

Appendix For Online Publication

Innovative Ideas and Gender (In)equality

Marlène Koffi

A Variables construction

A.1 Data Extraction

After a pre-processing step to include only full-length articles, the databases were merged using the Digital Object Identifier (DOI) or other attributes like title, journal name, publication year, and first-page number. To achieve this, the variables related to the journal name and article title were standardized to lowercase, and any punctuation and white spaces were removed.

Finally, duplicate articles were eliminated through a manual review process or by identifying similarities in titles and other attributes such as journal name, publication year, and first-page number. The initial sample spotlighted papers from 16 leading economics journals between 1985-2019, with rankings informed by several referenced studies and platforms like Laband and Piette [1994], Kalaitzidakis et al. [2003], Kodrzycki and Yu [2006], Engemann et al. [2009], Kalaitzidakis et al. [2011], Bornmann et al. [2018] as well as Thomson and Reuters Clarivate Analytics and IdeasRepec (IR). However, to show that the results do not depend on the restriction to 16 journals, I increase the sample size to encompass more than 100 economics journals.

A.2 Gender coding

Given that 48.13% of first names are missing in the Web of Science (WoS) database, merging databases offers a notable advantage in reducing the number of incomplete names. After using the algorithmic procedure, a team of undergraduate research assistants extensively searched for all the names that remained unassigned. This search encompassed 10,219 authors and involved referencing personal websites, web sources that explicitly referred to authors using pronouns, articles featuring authors' photographs, and publicly available sources. Subsequently, a second verification was carried out for observations not assigned by this initial verification. Finally, a third wave of verification involved re-verifying the assigned genders. This process resulted in the correction of 390 incorrect gender assignments.

A.3 References

The variable “References” pertains to the bibliographical references included in each published paper. The references were obtained from the Web of Science database and are identified by information

such as the authors’ names, the year of publication, the journal, the volume, and the DOI (when available). These details are used to match the references to the database initially created, enabling linking of related literature.

A.4 Institution Ranking

Author institutions were identified and extracted using the merged database, with each author potentially having multiple affiliations. These affiliations were then ranked based on the number of publications from journals in the database, as per Hengel [2022]. This more conservative method avoids dealing with changes in university rankings, while the resulting ranking aligns well with international rankings of economics departments. For example, institutions classified in the top tier in my data are (excluding the FED): Harvard University, University of California Berkeley, University of Chicago, Stanford University, MIT, Northwestern University, New York University, Columbia University, University of Pennsylvania, London School of Economics, Yale University, Princeton University, University of Michigan and Duke University. Of these, nine out of fifteen institutions also appear in the top 15 of the RePEc ranking and ten out of fifteen in the US News ranking. For comparison, the overlap between RePEc and US News is similarly nine out of fifteen.

The analysis utilizes dummies for institutional rankings. For papers with multiple co-authors, the highest-ranked institution among co-authors is considered. The granular set of dummies is used in our baseline estimations. For the heterogeneity analysis, these institutions are grouped into three: top tier (rank 1-15), middle tier (rank 16-100), and lower tier (rank above 100).

A.5 Primary Field

The primary field of each article is determined by employing a methodology similar to that utilized by Angrist et al. (2017). Specifically, the Journal of Economic Literature (JEL) codes are extracted from the Econlit database. Due to articles having multiple JEL codes, a classification algorithm is devised, drawing from articles where the first single-digit JEL code is identical for all listed JEL codes.

In more detail, the classification process encompasses articles published in the selected journals from 1985 to 2019. Areas of classification include: microeconomics, macroeconomics, finance, econometrics, international economics, labor, industrial organization, development, urban economics, economic history, law, and economics, among others.

A machine learning algorithm allocates each article to a specific field. The dataset is partitioned into two subsets: training (75% of the data) and test (25% of the data) datasets. The algorithm discerns distinctive category characteristics using the training dataset of articles bearing a single primary JEL code. For articles originating from journals with a well-established field (e.g., the Journal of Labor Economics), the journal’s field is assigned, as per Angrist et al. (2019).

The classification utilizes titles, keywords, and abstracts. The textual information was pre-

processed using a similar procedure as described in Appendix B below. Optimal hyperparameter values for classification were identified through grid search. Multiple classifiers were trained, including Logistic Regression, Random Forest Classifier, and Support Vector Machine (SVM), with the SVM performing best, achieving an 87.5% accuracy rate.

A.6 Determination of the style: Theory versus Empiric

I developed a machine learning classifier to categorize the methodology employed in articles as “theory,” “empirical,” or “econometrics.” For each article with accessible full text, This classification uses specific keywords within articles, following guidelines from Card et al. (2020). For example, “theorem” indicates a theoretical approach, while “randomized experiment” suggests empirical work.

In the dataset used for constructing and assessing the algorithm, articles for which more than 90% of the total words fall within a particular category are classified as belonging to that respective category. To distinguish between theoretical articles and theoretical econometrics articles, I identified articles with the primary JEL code “C” and categorized those related to theoretical econometrics.

Subsequently, the article characteristics are utilized for prediction, following a similar methodology as described in the previous section. The dataset was split into training and test samples, with the SVM model emerging as the best classifier, boasting a 91.5% accuracy.

As numerous articles frequently employ multiple methodologies simultaneously, the algorithm aids in estimating the probability that an article pertains to a particular category. These probability values will subsequently serve as a control variable in the empirical model.

A.7 Name disambiguation

The article database was converted to an author-level database to track each researcher’s publication history. So, the databases were merged not only at the article level but also at the author level. Specifically, when a match was identified at the article level, the author’s last name and position were employed for the author merging process.

The disambiguation process started by using MAG authors identifiers. Although MAG has consistent identifiers, an author might have multiple ones, which can distort publication counts. This can potentially lead to an underestimation of authors’ current publication levels. Conversely, excessive merging could artificially inflate authors’ current publication counts. To mitigate these issues, a disambiguation procedure akin to that outlined in Card et al. [2023] was conducted.

Names have been standardized, stripped of special characters and punctuation, converted to lowercase and non-standard characters are transformed into their standard English equivalents. For each name, a sequential process is initiated, involving the parsing of the string to extract the first name, middle name, and last name. Subsequently, at each step, thorough checks are conducted to ensure the generation of error-free results or to minimize errors wherever feasible.

When complete first, middle, and last names are available, they are considered to belong to the

same author. Subsequently, if both first and last names are present, along with an initial middle name, and only one candidate is identified, these names are merged. Following this, cases in which names comprise only the initials of the first and middle names are examined. Finally, cases in which middle names are absent are addressed. In such instances, the frequency of the first and last name in the database is analyzed, considering the relative proportion of the first name in the database and utilizing surname classifications from the US census data. For instance, a name like “Smith” which is very popular, is excluded from the process, and disambiguation was manually checked.

In cases where only the last name, along with initials for the first and/or middle names, was accessible and this information was not matched by MAG as the same author, I resorted to the manual gender verification process to establish author identity consistency. In this manual verification, each research assistant was tasked with providing a link to the page where they located the author’s gender information. This step facilitated a more efficient validation of the research assistants’ efforts and enabled cross-referencing instances where the same author possessed different identifiers.

Chinese names were systematically excluded during this process, and the disambiguation for those names occurred primarily via using MAG disambiguation and manual checks. Lastly, an additional round of manual verification involved assessing the accuracy of the matching process by reviewing the profiles of researchers with at least two publications (specifically targeting potential erroneous mergers) and sorting the database by last name and first name initial. The goal was to ensure that, for individuals with the same last name and the same initial for the first name, there were no instances of a single author being associated with multiple identifiers or wrongly associated with the same identifiers. While disambiguation procedures are inherently imperfect, the combination of MAG with this manual verification process enables the creation of a comprehensive and meaningful publication history for the authors.

After verification, a comprehensive publication history for each author was compiled yearly. This enabled the creation of variables reflecting authors’ publication trajectories, such as total lifetime publications or publications up to a specific date. At the article level, metrics like the most prolific author, average publications per author, and total publications by all authors were calculated, considering different journal tiers. Additionally, by tracking authors’ first publication date, the most senior author in the field was identified.

A.8 Mathematics and sociology: Data construction

Mathematics and Sociology were selected due to their distinct gender ratios. West et al. [2013] using JSTOR data from 1990-2011 showed 13.68% of economics authors were women, compared to 10.64% in mathematics and 41.44% in sociology. All three disciplines underwent an exactly identical processing to ensure a fair comparison.

The dataset primarily draws from Web of Science (WoS) and Microsoft Academic Graph (MAG). In the case of mathematics, additional data from ZbMath (zbMATH [1985-2019]) is incorporated to access specific sub-fields. Both mathematics and sociology’s datasets comprised top 16 journals from

1985 to 2019, such as *Acta Mathematica* for mathematics and *American Journal of Sociology* for sociology. It is noteworthy that many sociology articles, especially in prominent journals, are book reviews. A comprehensive list of the included journals is available in Appendix E.9 and E.10.

