.2 Decomposition Exercise

In this section, we document the substantial number of missing female PhD and analyze the driving forces behind this phenomenon through a simple decomposition exercise.

3.2.1 Methodology

We decompose the number of female PhD into three components. Firstly, the total number of PhD has changed over time, which is the *size effect*. This channel can be connected to the education policy that affects occupational choice as a PhD or production worker (Akcigit et al., 2022). Secondly, in each year, the relative shares of PhD in different fields also vary across years, which we call it *field-composition effect*. This effect aligns with the allocation of innovation and R&D resources across different fields (Acemoglu et al., 2016; Liu and Ma, 2022). Lastly, within a field, the gender ratio fluctuates, which is the *gender-composition effect*, closely connecting to Hsieh et al. (2019) about the

[7] Our results can be interpreted as a complement to Koning et al. (2021). Similar to our finding in Figure 2, they document a fact in commercial patenting that patents from all-female inventor teams are 35% more likely to focus on women's health than all-male teams. They further point out that the missing female-focused inventions since the 1970s might be driven by the inventor gender gap, which connects our observations in Figure 1 and Figure 2.

allocation of talents from a gendered perspective, which results in productivity differences at the aggregate level Lee (2016). Formally, we utilize the following accounting identity in Equation 1:

$$\#\text{ Female PhD}_t = \sum_s \underbrace{\frac{\#\text{ Female PhD}_{s,t}}{\#\text{ PhD}_{s,t}}}_{\text{gender composition}} \times \underbrace{\frac{\#\text{ PhD}_{s,t}}{\#\text{ PhD}_t}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_t}_{\text{size}}. \tag{1}$$

In practice, our field category is fine enough that there is no PhD in some field×year bin. Some zeros may appear in the denominator in Equation 1 (# $\text{PhD}_{s,t}$ or # PhD_t). It is particularly problematic if this happens to the reference year, which implies we do not have a valid gender composition as a reference for some of these fields. To mitigate this problem, we choose our base year as 2015 to better handle the emergence of new fields in the past three decades. However, there are still a small number of fields with no PhD in 2015 but with some PhDs in other years. We take care of this issue by adding an additional effect to capture this *extensive margin*. Another benefit of choosing 2015 as a reference is that we can interpret the gap between the actual change of female PhDs relative to the number of female PhD in 2015 as the amount of missing female PhDs.

For the total number of female-focus research, we use a very similar accounting identity as Equation 1, we write the total amount of female-related research as Equation 2:

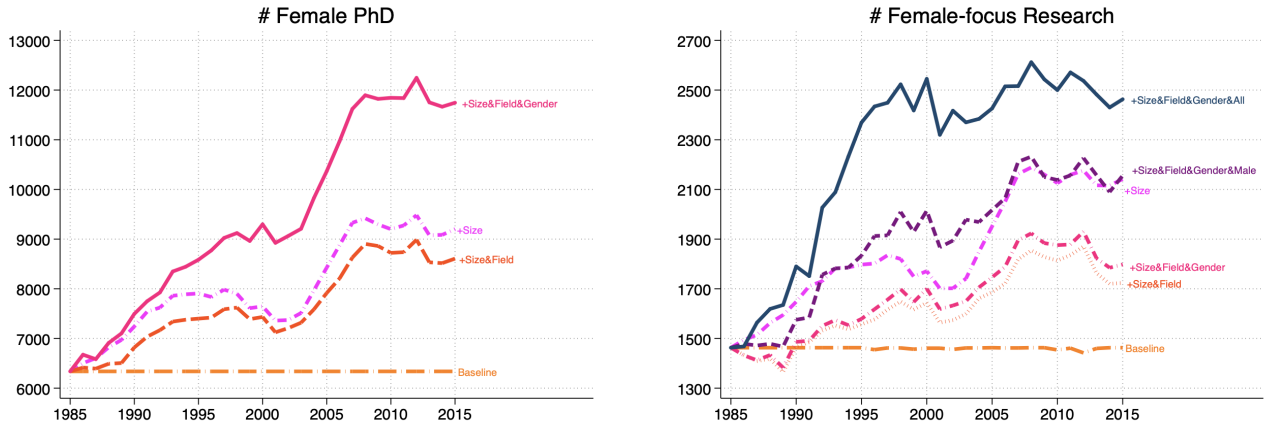
$$\begin{aligned}\#\text{ Female-focus Research}_t = \sum_s \sum_g \underbrace{\frac{\#\text{ Female-Focus}_{g,s,t}}{\#\text{ PhD}_{g,s,t}}}_{\text{inclination}} \times \underbrace{\frac{\#\text{ PhD}_{g,s,t}}{\#\text{ PhD}_{s,t}}}_{\text{gender composition}} \\ \times \underbrace{\frac{\#\text{ PhD}_{s,t}}{\#\text{ PhD}_t}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_t}_{\text{size}}.\end{aligned} \tag{2}$$

There are two main changes worth mentioning. Firstly, noticing that both females and males do female-focus research, we further decompose the total number of female-focus research in gender×field subgroups. Secondly, females and males have different *inclinations* (or propensities) of doing female-focus research. These inclination effects are measured as the ratio of the number of female-focus research to the total number of PhD research by each gender, which is exactly the additional term in Equation 2 relative to Equation 1. By the nature of our PhD dissertation dataset, the total number of PhD in a specific subgroup equals the total number of dissertations (or research papers) in that group. We call this additional term as *inclination effect*. In the decomposition exercise, we first add male-inclination and then the inclination of both genders to highlight the magnitude of this channel.

3.2.2 Decomposition Results

For both decomposition exercises, we start with the scenario that all channels are shut down and then add one additional channel once a time and show the counterfactual number in Figure 4. In

Figure 4: Decomposition of the Rise of Female PhD and Female-focus Research



Note: This figure presents a decomposition of the growing number of female PhDs and female-focus research, following the accounting identity in equation 1 and 2. Panel A shows the result for the total number of female PhD. The baseline shows the level in the starting year 1985. Keeping all other effects at the 1985 level, we gradually add the size effect, field-composition effect, and gender composition effect. Panel B shows the result for the total number of female-focus research. We add the size effect, field-composition effect, gender composition effect, male-inclination effect, and female-inclination effect once at a time to compute corresponding counterfactual total number of female PhD or Female-focus research.

Appendix A.2, we elaborate on how we construct these counterfactual numbers when different effects are kicking in.

The data we observe in the real world corresponds to the line named 'Actual Change'. From 1985 to 2015, the number of female PhD recipients grows from 6,338 to 11,744 in 2015, which is equivalent to an 85% increase. And the total number of female-focus dissertations has changed from 1,463 to 2,463 in the past three decades.

To quantitatively evaluate the contribution of missing Curie in each channel, we can make follow a similar approach as we plot Figure 4 by sequentially adding the economic changes. The contribution of a specific channel can be recognized as the change of counterfactual number between two consecutive steps. The result of this decomposition is order-dependent because there could be effects due to the interactions between different channels, and the interaction effects would be ambiguous. To overcome this issue, we follow Wu (2021) by investigating two polar cases: one is adding the channel when all channels have already been turned on, with potential interactions with all other groups. The other one is to add this channel when all other channels are turned off, without interaction with any other channels. We take the average contribution of these two polar cases, regarding it as the contribution of missing female PhD or female-focus research due to a particular channel. The details for this construction are elaborated in Appendix A.3.

Table 1 shows the decomposition results. The figures in the table are in percentage points [8]. Each row does not necessarily sum up to 100% due to the potential interactions among channels. Besides the baseline result for a time window from 1985 to 2017, we also consider a subsample 1985-2007 for female PhD and 1985-1995 for female-focus research before these two series are stabilized.

[8]Negative numbers can be interpreted as a particular channel that *mitigates* the missing female PhD or female-focus research or stimulates the growth of these two.

Table 1: Decomposition of Missing Female Ph.D and Female-focus Research

Panel A: Female Ph.D				Panel B: Female-focus Research					
Time Window	Gender	Field	Size	Time Window	Gender	Field	Size	Male	Female
1985-2015	49.33	-12.05	67.60	1985-2015	25.41	-14.06	159.06	46.14	-81.97
1985-2007	48.62	-11.38	67.51	1985-1995	15.77	-10.33	88.79	33.81	-13.15

Note: This table shows the contribution to total number of missing female Ph.D and female-focus research through different channel, following the methodology in Section 3.2.2 and in Appendix A.3. The figures in the table are in percentage points.

First, gender composition and size effects explain about 49.33% and 77.60% missing female PhD, and 25.41% and 159.06% of missing female-focus research in the past three decades. This coincides with a more equal gender ratio and an enlarging size of PhD in the past three decades. Secondly, we find that the field composition decreases the number of missing female PhD and female-focus research by about 12% and 14%, respectively. The field with a higher female ratio and more female-dominated grows faster from 1985 to 2015. Lastly, we find that male inclination contributes to 46.14% of missing female-focus research, while female inclination decreases this missing by 81.97%, consistent with our finding about the gender difference in conducting female-focus research in Figure 2. Our results indicate the female's inclination to do female-focus research is a main backward force to mitigate the missing female-focus research.

4 Peer Effect and Female-focus Research

One of the key motivations for scientific research is peer effects. On one hand, peer effects stimulate collaboration and joint projects, or stimulate booster idea-flow and knowledge diffusion through interaction (Akcigit et al., 2018), which we call it *diffusion effect*. On the other hand, peer effects can also be interpreted as peer pressure: a competition among researchers on the allocation of research resources, such as funding, coauthors, laboratories and equipment, etc (Azoulay et al., 2010, 2019; Lerner and Malmendier, 2013), which we dub it *competition effect*.

