Canadian Labour Economics Forum

WORKING PAPER SERIES

Innovative Ideas and Gender Inequality

Marlene Koffi (University of Toronto)

CLEF Award 2021 – Runner up

WP #35

Innovative Ideas and Gender Inequality*

Marlène Koffi
University of Toronto †

April 23, 2021

Abstract

This paper analyzes the recognition of women's innovative ideas. Bibliometric data from research in economics are used to investigate gender biases in citation patterns. Based on deep learning and machine learning techniques, one can (1) establish the similarities between papers (2) build a link between articles by identifying the papers citing, cited and that should be cited. This study finds that, on average, omitted papers are 15%-20% more likely to be female-authored than male-authored. This omission bias is more prevalent when there are only males in the citing paper. Overall, to have the same level of citation as papers written by males, papers written by females need to be 20 percentiles upper in the distribution of the degree of innovativeness of the paper.

1 Introduction

Context: Women face a lower entry rate and a higher exit rate than men in industries or fields that require mathematical skills and analytical abilities.[1] As a result, women are

*This is a short version of the paper "Innovative Ideas and Gender Inequality" available on my website https://sites.google.com/view/marlenekoffi/

†Email: marlene.koffi@utoronto.ca. I am deeply indebted to my advisor Vasia Panousi for her constant support and guidance. I am very grateful to Marti Mestieri, Dimitris Papanikolaou, Mar Reguant, Joshua Lewis, and Ismael Mourifié for extremely constructive feedback and discussions. I would like to thank Abel Brodeur, Sergio Salgado, Erin Hengel, Jonathan Guryan, the members of the Lab for Macroeconomic Policy, and the seminar participants of the University of Toronto, the Universite de Montreal, Cornell University, the Bank of Canada, Carleton College, the International Monetary Fund, and the World Bank, Analysis Group, Gender Online Seminars for useful comments.

[1]In 1999-2000, 13% of women received a bachelor degree in education versus 4% for men; 2% of women received a bachelor degree in engineering versus 12% for male (2001 Baccalaureate and Beyond Longitudinal Study, Zafar (2013)). Antecol and Cobb–Clark (2013) reach the same conclusion using survey data from the National Longitudinal Study of Adolescent Health over the period 1994-2008 on a survey database. Preston (1994) documents the higher exit rate of women in math-intensive fields. Hunt (2016) uses survey data from the National Survey of College Graduates to examine the difference in exit rates of women in science and engineering compared with other fields.

underrepresented in those fields, especially in top-ranked positions (Ceci et al. 2014; Ginther and Kahn 2004). Preferential choices such as family, risk aversion, and competitiveness, along with discriminatory factors, have been suggested as potential explanations for this gap. Yet one mechanism received less or no attention: the recognition of women's works. Knowing that ideas are at the core of the research and innovation process and that being recognized and valued for your work and ideas can be a motivation to start or continue in a field, it is necessary to consider the question of the credit given to women's work.

Contribution: This paper analyzes the state of intellectual property in academic research in economics, with an emphasis on the recognition of women's works. In this sense, the respect of intellectual property at the academic research level will work through the recognition of individual work and the acknowledgement of relevant prior literature.[2] Because academic research provides an ideal framework for analyzing ideas, we can test how women's ideas are perceived, used and referred to. Therefore, the paper explores whether articles by equally relevant and innovative female authors are listed as they should be in the references of the articles that follow. If not, we will then talk about a gender omission bias. The paper focuses on economics within academia for two main reasons. First, the representation gap is among the largest in economics (Bayer and Rouse 2016). Second, many voices have recently been raised against gender discrimination in economics research, which seems to be more prevalent than in other life sciences or engineering (Ginther and Kahn 2004, Wu 2018, Sarsons 2019). Thus, this paper sheds new light on the lack of recognition of women's works. Further, it contributes to the existing works by exhibiting the heterogeneous pattern in the omission bias, investigating some potential mechanisms, and discussing the policy implications.

Methods: To achieve the research objectives, this article uses bibliometric data on articles published in major economic journals. This data comes from Web of Science, Econlit, and Ideas RePEc. In the second step, textual analysis based on big data and machine learning tools adds key insights to the analysis. I infer the gender of the articles using the gender of the authors through gender name dictionaries, classification algorithms, and manual checking.

Then, I build a distance measure that compares different articles and establishes a link between the citing article, the articles cited, and the articles to be cited. At this level, it is worth mentioning how challenging establishing such a relationship is, especially when we know that in research, each article tries to differentiate itself from its predecessors (by way of writing, methodology, approach, ...). Recognizing this difficulty, Zhu et al. (2015) chose to

[2]A parallel can be drawn with the notion of intellectual property. According to the World Intellectual Property Office (WIPO), "Intellectual Property refers to creations of the minds...". The various processes for the protection of intellectual works (Trademarks, Copyrights, and Patents) aim to encourage authors to engage in innovative activities by guaranteeing them recognition, even exclusivity, over their production. Even if there is no regulation surrounding the protection of ideas in academic research in economics (and in many other fields), the fact remains that ideas are a work of the mind and therefore an intellectual property.

ask the authors which articles they think are the most important for their study. However, two drawbacks come with this approach. On the one hand, it would not make it possible to distinguish latent biases if they exist and, on the other hand, to have counterfactuals on which category of authors should be cited and is not. Beyond this difficulty, it seems also impossible to cite all the papers in the prior literature. The problem emerges when there is a discrepancy between authors based on their (observable) characteristics despite the relative closeness. Thus, this article proposes an objective method that allows us to relate different articles using textual similarities. Using natural language processing, the comparison is based on textual distance tools, extending to tools based on deep learning and neural networks for textual analysis, where words embeddings more extensively.

Then, two key indices are constructed from this analysis. The first is an omission index, which is the novelty and the methodological contribution of the current paper. It measures the propensity with which an article that is part of the relevant existing literature of certain articles is omitted from those articles' references. In other words, it captures the fact that an article that has several similarities to another in the future is not mentioned in the latter's references.[3] The second is the innovation index. This index offers an alternative way to assess the quality of an article. Unlike quotes, this metric is less likely to be biased.[4] Similar to Kelly et al. (2018) and Koffi and Panousi (2019), an article is considered very innovative (and therefore of high quality) if it is new and influences future research. The omission index coupled with the innovation index allows us to contrast what should be and what can be observed with the quotes.

While papers in the unsupervised learning literature and more recently those using texts as data in economics (see Gentzkow et al., 2017 for a review) contribute to the method's external validity, I carry out multiple internal validations to show that the proposed methodology in this article actually captures patterns in the data.

Finally, each article's observable characteristics are combined to build an author-level database to assess the effects of the omission for the authors in terms of future publications.

Findings: Turning to the findings, this paper presents the evidence for a gender omission bias in economics. Indeed, omitted articles are 15% to 30% more likely to be female-authored than male-authored. Mixed team papers (with both male and female authors) tend to fall in between both genders The papers most likely to be omitted are written by women (solo, mostly female team) working at mid-tier institutions, publishing in non-top journals. In a group of related papers, the probability of omission of those papers increases by 6 percentage points compared to men in similar affiliation when the citing authors are only males. Overall, for similar papers, having at least one female author reduces the probability of omitting other

[3]This omission can be intentional or unintentional.