Like the gender classification process in economics, the gender of authors in mathematics and sociology is established using a gender name dictionary and hand-coded names. Gender determination mirrored the economics approach, utilizing a gender name dictionary and manual reviews. Female names with under 99% probability and male names below 95% were manually checked. Names with Chinese-sounding origins are systematically manually reviewed regardless of their gender assignment likelihood. Overall, a manual search was done for 12,908 names in mathematics among which 6,160 were matching the chinese names dataset and 8310 names in sociology among which 1,484 were matching the chinese names dataset.

To identify the primary field in sociology, I employed various clustering algorithms to categorize articles based on thematic similarities.

In mathematics, to determine methodological styles, I manually annotated 1000 articles across categories: “pure mathematics,” “applied mathematics,” and “mathematical physics.” The distinction of mathematical physics as a separate category stemmed from the nuanced boundary between applied and theoretical aspects within this sample. In sociology, I examined methodological terminology, including “qualitative” (e.g., interviews, focus groups), “quantitative” (e.g., regression, estimation, logistic model), and “mathematics” (e.g., structural analysis, network theory). Subsequently, I used a machine learning algorithm, similar to the one employed in economics, to analyze the data. The top-performing classifiers in mathematics and sociology attain respective AUC scores of 87% and 92%.

A.9 Working papers date

I expanded my data collection to include information about working paper versions of the articles. I targeted highly influential platforms commonly used by researchers to disseminate working papers in economics and finance, such as NBER, CEPR, IZA, and SSRN. For each platform, I employed web scraping algorithms tailored to the specific structure of each site to download available working paper information. This approach enabled me to collect data on approximately 30,000 working papers from NBER, 16,000 from IZA, 15,000 from CEPR, and approximately 430,000 from SSRN. After removing duplicates and retaining the earliest available date for each paper, I downloaded data from 479,160 working papers across these four platforms.

I merged this database with the baseline one using article titles and ensuring that author names overlapped. Titles were standardized by removing special characters and punctuation and converting all letters to lowercase. Overall, I determined the publication date of a working paper version for 12,051 articles in the top 16. Subsequently, for each paper, I identified the most similar set by retaining articles for which either the journal publication date or the working paper’s publication date preceded the focal article’s publication date, with the cosine value exceeding the top 99 percentile.

B Similarity Algorithms

B.1 Data cleaning

A series of steps were implemented to effectively process the textual information and optimize the focus on informative content while eliminating noise and irrelevant elements.

The initial step involved the removal of stopwords. Stopwords are frequently occurring words in a language that typically lack specific contextual meaning within a given text. Examples of stopwords encompass “a,” “the,” “an,” and similar terms. This process was carried out using a compiled list of common stopwords, drawing from both online resources and the Natural Language Toolkit (NLTK) package. Next, punctuation was systematically eliminated, and all text was converted to lowercase to ensure consistency and prevent variations in capitalization from resulting in differential treatment of identical words.

The linguistic process of lemmatization was then employed, reducing words to their root form. For instance, verbs in various tenses were transformed into their infinitive form (Example: “producing” was simplified to “produce”). Post-lemmatization, the text was segmented into individual words through tokenization.

Furthermore, the lexical set encompasses individual words and collocations, often called n-grams, denoting combinations of multiple words. For instance, “public debt” is a bigram, while “capital income taxation” constitutes a trigram. These collocations are detected through unsupervised machine learning techniques by counting the relative occurrence of groups of words. This allows one to capture evolution over time in response to changes in the language used by economists. For instance, the collocation “idiosyncratic income risk” gained prominence after Aiyagari’s influential paper in 1994, while the collocation “unconventional monetary policy” emerged in the aftermath of the 2007 financial crisis in the United States.

While the preceding procedures align with conventional practices in natural language processing, an additional step was introduced to identify and decipher acronyms, which are commonplace in scholarly articles, serving as abbreviated representations of more complex technical terms or expressions. To find these acronyms, a common pattern frequently employed for their presentation was used. Typically, the full expression is initially provided, followed by the acronym enclosed in parentheses, as exemplified in instances such as “Value-at-Risk (VaR).” Leveraging the capabilities of regular expressions, a tool used to capture string patterns, all instances featuring such acronym constructions are detected. Subsequently, character counts and initial letters of each constituent word are utilized to ascertain the number of antecedent words that pertain to the respective acronym. It is important to note that while this approach is quite useful, it may not encompass acronyms that adhere to different definitional formats. To address this, a supplementary list of widely accepted acronyms prevalent in the field of economics, some of which may no longer necessitate explicit definitions (e.g., OLS, GMM), is incorporated.

B.2 More on the similarity algorithms

B.2.1 Collocation models

“Collocation” refers to word sequences that frequently appear together, offering a unique meaning than when considered individually. To identify these, the Pointwise Mutual Information (PMI) method is used. PMI evaluates how often two words co-occur compared to their separate occurrences. Formally, the PMI of two events x and y is expressed as:

$$\text{PMI}(x, y) = \log_2 \left(\frac{p(x, y)}{p(x) \times p(y)} \right)$$

Here, $p(x, y)$ denotes the joint probability of x and y appearing together, while $p(x)$ and $p(y)$ represent the individual probabilities of x and y , respectively. A high PMI value indicates that two words coexist more often than if they were independent. In this analysis, PMI values ranging from 0.1 to 0.4 are explored. It is worth noting that higher cutoffs reduce the likelihood of forming collocations. When combined with a threshold on the permissible occurrences of a collocation—essentially filtering out words and collocations below a certain frequency to minimize errors—24 distinct algorithms with collocation models are implemented.

B.2.2 Word embeddings

I use word embeddings, which are essentially vectorial representations of words derived from training on extensive text corpora. These models assign to each word a unique vector in a multi-dimensional space. The design of these vectors is such that they inherently reflect both the semantic and syntactic nuances of words. Consequently, words with akin meanings are represented by vectors in close proximity within this space. A notable characteristic of these embeddings is their static nature: each word is represented by an unchanging vector, independent of its contextual usage. I specifically employ pre-trained algorithms such as Word2Vec (Mikolov et al. [2013]) and GloVe (Pennington et al. [2014]) which are among the best in this family of models.

Word2Vec, using neural networks and large text datasets, learns word vectors under the premise that words in similar contexts have similar meanings. It has two primary architectures: Continuous Bag of Words (CBOW), which predicts the target word from its context, and Skip-Gram, which predicts the context from a given target word. The embeddings are essentially the trained parameters (or weights) of this neural model.

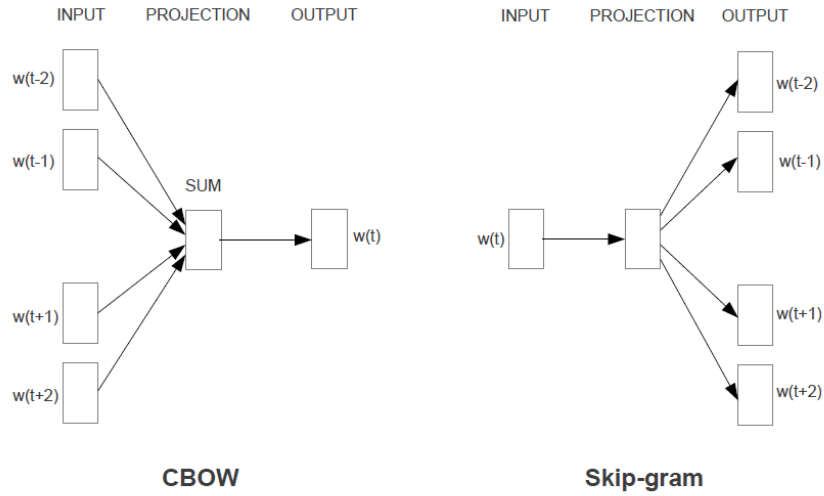
GloVe, short for Global Vectors for Word Representation, uses a word co-occurrence matrix from count-based models. It is founded on the idea that word pair statistics can provide insights into their semantic relationship. The method begins by building a co-occurrence matrix, where element X_{ij} denotes how often word i co-occurs with word j in a given dataset. Then, the objective is to find word vectors such that their dot product matches the logarithm of the words’ co-occurrence

probabilities. The objective function is:

$$\sum_{i,j=1}^V f(X_{ij}) \left(w_i^T \tilde{w}_j + b_i + \tilde{b}_j - \log X_{ij} \right)^2$$

Where w_i and $\tilde{w}_j$ are the word vectors for words i and j ; b_i and $\tilde{b}_j$ are bias terms; f is a weighting function.

The optimization is done using weighted least square method. Variants of those models are also considered based on the cleaning procedure: (1) no to limited cleaning and (2) stronger cleaning.



B.2.3 BERT

BERT utilizes the transformer architecture, which is particularly suited for handling a wide range of NLP tasks. A transformer is a deep learning architecture which at its core, uses a mechanism called “attention”, assigning weights to different segments of the input data according to their relative importance. This enables the model to emphasize pertinent words and recognize long-distance interdependencies in sentences.