The accounting identity for female-focus research, at best, captures the *within-gender* peer effect, through the female-inclination or male-inclination terms. Unfortunately, this decomposition may neglect the *across-gender* peer effect. In this section, we investigate the impact of peer effect on the production of FFR for male and female PhD, which speaks to within-gender and across-gender peer effects.

4.1 Intuition

We sketch a simple model to provide intuition how within-gender and across-gender peer effect affects the research direction for PhD with different gender identity, which can be investigated empirically.

An agent with gender g is endowed with a pairwise ability (z_g^*, z_g), where z_g^* is the research ability for FFR, and z_g for non-FFR. The research output depends on individual endowment in that research direction, captured by z_g or z_g^*, and the gender diversity in the department, measured by the female ratio s and male ratio $1-s$.

We denote the within gender and across gender peer effects as $\phi_=$ and $\phi_{\neq}$. Specifically, female ratio s affects FFR output for female (male) by $\phi_=(s)$ ($\phi_{\neq}(s)$) and male ratio affects non-FFR output for female (male) by $\phi_{\neq}(1-s)$ and $\phi_=(1-s)$. We consider two potential peer effects: $\phi_\cdot' > 0$ for the spillover effect, and $\phi_\cdot' < 0$ for competition effect.

We denote $\Gamma_m(s)$, a measure for the propensity of male PhD doing FFR, as

$$\Gamma_m(s) \equiv \frac{z_m^* \phi_{\neq}(s)}{z_m \phi_=(1-s)}, \tag{3}$$

and notice that when $\Gamma_m(s) > 1$, male PhD chooses to do FFR. Similarly, we can define similar measure for female PhD as

$$\Gamma_f(s) \equiv \frac{z_f^* \phi_=(s)}{z_f \phi_{\neq}(1-s)} \tag{4}$$

Our goal is to understand how the change of female ratio s affects research direction under potential within- and across-gender peer effects.

Lemma 1 (Diffusion Effect, Comparative Advantage and Research Focus)**.** *Under the case of* $\text{sgn}(\phi_=') = \text{sgn}(\phi_{\neq}')$*, for all* $g \in \{f, m\}$

1. *(Comparative Advantage) Higher* z_g^*/z_g *implies higher* $\Gamma_g(s)$
2. *(Spillover Effect) If* $\phi'(s) > 0$*, then* $\Gamma_g(s)' > 0$*;*
3. *(Competition Effect) If* $\phi'(s) < 0$*, then* $\Gamma_g(s)' < 0$*.*

The comparative advantage for doing FFR is captured by z_g^*/z_g. The higher z_g^*/z_g is, the higher the propensity of doing FFR, measured by $\Gamma_g(s)$, for $g \in \{f, m\}$.

Now, consider the case that both within- and across-gender peer effects are spillover effect, i.e., $\phi_=' > 0$ and $\phi_{\neq}' > 0$. For example, an increase of s will improve the the research outcome of FFR directly. Such increase in s implies a decrease of $1-s$ and thus doing non-FFR is less appealing. Taken together, we see all PhDs, regardless of their gender identity, have higher propensity of doing FFR. Similar analysis applies for competition effect. If we see in the data that higher female share results in a more (less) FFR for both female and male PhD, we can infer the diffusion channel is spillover (competition) effect.

If we see divergence in research focus, we can infer there should be different channels for within-gender diffusion or across-gender peer effect, that is the sign of $\phi_='$ and $\phi_{\neq}'$ are different. In Appendix A.4, we elaborate the imposed condition on $\phi_=$ and $\phi_{\neq}$ to speak to divergent research focus by gender. Intuitively, if an increase of female ratio creates competition pressure for male PhD and

encourages female PhD to do FFR through spillover effect. The decrease of male ratio implies smaller spillover effect. But if such change is much smaller than competition effect, we will see male PhD still switch to non-FFR. Following the similar logic, a decrease in male ratio implies less competition for female PhD doing non-FFR. But if the spillover effect is so strong, we can still witness a surge of FFR by female PhD.

Table 2: Within- and Across-gernder Peer Effects and Research Focus

	$\phi'_{=} > 0$ (Spillover)	$\phi'_{=} < 0$ (Competition)
$\phi'_{\neq} > 0$ (Spillover)	Convergence, More FFR for all PhD	Divergence, Unclear
$\phi'_{\neq} < 0$ (Competition)	Divergence, Unclear	Convergence, Less FFR for all PhD

Table 2 summarizes the model prediction. The change of research focus for female and male PhD following the change of female ratio help us test the mechanism of within- and across-gender peer effects, which is elaborated in the next section.

4.2 Empirical Strategy

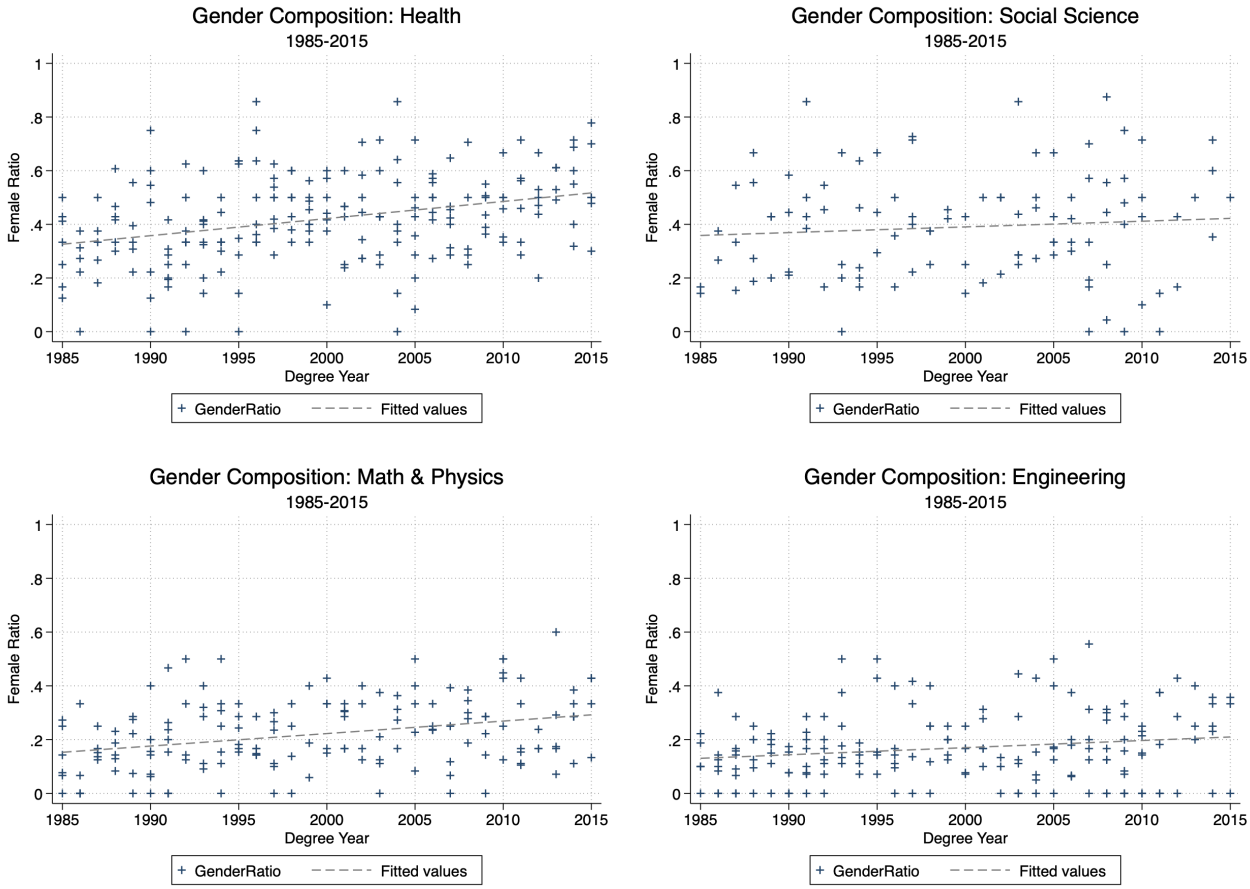
Our empirical strategy is essentially a difference-in-differences approach, comparing female and male PhD. recipients between cohorts with a high fraction of female students and cohorts with relatively low female students within a given doctoral program. The estimation equation is as below:

$$\begin{aligned} \text{Female-focus Research}_{i,u,s,t} = \beta_0 &+ \beta_1 \text{Male}_i + \beta_2 \text{Gender Composition}_{u,s,t} \\ &+ \beta_3 \text{Male}_i \times \text{Gender Composition}_{u,s,t} \\ &+ \lambda_{u,s} + \psi_t + \epsilon_{i,u,s,t}. \end{aligned} \tag{5}$$

where Female-focus Research$_{i,u,s,t} = 1$ indicates the student i's dissertation is female-focus and student i with field s is graduated from university u in year t. The primary variables of interest are Male$_i$, an indicator for student i's gender, and Gender Composition$_{u,s,t}$, an indicator for the gender composition, such as the ratio of female students, within the cohort, which is defined as all students in field s of university u graduating in year t. The variables $\lambda_{u,s}, \psi_t$ are university $\times$ field and year fixed effects, respectively.