[4]See Lampe (2012) and D'Ippoliti (2018), among others.

women's papers by up to 10 percentage points, whereas having only male authors increases the probability of being omitted by almost 4 percentage points. Moreover, the omission bias is twice as high in theoretical fields that involve mathematical economics than in applied fields such as education and health economics. In addition, even papers written by women published in top journals are not exempted from the omission bias. The baseline estimation includes only articles published in top-ranked economic journals. Provided that the journal in which a paper is published is a signal of quality, this ensures that the estimation does not capture doubt on the articles' quality. This indicates that we are more likely to take a lower bound by controlling the quality of an article using the journal publication. Indeed, if there is a bias in the standards imposed on men and women (Card et al. 2020, Engel and Moon (2020)), then the articles published by women in top-ranked economic journals are of better quality than those posted by men, and yet the bias still exists. Further, the regressions include observable characteristics for both the citing and the cited or omitted articles, such as the affiliation of the most prolific authors, the main field, the year of publication, and the gender structure of the articles in the most similar set, the number of references. Several robustness checks (including -but not limited to- controlling for the methodological style, the position of the authors' names, and the extension to more than 100 journals in economics) suggest that the estimates are not overly sensitive to the particular choice of control variables. To ensure that I am capturing patterns in the data, I validate the omission index in several steps. Finally, being omitted with respect to past publications reduces the probability of getting published in a top 5 journal in the future by up to 5%.

2 Related Literature

Overall, this article demonstrates that the lack of recognition of women's work is also noticeable through the non-reference of articles written by women. In addition, through the subjects discussed and the techniques used, this study builds on several areas of the economic literature.

First, the question of whether women get enough credit and therefore recognition for their research is at the core of this paper. In this sense, this paper is complementary to Sarsons (2019). Indeed, Sarsons (2019) tests the uncertainty about the individual contributions of co-authors favors men in terms of tenure rates compared to women. Here, I explicitly use article references to assess to whom credit is most often attributed and if this is done to the detriment of women. Moreover, the findings of Sarsons (2019) suggest that women are worse off when they collaborate with men. Similarly to Hengel and Moon (2019), I show that women also fare worse when they do not collaborate with men. In fact, mixed gender teams received treatment midway between that received by single-gender teams.

This paper is linked to the general literature on gender discrimination in academic research.

More specifically, three key points emerge from the recent literature.

The first one is the presence of stereotypes. Wu (2018) highlighted that female authors are most often associated with physical characteristics while male authors are most associated with intellectual characteristics. The second element is the difference in standards and evaluations between men and women. For example, Hengel (2019) shows that women experienced longer delays in the review process and are asked to make much more revisions before getting published. In the same line, Card et al. (2019) show that to publish in the same journal as males, females are required a citation premium by the referees, that is not corrected by the editors decision to ask for a revise and resubmit. As in Wu (2018), this paper uses textual analysis techniques to extract relevant information. This paper additionally constructs two indices revealing hidden patterns that traditional numerical data do not highlight. Further, it adds to this literature by arguing that beyond higher standards and stereotypes, women still face a lack of recognition of their work even when they publish high quality papers compared with their male colleagues. Moreover, in a general discrimination analysis, this paper also addresses a question raised by Hammermesh (2018), namely that merit may not always go to the rightful person.[5]

The last point is the existence of gender bias in citation patterns. Citations as well as the journal of publication (Hilmer, Ransom and Hilmer (2015), Heckman and Moktan (2019)) are commonly used measures to evaluate the quality of a paper. However, Fong and Wilhite (2012, 2017) show how citations may not necessarily reflect the merit of the cited article or are manipulated to increase the journal impact factor. Citations could therefore reflect a strategic decision (Lampe (2012)) or characterize a network (D'Ippoliti (2017)).[6]. At this general level, this paper departs from and complements the existing literature by first building a citation database over time. Second, this paper contrasts realized citations and expected citations. Third, this paper uses an alternative measure of the scientific quality of a paper.[7] Focusing on gender, Ferber (1986, 1988), Dion et al. (2018) and Koffi (2021) show that women's papers are mostly cited by women's papers. The current findings are in line with those of Ferber (1986, 1988) Dion et al. (2018) and Koffi (2021). Indeed, in an omission perspective, women's papers are more likely to be omitted by men's papers.

In addition, one key question is why we care about citations or missing citations. Ellison (2013), Hamermesh and Pfann (2012), Jensen et al. (2009) argue that citations are important

[5]For more literature on gender and academia, see Moss-Racusin et al. (2012), Chari and Goldsmith-Pinkham (2017), Teele and Thelen (2017), Ductor et al. (2018), Auriol et al. (2019), Lundberg and Stearns (2019), Hospido and Sanz (2019), Hofstra et al. (2020).

[6]Additional evidence of the networking effect can be found in this study by Colussi (2017), which demonstrated that publications in a journal are influenced by the social connections, faculty colleagues and Ph.D. students.

[7]This measure is based on a similar measure built in the patent literature. See Koffi and Panousi (2019), Kelly et al. (2019). In the same line, Hofstra et al. (2020) construct a measure of paper quality using textual information from Ph.D.. theses.

in determining labor market outcomes. They signal reputation and are important for hiring, salaries, tenure, and grants. In line with those findings, this paper further shows that being omitted influences an author's future publication possibilities. Missed authors tend to have lower chance of publishing in top economic journals.

Last, like Kelly et al. (2018), Koffi and Panousi (2019), and Hofstra et al. (2021), I rely on document topical content to build an innovation-related index. Finally, while Kelly et al. (2018) and Koffi and Panousi (2019) focuses on patents and Hofstra et al. (2021) on scientific theses, I look at publications in economic journals, which could better capture innovation in economics, by contrast to theses. In fact, theses need most of the time to be polished via publications before testifying the quality of a given paper. Finally, this paper makes an additional methodological contribution by constructing the omission index.

The remainder of the paper is organized as follows. First, the data are described and evidence of gender bias in omissions is provided. Second, the main empirical strategy is described. The paper ends with a discussion of implications of the empirical results.

3 Data description

The raw data are collected from two main websites, the Web of Science (WoS) database and IdeasRepec (IR). Together, these sites constitute the largest depository of academic research in economics. A web crawling algorithm is used to collect information from Ideas Repec (IR). This information is then organized into a novel database.

First, a corpus is created from all papers published in the top 16 journals in economics over the period 1991-2019. Details about journal ranking can be found in Laband and Piette (1994), Kalaitzidakis et al. (2003, 2011), Kodrzycki and Yu (2006), Engemann and Wall (2009), Bornmann et al. (2018), Thomson and Reuters Clarivate Analytics, and IR.[8] As is well known, published papers are submitted to a range of controls by reviewers so that they contain all relevant information concerning the prior literature. In all, the sample includes the five general-interest journals traditionally considered as the "top 5" (t5), i.e. American Economic Review, Econometrica, Journal of Political Economics, Quarterly Journal of Economics and Review of Economics Studies, as well as 11 renowned special-interest or field journals. The corpus excludes proceedings papers, comments, articles of less than three pages, book reviews, bibliographical items, articles without references and without abstracts, editorial material, letters, and corrections. The WoS database is merged with the IR database using the title of the article, the journal of publication, and the authors' last names. Because about 40% of authors' appellations on WoS consist of initials and last names, the authors' full names were validated

[8]The results do not depend on this selection. I am further increasing the set of papers to over 100 economy papers.

by the *Cited Reference API*. Overall, the merged database contains 24,033 papers and their associated information. There are also 914,371 references with an average of 38 references per article, where about 30% of the references (up to 50% in recent years) are to top 16 papers, and these are the ones used for comparisons.

Third, the gender of the authors is determined via a combination of automated algorithms and hand-collection efforts. First, *Genderize.io*, a built-in algorithm for gender attribution, is employed. The algorithms are based on an in-built large databases of names collected from the US census, from international dictionaries, and from social media. So the algorithm yields the probability that a certain first or last name is associated with the male or female gender. This library of names was then augmented in three ways. First, via the merging of a database of inventors' names from the World Intellectual Property Organization (WIPO), for an additional eight million names, as well as the merging of the IR list of names of the top 10% of female economists.[9] Second, I check several names manually via web searches.