Peculiar to the transformer is its use of multiple attention sets, termed “heads”. For instance, in sentence analysis, one head might discern adjective-noun relationships, another might detect verb-subject congruence, while a third could capture the overarching sentence structure. This multi-head system allows the model to concurrently pay “attention” various data facets.

BERT, distinctively, does not process words in their entirety. It employs WordPiece tokenization, segmenting words into smaller subword components, enabling effective handling of words not present in its vocabulary. During its pre-training phase, BERT utilizes the Masked Language Model (MLM) objective. A fraction of the input tokens are randomly obscured, challenging the model to predict these concealed tokens using context alone. This bidirectional contextual approach sets BERT apart from earlier unidirectional models. In addition to the MLM task, BERT is also trained on a next

sentence prediction task. Given two sentences, the model predicts if the second sentence in the pair is the subsequent sentence in the original text.

BERT’s input representations are a combination of token embeddings, segmentation embeddings (indicating if a token belongs to the first sentence or the second sentence), and positional embeddings. Each word w_i in the input sentence is initially represented as an embedding vector $E(w_i)$. BERT also utilizes position and segment embeddings.

$$\mathbf{X}_0 = [E(w_1), E(w_2), \dots, E(w_n)]$$

where $\mathbf{X}_0$ is the initial embedding for a sentence of n words.

To summarize, BERT applies L layers of Transformer blocks. Each block transforms its input representations $\mathbf{X}_{l-1}$ (from the previous layer) into a new set $\mathbf{X}_l$. Given an input $\mathbf{X}_{l-1}$, the attention score is computed. After attention scores are obtained and combined from all heads, the result passes through feed-forward neural networks. It is the final embeddings coming from the final Transformer layer (or a combination of intermediate layers) that capture the nuanced, contextual meaning of a word or token in its specific textual environment.

Bert is a very intensive algorithm with millions to billions of parameters, depending on the version. The performance of BERT has been evaluated in various contexts, as demonstrated by studies conducted by Arora et al. [2020]. Unlike BERT, which primarily focuses on word-level tasks, SBERT aims to capture entire sentences’ meaning and semantic similarity. SBERT has demonstrated improved speed compared to BERT while maintaining the accuracy achieved by BERT (Reimers and Gurevych [2019]).

The models used are the pre-trained SBERT algorithms publicly available: DistilBERT (smaller, faster and lighter version of the original BERT), MPNet (using a new pre-training objective by masking tokens and permute the order of the remaining tokens, allowing the model to predict the masked tokens in their original order; Song et al. [2020]), MiniLM (a smaller, more efficient variant of the transformer-based language models, designed to be even smaller than DistilBERT), Allenspecter (designed to create embeddings for scientific papers by using the citation graph of scientific papers; Cohan et al. [2020]). Variants of those models are also considered based on the cleaning procedure: (1) no to limited cleaning and (2) stronger cleaning.

B.2.4 Cross-references

I developed a similarity metric based on shared references, commonly employed by online platforms to establish links between articles.¹ The metric is defined as:

$$CrossReference(p, p') = \frac{\text{Common Citations to } p \text{ and } p'}{\text{Total papers cited in } p + \text{Total papers cited in } p' - \text{Common Citations to } p \text{ and } p'} \quad (1)$$

¹E.g., connectedpapers.com.

Or

$$CrossReference(p, p') = \min\left(\frac{\text{Common Citations to } p \text{ and } p'}{\text{Total papers cited in } p}, \frac{\text{Common Citations to } p \text{ and } p'}{\text{Total papers cited in } p'}\right) \quad (2)$$

This method has limitations compared to alternatives. First of all, references are subject to authors' willingness to cite each other. Second, the method also favors recent papers; for instance, a 2019 paper, X, is likelier to have more in common with a 2017 paper, Y, than a 2000 paper, Z. Therefore, cross-referencing introduces both a timing bias and citation biases.

B.3 Concrete example

Online Appendix Figure F.1 illustrates the procedure of calculating cosine similarity and the omission indicator using a TF-IDF-based vector representation. Let us consider three articles: p , p' , and p'' , with their contents represented by their titles, their abstracts and keywords.

Panel (a) first breaks each content into words. A matrix is created that marks the word occurrences in an article. In Panel (b), the relative word frequency (TF) within an article is found and multiplied by the Inverse Document Frequency (IDF) which considers all database articles. The cosine similarity between articles p and p' is expressed as:

$$\cos(p, p') = \frac{TfIdf(trade_credit, p) * TfIdf(trade_credit, p') + TfIdf(retailer, p) * TfIdf(retailer, p') + \dots}{(TfIdf(trade_credit, p)^2 + TfIdf(retailer, p)^2 + \dots)^{1/2} (TfIdf(trade_credit, p')^2 + TfIdf(retailer, p')^2 + \dots)^{1/2}}$$

After performing the necessary calculations for all articles in the database, I obtain the cosine similarity matrix, which represents the pairwise similarities between articles. Using this, the omission indicator is created. In this example, p is the citing article, and p' and p'' are part of the most similar set $\mathcal{P}_p$, for paper p . Meaning, similarities between p , p' , and p'' rank high.

Panel (c) suggests that p and p' share many similarities, as do p and p'' . However, p did not cite p' , but it did cite p'' . Consequently, the omission indicator between p and p' is 1 ($Omit_{p,p'} = 1$), indicating that p omitted citing p' , whereas the omission indicator between p and p'' is 0 ($Omit_{p,p''} = 0$), indicating that p cited p'' . Notably, p' is female-authored and p'' male-authored, highlighting the type of information captured by the omission indicator, on average.

B.4 Focus on top percentile

The analysis targets articles with the highest cosine similarity. However, taking directly the top percentile of the overall cosine distribution might introduce bias by over-representing specific articles. To mitigate this, each article is linked to its 200 most similar counterparts, retaining only the top 1% (alternatively the top 2%, 3%...) from this pool. Although this narrows the sample, it emphasizes the most similar articles. Yet, in order to uphold a sample that accurately reflects the broader population and reduce the possibility of capturing patterns exclusive to particular fields of study or

specific areas within the literature, an “ensemble method” approach is used. This combines the top 1% most similar pairs from 42 algorithms across 4 families (listed above) to offset potential biases.

B.5 Audit study detailed

For the audit study, I began by reviewing hundreds of articles, which helped develop classification criteria, design an audit template on Google Forms, and pinpoint key textual information to be read by the auditors.

Later, with a research assistant, we assessed around 700 text pairs. This helped address challenges that arose during the evaluation process, refine the rating scale, clarify score definitions, and identify shortcomings in some similarity models that had been employed (that was subsequently removed when evaluated to be too noisy for the task).²

Before the main audit, a team of undergraduates downloaded articles and extracted abstracts, introductions, and conclusions. Given the extensive literature on biases, it was essential to mitigate any conscious or unconscious biases that could potentially influence the evaluation of documents during the audit study. These biases could manifest in different ways, particularly when perceiving the gender associations within the documents. Notably, as the present article is situated within this literature, it was important to avoid any potential biases in the evaluation process. This task also aimed to prevent biases in the evaluations, ensuring that they remained independent rather than inadvertently mirroring the authors’ perspectives. So, the students were instructed to make texts (gender-)neutral, removing author details and changing pronouns neutral alternatives such as “we” and “their.”. However, well-known terms like “Melitz model” or “Dickey-Fuller test” remained unchanged so as not to distort the content of the articles.

For the full-scale audit, a shared understanding of the rating scale was crucial. I trained new PhD research assistants, starting with an audit guide review and followed by trial evaluations. Feedback was given, and meetings were held as needed for clarity and consistency.

I selected articles in three stages: an initial random sample of 200 articles, a refined selection of 150 from this group, and an addition of 50 to balance underrepresented fields. To ensure auditor thoroughness, I introduced placebo papers not suggested by algorithms. These could be entirely unrelated or within the same field as the focal article, testing the auditors’ attentiveness and accuracy in distinguishing between relevant and irrelevant articles.

To improve quality, auditors mainly reviewed articles in their areas of expertise. Furthermore, it was essential to maintain the integrity of the audit process by ensuring that the auditors remained unaware of the study’s hypotheses at every stage of the full-scale audit. They were neither informed that the generated texts were derived from algorithmic suggestions nor made aware of the study’s

²The initial models featured Doc2Vec, an extension of Word2Vec that creates vectors for larger text units like paragraphs by associating a unique tag with each document. This tag, treated as a word during training, predicts surrounding words. However, possibly because of data size, its performance in detecting similarities was subpar, with a mean score of around 1.

connection to gender research. Post-assessment, they indicated their confidence level. Uncertain ratings, approximately 200 pairs, were re-evaluated by another auditor.