The coefficient β_1 indicates the percentage point difference in probabilities of doing female-focus research for a male PhD. with no female peers versus women in a cohort with no female peers. The coefficient β_2 can be interpreted as the difference in probabilities of doing female-focus research for women in a cohort with all female peers (gender composition equals 1) versus women in cohorts with no female peers (gender composition equals 0). Finally, the coefficient β_3 is the differential effect on women versus men of being in a cohort with all female peers (gender composition equals 1). We estimate this specification using OLS and probit model with fixed effects. When estimated using the probit model, we report the average marginal effects corresponding to the descriptions above and are evaluated at the means of all covariates.

Figure 5: Trends of Cohort Gender Composition: Cornell University



Note: This figure presents the trend of cohort gender composition in four fields in one university.

Our identification strategy relies on the assumption that within a particular doctoral program (university-field pairs), year-to-year variation in cohort gender composition is quasi-random and not correlated with other unobservables influencing the research focus of the PhD. students within that cohort. One potential violation of this method could be some omitted variables that affect the gender ratio of a program and the research focus of the PhD. students simultaneously. For example, a new female faculty member may attract more female students to the program while encouraging students to do female-focus research at the same time. A telling signal of this type of endogeneity would be any evidence of time trends in the cohort gender composition within programs.

Figure 5 shows the trends of cohort female ratio from 1985 to 2015, as we can notice there are no clear upward or downward trends in cohort gender composition, especially in a relatively shorter time window, even though the fitted lines are slightly upward.

To verify the exogeneity of the gender ratio of a program across years, we perform two exercises: First, a first-order autoregressive model of gender composition with university times field and year fixed effects reveals no evidence of path dependence in female ratio, as shown in Table 11. Second, through a Monte Carlo simulation exercise (Lavy and Schlosser, 2011; Bostwick and Weinberg, 2022), we show that the observed within-program variation in gender composition in our data closely resembles the randomly generated variation from a binomial distribution. Specifically,

for each doctoral program, we randomly generate the gender of the students in each cohort using a binomial distribution function Binomial(n, p). The parameter n equals the actual cohort size and p equals the average proportion of females in that program across all years. Then, we compute the within-program simulated standard deviation of the proportion of females over all years. We repeat this process over 1,000 iterations to obtain an empirical confidence interval for the standard deviation for each program. Our observed within-program standard deviation lies within the empirical 90% confidence interval for 91% of PhD programs in our sample, which further supports our assumption that the within-program, year-to-year variation in cohort gender composition is in fact quasi-random.

Moreover, we add the gender and the research focus of PhD students' advisors as control variables to further address the omitted variable bias from the faculty level. As we know, the advisor matters a lot for a PhD's research topics, and the gender and research focus of the advisor is a main source of omitted variable bias. When a PhD comes to decide which program to choose, they should take the research of their potential advisors into consideration. In fact, in many natural science fields, admission is mainly a process of advisors and students choosing each other. As a result, controlling the gender and the research focus of PhD students' advisors can help us to circumvent much omitted variable bias caused by the selection process during admission.

4.3 Empirical Results: Short-Term Effects

Table 3 presents the estimation results using OLS with fixed effects. In the first three columns, we use the thesis abstracts to define the female-focus research, and in the last three columns, we use thesis titles. Column 1 compares the difference in doing female-focus research between women and men, suggesting that male Ph.D. recipients are 6.04 percentage points less likely to do female-focus research. This result is consistent with the literature that males pay less attention to female-related work (Koning et al., 2020). In column 2, we find that when the cohort female ratio increases 10 percentage points, the PhD recipients within that cohort are 0.69 percentage points more likely to do female-focus research. Considering female-focus research accounts for less than 10 percent of all research, this effect is equivalent to an 8 percent increase in female-focus research.

The results in column 3 are multifaceted: First, men in a cohort with no female peers are 1.7 percentage points less likely than their female peers to do female-focus research. Second, for each additional 10 percentage points of female students in a cohort, women are 0.61 percentage points more likely to do female-focus research. However, the differentiation effect on men is -9.4 percentage points and is statistically significant, which indicates the effect of additional 10 percentage points of female students in a cohort for a male is 0.32 percentage points less likely to do female-focus research. Column 4 to column 6 present similar patterns. In general, our results show higher female ratio has opposite effects on female and male PhD recipients.

The different effect on female and male PhDs' research focus indicates having more females in a cohort stimulates the female PhDs' interest to do female-focus research while crowding out the male from doing female-focus research. There are at least two mechanisms that can explain this

Table 3: Female-focus Research and Gender Ratio: 1985-2015

	Dependent Variable: Female-focus Research					
	Thesis Abstract			Thesis Title		
	(1)	(2)	(3)	(4)	(5)	(6)
Male (β_1)	-0.0604***		-0.0172	-0.0439***		-0.0103
	(0.0091)		(0.0110)	(0.0072)		(0.0083)
% Female Peers (β_2)		0.0694***	0.0614***		0.0521***	0.0487***
		(0.0071)	(0.0105)		(0.0059)	(0.0080)
Male×% Female Peers (β_3)			-0.0936***			-0.0722***
			(0.0140)			(0.0103)
β_2 + β_3			-0.0323***			-0.0236***
			(0.0066)			(0.0051)
Constant	0.1303***	0.0649***	0.0954***	0.0691***	0.0214***	0.0416***
	(0.0055)	(0.0028)	(0.0087)	(0.0044)	(0.0023)	(0.0066)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.163	0.158	0.164	0.118	0.112	0.120
Observations	739751	767146	739751	739751	767146	739751

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in their dissertations. Column 1 to 3 use the key words in abstract to define the research focus of a thesis, and Column 4 to 6 use the key words in title to define the research focus of a thesis. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

phenomenon: First, having more females in a cohort encourages female PhDs to find more female coauthors instead of collaborating with male PhDs, which is called homophily. In such case, female peers crowd out men's potential collaboration with women. Second, even though the collaboration pattern does not change, males may feel they don't have a comparative advantage in doing female-focus research. In section 6, we will discuss the mechanisms in detail.

Table 4 shows the effect of gender composition on PhDs' research focus in different fields. We find there is large heterogeneity across different fields. For the health and biology field, ten additional percentage points increase in the female ratio will increase the female PhD's probability to do female-focus research by around 0.5 percentage points, and for social science, the effect is around 0.4 percentage points. However, for fields like math and physics, engineering, and psychology, the effects are not significant, while for education fields, the effect is negative. The heterogeneity across different fields is within our expectation since it's very hard to find an example of female-focus research in some fields like math and physics, while in other fields, female-focus research is too prevalent.

Table 5 shows the results when we add the gender and research focus of PhDs' advisors as control variables. We tried different specifications of advisors' research focus: whether they do female focus research both in their publication history and before the graduation of the student, and the ratio of their female focus research both in their publication history and before the graduation of the student. In general, the results are quite consistent with our main regression. From now on, we add the advisor's gender and whether the advisor does female-focus research as control variables.

Table 4: Female-focus Research and Gender Ratio: 1985-2015 (Different Fields)

	Dependent Variable: Female-focus Research						
	All fields	Engineering	Education	Health & Bio	Math & Physics	Psychology	Social Science
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Male (β_1)	-0.0171	-0.0007	-0.0720***	-0.0038	-0.0006	-0.0857**	-0.0557***
	(0.0110)	(0.0012)	(0.0156)	(0.0121)	(0.0011)	(0.0191)	(0.0124)
% Female Peers (β_2)	0.0613***	0.0091*	-0.0200***	0.0482***	0.0010	0.0326	0.0379**
	(0.0105)	(0.0048)	(0.0062)	(0.0140)	(0.0017)	(0.0125)	(0.0151)
Male×% Female Peers (β_3)	-0.0937***	-0.0124*	0.0218	-0.0659***	-0.0015	-0.0388	-0.0661***
	(0.0140)	(0.0065)	(0.0124)	(0.0188)	(0.0032)	(0.0299)	(0.0220)
$\beta_2 + \beta_3$	-0.0323***	-0.0033	0.0018	-0.0177	-0.0004	-0.0062	-0.0282**
	(0.0066)	(0.0025)	(0.0096)	(0.0109)	(0.0026)	(0.0214)	(0.0125)
Constant	0.0955***	0.0064***	0.1532***	0.1029***	0.0055***	0.2730***	0.1324***
	(0.0087)	(0.0012)	(0.0063)	(0.0097)	(0.0008)	(0.0098)	(0.0093)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.164	0.033	0.031	0.123	0.003	0.059	0.143
Observations	739771	123025	65350	151972	115233	31055	85328

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in different fields. Gender composition is measured by the female ratio of a cohort. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

4.4 Empirical Results: Probit Model

For the short-term effects, our dependent variable is a dummy variable indicating whether the PhD's thesis is female-focus research or not. We use the OLS model with fixed effects to estimate the peer effects. Considering the dummy variable nature of the dependent variable, we report estimation results using the probit model in this part.

Table 6 shows the estimation results using the probit model with program (university×field), and year fixed effects. In addition to fixed effects, we add the gender of the advisor and research focus of the advisor as control variables. We find the estimation results using the probit model reveal similar patterns as those using the OLS model, that is, higher gender composition will encourage female PhDs to do more female-focus research, yet has smaller effects on male PhDs, even though the difference is not significant.

4.5 Empirical Results: Long-Term Effects

Table 7 and Table 12 show estimation results using linked PhD and MAG data. Column 1 of Table 7 is the baseline results which we use only the ProQuest PhD thesis data. Column 2 to 5 of Table 7 display estimated effects of cohort gender composition on whether the PhDs do female-focus research for a given time period. Table 12 is a continuation of Table 7. Column 1 to 4 of Table 12 and Column 5 to 8 of Table 12 display estimated effects of cohort gender composition on the number and ratio, respectively, of the PhD's female-focus publications for a given time period.