Fifth, the "gender composition" of each team of coauthors is identified. A paper is identified as male-authored, if all the coauthors are men (75% of the articles in the database). A paper is identified as female-authored, if all the coauthors are women (5%). A paper is identified as mostly-male-authored, if most of the coauthors are men. A paper is identified as mostly-female-authored, if most of the coauthors are women. In the early 1990s, papers with at least one female author constituted only 10% of published papers, whereas in recent years this number is closer to 30%. However, the share of female-authored articles has remained constant since 2010. The gender composition of the authors differs systematically across fields. For example, in labor economics and in the economics of education, about 8% of the papers are female-authored and about 23% have at least one female coauthor, compared to 3% and 15%, respectively, in the fields of theory, finance, and macroeconomics.[10]

4 Similarity and omission indexes

This section presents the construction of the omission index based on the text analysis. Each paper is linked to the chronological sets of pre-existing and subsequent papers using commonalities in the topical content of each pair of papers. In turn, the topical content is culled from titles, abstracts, keywords and, in more advanced analysis, from the text of the paper. These so-called textual data are cleaned and then taxonomized into sets of words. The set of words includes individual words as well as word expressions (collocations or n-grams).

[9] IR, January 2019, https://ideas.repec.org/top/top.women.html

[10] These findings are similar to those in Card et al.(2019).

4.1 Term frequency-Inverse document frequency

The Term frequency - Inverse document frequency (TFIDF) is a metric often used in machine learning to identify the relative frequency of a word in a corpus or collection of documents. For each word w in each paper p, the TF is therefore computed as:

$$TF(w,p) = \frac{Card(w \in p)}{Card(p)} \tag{1}$$

where $Card(w \in p)$ is the number of times the word w appears in paper p, and $Card(p)$ is the cardinal of p or the number of words in paper p.

The inverse document frequency (IDF) is defined as the logarithm of the inverse ratio of the number of documents in which a word appears over the total number of documents in the corpus. Let C be the corpus or the set of all documents in the database and $Card(C)$ the cardinal or the number of papers in the corpus C. The IDF is then computed as:

$$IDF(w) = -\log\left(\frac{\sum_P \mathbb{1}_{w \in P}}{Card(C)}\right) \tag{2}$$

Thus, the words that appear in every document will have IDF = 0, whereas the words that occur less frequently in the corpus will have a high IDF, because they are more informative for assessing similarities across documents.

The TFIDF of a word is then the product of the TF of the word times the IDF of the word, or $TFIDF = TF \cdot IDF$.

4.2 Similarity index

The similarity index, which measures the textual or conceptual similarities across two papers, is basically a cosine similarity distance measure. The cosine similarity is a measure of similarity between two non-zero vectors of an inner product space. It measures the cosine of the angle between the vectors, where the cosine of a 0-degree angle is 1, and the cosine of a 90-degree angle is 0. Each of the two papers to be compared is represented by a vector based on the TFIDF of each word. Let U and V be the respective vector representations of papers p and p':

$$\lambda_{p,p'} = cos(p,p') = \frac{U.V}{\|U\|\|V\|} \tag{3}$$

Clearly, $cos(p,p') \in [0,1]$.

4.3 Omission index

Next, the relevant prior literature of paper p, denoted by $\mathcal{P}_p$, is defined as the n-papers, denoted by p_i, with the highest cosine:

$$\mathcal{P}_p = \{p_1, p_2, ..., p_n\} \text{ such that for } i \in [0, n], \lambda_{p,p_i} > \lambda_{p,p'}, \quad \forall p' \in C \backslash \{p_1, p_2, ..., p_n\} \tag{4}$$

The preferred specification uses $n = 5$, thereby examining which out of the top-five most related prior papers are omitted from the references of a current paper. However, the qualitative results are robust to higher values of n. Next, the omission index for the comparison between similar papers p and p', of which p' was published first, is a binary variable, denoted by $omit_{p,p'}$, which takes the value of 1 if paper p cites paper p' in its references, and 0 otherwise:

$$omit_{p,p'} = \begin{cases} 1 \text{ if p does not cite p' conditional on p' in } \mathcal{P}_p \\ \\ 0 \text{ if p cites p' conditional on p' in } \mathcal{P}_p \end{cases}$$

This index therefore determines if the relevant prior literature is included in the references of a current paper or not, and to what extent.

5 Omission and gender: Empirical analysis

The previous section presented empirical evidence suggesting that there is a potential role of gender for determining omissions of prior related literature from the references of a current paper. This section proceeds with a rigorous empirical analysis that controls for several factors that may influence the observed omission patterns in economic publications.

5.1 Benchmark probability model

Assume that paper i was published in year t and that paper j was published in year t'. Assume that paper j belongs in the relevant prior literature of paper i, according to the similarity index. Let $gender_j$ be the variable that defines the gender of paper j's authors. This is the variable of interest. To make the papers as similar as possible, the estimation equation includes a wide range of control variables. Let Z_j^1 and Z_i^2 be sets of controls for papers j and i, such as the journal of publication, the authors' affiliation, the number of authors, and the number of references. Controlling for the journal is a way of conditioning on the quality of the paper. Similarly, the affiliations of the authors make it possible to exclude the fact that a potential bias could be because women have less visibility if they are more likely to be affiliated to lower-ranked institutions. The number of authors can also affect the probability to cite other authors

because of the increasing size of the network as the number of authors grows. The number of references of an article makes it possible to exclude the fact that the bias is systematic with articles with few references which will, therefore, choose just a handful of articles to be cited. Let $Z^3_{i,j}$ be a set of controls about observed commonalities across papers i and j, such as the primary field of study. Let $Z^4_{t,t'}$ be a set of controls about the year of publication of each paper. The number of years between the cited and the citing papers can also affect the likelihood of a paper being cited. The determinant of a paper omission are investigated given that this paper is similar to the citing one. Then, the probability of paper i omitting paper j, when it should have cited it according to the similarity index, termed $omit_{ij}$, is given by the following logit model (or a corresponding linear probability model):

$$omit_{it,jt'} = \beta_0 + \beta_1 gender_j + \beta_2 Z^1_j + \beta_3 Z^2_i + \beta_4 Z^3_{i,j} + \beta_5 Z^4_{t,t'} + \epsilon_{it,jt'} \tag{5}$$

where standard errors are clustered at the paper level.[11] The coefficient of interest is β_0 which captures the relative omission of female-authored paper (or any variable capturing the propensity of female authors in an article) compared with male-authored paper. In what follows, β_0 could be interpreted in percentages when the logarithm of the odds ratio is considered or in percentage point when the marginal probability is considered.

The results of this estimation are presented in Table 1, for a number of different specifications and controls. The dependent or outcome variable is the probability of omission. It captures the probability that paper i cites paper j in the data, given that j is in the relevant prior literature of i, according to the similarity index.

The variable "female" in the first row refers to an all-female author team (solo or multiple authors). The associated coefficient is the odds ratio that prior relevant paper j is omitted from the citations of paper i, when the author team of paper j is all female, compared to all male. On average, the coefficient is estimated between 20% and 30% and and it is always statistically significant at 1% level. In other words, the odds of being omitted from the references are 20% to 30% higher for papers written by all-female teams, compared to those by all male-teams.