B.6 Example of similar papers identified by the algorithm

Cited paper	Citing paper
Auer, R. A., & Schoenle, R. S. (2016). Market structure and exchange rate pass-through. <i>Journal of International Economics</i> , 98, 60-77.	Fitzgerald, D., & Haller, S. (2014). Pricing-to-market: evidence from plant-level prices. <i>Review of Economic Studies</i> , 81(2), 761-786.
Basu, A. K., & Chau, N. H. (2004). Exploitation of child labor and the dynamics of debt bondage. <i>Journal of Economic Growth</i> , 9(2), 209-238.	Baland, J. M., & Robinson, J. A. (2000). Is child labor inefficient?. <i>journal of Political Economy</i> , 108(4), 663-679.
Beltratti, A., & Morana, C. (2006). Breaks and persistency: macroeconomic causes of stock market volatility. <i>Journal of econometrics</i> , 131(1-2), 151-177.	Campbell, J. Y., Lettau, M., Malkiel, B. G., & Xu, Y. (2001). Have individual stocks become more volatile? An empirical exploration of idiosyncratic risk. <i>The journal of finance</i> , 56(1), 1-43.
Bessembinder, H., & Seguin, P. J. (1992). Futures-trading activity and stock price volatility. <i>the Journal of Finance</i> , 47(5), 2015-2034.	Daigler, R. T., & Wiley, M. K. (1999). The impact of trader type on the futures volatility-volume relation. <i>The Journal of Finance</i> , 54(6), 2297-2316.
Bombardini, M., Orefice, G., & Tito, M. D. (2019). Does exporting improve matching? Evidence from French employer-employee data. <i>Journal of International Economics</i> , 117, 229-241.	Krishna, P., Poole, J. P., & Senses, M. Z. (2014). Wage effects of trade reform with endogenous worker mobility. <i>Journal of International Economics</i> , 93(2), 239-252.
Borjas, G. J., & Hilton, L. (1996). Immigration and the welfare state: Immigrant participation in means-tested entitlement programs. <i>The quarterly journal of economics</i> , 111(2), 575-604.	Bratsberg, B., Raaum, O., & Røed, K. (2010). When minority labor migrants meet the welfare state. <i>Journal of Labor Economics</i> , 28(3), 633-676.
Brown, J., Duggan, M., Kuziemko, I., & Woolston, W. (2014). How does risk selection respond to risk adjustment? New evidence from the Medicare Advantage Program. <i>American Economic Review</i> , 104(10), 3335-64.	Aizawa, N., & Kim, Y. S. (2018). Advertising and risk selection in health insurance markets. <i>American Economic Review</i> , 108(3), 828-67.
Callaghan, S. R., Saly, P. J., & Subramaniam, C. (2004). The timing of option repricing. <i>The Journal of Finance</i> , 59(4), 1651-1676.	Carter, M. E., & Lynch, L. J. (2001). An examination of executive stock option repricing. <i>Journal of Financial Economics</i> , 61(2), 207-225.
Chernozhukov, V., & Hansen, C. (2006). Instrumental quantile regression inference for structural and treatment effect models. <i>Journal of Econometrics</i> , 132(2), 491-525.	Horowitz, J. L., & Lee, S. (2007). Nonparametric instrumental variables estimation of a quantile regression model. <i>Econometrica</i> , 75(4), 1191-1208.
Choi, Y. S., & Krishna, P. (2004). The factor content of bilateral trade: An empirical test. <i>Journal of Political Economy</i> , 112(4), 887-914.	Lai, H., & Zhu, S. C. (2007). Technology, endowments, and the factor content of bilateral trade. <i>Journal of International Economics</i> , 71(2), 389-409.
Damm, A. P. (2009). Ethnic enclaves and immigrant labor market outcomes: Quasi-experimental evidence. <i>Journal of Labor Economics</i> , 27(2), 281-314.	Edin, P. A., Fredriksson, P., & Åslund, O. (2003). Ethnic enclaves and the economic success of immigrants—Evidence from a natural experiment. <i>The quarterly journal of economics</i> , 118(1), 329-357.

Cited paper	Citing paper
Dubois, P., & Lasio, L. (2018). Identifying industry margins with price constraints: Structural estimation on pharmaceuticals. <i>American Economic Review</i> , 108(12), 3685-3724.	Nevo, A. (2001). Measuring market power in the ready-to-eat cereal industry. <i>Econometrica</i> , 69(2), 307-342.
Dupor, B., & Guerrero, R. (2017). Local and aggregate fiscal policy multipliers. <i>Journal of Monetary Economics</i> , 92, 16-30.	Barro, R. J., & Redlick, C. J. (2011). Macroeconomic effects from government purchases and taxes. <i>The Quarterly Journal of Economics</i> , 126(1), 51-102.
Edmonds, E. V., & Pavcnik, N. (2005). Child labor in the global economy. <i>Journal of Economic Perspectives</i> , 19(1), 199-220.	Baland, J. M., & Robinson, J. A. (2000). Is child labor inefficient?. <i>Journal of Political Economy</i> , 108(4), 663-679.
Geronimus, A. T., & Korenman, S. (1992). The socioeconomic consequences of teen childbearing reconsidered. <i>The Quarterly Journal of Economics</i> , 107(4), 1187-1214.	Hotz, V. J., Mullin, C. H., & Sanders, S. G. (1997). Bounding causal effects using data from a contaminated natural experiment: Analysing the effects of teenage childbearing. <i>The Review of Economic Studies</i> , 64(4), 575-603.
Harrigan, J., Ma, X., & Shlychkov, V. (2015). Export prices of US firms. <i>Journal of International Economics</i> , 97(1), 100-111.	Manova, K., & Zhang, Z. (2012). Export prices across firms and destinations. <i>The Quarterly Journal of Economics</i> , 127(1), 379-436.
Hatfield, J. W., & Milgrom, P. R. (2005). Matching with contracts. <i>American Economic Review</i> , 95(4), 913-935.	Kojima, F., & Pathak, P. A. (2009). Incentives and stability in large two-sided matching markets. <i>American Economic Review</i> , 99(3), 608-27.
Kang, J. K., & Shivdasani, A. (1995). Firm performance, corporate governance, and top executive turnover in Japan. <i>Journal of financial economics</i> , 38(1), 29-58.	Denis, D. J., Denis, D. K., & Sarin, A. (1997). Ownership structure and top executive turnover. <i>Journal of financial economics</i> , 45(2), 193-221.
Keane, M., & Stavrunova, O. (2016). Adverse selection, moral hazard and the demand for Medigap insurance. <i>Journal of Econometrics</i> , 190(1), 62-78.	Fang, H., Keane, M. P., & Silverman, D. (2008). Sources of advantageous selection: Evidence from the Medigap insurance market. <i>Journal of political Economy</i> , 116(2), 303-350.
Klette, T. J., & Kortum, S. (2004). Innovating firms and aggregate innovation. <i>Journal of political economy</i> , 112(5), 986-1018.	Klepper, S. (1996). Entry, exit, growth, and innovation over the product life cycle. <i>The American economic review</i> , 562-583.
Knoeber, C. R., & Thurman, W. N. (1994). Testing the theory of tournaments: An empirical analysis of broiler production. <i>Journal of labor economics</i> , 12(2), 155-179.	Levy, A., & Vukina, T. (2004). The league composition effect in tournaments with heterogeneous players: An empirical analysis of broiler contracts. <i>Journal of labor economics</i> , 22(2), 353-377.
Levitt, S. D. (2004). Understanding why crime fell in the 1990s: Four factors that explain the decline and six that do not. <i>Journal of Economic perspectives</i> , 18(1), 163-190.	Engelhardt, B. (2010). The effect of employment frictions on crime. <i>Journal of Labor Economics</i> , 28(3), 677-718.