The results in column 2 of Table 7 show that the cohort gender composition does not affect on the research focus of the pre-PhD publications, which are defined as publications that are 5 years before graduation after we control university-field and year fixed effects. This evidence further supports the randomness of our research design. Column 1 and Column 5 in Table 12 show similar findings.

Table 5: Female-focus Research and Gender Ratio: 1985-2015 (Add Advisor Info)

	Dependent Variable: Female-focus Research					
	Thesis Abstract					
	(1)	(2)	(3)	(4)	(5)	(6)
Male (β_1)	-0.0171	-0.0173	-0.0121	-0.0125	-0.0128	-0.0131
	(0.0110)	(0.0113)	(0.0104)	(0.0104)	(0.0099)	(0.0102)
% Female Peers (β_2)	0.0613***	0.0592***	0.0551***	0.0544***	0.0515***	0.0534***
	(0.0105)	(0.0110)	(0.0104)	(0.0103)	(0.0098)	(0.0102)
Male×% Female Peers (β_3)	-0.0937***	-0.0876***	-0.0814***	-0.0801***	-0.0763***	-0.0783***
	(0.0140)	(0.0145)	(0.0139)	(0.0137)	(0.0132)	(0.0137)
$\beta_2 + \beta_3$	-0.0323***	-0.0284***	-0.0263***	-0.0257***	-0.0247***	-0.0250***
	(0.0066)	(0.0066)	(0.0065)	(0.0064)	(0.0061)	(0.0063)
Female Advisor		0.0505***	0.0429***	0.0430***	0.0277***	0.0352***
		(0.0063)	(0.0057)	(0.0055)	(0.0040)	(0.0047)
Adv. Fem. Res.			0.0634***			
			(0.0057)			
Adv. Fem. Res. b. Grad.				0.0795***		
				(0.0061)		
Adv. Fem. Res. Ratio					0.6829***	
					(0.0532)	
Adv. Fem. Res. b. Grad. Ratio						0.7425***
						(0.0710)
Constant	0.0955***	0.0884***	0.0579***	0.0604***	0.0665***	0.0714***
	(0.0087)	(0.0081)	(0.0075)	(0.0077)	(0.0079)	(0.0079)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.164	0.172	0.179	0.181	0.196	0.186
Observations	739771	599407	500955	500955	500955	500955

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in their dissertations with advisors' gender and advisors' research focus as control variables. Column 1 is the baseline estimation without advisors' information as control variables. Column 2 to 6 includes advisors' gender as control variable. Column 3 to 6 includes whether the advisor does female-focus research, whether the advisor does female-focus research before the PhD's graduation, the ratio of the advisor's female-focus research and the ratio of the advisor's female-focus research before the PhD's graduation, respectively, as control variable. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, and year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

We find that for each additional 10 percentage points of female students in a cohort, women are 0.17 percentage points more likely to do female-focus research within 5 years after graduation, while the male is 0.2- percentage points less likely to do female-focus research after graduation compared with the female, as shown in column 4 of Table 7. The magnitude is smaller than that in column 1, which indicates for each additional 10 percentage points of female students in a cohort, women are 0.55 percentage points more likely to do female-focus research. This is because the definition of female-focus research in the classification of female-focus research in MAG data is stricter than the key word approach in our data and doesn't include research that uses females as objects as female-focus research.

Moreover, we find that the long-term effect of gender composition is decreasing over time. Within 5 years after graduation, women are 0.17 percentage points more likely to do female-focus research for each additional 10 percentage points of female students in a cohort, which is smaller than that

Table 6: Female-focus Research and Gender Ratio: 1985-2015 (Probit Model)

	Dependent Variable: Female-focus Research		
	(1)	(2)	(3)
Male (β_1)	-0.0443***		-0.0422***
	(0.0008)		(0.0019)
% Female Peers (β_2)		0.0493***	0.0127***
		(0.0016)	(0.0037)
Male×% Female Peers (β_3)			-0.0058
			(0.0040)
$\beta_2 + \beta_3$			0.0069***
			(0.0017)
University FE	Yes	Yes	Yes
Field FE	Yes	Yes	Yes
Year FE	Yes	Yes	Yes
Observations	497952	515905	497952

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in their dissertations using probit model. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

before graduation. However, for more than 5 years but less than 10 after graduation, the effect of gender composition on whether the PhD does female-focus research becomes insignificant, as that in column 5 of Table 7. This does not indicate that the effects of gender composition disappear, in fact, column 7 to column 8 in Table 12 show that the gender composition has persistent effects on the PhD's ratio of female-female focus research even though the magnitude is decreasing. These results suggest that the effects persist but decrease in the long run. Figure 6 shows the effects of gender composition on female-focus research over time. Consistent with Table 7, we find the peer effects are decreasing over time.

One thing that is worthwhile to mention is that in column 1 to 4 of Table 12, we use the Poisson Quasi-Maximum Likelihood model following Azoulay et al. (2019) and Truffa and Wong (2022) since our dependent variable is the number of female-focus research papers, and the results are not significant. One explanation is that our data contains many zero values (around 95%), this is understandable since more than 80 percent of PhD students leave academia upon graduation and female-focus research only accounts for less than 2 % of our publication data sample. However, this kind of distribution is not consistent with the Poisson distribution.

Our estimation of long-term effects not only documents the existence of long-term peer effects but also finds the diminishing nature of peer effects. Existing literature documents either short-run peer effects or long-run peer effects, without enough attention to the evolution of peer effects over time. Our results show the dynamic nature of peer effects and attest to the necessity of continuous support for female researchers.

Table 7: Gender Ratio's Effect on Doing Female-focus Research or Not (MAG)

	Baseline	Female Res. or Not			
	Baseline	Pre-PhD	Pre-Grad.	Post-Grad. (< 5)	Post-Grad. (5-10)
	(1)	(2)	(3)	(4)	(5)
Male (β_1)	-0.0121	0.0003	-0.0001	0.0008	-0.0033
	(0.0104)	(0.0006)	(0.0015)	(0.0027)	(0.0028)
% Female Peers (β_2)	0.0551***	-0.0004	0.0059**	0.0165***	0.0044
	(0.0104)	(0.0011)	(0.0029)	(0.0055)	(0.0054)
Male×% Female Peers (β_3)	-0.0814***	-0.0000	-0.0100**	-0.0195***	-0.0002
	(0.0139)	(0.0015)	(0.0038)	(0.0071)	(0.0071)
$\beta_2 + \beta_3$	-0.0263***	-0.0004	-0.0040	-0.0031	0.0042
	(0.0065)	(0.0011)	(0.0027)	(0.0035)	(0.0043)
Constant	0.0579***	0.0022***	0.0092***	0.0155***	0.0221***
	(0.0075)	(0.0006)	(0.0022)	(0.0048)	(0.0061)
Year FE	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.179	0.019	0.097	0.150	0.161
Observations	500955	285947	285947	285947	235439

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in their academic life. Column 1 is the baseline estimation with whether the PhD's thesis is female-focus as dependent variable. Column 2 to 5 uses the data from Microsoft Academic Graph (MAG). Column 2 shows the effects on the whether the PhD does female-focus research before PhD, we define the publication more than 5 years before graduation as pre-PhD publication. Column 3 shows the effects on the whether the PhD does female-focus research within 5 years before graduation before graduation. Column 4 shows the effects on the whether the PhD does female-focus research within 5 years after graduation. Column 5 shows the effects on the whether the PhD does female-focus research within 10 years but more than 5 years after graduation. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

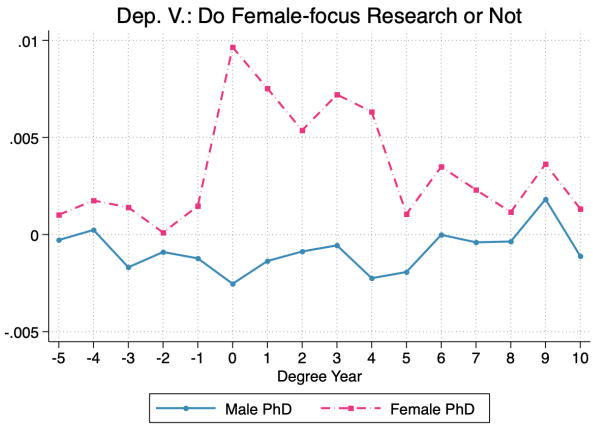
5 Robustness Check

5.1 Gender Identification

Since ProQuest doesn't contain gender information, we use the genderize.io to identify the gender of PhD recipients using their first names. The same approach has been widely used in the innovation literature (Topaz and Sen (2016), Huang et al. (2020)). Despite reporting the identified gender for each first name, genderize.io also reports the probability of that name being the assigned gender. For example, the first name "Paul" has a 0.99 probability to be a male, and the first name "Hongyuan" has a 0.86 probability to be a male. To alleviate any concerns attributed to the way we assign the gender, we check the robustness of our results to different probability thresholds above which we take the assigned gender as reliable.

Table 13 shows the results if we set different standards on assigned gender. For column 2 to column 4, We calculate the gender composition corresponding to the standard in each column again. Our results are quite robust not only in significance level but also in magnitude across different thresholds.

Figure 6: Peer Effects over Time



Note: This figure presents the peer effects of cohort gender composition on PhD's research focus from five years before graduation to ten years after graduation.

5.2 Different Definitions of Gender Composition

We use the cohort female ratio to measure the gender composition, in this section, we try two alternative measures: the number of cohort female PhDs and the female to male ratio in the cohort.