Overall, the results for the main variables of interest are significant and quantitatively similar across specifications. This finding is robust to the inclusion of fixed effects for the field, for the institutional affiliation of the authors of i and j, and for the journal and year of publication of paper i.[12] The other variables have the expected sign. For example, publishing in a top 5 journal reduces the odds of omission by 60% on average.

Column (1) of table 2 shows the marginal probability of omission for all-female teams which is 2.6 percentage points (pp). The probability to get cited conditional on being in the most

[11] We cluster standard errors at the citing paper level and/or at the cited/omitted paper level.

[12] When there are multiple coauthors, the paper's affiliation is taken to be the affiliation of the coauthor at the highest-ranked institution.

similar set is reduced by approximately 3 pp for female-authored papers compared to male-authored papers. This represents almost 15% of the mean of citation conditional on being in the most similar set of 18%. A simple back of the envelope exercise reveals that for one citation of an all-male team, an all-female team will earn 0.84 citations.[13]

5.2 Two-sided gender

This section examines in more detail the role of the gender structure of the citing and cited papers on the probability of omission. The dependent variable is the same as before. The controls now include a number of cross-gender variables:

$$omit_{it,jt'} = \tilde{\beta}_0 + \tilde{\beta}_1 gender_j + \tilde{\beta}_2 Z_j^1 + \tilde{\beta}_3 Z_i^2 + \tilde{\beta}_4 Z_{i,j}^3 + \tilde{\beta}_5 Z_{t,t'}^4 + \tilde{\beta}_6 \cdot gender_i + \tilde{\beta}_7 \cdot gender_i \cdot gender_j + \epsilon_{it,jt'} \tag{6}$$

In an ideal setting, the interaction effect reflects the difference-in-differences in the relative omission bias for a paper j written by an all-female team versus a paper j written by an only-male team, when paper i is written by all-women relative to when paper i is written by all-men.

The results are presented in Table 2. The variable *female j* indicates only female authors in the cited paper (solo or all-female team). The variable *female i* indicates only female authors in the citing paper (solo or all-female team). The variable $A1f_j$ indicates at least one female author in the cited paper. The variable $A1f_i$ indicates at least one female author in the citing paper. The variables *female j* $\cdot$ *female i*, *female j* $\cdot$ $A1f_i$, $A1f_j$ $\cdot$ *female i* and $A1f_j \cdot A1f_i$ are cross-variables for citing and cited/omitted papers.

Let us consider column (2), in which the citing paper i has an all female team, *female i*, and paper j has an all female team, *female j*. First, the coefficient on *female j* corresponds to $\tilde{\beta}_1 = 0.046$ and it is statistically significant. It means that having only male authors in citing paper i increases the probability to get omitted for an all-female relevant paper j by 4.6 pp, compared to an all-male paper j. In other words, conditional on an all-male citing team, the probability of omission is 5 pp higher for female papers than for male papers.

Second, the coefficient on *female i* corresponds to $\tilde{\beta}_6 = 0.009$, which however is not statistically significant. It means that, for an all-male relevant paper j, the probability to be omitted is the same, regardless of whether citing team i is all-female or all-male.

Third, the coefficient on *female j* $\cdot$ *female i* corresponds to $\tilde{\beta}_7 = -0.103$ and it is statistically

[13]Even if the study is not about wage, it is convenient to make a parallel with the gender pay gap. Recent news coverage in ABC news shows that women earned 84.7 cents for every dollar earned by their male counterparts in 2019: https://abcnews.go.com/Business/gender-pay-gap-persists-executive-level-study-finds/story?id=75945000. Goldin (2014) shows that a woman earns approximately 0.77 dollars for every 1 dollar earned by a man. Freund et al. (2016) find in a sample of faculty followed over 17 years that women continued to earn 90 cents for every dollar that a man earned.

significant. It means that having only female authors in citing paper i decreases the probability to get omitted for a relevant all-female paper j by around 10 pp, compared to an all-male paper i. In other words, a complete change in the gender structure of the authors of i from male to female is associated with a substantial, statistically and economically, increase of 10 pp in the probability of citation for relevant female-authored papers.

5.3 Heterogeneous effects

Tables 3 to 4 report the estimates for different subgroups.

Top 5 versus non top 5: Table 3 shows that switching from only males to only females in i reduces the probability to be omitted for paper j published in a top 5 (respectively non top 5) written by only women by 10 percentage points (respectively 10 percentage points). The omission bias is present in top 5 and in non top 5 journal. Thus, even females publishing in top 5 journals are not exempted from the omission bias.

Split by affiliation types: Furthermore, switching from only males to only females in i reduces the probability to be omitted for paper j from a top tier affiliation written by only women by 9 pp. Overall, switching from only males to only females in paper j from a top affiliation, increases the probability to get omitted by 3 pp when i is written by only males. Similarly, switching from only males to only females in i reduces the probability to be omitted for paper j from a mid tier affiliation written by only women by 12 pp. Overall, switching from only males to only females in paper j from a mid-tier affiliation, increases the probability to get omitted by 8 pp when i is written by only males. Finally, switching from only males to only females in i reduces the probability to be omitted for paper j from a low tier affiliation written by only women by 10 pp. Overall, the omission of females relative to males tends to be bigger when comparing papers written by males and those written by female from mid-tier institutions.

Split by field: Table 4 examines the robustness of the qualitative result $\tilde{\beta}_0 > 0$ and significant, $\tilde{\beta}_6 < 0$ and significant, across different fields in economics. As can be seen, the pattern is especially strong in columns (1)-mathematical economics and econometrics, (2)-microeconomics, (3)-macroeconomics, (4)-international economics, and (5)-finance. Those fields show higher probability of omission of relevant papers that include at least one female, when the citing team consists mostly of men, compared to mostly women. By contrast, the omission of teams with at least one female is not as strong in the fields of labor and education (column (6)), industrial organization (column (7)). Overall, fields that are more theoretical and mathematical display a higher level of female-authored papers' omission when we have only male in the citing paper, while field that are more empirical display a lower level of female-authored papers'

omission when we have only male in the citing paper.[14]

5.4 Citations and Innovativeness index

The existing literature has suggested that citations are a noisy signal of quality, for example in the case of patents. The analysis above also indicates that citations may not accurately reflect the quality of a published paper in economics, as they tend to systematically omit the contributions of female economists and groups of female economists. Therefore, this section, following Kelly et al.(2018) and Koffi and Panousi(2019), constructs an alternative index for measuring the quality of a publication in economics, the innovativeness index, using a textual and linguistic comparison across different papers. By constructing a measure of the quality which ignores the authors' willingness to refer to articles (therefore without bias in this sense), we can assess the differential relationship that exists between men and women by comparing this measure of quality with no inherent gender bias and citations where gender bias has been highlighted in the previous section.

Construction of the innovativeness index: Specifically, the new quality index, denoted by q, has two dimensions, which together capture the degree of innovativeness of a paper. First, more innovative papers are more distinct from prior related papers, in that they offer a novel idea or method to the pre-existing stock of knowledge. Second, more innovative papers are more likely to influence the framework or the methodology of future papers. In other words, the concept of innovation used here reflects the novelty as well as the influence of a publication. Papers with high innovativeness index are both novel (distinct from prior papers) and influential (similar to future papers). The "novelty" of a paper is captured by a backward-similarity (BS) index, which is the sum of pairwise relative cosine similarities of paper p, published in t, with papers p' published in $t-T$: $BS^0_{-T}(p) = \Sigma_{p'}\lambda_{p,p'}$. The "influence" of a paper is captured by a forward-similarity (FS) index. The forward similarity is the sum of pairwise cosine similarities of paper p, published in t, with papers p' published in $t+T$: $FS^T_0(p) = \Sigma_{p'}\lambda_{p,p'}$.