Cited paper	Citing paper
Lewbel, A. (1991). The rank of demand systems: theory and nonparametric estimation. <i>Econometrica: Journal of the Econometric Society</i> , 711-730.	Lewbel, A., & Pendakur, K. (2009). Tricks with Hicks: The EASI demand system. <i>American Economic Review</i> , 99(3), 827-63.
Niederle, M., & Vesterlund, L. (2010). Explaining the gender gap in math test scores: The role of competition. <i>Journal of Economic Perspectives</i> , 24(2), 129-44.	Ors, E., Palomino, F., & Peyrache, E. (2013). Performance gender gap: does competition matter?. <i>Journal of Labor Economics</i> , 31(3), 443-499.
Petersen, M. A., & Rajan, R. G. (1994). The benefits of lending relationships: Evidence from small business data. <i>The journal of finance</i> , 49(1), 3-37.	Brown, J. D., & Earle, J. S. (2017). Finance and growth at the firm level: evidence from SBA loans. <i>The Journal of Finance</i> , 72(3), 1039-1080.
Rivkin, S. G., Hanushek, E. A., & Kain, J. F. (2005). Teachers, schools, and academic achievement. <i>Econometrica</i> , 73(2), 417-458.	Aaronson, D., Barrow, L., & Sander, W. (2007). Teachers and student achievement in the Chicago public high schools. <i>Journal of labor Economics</i> , 25(1), 95-135.
Romano, J. P., & Wolf, M. (2001). Subsampling intervals in autoregressive models with linear time trend. <i>Econometrica</i> , 69(5), 1283-1314.	Mikusheva, A. (2007). Uniform inference in autoregressive models. <i>Econometrica</i> , 75(5), 1411-1452.
Servaes, H. (1996). The value of diversification during the conglomerate merger wave. <i>The Journal of Finance</i> , 51(4), 1201-1225.	Lins, K., & Servaes, H. (1999). International evidence on the value of corporate diversification. <i>The Journal of Finance</i> , 54(6), 2215-2239.
Urquiola, M., & Verhoogen, E. (2009). Class-size caps, sorting, and the regression-discontinuity design. <i>American Economic Review</i> , 99(1), 179-215.	Hoxby, C. M. (2000). The effects of class size on student achievement: New evidence from population variation. <i>The Quarterly Journal of Economics</i> , 115(4), 1239-1285.
Van Zandt, T. (1999). Real-time decentralized information processing as a model of organizations with boundedly rational agents. <i>The Review of Economic Studies</i> , 66(3), 633-658.	Vayanos, D. (2003). The decentralization of information processing in the presence of interactions. <i>The Review of Economic Studies</i> , 70(3), 667-695.
Whang, Y. J., & Andrews, D. W. (1993). Tests of specification for parametric and semiparametric models. <i>Journal of Econometrics</i> , 57(1-3), 277-318.	Hong, Y., & White, H. (1995). Consistent specification testing via nonparametric series regression. <i>Econometrica: Journal of the Econometric Society</i> , 1133-1159.
Zlate, A. (2016). Offshore production and business cycle dynamics with heterogeneous firms. <i>Journal of International Economics</i> , 100, 34-49.	Bergin, P. R., Feenstra, R. C., & Hanson, G. H. (2011). Volatility due to offshoring: Theory and evidence. <i>Journal of International Economics</i> , 85(2), 163-173.

B.7 Is topic similarity a sufficient measure for citation?

To validate the hypothesis that topic similarity is a key factor in citing an article, I analyzed a random sample of 100 articles and manually categorize their 3831 references into subject-related or methodological, with some overlapping both categories. Of the 3831 references annotated, 68% were subject-related, with no significant variation across disciplines. Even in theoretical fields, citations are often linked to articles on the same topic. Non-topic-related citations were typically theorem proofs or very different mathematical methods from other works. This confirms that a substantial part of the citations are topic-driven.

Considering the potential impacts of sample size and subjective views on what a “similar topic” is, especially in multi-subfield articles, I explored the three-digit JEL codes that could, to some extent, characterize topic-related matches. Analysis reveals that 60% of references with available JEL codes share at least a three-digit code with citing articles. Despite variations across disciplines, each field shows at least 50% of articles with related JEL codes.

C Theoretical model

This section builds on models by Bohren et al. [2019] and Garg et al. [2020] and applies them to the current setup to help understand the underlying mechanisms of the gender omission bias.

Set-up: Consider an author who has an observable group identity (gender) $g \in \{F, M\}$ and unobservable ability $q \sim N(\mu_g, \sigma_q^2)$, with mean $\mu_{q,g} \in \mathbb{R}$.

There is a set of potential citing authors (pca) who decide to cite or omit a paper written by author of type g . For simplicity, I assume that there is one pca.

The author is characterized by a set of features $\mathbf{s} = (s_1, s_2, \dots, s_K)$ and each feature is an imperfect signal of the author's ability as follows: $s_k = q + \epsilon_k$ where ϵ_k is an independent random shock (independent across features and authors) such that $\epsilon_k \sim N(0, \sigma_{k,g}^2)$. The features represent what a pca sees on this author. For example, when $k = 1$, one observes the current publication of the author. For $k = 2$, one can observe a second published paper; $k = 3$ a third published paper; $k = 4$, one of the existing paper is a top five publication; $k = 5$, the author is an editor of a famous journal, and so on.

The pca evaluates the current paper of the author before citing it. For this purpose, she/he makes an evaluation $v \in \mathbb{R}$ and receives a payoff: $-(v - (q - \kappa_g))^2$, where κ_g is a gender-specific taste parameter. In equilibrium, the pca maximizes her/his expected payoff following:

$$\hat{v}(\mathbf{s}, g) = \mathbb{E}(q|\mathbf{s}, g) - \kappa_g \quad (3)$$

Let $\tilde{q}(\mathbf{s}, g) = \mathbb{E}(q|\mathbf{s}, g)$.

Gender bias: A gender bias is defined as:

$$B(\mathbf{s}) = \hat{v}(\mathbf{s}, F) - \hat{v}(\mathbf{s}, M) \quad (4)$$

$$= (\tilde{q}(\mathbf{s}, F) - \tilde{q}(\mathbf{s}, M)) - (\kappa_F - \kappa_M) \quad (5)$$

There a gender bias in omission from references against women if $B(\mathbf{s}) < 0$. There is a gender bias in omission from references against men if $B(\mathbf{s}) > 0$. Equation 5 enables to assess two sources of discrimination. The first one is due to differences in the perceived ability given the same signal, which commonly refers to a belief-based discrimination set-up. The second one is related to differences in gender-specific preference parameters, which commonly refers to a taste-based discrimination (Becker [2010]). Without loss of generality, we assume that $\kappa_M = 0$.

Let us define the difference in omission probability as: $\Delta(q) = P(\text{Omission} = 1|q, F) - P(\text{Omission} = 1|q, M)$

Distribution of the signal given true ability ($s|q, g$): $s|q, g \sim N(q, \sigma_{k,g}^2)$

Distribution of the true ability given signal and gender ($q|\mathbf{s}, g$): Following DeGroot

[2005], we have $q \sim N(\mu_{q,g}, \sigma_q^2)$ and $s_k|q, g \sim N(q, \sigma_{k,g}^2)$, then:

$$q|s_k, g \sim N(\tilde{q}(s_k, g), \text{Var}[q|s_k, g]) \quad (6)$$

with $\tilde{q}(s_k, g) = E[q|s_k, g] = \frac{\sigma_q^2}{\sigma_q^2 + \sigma_{k,g}^2} s_k + \frac{\sigma_{k,g}^2}{\sigma_q^2 + \sigma_{k,g}^2} \mu_{q,g}$ and $\text{Var}[q|s, g] = \frac{1}{\sigma_q^{-2} + \sigma_{k,g}^{-2}}$.

To have the generalize expression, let us proceed recursively.

For $K = 1$, $q|s_1, g \sim N(\frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sigma_{1,g}^{-2}} \mu_{q,g} + \frac{\sigma_{1,g}^{-2}}{\sigma_q^{-2} + \sigma_{1,g}^{-2}} s_1, \frac{1}{\sigma_q^{-2} + \sigma_{1,g}^{-2}})$.

For $K = 2$, $q|s_2, s_1, g \sim N(\frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sigma_{1,g}^{-2} + \sigma_{2,g}^{-2}} \mu_{q,g} + \frac{\sigma_{1,g}^{-2} s_1 + \sigma_{2,g}^{-2} s_2}{\sigma_q^{-2} + \sigma_{1,g}^{-2} + \sigma_{2,g}^{-2}}, \frac{1}{\sigma_q^{-2} + \sigma_{1,g}^{-2} + \sigma_{2,g}^{-2}})$.

Finally, $q|s_K, \dots, s_1, g \sim N(\frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}} \mu_{q,g} + \frac{\sum_k \sigma_{k,g}^{-2} s_k}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}}, \frac{1}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}})$.

Then, $\tilde{q}(\mathbf{s}, g) = E[q|\mathbf{s}, g] = \frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}} \mu_{q,g} + \frac{\sum_k \sigma_{k,g}^{-2} s_k}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}}$.

Distribution of the perceived ability given true ability and gender ($\tilde{q}|q, g$):

Replacing s_k by $q + \epsilon_k$, we have that: $\tilde{q}|q, g \sim N(E[\tilde{q}|q, g], \text{Var}[\tilde{q}|q, g])$ where:

$$E[\tilde{q}|q, g] = \frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}} \mu_{q,g} + \frac{\sum_k \sigma_{k,g}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}} q \quad (7)$$

and:

$$\text{var}[\tilde{q}|q, g] = \frac{\sum_k \sigma_{k,g}^{-2}}{(\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2})^2} \quad (8)$$

Assume that, $\sigma_{k,g} = \sigma_k$, $\sigma_q^2 = \sigma_q^2$ **and** $\mu_g = \mu$: In this case, $B(q) < 0$, whatever q as long as $\kappa_F > 0$. When there is no statistical discrimination and there is a taste-based discrimination, $B(\mathbf{s}) = -\kappa_F$. In other words, whatever the signal, one should always observe a lower valuation for female-authored papers than for male-authored papers. The analysis by institutions shows that the results are not consistent with such a framework. In particular, the gender omission bias is lower, even insignificant for low-tier institutions.