Moreover, since we use the female ratio in the graduation cohort, one may worry that the gender composition in the graduation cohort is not the actual environment where the PhD studies and implements research. To deal with this problem, we calculate the female ratio including the neighbor years. For example, if the PhD graduated in 2014, we calculate the female ratio using the data in 2013, 2014, and 2015 to proxy the diversity of the environment.

Table 14 represents the results using different definitions of cohort gender composition. We find the effects of cohort gender composition on whether the female PhD does female-focus research change from 0.55 percentage points increase to 0.87 percentage points increase for additional 10 percentage points increase in gender composition if we include neighbor cohorts when calculating the female ratio. This finding is consistent with the nature that PhDs' research, which involves discussion with neighbor cohort peers. Using the female-to-male ratio to measure cohort gender composition reveals a similar phenomenon as that in our baseline estimation. However, the results are not significant if we use the number of female peers as a measure of gender composition. This finding is not inconsistent with the main results and merely indicates that the effect of an additional female peer interacts with the cohort size (e.g., one additional female peer has a large effect in a small cohort and little to no effect in a very large cohort). This interaction is better captured by using the percent female measure in the main specification.

5.3 Cohort Size

There is concern that in a very small cohort, PhDs are more likely to be affected by other factors instead of their peers. To address this concern, we try to add cohort size as control variables and drop observations with a very small cohort size.

Table 15 displays the estimation results across different kinds of cohorts. We add cohort size as a control variable in column 2 and notice that the changes in significance level and magnitude are quite small. Moreover, we restrict our sample to cohort size larger than 5 and cohort size larger than 10, we find that (1). many observations are in a small cohort, with more than half PhDs in cohorts with no more than 10 students. (2). the peer effects are even larger in larger cohorts, indicating in larger cohorts the peer effects are less likely to be disturbed by other factors.

5.4 Keyword Approach

To define the research focus of a dissertation, we use keyword approach following Truffa and Wong (2022). However, it's reasonable to doubt that our results are driven by only a certain set of keywords. In this section, we use different keyword lists to show that our results are robust in terms of the keyword list, as shown in Table 16.

First, column 1, column 3, and column 5 report results using only abstracts to extract keywords, and column 2, column 4, and column 6 results using both abstracts and titles to extract keywords. We find there is no systematic difference across the two definitions despite that the magnitudes in columns using both abstracts and titles are larger than those using only abstracts. Second, we attempt to use a short keyword list, which only includes four words "female", "woman", "gender", "sex", and a long keyword list, which includes not only the keywords in the baseline estimation but also contains medical terminology pertaining to female-specific diseases [9]. The keyword list is provided in A.1. The robustness check results show that our results are mainly driven by the short keyword list and the variation across different keywords should not be a huge concern.

6 Mechanism

6.1 Female-dominant Team or Male-dominant Team

What explains the treatment effect of gender composition on PhD students' research focus? How to understand the different effects on female and male PhDs' research focus? We propose two potential mechanisms: First, having more female peers who have the potential to do female-focus research makes male peers re-think their competitive advantage and decide not to do female-focus research. Second, having more females in a cohort encourages female PhDs to find more female coauthors to do female-focus research instead of coauthoring with male PhDs, which is called homophily.

Table 8 represents results on female-dominant teams and male-dominant teams within 5 years after graduation. Column 1 to 3 use whether the PhD do female-focus research as the dependent variable. Column 1 of Table 8 is the same as column 4 of Table 7, which displays the peer effects on whether the PhD does female-focus research within 5 years after graduation. Column 2 and Column 3 show the cohort gender composition effects on whether the PhD does more female research in

[9]The female-specific diseases can be found on the NIH website: https://www.nichd.nih.gov/health/topics/womenshealth/conditioninfo/whatconditions

Table 8: Gender Ratio's Effect on Different Research Teams

	Female Res. or Not			Female Res. Ratio		
	All	F. Dom. Team	M. Dom. Team	All	F. Dom. Team	M. Dom. Team
	(1)	(2)	(3)	(4)	(5)	(6)
Male (β_1)	0.0008	0.0020	-0.0021	0.0002	0.0014	-0.0019
	(0.0027)	(0.0052)	(0.0037)	(0.0007)	(0.0019)	(0.0012)
% Female Peers (β_2)	0.0165***	0.0583***	-0.0392***	0.0052***	0.0213***	-0.0098***
	(0.0055)	(0.0090)	(0.0066)	(0.0014)	(0.0037)	(0.0026)
Male×% Female Peers (β_3)	-0.0195***	-0.0817***	0.0647***	-0.0065***	-0.0292***	0.0187***
	(0.0071)	(0.0140)	(0.0110)	(0.0019)	(0.0057)	(0.0045)
$\beta_2 + \beta_3$	-0.0031	-0.0234***	0.0255***	-0.0013	-0.0078***	0.0089***
	(0.0035)	(0.0069)	(0.0068)	(0.0012)	(0.0030)	(0.0027)
Constant	0.0155***	0.0020	0.0138***	0.0026***	0.0014	0.0057***
	(0.0048)	(0.0043)	(0.0052)	(0.0007)	(0.0017)	(0.0015)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.150	0.140	0.082	0.088	0.103	0.064
Observations	285947	285947	285947	285947	285947	285947

Note: This table shows the estimation effects on different research teams. Column 1 to 3 use whether the PhD do female-focus research as dependent variable. Column 4 to 6 use the ratio of female-focus research as dependent variable. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

female-dominant teams and male-dominant teams, respectively. We find the peer effects are different in female-dominant teams and male-dominant teams. Higher cohort gender composition causes the female to do more female-focus research in female-dominant teams, with 0.58 percentage points increase for additional 10 percentage points increase in the female ratio, but discourages females to do more female-focus research in male-dominant teams. While for males, the effects are the opposite. Column 4 to 6 show similar results using the ratio of female-focus research as the dependent variable.

Homophily theory can be used to explain the phenomena mentioned above. Having more females in a cohort encourages female PhDs to find more female coauthors to do female-focus research instead of coauthoring with male PhDs. However, for male PhD who are interested in female-related or gender-related topics, they are more difficult to find a suitable female coauthor, instead, they are more likely to collaborate with a male coauthor to investigate these topics.

The analysis leads us to further investigate the characteristics of research teams as a mechanism to understand the peer effects. More specifically, does the cohort gender composition affect the formation of the research teams and the collaboration within research teams?

6.2 Width and Depth of Female Coauthorship

In the last section, we show that peer effects are mainly driven by female-dominant research teams. Here, we provide direct evidence on the second mechanism: homophily and the width and depth of female coauthorship. We show that female PhDs within a higher female composition cohort are more likely to coauthor with other female researchers, both in the short run and long run. Moreover, they tend to maintain a longer and more productive academic relationship with other female researchers,

which indicates a deeper collaboration.

Table 9: Gender Ratio's Effect on The Width of Female Coauthorship

	Female Coauthor or Not		Female Coauthor Ratio	
	5 Years around Grad.	All Coauthor	5 Years around Grad.	All Coauthor
	(1)	(2)	(3)	(4)
Male (β_1)	-0.0162***	-0.0002	-0.0170***	-0.0212***
	(0.0050)	(0.0043)	(0.0037)	(0.0045)
% Female Peers (β_2)	0.0091	0.0201***	0.0357***	0.0421***
	(0.0071)	(0.0067)	(0.0066)	(0.0072)
Male×% Female Peers (β_3)	-0.0012	-0.0201**	-0.0550***	-0.0545***
	(0.0096)	(0.0082)	(0.0095)	(0.0099)
$\beta_2 + \beta_3$	0.0079	0.0000	-0.0193***	-0.0125***
	(0.0058)	(0.0048)	(0.0048)	(0.0047)
Constant	0.6895***	0.7508***	0.1878***	0.2624***
	(0.0043)	(0.0047)	(0.0047)	(0.0039)
Year FE	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.211	0.167	0.247	0.328
Observations	346636	346636	346636	346636

Note: This table shows the estimation effects of cohort gender composition on the width of PhDs' collaboration with female researchers. Column 1 and 2 reports estimation results using whether the PhD has a female coauthor within 5 years before or after graduation and in the whole academic life, respectively, as dependent variable. Column 3 and 4 reports estimation results using the ratio of the PhD's female coauthors within 5 years before or after graduation and in the whole academic life, respectively, as dependent variables. All the regressions include number of coauthors and number of papers as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 9 shows the peer effects on the width of female coauthorship. We measure the width of female coauthorship using two measures: (i). whether the PhD has a female coauthor? (ii). The ratio of female coauthors. We add the total number of coauthors and the total number of papers as control variables in all the regression estimations.

Column 1 and Column 2 estimate the gender composition effects on whether the PhD has a female coauthor within 5 years around graduation and in the whole academic life, respectively. We find ten percentage points increase in cohort female ratio will increase female PhDs' probability to have a female coauthor by 0.20 percentage points, yet there is no significant gender difference.

Column 3 and Column 4 estimate the gender composition effects on the ratio of female coauthors within 5 years around graduation and in the whole academic life, respectively. The ratio is a number between 0 to 1. We find there is 0.42 percentage points increase in the female coauthor ratio when female composition increase 10 percentage points. Moreover, we notice there is a significant difference between female and male PhD students, the effects on male PhD students are 5.45 percentage points lower than that on female PhD students.