Therefore, the innovativeness index q will be a combination of the novelty and of the impact of a paper, as measured, respectively, by the backward and the forward similarity

$$q^T(p) = \frac{FS^T_0(p)}{BS^0_{-T}(p)}$$

In the benchmark specifications, $T = 5$, but the results are robust to alternative windows. The q-index is a measure of the underlying scientific innovativeness of a paper. If a paper has a

[14]The full paper discusses robustness of the methods, increases the set of journal, controls for the methodological style, the distance metric, differential effect over time, the share of females by field, 3-digit level,...

high forward similarity (high numerator) and a high backward similarity (high denominator), this could mean that the paper is a follower among other followers in a research area. Hence, it will have a low q-index, compared to a paper with a high forward similarity and a relatively low backward similarity. In that respect, it operates like the citations measure.

Innovativeness, citations and gender This section examines the relationship between the number of citation, the innovativeness index, and the gender of the paper.

$$C_{pt} = a_1 \cdot Q_{pt} + a_2 \cdot gend_{pt} + a_3 \cdot gend_{pt} \cdot Q_{pt} + a_4 \cdot Z_p + \theta_t + \epsilon_{pt} \tag{7}$$

C_{pt} is the logarithm of the number of citations of paper p published in year t; Q_{pt} is the innovativeness index of paper p published in year t; $\hat{Z}_p^1$ is a set of paper-level controls, such as the number of coauthors, the coauthors' affiliation,the field of the paper,the journal of publication, and NBER membership; $\hat{Z}_t^2$ captures publication-year fixed effects. Field- and journal- fixed effects are included. The standard errors are clustered by publication year and journal. The variable *gend* takes the value 1 if there is at least one female author, and 0 if the paper is written by all-male teams; or a dummy variable that take 1 if the paper is written by only women and 0 if the paper is written all-male teams (solo and co-authored).[15] Overall, there is a positive and statistically significant correlation between the innovativeness index and the number of citations received by an article. However, there are strong heterogeneous effect depending on the gender of the authors on the paper the paper as shown in figure 2. Plot (a) shows that for the same innovativeness index value, the male paper will get more citations than the female paper. For the same number of citations (let us say the mean value of citation), the innovativeness of the female paper is higher by, on average, 0.1 unit. In fact, the innovativeness index is near 0.8 (close to the 50th percentile) for male and near 0.85 (close to the 75th percentile). Panel (b) focuses only on top 5 publications. For the same number of citations at the mean, the male-authored papers are at the 70th percentile of the quality distribution, the female-authored papers are at the 90th percentile. The gap is persistent even considering top 5 publications.

Counterfactual Analysis: Compensating citations with Omissions Panel (c) of figure 2 is a key result showing the effect of the bias generated by the omissions. I realize a counterfactual analysis, in which the total number of omissions is added to the number of citations. This is interpreted as the number of citations an article would have received if all the papers with which it shares the most similarities had cited it. The gap corresponding

[15]The innovativeness index is here standardized. For each value, the mean is subtracted and the result is divided by the standard deviation. This is helpful for the interpretation of the regression coefficient in estimation with a cross variable and a continuous variable. The results are not dependant of the standardisation.

to more than 20 percentile in terms of innovativeness index disappears completely with this compensation. In other words, if it were not because of the omission bias, the standards to be cited would have been the same for women and men.

6 Conclusion

Women are still underrepresented in math-intensive fields. But very few studies have tried to analyze whether the potential problem lay in the lack of recognition of their work. This paper has, therefore, addressed the issue using data on Economics. It shows that women have a higher probability of being omitted from references. This problem is persistent, even for women publishing in top-journal in the same way as men. However, the most vulnerable population appears to be women of mid-tier institutions. Indeed, the bias is lower comparing female-authored papers from top-tier institutions to similar male-authored papers.

However, what drives the gender omission pattern? The empirical results give insights to this question. First, the analysis of "peer effects" and androgynous first names versus female first names clearly shows that the results are not driven by a lack of information on the existence of articles written by women. Rather, they are related to the fact that women write these articles. So this rules out the assumption of lack of information. Second, the two-sided gender effect also clarifies that men and women have different preferences when citing other women. Moreover, switching from men to women in the citing paper does not reduce men's omission. Therefore, women's behavior is not that consistent with a homophilic pattern. The absence of a strong effect regarding the link between the history of publications and the female's omissions impedes to give a statistical discrimination interpretation. At the same time, the relatively lower omission bias in female-dominated sub-fields hides both a compliance and a recognition pattern.[16]

Overall, the results of this paper have strong implications on the policy side. By highlighting this uncovered phenomenon, it raises the awareness of the scholars on this issue. For example, editors and referees may pay more attention to the omission from references when receiving a paper to evaluate. Institutions may take into account this relative under-citation of female-authored papers in taking tenure decisions or attributing grants. The omission index could constitute a barometer in many instances.

Finally, the conclusion of this study is not only restricted to the economic field. It aims at a more general horizon by explaining how discriminating factors can influence the perception of individual works even when they are more deserving than others. Besides recognizing how inequality issues can affect socio-economic factors, the current paper advocates for better inclusion of minorities to increase overall productivity.

[16]See the additional tables in the appendix.

References

Abrevaya, J., Hamermesh, D. S., 2012. Charity and favoritism in the field: Are female economists nicer (to each other)?. Review of Economics and Statistics, 94(1), 202-207.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., Feng Lu, S., 2017. Economic Research Evolves: Fields and Styles. American Economic Review: Papers and Proceedings 107(5). 293-297.

Angrist, J., Azoulay, P., Ellison, G., Hill, R., Feng Lu, S., 2017. Inside job or deep impact? Using extramural citations to assess economic scholarship. National Bureau of Economic Research.

Antecol, H., Bedard, K., Stearns, J., 2018. Equal but inequitable: who benefits from gender-neutral tenure clock stopping policies?. American Economic Review, 108(9), 2420-41.

Antecol, H., Cobb-Clark, D. A., 2013. Do psychosocial traits help explain gender segregation in young people's occupations?. Labour Economics, 21, 59-73.

Auriol, E., Friebel, G., Wilhelm, S., 2019. Women in European Economics. Mimeo.

Bayer, A., Rouse, C., 2016. Diversity in the Economics Profession: A New Attack on an Old Problem. Journal of Economic Perspectives, 30(4): 221-42.

Blau, F. D., DeVaro, J., 2007. New evidence on gender differences in promotion rates: An empirical analysis of a sample of new hires. Industrial Relations: A Journal of Economy and Society, 46(3), 511-550.

Bornmann, L., Butz, A., Wohlrabe, K., 2018. What are the top five journals in economics? A new meta-ranking. Applied Economics, 50(6), 659-675.

Card, D., DellaVigna, S., 2013. Nine Facts About Top Journals in Economics. Journal of Economic Literature 51(1):144-161.

Card, D., DellaVigna, S., Funk, P., Iriberri, N., 2019. Are referees and editors in economics gender neutral? Quarterly Journal of Economics (forthcoming).

Ceci, S. J., Ginther, D. K., Kahn, S., Williams, W. M., 2014. Women in academic science: A changing landscape. Psychological Science in the Public Interest, 15(3), 75-141.

Chari, A., Goldsmith-Pinkham, P., 2017. Gender representation in economics across topics and time: Evidence from the NBER summer institute (No. w23953). National Bureau of Economic Research.

Colussi, T., 2017. Social Ties in Academia: A Friend is a Treasure. Review of Economics and Statistics , forthcoming.

D'Ippoliti, C. (2017). 'Many-Citedness': Citations Measure More Than Just Scientific Impact. Institute for New Economic Thinking Working Paper Series, (57).