Assume that $\kappa_F = 0$ (**no taste based discrimination**), $\sigma_{k,g} = \sigma_k$: Let $\tilde{q}^*$ be the perceived ability below which one decides to omit a given paper from her/his paper's references.

$$\begin{aligned} \Delta(q) &= P(\text{Omission} = 1|q, F) - P(\text{Omission} = 1|q, M) \\ &= P(\tilde{q} < \tilde{q}^*|q, F) - P(\tilde{q} < \tilde{q}^*|q, M) \\ &= \Phi\left(\frac{\sigma_q^{-2} + \sum_k \sigma_k^{-2}}{\sqrt{\sum_k \sigma_k^{-2}}}(\tilde{q}^* - \frac{\mu_F \sigma_q^{-2} + q \sum_k \sigma_{k,g}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}})\right) - \Phi\left(\frac{\sigma_q^{-2} + \sum_k \sigma_k^{-2}}{\sqrt{\sum_k \sigma_k^{-2}}}(\tilde{q}^* - \frac{\mu_M \sigma_q^{-2} + q \sum_k \sigma_{k,g}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,g}^{-2}})\right) \end{aligned}$$

Where Φ is the cumulative distribution function of the standard normal distribution. Therefore, $\Delta(q) > 0$ implies by the monotonicity of Φ that $\mu_F < \mu_M$. In other words, the probability of omission of women is higher than the probability of omission of men for the same ability q , as long as there is a difference in the average ability where the average ability of women is lower than the

average ability of men.

We also have:

$$\begin{aligned}
B(s) &= \tilde{q}(s, F) - \tilde{q}(s, M) \\
&= \frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_k^{-2}} \mu_{q,F} + \frac{\sum_k \sigma_k^{-2} s_k}{\sigma_q^{-2} + \sum_k \sigma_k^{-2}} - \frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_k^{-2}} \mu_{q,M} - \frac{\sum_k \sigma_k^{-2} s_k}{\sigma_q^{-2} + \sum_k \sigma_k^{-2}} \\
&= \frac{\sigma_q^{-2}}{\sigma_q^{-2} + \sum_k \sigma_k^{-2}} (\mu_F - \mu_M)
\end{aligned}$$

Then $B(s) < 0$ if and only if $\mu_F < \mu_M$. In other words, the valuation of women's paper is lower than the valuation of men's paper for the same signal s , as long as there is a difference in the average ability where the average ability of women is lower than the average ability of men. Statistical discrimination on the mean (either correct or incorrect) always implies that for the same signal, the valuation for women is lower and the probability to omit a woman-authored is higher for the same true ability.

Assume that $\kappa_F = 0$ (no taste based discrimination), $\sigma_{k,g} = \sigma_k$, $\sigma_{q,g}^2 = \sigma_q^2$, what is the effect of additional information?

Consider the case where $K = 1$ versus the case where $K = 2$. We have that:

$$\begin{aligned}
B(s_1) - B(s_1, s_2) &= (\mu_F - \mu_M) \sigma_q^{-2} \left(\frac{1}{\sigma_q^{-2} + \sigma_1^{-2}} - \frac{1}{\sigma_q^{-2} + \sigma_1^{-2} + \sigma_2^{-2}} \right) \\
&= (\mu_F - \mu_M) \sigma_q^{-2} \frac{\sigma_2^{-2}}{(\sigma_q^{-2} + \sigma_1^{-2})(\sigma_q^{-2} + \sigma_1^{-2} + \sigma_2^{-2})}
\end{aligned}$$

Then, for $\mu_F < \mu_M$, $D(s_1) < D(s_1, s_2)$. Given that for $\mu_F < \mu_M$, $D(s_1) < 0$, it means that the difference in valuation with just one feature is more negative (higher) than the difference in valuation with two features. In other words, increasing the number of features reduces the gender bias when we have statistical discrimination based on the mean. We can generalize this to k features.

Assume that $\kappa_F = 0$ (no taste based discrimination), $\sigma_q^2 = \sigma_q^2$, $\mu_g = \mu$ compute $\Delta(q)$.

$$\begin{aligned}
\Delta(q) &= P(\text{Omission} = 1|q, F) - P(\text{Omission} = 1|q, M) \\
&= P(\tilde{q} < \tilde{q}^*|q, F) - P(\tilde{q} < \tilde{q}^*|q, M) \\
&= \Phi\left(\frac{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}}{\sqrt{\sum_k \sigma_{k,F}^{-2}}}(\tilde{q}^* - \frac{\mu\sigma_q^{-2} + q\sum_k \sigma_{k,F}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}})\right) - \Phi\left(\frac{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}}{\sqrt{\sum_k \sigma_{k,M}^{-2}}}(\tilde{q}^* - \frac{\mu\sigma_q^{-2} + q\sum_k \sigma_{k,M}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}})\right)
\end{aligned}$$

Then, $\Delta(q) > 0$ implies by the monotonicity of Φ that:

$$\begin{aligned}
\frac{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}}{\sqrt{\sum_k \sigma_{k,F}^{-2}}} (\tilde{q}^* - \frac{\mu_F \sigma_q^{-2} + q \sum_k \sigma_{k,F}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}}) &> \frac{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}}{\sqrt{\sum_k \sigma_{k,M}^{-2}}} (\tilde{q}^* - \frac{\mu \sigma_q^{-2} + q \sum_k \sigma_{k,M}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}}) \\
\Leftrightarrow \frac{\sigma_q^{-2}(\tilde{q}^* - \mu) + (\tilde{q}^* - q) \sum_k \sigma_{k,F}^{-2}}{\sqrt{\sum_k \sigma_{k,F}^{-2}}} &> \frac{\sigma_q^{-2}(\tilde{q}^* - \mu) + (\tilde{q}^* - q) \sum_k \sigma_{k,M}^{-2}}{\sqrt{\sum_k \sigma_{k,M}^{-2}}} \\
\Leftrightarrow q > \tilde{q}^* - \frac{\sigma_q^{-2}(\tilde{q}^* - \mu)}{\sqrt{\sum_k \sigma_{k,F}^{-2}} \sqrt{\sum_k \sigma_{k,M}^{-2}}}
\end{aligned}$$

Let $q_b = \tilde{q}^* - \frac{\sigma_q^{-2}(\tilde{q}^* - \mu)}{\sqrt{\sum_k \sigma_{k,F}^{-2}} \sqrt{\sum_k \sigma_{k,M}^{-2}}}$. We observe a gender bias in omission from references not for all q , but starting at a certain value of q , q_b . By contrast to statistical discrimination on the mean, when agents care about the variance, we do not observe discrimination for all ability levels. In particular, for a value of q_b above a certain quantile of the distribution of q , medium or high skill individual may suffer the most.

Assume that $\kappa_F = 0$ (no taste based discrimination), $\sigma_q^2 = \sigma_q^2$, variation of $\Delta(q)$. Let us take the derivative of $\Delta(q)$ with respect to q .

$$\begin{aligned}
\frac{\partial \Delta(q)}{\partial q} = & -\sqrt{\sum_k \sigma_{k,F}^{-2}} \phi \left(\frac{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}}{\sqrt{\sum_k \sigma_{k,F}^{-2}}} (\tilde{q}^* - \frac{\mu \sigma_q^{-2} + q \sum_k \sigma_{k,F}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,F}^{-2}}) \right) + \\
& \sqrt{\sum_k \sigma_{k,M}^{-2}} \phi \left(\frac{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}}{\sqrt{\sum_k \sigma_{k,M}^{-2}}} (\tilde{q}^* - \frac{\mu \sigma_q^{-2} + q \sum_k \sigma_{k,M}^{-2}}{\sigma_q^{-2} + \sum_k \sigma_{k,M}^{-2}}) \right) \quad (9)
\end{aligned}$$

Where ϕ is the probability density function of the standard normal distribution.