Table 10 represents the peer effects on the average year of coauthorship per coauthor and the average number of coauthor papers per coauthor. We control the total number of coauthors and the total number of papers as above. Our results show that female PhDs in a higher female composition cohort maintain a longer and more productive collaboration with their female coauthors. Besides,

Table 10: Gender Ratio's Effect on The Depth of Female Coauthorship

	Avg. Year of Coauthorship		Avg. # of Papers	
	w. Female	w. Male	w. Female	w. Male
	(1)	(2)	(3)	(4)
Male (β_1)	-0.0216**	0.0359*	-0.0139	0.0583***
	(0.0103)	(0.0199)	(0.0130)	(0.0197)
% Female Peers (β_2)	0.0509***	-0.0381	0.0656***	-0.0118
	(0.0186)	(0.0299)	(0.0189)	(0.0261)
Male×% Female Peers (β_3)	-0.0368	0.0655	-0.0442	0.0380
	(0.0222)	(0.0406)	(0.0303)	(0.0357)
β_2 + β_3	0.0273	0.0211	0.0357	0.0145
	(0.0184)	(0.0241)	(0.0299)	(0.0309)
Constant	0.6577***	1.0566***	1.2913***	1.8071***
	(0.0096)	(0.0178)	(0.0293)	(0.0260)
Year FE	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.041	0.040	0.451	0.486
Observations	346636	346636	346636	346636

Note: This table shows the estimation effects of cohort gender composition on the depth of PhDs' collaboration with female researchers. Column 1 and 2 reports estimation results using average year of coauthorship with female coauthors and male coauthors, respectively, as dependent variables. Column 3 and 4 reports estimation results using average number of papers with female coauthors and male coauthors, respectively, as dependent variables. All the regressions include number of coauthors and number of papers as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

there is no significant effect on collaboration with male coauthors, as shown in column 2 and column 4. However, the signs of the effects on the average year of coauthorship and the average number of papers with male coauthors are both negative, indicating deeper collaboration with female coauthors crowds out collaboration with male coauthors.

Homophily theory describes the observed tendency of "like to associate with like", similarity of attributes and experience arguably simplifies the process of evaluating, communicating with, and even predicting the behavior of others (McPherson et al., 2001; Kossinets and Watts, 2009). However, an individual's choice of relations is heavily constrained by other aspects of his or her life. Our results show that the under-representation of women in academia is a huge obstacle for female scientists to find similar collaborators who are enthusiastic about female-related topics.

7 Conclusion

The under-representation of women in academia is a topic of great interest in economics and public policy today. Besides the concern for "missing Curie", little is known about how the under-representation of women in academia affects "who" and "what" gets studied in research. This paper presents novel estimates on the short-term and long-term impacts of cohort gender composition on the focus of academic research. We find for each additional 10 percentage points of female students in a cohort, women are 0.61 percentage points more likely to do female-focus research in the dissertation, while men are 0.32 percentage points less likely to do female-focus research. These effects last

for at least 5 years after graduation. Further investigation shows these effects are mainly driven by female-dominant teams.

We provide evidence for the homophily mechanism. We show female PhDs within a higher female composition cohort are more likely to coauthor with other female researchers, both in the short run and long run. Moreover, they tend to maintain a longer and more productive academic relationship, measured by the average length of coauthor years and the average number of coauthored papers, respectively, with other female researchers, which indicates a deeper collaboration.

Our analysis is relevant to policy-makers, given the emphasis they are placing on the diverse environment and the production of new ideas. Our results reveal that building a diverse and inclusive academic environment can encourage females to do more female-focus research and help to close the knowledge gap related to the female. However, the diverse environment has a heterogeneous effect on female and male researchers.

We see two avenues for future research. First, this paper suggests the importance of female peers on coauthor networks, despite the preliminary evidence on the width and depth of coauthorship provided in this paper. Little is known about how a diverse environment in the early stage of academic life affects female and male scientists' further teamwork differently.

Second, this study focuses mainly on the supply side, the production of female-focus research. However, the demand side is also of first-order importance in understanding the knowledge production process. Whether the researchers doing female-focus research be penalized through worse placement, a lower probability of commercialization requires further investigation.

References

ACEMOGLU, D., U. AKCIGIT, AND W. R. KERR (2016): "Innovation network," *Proceedings of the National Academy of Sciences*, 113, 11483–11488.

AGRAWAL, A., D. KAPUR, AND J. MCHALE (2008): "How do spatial and social proximity influence knowledge flows? Evidence from patent data," *Journal of urban economics*, 64, 258–269.

AKCIGIT, U., S. CAICEDO, E. MIGUELEZ, S. STANTCHEVA, AND V. STERZI (2018): "Dancing with the stars: Innovation through interactions," .

AKCIGIT, U., J. G. PEARCE, AND M. PRATO (2022): "Tapping into talent: Coupling education and innovation policies for economic growth," *NBER Working Paper*.

ALON, T., D. CAPELLE, AND K. MATSUDA (2022): "University Research and the Market for Higher Education," *Working Paper*.

AZOULAY, P., C. FONS-ROSEN, AND J. S. GRAFF ZIVIN (2019): "Does science advance one funeral at a time?" *American Economic Review*, 109, 2889–2920.

AZOULAY, P., J. S. GRAFF ZIVIN, AND J. WANG (2010): "Superstar extinction," *The Quarterly Journal of Economics*, 125, 549–589.

AZOULAY, P., R. MICHIGAN, AND B. N. SAMPAT (2007): "The anatomy of medical school patenting," *New England Journal of Medicine*, 357, 2049–2056.

BLOOM, N., C. I. JONES, J. VAN REENEN, AND M. WEBB (2020): "Are ideas getting harder to find?" *American Economic Review*, 110, 1104–44.

BOSTWICK, V. K. AND B. A. WEINBERG (2022): "Nevertheless she persisted? Gender peer effects in doctoral STEM programs," *Journal of Labor Economics*, 40, 397–436.

DING, W. W., F. MURRAY, AND T. E. STUART (2006): "Gender differences in patenting in the academic life sciences," *science*, 313, 665–667.

DUPAS, P., A. S. MODESTINO, M. NIEDERLE, J. WOLFERS, ET AL. (2021): "Gender and the dynamics of economics seminars," Tech. rep., National Bureau of Economic Research.

ERTUG, G., J. BRENNECKE, B. KOVÁCS, AND T. ZOU (2022): "What does homophily do? A review of the consequences of homophily," *Academy of Management Annals*, 16, 38–69.

FREEMAN, R. B. AND W. HUANG (2015): "Collaborating with people like me: Ethnic coauthorship within the United States," *Journal of Labor Economics*, 33, S289–S318.

GOLDIN, C. (2014): "A grand gender convergence: Its last chapter," *American Economic Review*, 104, 1091–1119.

HOFSTRA, B., V. V. KULKARNI, S. M.-N. GALVEZ, B. HE, D. JURAFSKY, AND D. A. MCFARLAND (2020): "The diversity–innovation paradox in science," *Proceedings of the National Academy of Sciences*, 117, 9284–9291.

HOXBY, C. M. (2000): "Peer effects in the classroom: Learning from gender and race variation," .

HSIEH, C.-T., E. HURST, C. I. JONES, AND P. J. KLENOW (2019): "The allocation of talent and us economic growth," *Econometrica*, 87, 1439–1474.

HUANG, J., A. J. GATES, R. SINATRA, AND A.-L. BARABÁSI (2020): "Historical comparison of gender inequality in scientific careers across countries and disciplines," *Proceedings of the National Academy of Sciences*, 117, 4609–4616.

HYLAND, M., S. DJANKOV, AND P. K. GOLDBERG (2020): "Gendered laws and women in the workforce," *American Economic Review: Insights*, 2, 475–90.

KIM, S. D. AND P. MOSER (2021): "Women in Science. Lessons from the Baby Boom," Tech. rep., National Bureau of Economic Research.

KOFFI, M. (2021): "Innovative ideas and gender inequality," .

KONING, R., S. SAMILA, AND J.-P. FERGUSON (2020): "Inventor Gender and the Direction of Invention," in *AEA Papers and Proceedings*, vol. 110, 250–54.

——— (2021): "Who do we invent for? Patents by women focus more on women's health, but few women get to invent," *Science*, 372, 1345–1348.

KOSSINETS, G. AND D. J. WATTS (2009): "Origins of homophily in an evolving social network," *American journal of sociology*, 115, 405–450.

KWIEK, M. AND W. ROSZKA (2021): "Gender-based homophily in research: A large-scale study of man-woman collaboration," *Journal of Informetrics*, 15, 101171.

LAVY, V. AND A. SCHLOSSER (2011): "Mechanisms and impacts of gender peer effects at school," *American Economic Journal: Applied Economics*, 3, 1–33.

LEE, M. (2016): "Allocation of female talent and cross-country productivity differences," *Working Paper*.

LERCHENMUELLER, M. J. AND O. SORENSON (2018): "The gender gap in early career transitions in the life sciences," *Research Policy*, 47, 1007–1017.

LERNER, J. AND U. MALMENDIER (2013): "With a little help from my (random) friends: Success and failure in post-business school entrepreneurship," *The Review of Financial Studies*, 26, 2411–2452.

LIU, E. AND S. MA (2022): "Innovation Networks and R&D Allocation," *NBER Working Paper*.

MCPHERSON, M., L. SMITH-LOVIN, AND J. M. COOK (2001): "Birds of a feather: Homophily in social networks," *Annual review of sociology*, 415–444.

MOUGANIE, P. AND Y. WANG (2020): "High-performing peers and female STEM choices in school," *Journal of Labor Economics*, 38, 805–841.