Dahl, G., Kotsadam, A., Rooth, D. O., 2018. Does integration change gender attitudes? The effect of randomly assigning women to traditionally male teams (No. w24351). National Bureau of Economic Research.

Dietz, L., Bickel, S., Scheffer, T., 2007. Unsupervised prediction of citation influences. In Proceedings of the 24th International Conference on Machine Learning , pp. 233–240.

Dion, M., Sumner, J., Mitchell, S. M., 2018. Gendered citation patterns across political science and social science methodology fields. Political Analysis 26(3)312–327.

Ductor, L., Goyal, S., Prummer, A., 2018. Gender and collaboration. Working Paper, School of Economics and Finance, Queen Mary University of London, No. 856.

Ellison, G., 2002. The slowdown of the economics publishing process. Journal of Political Economy 110, 947-993.

Ellison, G., 2002. Evolving standards for academic publishing: A q-r theory. Journal of Political Economy 110, 994-1034.

Ellison, G., 2011. Is Peer Review in Decline?. Economic Inquiry 49(3), 635-657.

Ellison, G., 2013. How Does the Market Use Citation Data? The Hirsch Index in Economics. American economic Journal: Applied Economics 5(3), 63-90.

Engemann, K., Wall, H., 2009. A journal ranking for the ambitious economist. Federal Reserve Bank of St. Louis Review, 91(3), 127-139

Ferber, M., 1986. Citations: Are They an Objective Measure of Scholarly Merit? Signs 11 (2), pp. 381–389.

Ferber, M., 1988. Citations and Networking. Gender and Society 2 (1), pp. 82–89.

Fong, E.A., Wilhite, A.W., 2017. Authorship and citation manipulation in academic research. PLoS ONE 12(12): e01897394.

Freund, K. M., Raj, A., Kaplan, S. E., Terrin, N., Breeze, J. L., Urech, T. H., Carr, P. L., 2016. Inequities in academic compensation by gender: a follow-up to the National Faculty Survey Cohort Study. Academic medicine: journal of the Association of American Medical Colleges, 91(8), 1068.

Gibson, J., Anderson, D. L., Tressler, J., 2017. Citations or Journal Quality: Which Is Rewarded More in the Academic Labor Market?. Economic Inquiry 55 (4): 1945–65.

Ginther, D. K., Kahn, S., 2004. Women in economics: moving up or falling off the academic career ladder?. Journal of Economic perspectives, 18(3), 193-214.

Hamermesh, D., 2018. Citations In Economics: Measurement, Uses, and Impacts. Journal of Economic Literature 56, 115-156.

Hamermesh, D., Pfann A., 2012. Reputation and Earnings: The Roles of Quality and Quantity in Academe. Economic Inquiry 50 (1):1–16.

Heckman, J., Moktan, S., 2018. Publishing and Promotion in Economics: The Tyranny of the Top Five.Journal of Economic Literature (forthcoming).

Hengel, E., 2017. Publishing while Female. Are women held to higher standards? Evidence from peer review.

Hengel, E., Moon, E., 2019. Gender and quality at top economics journals.

Hilmer, M., Ransom, M., Hilmer, C., 2015. Fame and the Fortune of Academic Economists: How the Market Rewards Influential Research in Economics. Southern Economic Journal 82 (2): 430–52.

Hofstra, B., Kulkarni, V. V., Galvez, S. M. N., He, B., Jurafsky, D., McFarland, D. A., 2020. The Diversity–Innovation Paradox in Science. Proceedings of the National Academy of Sciences, 117(17), 9284-9291.

Hospido, L., Sanz, C., 2019. Gender gaps in the evaluation of research: evidence from submissions to economics conferences.

Hunt, J., 2016. Why do women leave science and engineering?. ILR Review, 69(1), 199-226.

Hunter, L., Leahey, E., 2008. Collaborative research in sociology: Trends and contributing factors. The American Sociologist, 39(4), 290-306.

Jensen, P., Rouquier, J-B., Croissant,Y., 2009. Testing Bibliometric Indicators by Their Prediction of Scientists Promotions. Scientometrics 78 (3): 467–79.

Kalaitzidakis, P., Mamuneas, T., Stengos, T., 2003. Rankings of Academic Journals and Institutions in Economics. Journal of the European Economic Association 1(6): 1346-1366.

Kalaitzidakis, P., Mamuneas, T., Stengos, T., 2011. An updated ranking of academic journals in economics. Canadian Journal of Economics, 44(4), 1525-1538.

Kelly, B., Papanikolaou, D., Seru, A., Taddy, M., 2018. Measuring technological innovation over the long run (No. w25266). National Bureau of Economic Research.

Kodrzycki, Y., Yu, P., 2006. New approaches to ranking economics journals. The BE Journal of Economic Analysis and Policy, 5(1).

Koffi, M., Panousi, V., 2019. Patents, Innovation and Growth in Canadian Pharmaceuticals.

Kosnik, L. R., 2015. What have economists been doing for the last 50 years? A text analysis of published academic research from 1960–2010. working paper.

Laband, D. N., 2013. On the Use and Abuse of Economics Journal Rankings. Economic Journal 123(570): F223–54.

Laband, D. N., Piette, M. J., 1994. The relative impacts of economics journals: 1970-1990. Journal of economic Literature, 32(2), 640-666.

Lampe, R., 2012. Strategic citation. Review of Economics and Statistics, 94(1), 320– 333.

Le, Q., Mikolov, T., 2014. Distributed representations of sentences and documents. In International conference on machine learning.1188-1196.

Lundberg, S. J., Stearns, J, 2019. Women in Economics: Stalled Progress. Journal of Economic Perspectives 33 (1), pp. 3–22.

Mauleon E, Bordons M. Authors and Editors in Mathematics Journals: a gender perspective. International Journal of Gender, Science and Technology. 2012; 4(3):267-293.

Moss-Racusin, C., Dovidio, J., Brescoll, V., Graham, M., Handelsman, J., 2012. Science faculty's subtle gender biases favor male students. Proceedings of the National Academy of

Sciences, 109(41), 16474-16479.

Pennington, J., Socher, R., Manning, C. D., 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 1532-1543.

Preston, A. E., 1994. Why have all the women gone? A study of exit of women from the science and engineering professions. The American Economic Review, 84(5), 1446-1462.

Rossiter, M. W. (1993). The Matthew Matilda effect in science. Social studies of science, 23(2), 325-341.

Sarsons, H., 2019. Gender differences in recognition for group work. working paper.

Kusner, M., Sun, Y., Kolkin, N., Weinberger, K., 2015. From word embeddings to document distances. In International conference on machine learning (pp. 957-966).

Teele, D. L., Thelen, K., 2017. Gender in the journals: Publication patterns in political science. PS: Political Science and Politics, 50(2), 433-447.

West, J. D., Jacquet, J., King, M. M., Correll, S. J., Bergstrom, C. T,2013. The Role of Gender in Scholarly Authorship. PLOS ONE 8(7): e66212.

Wilhite, A. W., Fong, E., 2012. Coercive citation in academic publishing. Science 335:542–543.

Wu, A., 2018. Gendered Language on the Economics Job Market Rumors Forum. AEA Papers and Proceedings 108: 175–79.

Zafar, B., 2013. College major choice and the gender gap. Journal of Human Resources, 48(3), 545-595.