This implies that:

$$\begin{aligned}
2 \left(\ln \left(-\frac{\partial P(Omission = 1|q, F)}{\partial q} \right) - \ln \left(-\frac{\partial P(Omission = 1|q, M)}{\partial q} \right) \right) &= \ln \left(\sum_k \sigma_{k,F}^{-2} \right) - \ln \left(\sum_k \sigma_{k,M}^{-2} \right) - \\
& \frac{((\tilde{q}^* - \mu) \sigma_q^{-2} + (\tilde{q}^* - q) \sum_k \sigma_{k,F}^{-2})^2}{\sigma_{k,F}^{-2}} + \frac{((\tilde{q}^* - \mu) \sigma_q^{-2} + (\tilde{q}^* - q) \sum_k \sigma_{k,M}^{-2})^2}{\sigma_{k,M}^{-2}} \\
&= \ln \left(\sum_k \sigma_{k,F}^{-2} - \sum_k \sigma_{k,M}^{-2} \right) + (\tilde{q}^* - \mu)^2 \sigma_q^{-4} \left(\frac{1}{\sigma_{k,F}^{-2}} - \frac{1}{\sigma_{k,M}^{-2}} \right) + (\tilde{q}^* - q) \left(-\sum_k \sigma_{k,F}^{-2} + \sum_k \sigma_{k,M}^{-2} \right)
\end{aligned}$$

This second order equation in q has two solutions:

$$\underline{q} = \tilde{q}^* - \sqrt{\frac{(\tilde{q}^* - \mu)^2 \sigma_q^{-4}}{\sum_k \sigma_{k,F}^{-2} \sum_k \sigma_{k,M}^{-2}} + \frac{\ln(\sum_k \sigma_{k,F}^{-2} - \sum_k \sigma_{k,M}^{-2})}{-\sum_k \sigma_{k,F}^{-2} + \sum_k \sigma_{k,M}^{-2}}} \quad (10)$$

$$\bar{q} = \tilde{q}^* + \sqrt{\frac{(\tilde{q}^* - \mu)^2 \sigma_q^{-4}}{\sum_k \sigma_{k,F}^{-2} \sum_k \sigma_{k,M}^{-2}} + \frac{\ln(\sum_k \sigma_{k,F}^{-2} - \sum_k \sigma_{k,M}^{-2})}{-\sum_k \sigma_{k,F}^{-2} + \sum_k \sigma_{k,M}^{-2}}} \quad (11)$$

Therefore, $\frac{\partial \Delta(q)}{\partial q} > 0$ if $\underline{q} < q < \bar{q}$, i.e. $\Delta(q)$ is increasing in q . $\frac{\partial \Delta(q)}{\partial q} < 0$ if $q \leq \underline{q}$ or $q \geq \bar{q}$, i.e. $\Delta(q)$ is decreasing in q .

Further, one can show that $\underline{q} < q_b < \bar{q}$ and $\lim_{q \rightarrow \infty} \Delta(q) = 0$. It means that, for ability between q_b and $\bar{q}$, there is a gender bias against women in the omission from references and it is increasing as q increases. After $\bar{q}$, the bias start decreasing, but can still be positive. For sufficiently high q , the bias tend to 0.

Effect of having more features:

$$\tilde{q}(\mathbf{s}, g) = E[q|\mathbf{s}, g] = \frac{\sigma_{q,g}^{-2}}{\sigma_{q,g}^{-2} + \sum_k \sigma_{k,g}^{-2}} \mu_{q,g} + \frac{\sum_k \sigma_{k,g}^{-2} s_k}{\sigma_{q,g}^{-2} + \sum_k \sigma_{k,g}^{-2}}$$

When $\sum_k \sigma_{k,g}^{-2}$ increase, this gives more weights to q . Therefore, more features enable the perceived ability to depart from the average ability and to be closer of the true ability.

Overall, I can further derive the following proposition.

Proposition. *If there is a difference only in the mean ability between men and women, where women mean ability is lower than men mean ability, $\mu_F < \mu_M$, the following statements hold:*

1. *There is a gender bias against women in the omission from references whatever the signal: $B(\mathbf{s}) > 0, \forall \mathbf{s}$.*
2. *Adding additional features reduce the bias: $|B(s_1, \dots, s_p)| > |B(s_1, \dots, s_p, s_{p+1})|$ for $p \in 1, \dots, K - 1$.*

The first part of proposition 1 implies that the probability of omission of women is higher than the probability of omission of men for the same ability q , as long as there is a difference in the average ability where the average ability of women is lower than the average ability of men. Similarly, the valuation of women's paper is lower than the valuation of men's paper for the same signal $\mathbf{s}$. Interestingly, the second part of the proposition implies an updating effect, where increasing the number of features reduces gender bias. This is especially the case because increasing the number of features increases $\sum_k \sigma_{k,g}^{-2}$. Then, this gives more weights to $\mathbf{s}$ in the expression of the perceived ability. Therefore, more features enable the perceived ability to depart from the average ability and be closer to the signal.

Proposition. *If there is a difference only in the variance of the signals of the ability between men and women, $\sum_k \sigma_{k,M}^{-2} > \sum_k \sigma_{k,F}^{-2}$ where the precision of the signal is higher for men than for women, the following statements hold:*

1. *There is a value q_b of the true ability above which there is a gender bias against women in the omission from references.*
2. *There exist thresholds $\underline{q}$ and $\bar{q}$ such that the difference in the probability of omission is increasing for $\underline{q} < q < \bar{q}$ and decreasing for $q > \bar{q}$.*
3. $\lim_{q \rightarrow \infty} \Delta(q) = 0$ with $\Delta(q) = P(\text{Omission} = 1|q, F) - P(\text{Omission} = 1|q, M)$.

Proposition 2 offers interesting insights. The first part shows that a gender bias in omission from references appears after a specific value of q , q_b . Further, the second part means that, for ability between q_b and $\bar{q}$, there is a gender bias against women in the omission from references, and it is increasing as q increases. After $\bar{q}$, the bias starts decreasing but can still be positive. For sufficiently high q , the bias tends to 0. In other words, for a value of q_b above a certain quantile of the distribution of q , medium-skill individuals may suffer the most. This could be viewed as a hump-shaped response where gender bias in omission from references increases for $q < \bar{q}$ and start decreasing after, even almost disappearing for very high-skilled females. This pattern is pretty consistent with the overall results, especially taking the split with the institutions. Overall, agents seem to care about the variance.

Lastly, an additional argument could be made regarding homophily. The two-sided gender effect could let someone argues in favor of a homophilic pattern. Zeltzer [2020], for instance, shows evidence of homophilic behaviors in the medical practice where male/female physicians were more likely to refer patients to a surgeon of the same gender. Zeltzer [2020] additionally shows that homophily is higher as the fraction of the women increases in the population. Therefore, one should expect homophily to be higher in fields with more women. However, the field decomposition suggests a lower two-sided gender effect in labor and education that traditionally have more females.

D Citations and innovativeness index

Following Kelly et al. [2021], the “novelty” of a paper is captured by a backward-similarity index, which is the sum of pairwise relative cosine similarities of paper p , published in t , to paper p' , published in $t - T$:

$$BS_{-T}^0(p) = \sum_{p'} \tilde{\lambda}_{p,p'}$$

Papers with low backward similarity are dissimilar from the past literature, and hence they are innovative compared with the existing literature.

The “influence” of a paper is captured by a forward-similarity index. The forward similarity is the sum of pairwise cosine similarities of paper p , published in t , to paper p' , published in $t + T$:

$$FS_0^T(p) = \sum_{p'} \tilde{\lambda}_{p,p'}$$

Papers with high forward similarity have a higher impact on future publications. For example, they may open up a rich future line of literature, or they may propose an empirical methodology that many future papers will use.

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

$$q^T(p) = \frac{FS_0^T(p)}{BS_{-T}^0(p)}$$

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 high forward similarity (high numerator) and 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 with a paper with high forward similarity and relatively low backward similarity. In that respect, it operates like the citations measure. In the baseline, the horizon chosen to compute the q -index is 5 years.

E Additional Tables

E.1 Gender Audit

Panel A: Economics

		Assigned authors based on gender coding procedure				Total
		All Male	All Female	Mixed-Gender	Undetermined	
True authors gender	All Male	1,192	0	1	1	1,194
	All Female	1	65	1	0	67
	Mixed-Gender	4	0	211	0	215
	Undetermined	1	0	1	7	9
	Total	1,198	65	214	8	1,485

Panel B: Mathematics

		Assigned authors based on gender coding procedure				Total
		All Male	All Female	Mixed-Gender	Undetermined	
True authors gender	All Male	1,245	0	1	2	1,248
	All Female	0	37	0	0	37
	Mixed-Gender	3	0	231	5	239
	Undetermined	6	0	3	52	61
	Total	1,254	37	235	59	1,585

Panel C: Sociology

		Assigned authors based on gender coding procedure				Total
		All Male	All Female	Mixed-Gender	Undetermined	
True authors gender	All Male	426	0	4	1	431
	All Female	0	274	1	0	275
	Mixed-Gender	0	1	280	2	283
	Undetermined	0	0	0	12	12
	Total	426	275	285	15	1,001

This table provides a confusion matrix highlighting gender assignment errors in our sample. There are 1,485 articles from economics, 1,585 from mathematics, and 1,001 from sociology, each accounting for about 5% of their total databases.

E.2 Cosine similarity

E.2.1 Distribution of cosine similarity

	TF-IDF no collocation	TF-IDF with collocation collocation (5,0.4)	Word embeddings Word2Vec	BERT Distill-Roberta
	(1)	(2)	(3)	(4)
Mean	0.093	0.089	0.577	0.219
Standard deviation	0.053	0.048	0.101	0.097
Median	0.075	0.073	0.583	0.208
Top 75 percentile	0.107	0.101	0.648	0.279
Top 80 percentile	0.118	0.111	0.664	0.298
Top 90 percentile	0.155	0.144	0.703	0.351
Top 99 percentile	0.303	0.278	0.788	0.489

This table shows the distribution of the cosine similarity measures across different families of algorithms for non-zeros values of the cosine.