NIELSEN, M. W., J. P. ANDERSEN, L. SCHIEBINGER, AND J. W. SCHNEIDER (2017): "One and a half million medical papers reveal a link between author gender and attention to gender and sex analysis," *Nature human behaviour*, 1, 791–796.

SINHA, A., Z. SHEN, Y. SONG, H. MA, D. EIDE, B.-J. HSU, AND K. WANG (2015): "An overview of microsoft academic service (mas) and applications," in *Proceedings of the 24th international conference on world wide web*, 243–246.

TOPAZ, C. M. AND S. SEN (2016): "Gender representation on journal editorial boards in the mathematical sciences," *PLoS One*, 11, e0161357.

TRUFFA, F. AND A. WONG (2022): "Undergraduate Gender Diversity and Direction of Scientific Research," *Working paper*.

WU, A. H. (2018): "Gendered language on the economics job market rumors forum," in *AEA Papers and Proceedings*, vol. 108, 175–79.

——— (2020): "Gender bias among professionals: an identity-based interpretation," *Review of Economics and Statistics*, 102, 867–880.

WU, C. (2021): "More unequal income but less progressive taxation," *Journal of Monetary Economics*, 117, 949–968.

A Additional Results

A.1 Keywords and Example of female-focus research

1. Keywords List:

    - Baseline List: woman, female, lady, feminism, feminine, femininity, girl, pregnancy, pregnant, gender, sex , wife, daughter, mother;
    - Short List: woman, female, sex, gender;
    - Long List: woman, female, lady, feminism, feminine, femininity, girl, pregnancy, pregnant, gender, sex , wife, daughter, mother, gynecological, gynecology, menopause, menstruation, menstrual, urinary tract, vaginosis, vaginitis, uterine fibroids, pregnant, pregnancy, preterm, endometriosis, ovary, ovarian, cervical, turner syndrome, rett syndrome.

2. Example 1:

    - Title: Welfare waivers and women's non-marital fertility decisions
    - Abstract: A large body of literature establishes a negative correlation between non-marital childbearing and outcomes for both mother and child. Concern about the negative consequences of non-marital childbearing led policymakers to implement policies designed to decrease the incidence of non-marital childbearing. Specifically in the late 1980's states began applying to the federal government for waivers granting permission to implement state level welfare policies that differed from existing federal policy. An explicit goal of state policymakers in drafting the waivers was to alter incentives in order to influence unmarried women's fertility decisions.

3. Example 2:

    - Title: Development of novel antiestrogens for the treatment of tamoxifen-resistant breast cancer
    - Abstract: The antiestrogen tamoxifen is the most widely prescribed chemotherapeutic agent for the treatment of estrogen receptor (ER)-positive breast cancers. Although this compound has had a tremendous impact on overall mortality and morbidity in metastatic breast cancer patients, its utility is limited by the eventual development of resistance. Recent insights into the pharmacology of tamoxifen indicate that this compound is not a pure antiestrogen, but rather a Selective Estrogen Receptor Modulator (SERM) whose relative agonist/antagonist activity varies in a cell-dependent manner. Therefore, resistance most

likely results when breast cancer cells start to recognize tamoxifen as an agonist instead of an antagonist. Our laboratory and others have shown that SERMs exert their differential effects on ER transcriptional activity by inducing distinct conformational changes within the ligand-binding domain of the receptor, an activity which influences the ability of ER to interact with the transcriptional machinery. My hypothesis, therefore, was that compounds which were mechanistically distinct from tamoxifen would inhibit the growth of tamoxifen-resistant tumors.

A.2 Decomposition in Figure 4

In Section, we decompose the rising number of female-focus research into different channels. We specify the construction of each counterfactual total number of female-focus research for when different effects kick in

$$\#\text{ Female-focus Research}_t^{\text{size}} = \sum_s \sum_g \underbrace{\frac{\#\text{ Female-Focus}_{g,s,t}}{\#\text{ PhD}_{g,s,t}}}_{\text{inclination}} \times \underbrace{\frac{\#\text{ PhD}_{g,s,t}}{\#\text{ PhD}_{s,t}}}_{\text{gender composition}} \times \underbrace{\frac{\#\text{ PhD}_{s,t}}{\#\text{ PhD}_{t}}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_{2015}}_{\text{size}} \tag{6}$$

$$\#\text{ Female-focus Research}_t^{\text{size+field}} = \sum_s \sum_g \underbrace{\frac{\#\text{ Female-Focus}_{g,s,t}}{\#\text{ PhD}_{g,s,t}}}_{\text{inclination}} \times \underbrace{\frac{\#\text{ PhD}_{g,s,t}}{\#\text{ PhD}_{s,t}}}_{\text{gender composition}} \times \underbrace{\frac{\#\text{ PhD}_{s,2015}}{\#\text{ PhD}_{2015}}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_{2015}}_{\text{size}} \tag{7}$$

$$\#\text{ Female-focus Research}_t^{\text{size+field+gender}} = \sum_s \sum_g \underbrace{\frac{\#\text{ Female-Focus}_{g,s,t}}{\#\text{ PhD}_{g,s,t}}}_{\text{inclination}} \times \underbrace{\frac{\#\text{ PhD}_{g,s,2015}}{\#\text{ PhD}_{s,2015}}}_{\text{gender composition}} \times \underbrace{\frac{\#\text{ PhD}_{s,2015}}{\#\text{ PhD}_{2015}}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_{2015}}_{\text{size}} \tag{8}$$

$$\#\text{ Female-focus Research}_t^{\text{size+field+male}} = \sum_s \left(\begin{array}{c} \underbrace{\frac{\#\text{ Female-Focus}_{f,s,t}}{\#\text{ PhD}_{f,s,2015}}}_{\text{female inclination}} \times \underbrace{\frac{\#\text{ PhD}_{f,s,2015}}{\#\text{ PhD}_{s,2015}}}_{\text{female ratio}} \\ \times \underbrace{\frac{\#\text{ PhD}_{s,2015}}{\#\text{ PhD}_{2015}}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_{2015}}_{\text{size}} \\ + \underbrace{\frac{\#\text{ Female-Focus}_{m,s,2015}}{\#\text{ PhD}_{g,s,t}}}_{\text{male inclination}} \times \underbrace{\frac{\#\text{ PhD}_{m,s,2015}}{\#\text{ PhD}_{s,2015}}}_{\text{male ratio}} \\ \times \underbrace{\frac{\#\text{ PhD}_{s,2015}}{\#\text{ PhD}_{2015}}}_{\text{field composition}} \times \underbrace{\#\text{ PhD}_{2015}}_{\text{size}} \end{array} \right) \tag{9}$$

Similar decomposition applies for total number of female PhD.

A.3 Quantification of Decomposition

In order to quantify the contribution to missing female Ph.D, we consider the total number of missing Curies in a time window: a full sample from 1985 to 2015 and a subsample from 1985 to 2007, because the number of female Ph.D is stabilized since 2007.

we consider

$$\underbrace{\left(\sum_{t=1985}^{T} \text{Data}_{2015} - \text{Counterfactual}_t^{\text{w/o}}\right)}_{\text{Counterfactual \# Missing (w/o interaction)}} / \underbrace{\left(\sum_{t=1985}^{T} \text{Data}_{2015} - \text{Data}_t\right)}_{\text{\# Missing}} \tag{10}$$

as the contribution of missing female Ph.D from a particular channel when this channel does not interact with all other channels. We compute

$$\underbrace{\left(\sum_{t=1985}^{T} \text{Counterfactual}_t^{\text{w}} - \text{Data}_{1985}\right)}_{\text{Counterfactual \# Missing (w/ all interaction)}} / \underbrace{\left(\sum_{t=1985}^{T} \text{Data}_{2015} - \text{Data}_t\right)}_{\text{\# Missing}} \tag{11}$$

as the contribution of missing female Ph.D from a particular channel when this channel potentially interacts with all other channels. We interpret the mean of these two ratios as the contribution of missing female Ph.D due to a particular channel. Similar approach applies for decomposition of missing female-focus research.

A.4 Discussion of Simple Model

Suppose there is no heterogeneous diffusion effects across genders, i.e., $\text{sgn}(\phi'_{=}) \neq \text{sgn}(\phi'_{\neq})$, and denote

$$\Phi(s) \equiv \frac{\phi_{\neq}(s)}{\phi_{=}(1-s)}, \tag{12}$$

If $\Phi(s)' \lesseqgtr 0$ for all $x \in (0, \frac{1}{2})$ and $\Phi(s)' \gtrless 0$ for all $x \in (\frac{1}{2}, 1)$, then $\Gamma_m(s)' \lesseqgtr 0$ and $\Gamma_f(s)' \gtrless 0$.