Zhu, X., Turney, P., Lemire, D., Vellino, A., 2015. Measuring academic influence: Not all citations are equal. Journal of the Association for Information Science and Technology, 66 (2) (2015), pp. 408-427

Appendix

Tables

Table 1: Relationship between omission and gender

	Outcome variable: Omission				
	(1)	(2)	(3)	(4)	(5)
female j	0.289***	0.253***	0.257***	0.215***	0.215***
	(0.041)	(0.042)	(0.043)	(0.043)	(0.081)
Top 5 j		-0.497***	-0.558***	-0.692***	-0.692***
		(0.018)	(0.019)	(0.020)	(0.032)
Primary field		-0.586***	-0.564***	-0.503***	-0.503***
		(0.019)	(0.019)	(0.019)	(0.024)
Years lag		0.008***	0.023***	0.028***	0.028***
		(0.001)	(0.002)	(0.002)	(0.003)
Gender Structure			-0.142***	-0.168***	-0.168***
			(0.043)	(0.045)	(0.049)
Number of Reference i			-0.050***	-0.048***	-0.048***
			(0.001)	(0.002)	(0.002)
Number of Authors i				-0.028***	-0.028***
				(0.011)	(0.010)
Institution of j FE			Y	Y	Y
Institution of i FE				Y	Y
Journal of i FE				Y	Y
Year of publication of i FE				Y	Y
Field FE				Y	Y
N	110767	110767	110767	110763	110763
R-sqr	0.002	0.021	0.064	0.075	0.075

This table shows the relationship between the omission and the gender of the omitted paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 4. *female j* represents papers written by only women. The reference variable is *male*, which represents papers written by only men (the two other gender structure -Mixed and undetermined- are added but not shown in the table to ease the reading. See appendix for more details). *Top 5 j* is binary and indicates if paper j is published in a top 5 journal or not. *Primary field* is binary and indicates if paper i and paper j have the same primary field. *Years Lag* is the difference between the publication year of paper i and the publication year of paper j. *Gender Structure* is the share of paper written by at least one female author in the relevant prior literature. *Number of Reference i* is the number of references recovered of paper i from the database. *Number of Authors i* is the number of authors in paper i. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by the citing papers, except in column (5) where the cluster

Table 2: Omission and two-sided gender

	Outcome variable: Omission					
	(1)	(2)	(3)	(4)	(5)	(6)
female j	0.026***	0.046***			0.046***	
	(0.005)	(0.006)			(0.006)	
female i		0.009				0.011
		(0.007)				(0.007)
female j · female i		-0.103***				
		(0.021)				
$A1f_j$			0.011***	0.031***		0.032***
			(0.003)	(0.003)		(0.003)
$A1f_i$				0.016***	0.015***	
				(0.003)	(0.003)	
$A1f_j \cdot A1f_i$				-0.070***		
				(0.007)		
female j ·$A1f_i$					-0.066***	
					(0.011)	
$A1f_j$· female i						-0.060***
						(0.013)
N	92105	72546	107301	103377	88759	83598
R-sqr	0.067	0.069	0.069	0.069	0.068	0.069

This table shows the relationship between the omission and the gender of the omitted paper emphasizing the gender of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 4. *female_x* represents paper x written by only women. $A1f_x$ represents paper x with at least one female author. All the specifications include controls for paper j published in a top 5 journal; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a linear probability model. The size of the sample varies because of the selection in the specification considered. For example, column (2) includes only citing papers and cited papers that are written only by females or only males. and The table displayed the marginal probabilities. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table 3: Omission and gender: By Journals and Institutions

	Outcome variable: Omission				
	Journal j		Institution j		
	(1)	(2)	(3)	(4)	(5)
	Top5	Non Top5	Top tier	Mid tier	Low tier
female j	0.053***	0.039***	0.028**	0.080***	0.029**
	(0.011)	(0.006)	(0.014)	(0.011)	(0.013)
female i	0.013	0.013*	0.031***	0.001	0.024*
	(0.012)	(0.008)	(0.012)	(0.013)	(0.012)
female j · female j	-0.101***	-0.099***	-0.090**	-0.123***	-0.098**
	(0.039)	(0.024)	(0.040)	(0.040)	(0.044)
N	25433	47113	24543	18306	13370
R-sqr	0.130	0.094	0.117	0.094	0.079

This table shows the relationship between the omission and the gender of the omitted paper emphasizing the gender of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 4. *female* i represents paper x written by only women. All the specifications include controls for paper j published in a top 5 journal; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a linear probability model. The table displayed the marginal probabilities. The total number of observations in the institution section does not add up to 72546 (total number of observations in two-sided case with only females in the citing and the cited/omitted) because some affiliations are missing in the database. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table 4: Omission and gender: Field of study

	Outcome variable: Omission							
	(1) Mathe-matical	(2) Micro	(3) Macro	(4) International Economics	(5) Finance	(6) Labour - Education	(7) IO	(8) Other fields
$A1f_j$	0.030***	0.035***	0.043***	0.040***	0.023***	0.000	0.014	0.029***
	(0.008)	(0.009)	(0.011)	(0.010)	(0.007)	(0.010)	(0.030)	(0.009)
$A1f_i$	0.017**	0.017*	-0.005	0.024**	0.020***	0.008	-0.001	0.012
	(0.008)	(0.009)	(0.012)	(0.010)	(0.007)	(0.009)	(0.036)	(0.010)
$A1f_j \cdot A1f_i$	-0.092***	-0.082***	-0.072***	-0.075***	-0.078***	-0.011	-0.004	-0.034*
	(0.018)	(0.020)	(0.025)	(0.018)	(0.014)	(0.016)	(0.065)	(0.019)
N	17519	17119	8549	9851	27881	10858	998	10602
R-sqr	0.135	0.107	0.116	0.127	0.089	0.115	0.098	0.117

This table shows the relationship between the omission and the gender of the omitted paper emphasizing the gender of the citing paper and splitting by primary field of the citing paper. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 4. $A1f_x$ represents paper x with at least one female author. All the specifications include controls for paper j published in a top 5 journal; the relative cosine; paper i and paper j having the same primary field; difference between the publication year of paper i and the publication year of paper j; the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The field section is defined based on the Journal of Economic Literature (JEL) codes. The category other fields includes public economics, agricultural economics, general economics, urban economics, law and economics, business administration, economic history, economics systems. The equations are estimated using a linear probability model. The table displayed the marginal probabilities. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Table 5: Omission and gender: Peer effects

	Outcome variable: Omission				
	(1)	(2)	(3)	(4)	(5)
female j	0.021***	0.021***			0.017***
	(0.005)	(0.005)			(0.005)
Same Affiliation	-0.075***	-0.075***	-0.076***	-0.075***	
	(0.010)	(0.010)	(0.009)	(0.010)	
female j · Same Affiliation		-0.008			
		(0.052)			
At least one female j			0.010***	0.010***	
			(0.003)	(0.003)	
At least one female j · Same Affiliation				-0.003	
				(0.022)	
Connection					-0.098***
					(0.005)
female j · Connection					0.012
					(0.027)
N	88,175	88,175	102,664	102,664	88,175
R-sqr	0.066	0.066	0.068	0.068	0.071

This table shows the relationship between the omission and the gender of the omitted paper emphasizing the effect of being in the same affiliation. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The relevant prior literature is defined by equation 4. f_j represents paper j written by only women. The control variables include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. This regression excludes the self-citation. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

Table 6: Omission and gender: Androgynous versus female

	Outcome variable: Omission				
	Adrogynous	Adrogynous	Adrogynous	Adrogynous	"non typical white"
	(1)	(2)	(3)	(4)	(5)
	Proba <0.4	Proba <0.5	Proba <0.6	Proba <0.7	
Baseline controls	-0.319**	-0.270*	-0.252*	-0.478***	-0.166
	(0.152)	(0.145)	(0.138)	(0.124)	(0.103)
N	5,300	5,300	5,300	5,300	5,300
R-sqr	0.142	0.142	0.142	0.144	0.142