Proof. It is not hard to find

$$\Gamma_m(s) = \frac{z_m^*}{z_m}\Phi(s), \quad \Gamma_f(s) = \frac{z_f^*}{z_f}\frac{1}{\Phi(1-s)}. \tag{13}$$

Since for all $s \in (0,1)$, we have $1-s \in (0,1)$, $\Phi(s)' \gtrless 0$, i.e.

$$\frac{\mathrm{d}}{\mathrm{d}x}\frac{1}{\Phi(1-s)} = \frac{\Phi'(1-s)}{\Phi(1-s)^2} \gtrless 0. \tag{14}$$

The proof is done by noticing when $s \in (0, 1/2)$, then $1-s \in (1/2, 1)$. ∎

A.5 Supplementary Tables

Table 11: Gender Composition: First-order Auto-Regression

	% Female Peers
	(1)
% Female Peers in Last Year	-0.0053
	(0.0047)
Constant	-0.3943***
	(0.0018)
Year FE	Yes
University×Field FE	Yes
Adjusted R-squared	0.581
Observations	41913

Note: This table shows the estimation results of first-order auto-regression. The regression includes university-field fixed effects, year fixed effects. Cohorts that are with few than 5 PhD students are dropped from the the sample. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 12: Gender Ratio's Effect on The Number & Ratio Female-focus Research (MAG)

	Female Res. #				Female Res. Ratio			
	Pre-PhD	Pre-Grad.	Post-Grad. (< 5)	Post-Grad. (5-10)	Pre-PhD	Pre-Grad.	Post-Grad. (< 5)	Post-Grad. (5-10)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Male (β_1)	0.2185	-0.2322**	-0.0960	-0.0828	-0.0002	-0.0001	0.0002	0.0002
	(0.2241)	(0.0927)	(0.0749)	(0.0812)	(0.0001)	(0.0004)	(0.0007)	(0.0004)
% Female Peers (β_2)	0.0361	-0.1407	0.1575*	0.0044	-0.0002	0.0021***	0.0052***	0.0022***
	(0.2803)	(0.0982)	(0.0846)	(0.0892)	(0.0003)	(0.0008)	(0.0014)	(0.0008)
Male×% Female Peers (β_3)	-0.2640	0.2969**	0.0399	0.0608	0.0002	-0.0032***	-0.0065***	-0.0032***
	(0.3837)	(0.1502)	(0.1327)	(0.1529)	(0.0004)	(0.0010)	(0.0019)	(0.0010)
β_2 + β_3	-0.2279	0.1562	0.1974***	0.0652	0.0001	-0.0010	-0.0013	-0.0010**
	(0.2677)	(0.1213)	(0.0852)	(0.0972)	(0.0003)	(0.0007)	(0.0012)	(0.0005)
Constant	-4.1746***	-3.2825***	-2.4498***	-1.7519***	0.0005***	0.0022***	0.0026***	0.0017***
	(0.1883)	(0.1325)	(0.1108)	(0.0983)	(0.0001)	(0.0004)	(0.0007)	(0.0004)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	Poisson	Poisson	Poisson	Poisson	0.022	0.065	0.088	0.099
Observations	81674	173180	222126	189754	285947	285947	285947	235439

Note: This table shows the estimation effects of cohort gender composition on PhDs' research focus in their academic life. Column 1 to 4 uses the number of female-focus papers as dependent variable and use Poisson Quasi-Maximum Likelihood model to implement estimation. Column 5 to 8 uses the ratio of female-focus papers as dependent variable. Column 1 and 5 shows the effects on the number and ratio of female-focus papers before PhD, we define the publication more than 5 years before graduation as pre-PhD publication. Column 2 and 6 shows the effects on the number and ratio of female-focus papers within 5 years before graduation before graduation. Column 3 and 7 shows the effects on the number and ratio of female-focus papers within 5 years after graduation. Column 4 and 8 shows the effects on the number and ratio of female-focus papers within 10 years but more than 5 years after graduation. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 13: Robustness Check: Change Gender Assignment Standard

	Dependent Variable: Female-focus Research			
	(1) Baseline	(2) 85 %	(3) 90 %	(4) 95 %
Male (β_1)	-0.0121	-0.0173	-0.0179	-0.0185
	(0.0104)	(0.0115)	(0.0116)	(0.0120)
% Female Peers (β_2)	0.0551***	0.0529***	0.0527***	0.0540***
	(0.0104)	(0.0110)	(0.0112)	(0.0116)
Male×% Female Peers (β_3)	-0.0814***	-0.0821***	-0.0818***	-0.0836***
	(0.0139)	(0.0146)	(0.0147)	(0.0149)
$\beta_2 + \beta_3$	-0.0263***	-0.0292***	-0.0291***	-0.0296***
	(0.0065)	(0.0069)	(0.0068)	(0.0068)
Constant	0.0579***	0.0643***	0.0653***	0.0654***
	(0.0075)	(0.0084)	(0.0086)	(0.0089)
Year FE	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.179	0.179	0.178	0.178
Observations	500955	450442	433989	407310

Note: This table shows the estimation effects of cohort gender composition on research focus. Column 1 is the baseline estimation where we fully rely on the assigned gender by Genderize.io. Column 2 only takes the assigned gender with probability no smaller than 85 % as assigned gender. Column 3 and 4 only takes the assigned gender with probability no smaller than 90 % and 95 %, respectively, as assigned gender. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 14: Robustness Check: Different Definitions of Gender Composition

	Dependent Variable: Female-focus Research			
	(1) Baseline	(2) Include Neighbor Cohorts	(3) # Female Peers	(4) Female/Male
Male (β_1)	-0.0121	0.0048	-0.0500***	-0.0342***
	(0.0104)	(0.0102)	(0.0072)	(0.0092)
GenderComposition (β_2)	0.0551***	0.0870***	0.0005	0.0070***
	(0.0104)	(0.0132)	(0.0005)	(0.0016)
Male×GenderComposition (β_3)	-0.0814***	-0.1284***	0.0002	-0.0151***
	(0.0139)	(0.0184)	(0.0009)	(0.0028)
$\beta_2 + \beta_3$	-0.0263***	-0.0414***	0.0007	-0.0081***
	(0.0065)	(0.0091)	(0.0005)	(0.0016)
Constant	0.0579***	0.0462***	0.0846***	0.0742***
	(0.0075)	(0.0079)	(0.0032)	(0.0049)
Year FE	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.179	0.180	0.178	0.168
Observations	500955	500955	500955	475092

Note: This table shows the estimation effects of cohort gender composition on research focus using different definitions of cohort composition. Column 1 uses the female ratio in the graduation cohort as cohort gender composition. Column 2 includes neighbor cohorts' PhD students and calculate the overall female ratio in the three years. Column 3 and column 4 uses number of female peers and female-to-male ratio, respectively, as the measure of cohort gender composition. Column 3 include cohort size as control variable. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 15: Robustness Check: Cohort Size

	Dependent Variable: Female-focus Research			
	(1) Baseline	(2) + Cohort Size	(3) CohortSize > 5	(4) CohortSize > 10
Male (β_1)	-0.0121	-0.0111	0.0023	0.0073
	(0.0104)	(0.0103)	(0.0110)	(0.0116)
% Female Peers (β_2)	0.0551***	0.0560***	0.0772***	0.0792***
	(0.0104)	(0.0103)	(0.0128)	(0.0153)
Male×% Female Peers (β_3)	-0.0814***	-0.0837***	-0.1130***	-0.1195***
	(0.0139)	(0.0136)	(0.0177)	(0.0212)
$\beta_2 + \beta_3$	-0.0263***	-0.0276***	-0.0357***	-0.0403***
	(0.0065)	(0.0064)	(0.0074)	(0.0092)
Constant	0.0579***	0.0510***	0.0430***	0.0335***
	(0.0075)	(0.0072)	(0.0076)	(0.0079)
Year FE	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes
Adjusted R-squared	0.179	0.179	0.174	0.164
Observations	500955	500955	361982	239726

Note: This table shows the estimation effects of cohort gender composition on research focus across different cohorts. Column 1 reports baseline estimation and column 2 includes cohort size as a control variable. Column 3 and column 4 drops observations with cohort size no more than 5 and no more than 10, respectively. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.

Table 16: Robustness Check: Different Keywords

	Dependent Variable: Female-focus Research					
	Baseline		Short List		Long List	
	(1)	(2)	(3)	(4)	(5)	(6)
	Abstract	Abstract+Title	Abstract	Abstract+Title	Abstract	Abstract+Title
Male (β_1)	-0.0121	-0.0130	-0.0125	-0.0133	-0.0119	-0.0128
	(0.0104)	(0.0109)	(0.0098)	(0.0102)	(0.0102)	(0.0108)
% Female Peers (β_2)	0.0551***	0.0581***	0.0479***	0.0505***	0.0548***	0.0576***
	(0.0104)	(0.0109)	(0.0094)	(0.0098)	(0.0102)	(0.0108)
Male×% Female Peers (β_3)	-0.0814***	-0.0844***	-0.0680***	-0.0710***	-0.0818***	-0.0847***
	(0.0139)	(0.0144)	(0.0120)	(0.0124)	(0.0137)	(0.0143)
β_2 + β_3	-0.0263***	-0.0263***	-0.0201***	-0.0205***	-0.0270***	-0.0270***
	(0.0065)	(0.0067)	(0.0055)	(0.0058)	(0.0065)	(0.0067)
Constant	0.0579***	0.0601***	0.0502***	0.0522***	0.0627***	0.0652***
	(0.0075)	(0.0078)	(0.0066)	(0.0069)	(0.0075)	(0.0078)
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
University×Field FE	Yes	Yes	Yes	Yes	Yes	Yes
Adjusted R-squared	0.179	0.185	0.166	0.172	0.170	0.176
Observations	500955	500955	500956	500956	500956	500956

Note: This table shows the estimation effects of cohort gender composition on research focus defined using different keywords. Column 1 and Column 2 use the baseline keyword list to define female-focus research, Column 3 and Column 4 use the short keyword list to define female-focus research. Column 5 and Column 6 use the long keyword list including medical terminology pertaining to female-specific diseases to define female-focus research. All the regressions include the gender of the advisor and whether the advisor does female-focus research as control variables. Gender composition is measured by the female ratio of a cohort. All regressions include university-field fixed effects, year fixed effects. Standard errors in parentheses are clustered at the program (university × field) level. * p < 0.1, ** p < 0.05, *** p < 0.01.