This table shows the relationship between the omission and the gender of the omitted paper, emphasizing the effect of the name connotation. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The estimation focuses on cases with only females omitted, distinguishing between names known as females with a higher probability and names known as females with a relatively lower likelihood. For example, column (1) presents the results for female names with a probability lower than 0.4, putting names known as female with probability greater than 0.4 as a reference. Therefore, the coefficient has to be read relative to the latter category. The last column presents a similar analysis comparing typical non-white names to white names. For example, Asian names will be in the typical non-white sample and European and American names in the typical white sample. The control variables are the same as in the baseline and include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect, cosine. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$* * * = p < 0.01$)

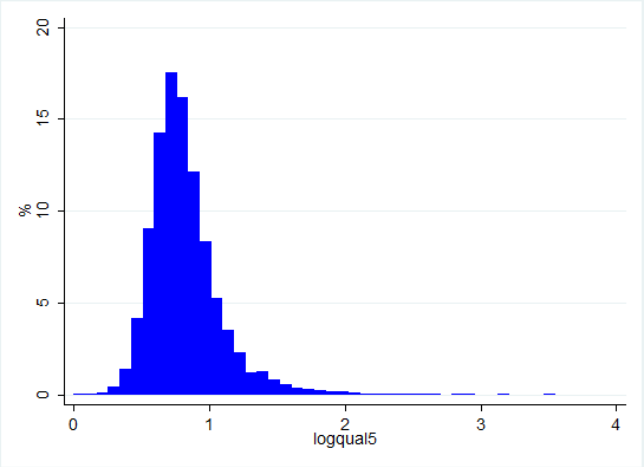
Table 7: Omission and gender: Seniority

	Outcome variable: Omission						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
female j	0.205***	0.177***	0.145***	0.204***	0.185***	0.154***	0.144***
	(0.043)	(0.043)	(0.043)	(0.043)	(0.044)	(0.044)	(0.055)
NBER_j	-0.181***					-0.167***	
	(0.026)					(0.026)	
max top5_j		-0.021***					
		(0.002)					
max papers_j			-0.014***			-0.014***	-0.013***
			(0.001)			(0.001)	(0.001)
superstar_j				-0.222***		-0.128***	
				(0.033)		(0.035)	
senior age_j					-0.008***	0.006***	
					(0.002)	(0.002)	
N	110763	110763	110763	110763	110763	110763	92100
R-sqr	0.076	0.076	0.077	0.076	0.075	0.078	0.075

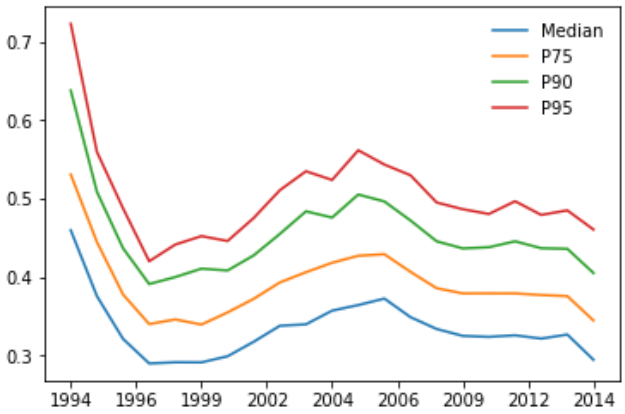
This table shows the relationship between the omission and the gender of the omitted paper, emphasizing the effect of the seniority. The dependent variable, omission, is binary and indicates whether a paper i cites a paper j in the database given that j is in the relevant prior literature of i. The other control variables are the same as in the baseline and include the share of paper written by at least one female author in the relevant prior literature; the number of references recovered from the database; the number of authors writing the paper; field fixed effect, journal fixed effects, year fixed effect, institutions fixed effect. The equations are estimated using a logit model. The odds ratio for a variable is the exponential of its given coefficient. Standard errors are clustered by citing papers and reported in parentheses. ($* = p < 0.10$, $** = p < 0.05$,$*** = p < 0.01$)

Figures

Figure 1: **Distribution of innovativeness (quality) index**



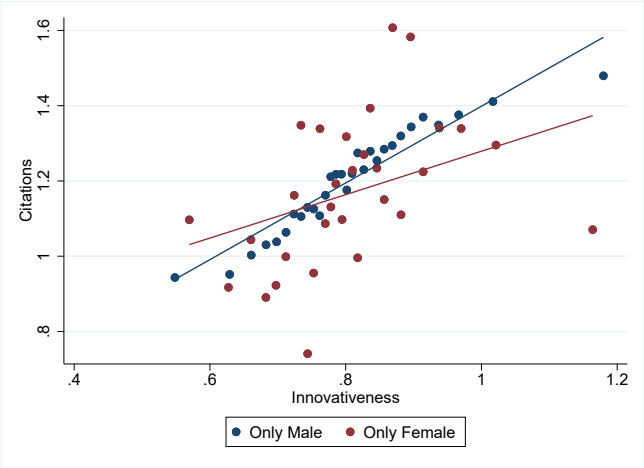
(a) Overall distribution



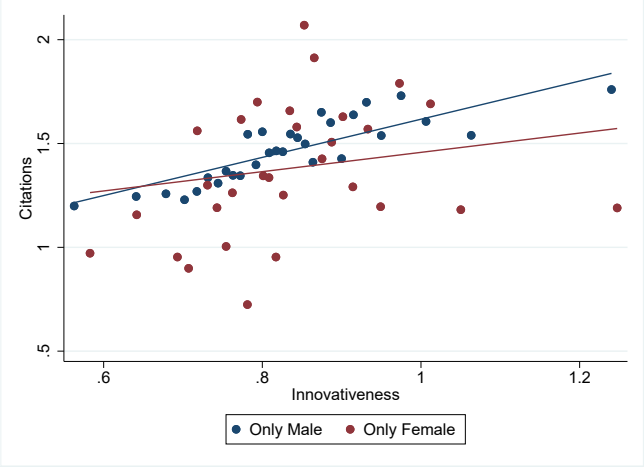
(b) Distribution over time

Panel (a) shows the overall distribution of the innovativeness index (q-index). Panel (b) shows the distribution of this index over time. The index is built following equation 5.4.

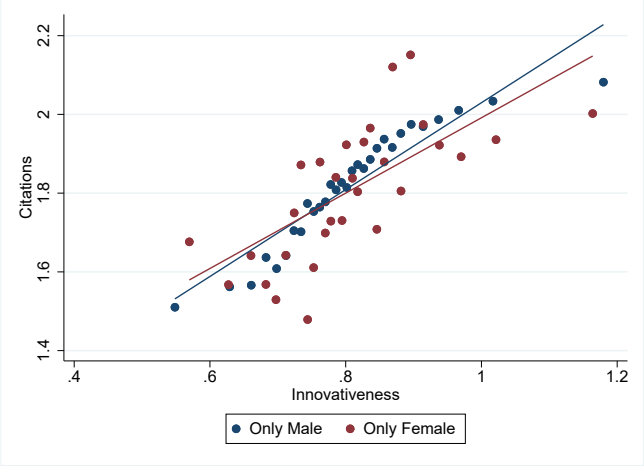
Figure 2: **Distribution of innovativeness (quality) index**



(a) Overall



(b) Only Top 5 Publications



(c) Counterfactual: citation compensating with omission cases

The figure plots the link between the number of citations and the innovativeness index for papers written by males and females. The binned scatter plot controls for journals, field, institutions, year of publications, number of authors, maximum number of publications.