E.2.2 Cosine distribution and field (full table)

	Math	Micro	Macro	Int	Fin	LabEduc	Other emp	Miscall	chi2 test of group means equality
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Distilroberta 2	0.708	0.755	0.733	0.736	0.704	0.721	0.737	0.728	167.74
Minilm l12 2	0.763	0.734	0.739	0.773	0.739	0.759	0.735	0.733	92.07
Minilm l6 2	0.735	0.707	0.736	0.777	0.812	0.705	0.719	0.743	160.11
Mpnet-base 2	0.779	0.785	0.797	0.796	0.77	0.784	0.772	0.852	101.37
Allen AI 2	0.955	0.952	0.954	0.96	0.954	0.948	0.949	0.953	514.98
Distilroberta	0.667	0.671	0.723	0.703	0.736	0.68	0.668	0.724	104.1
Minilm l12	0.731	0.734	0.7	0.709	0.697	0.698	0.76	0.723	61.03
Minilm l6	0.726	0.72	0.759	0.711	0.715	0.689	0.734	0.774	133.8
Mpnet-base	0.803	0.8	0.756	0.753	0.745	0.771	0.737	0.8	128.03
Allen AI	0.942	0.949	0.948	0.947	0.942	0.944	0.942	0.944	540.54
Collocation (10,0.1)	0.335	0.309	0.403	0.401	0.45	0.394	0.318	0.352	39.78
Collocation (10,0.2)	0.398	0.462	0.399	0.36	0.377	0.436	0.373	0.386	51.55
Collocation (10,0.3)	0.523	0.511	0.419	0.685	0.526	0.422	0.405	0.619	56.82
Collocation (10,0.4)	0.45	0.472	0.517	0.554	0.52	0.579	0.501	0.594	95.79
Collocation (15,0.1)	0.348	0.545	0.4	0.35	0.487	0.352	0.383	0.414	55.12
Collocation (15,0.2)	0.53	0.451	0.48	0.374	0.385	0.374	0.368	0.438	78.36
Collocation (15,0.3)	0.497	0.434	0.521	0.531	0.45	0.441	0.458	0.466	72.89
Collocation (15,0.4)	0.477	0.494	0.54	0.561	0.451	0.483	0.453	0.624	83.12
Collocation (20,0.1)	0.504	0.361	0.361	0.371	0.363	0.366	0.397	0.412	78.03
Collocation (20,0.2)	0.461	0.658	0.383	0.392	0.449	0.5	0.392	0.424	101.06
Collocation (20,0.3)	0.505	0.492	0.431	0.539	0.506	0.506	0.418	0.52	76.47
Collocation (20,0.4)	0.513	0.474	0.508	0.461	0.612	0.539	0.455	0.555	81.45
Collocation (30,0.1)	0.451	0.389	0.467	0.397	0.471	0.399	0.376	0.384	78.62
Collocation (30,0.2)	0.462	0.456	0.408	0.414	0.416	0.423	0.393	0.418	52.96
Collocation (30,0.3)	0.549	0.422	0.497	0.498	0.514	0.496	0.543	0.495	70.34
Collocation (30,0.4)	0.467	0.538	0.466	0.554	0.457	0.598	0.581	0.58	93.15
Collocation (40,0.1)	0.594	0.483	0.435	0.415	0.485	0.481	0.425	0.416	64.51
Collocation (40,0.2)	0.451	0.45	0.497	0.503	0.788	0.446	0.554	0.406	56.13
Collocation (40,0.3)	0.445	0.433	0.432	0.457	0.436	0.43	0.635	0.518	65.7
Collocation (40,0.4)	0.496	0.46	0.478	0.492	0.456	0.469	0.51	0.533	95.63
Collocation (5,0.1)	0.266	0.285	0.265	0.322	0.338	0.286	0.393	0.328	20.22
Collocation (5,0.2)	0.387	0.333	0.372	0.398	0.434	0.397	0.374	0.344	18.11
Collocation (5,0.3)	0.54	0.399	0.397	0.566	0.424	0.456	0.403	0.422	57.53
Collocation (5,0.4)	0.498	0.466	0.471	0.475	0.499	0.475	0.521	0.517	91
Glove 42b 2	0.955	0.958	0.962	0.963	0.955	0.957	0.954	0.972	134.32
Glove 6b 2	0.931	0.913	0.909	0.906	0.909	0.912	0.908	0.936	136.09
Word2vec 2	0.895	0.904	0.885	0.901	0.903	0.939	0.885	0.907	169.93
No coloc 2	0.513	0.52	0.551	0.536	0.565	0.517	0.537	0.526	472.89
Glove 42b	0.954	0.976	0.959	0.954	0.954	0.951	0.952	0.954	175.12
Glove 6b	0.913	0.898	0.909	0.906	0.912	0.904	0.908	0.915	198.79
Word2vec	0.879	0.879	0.876	0.871	0.871	0.875	0.908	0.87	319.99
No coloc	0.528	0.51	0.492	0.521	0.571	0.529	0.546	0.674	613.74

This table shows the distribution of the cosine similarity measures across different families of algorithms for non-zeros values of the cosine.

E.3 Algorithm audit Study: results

E.3.1 Audit Study: descriptive table

	All Male	All Female	Mixed	Undetermined	Total
N	1,129	80	156	9	1,374
Audit score	2.371 (1.053)	2.331 (0.977)	2.317 (0.966)	2.222 (0.972)	2.362 (1.038)
Audit score greater than 2	0.784	0.825	0.763	0.778	0.784
Audit score greater than 3	0.453	0.425	0.404	0.333	0.445
Omitted with score greater than 2	0.794	0.758	0.731	1	0.786
Omitted with score greater than 3	0.748	0.735	0.651	1	0.738
Omitted with score greater than 4	0.735	0.778	0.55	1	0.72

The table describes the papers used for the audit study to evaluate the algorithmic prediction. Papers were rated on a 0-4 scale, with scores of 2 or higher indicating potential citability due to shared fields and topics.

E.3.2 Audit study: overall results

Outcome variable: Audit scores							
	Audit score		Audit score (dummy)	Audit score	Audit score (dummy)	Audit score	Audit score (dummy)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All Female	-0.041 (0.125)	-0.017 (0.079)	0.047 (0.049)	0.028 (0.149)	0.075 (0.055)	-0.030 (0.151)	0.055 (0.059)
Mixed	-0.009 (0.093)	-0.004 (0.058)	0.003 (0.036)	0.012 (0.108)	-0.015 (0.040)	-0.085 (0.110)	-0.025 (0.043)
Undetermined	-0.385 (0.367)	-0.176 (0.251)	-0.041 (0.144)	-0.305 (0.406)	-0.007 (0.150)	-0.438 (0.588)	-0.136 (0.231)
Observations	1214	1214	1214	825	825	738	738
R-sqr	-0.001	0.0001	-0.002	-0.003	-0.001	-0.003	-0.002
Mean of DV	2.379	2.379	0.794	2.473	0.842	2.423	0.802

The table presents a regression analysis correlating a paper's audit score with the gender of the authors on the suggested paper. Papers were rated on a 0-4 scale, with scores of 2 or higher indicating potential citability due to shared fields and topics. Column (1) uses a linear regression, while column (2) employs a Poisson regression with the same dependent variable. Column (3) adopts a linear probability model where the dependent variable equals 1 for scores over 2. Columns (4)-(5) and (6)-(7) repeat the methodologies in columns (1) and (3) but for papers sharing a three-digit JEL code and those ranking in the top 10 for similarity, respectively. The gender reference category is *All Male*, denoting papers by male-only authors.

E.3.3 Audit Study: alternative algorithms

Panel A: Family of algorithms

	Outcome variable: Audit scores							
	TF-IDF		TF-IDF		Word embeddings		BERT	
	no collocation		with collocation					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female	-0.097 (0.269)	0.142 (0.095)	-0.187 (0.217)	0.036 (0.087)	0.187 (0.321)	0.069 (0.130)	0.013 (0.157)	0.013 (0.057)
Mixed	0.129 (0.195)	0.068 (0.068)	0.078 (0.161)	0.050 (0.065)	0.333 (0.220)	0.171 (0.089)	0.047 (0.117)	-0.043 (0.043)
Undetermined	0.018 (0.662)	0.142 (0.233)	0.276 (0.764)	0.258 (0.307)	-0.267 (1.040)	0.251 (0.421)	-0.723 (0.409)	-0.181 (0.149)
Observations	190	190	369	369	244	244	801	801
R-sqr	-0.013	0.001	-0.005	-0.004	-0.002	0.005	0.000	-0.001
Mean of DV	2.494	0.879	2.223	0.753	2.309	0.770	2.557	0.843

Panel B: Alternative specifications

	Outcome variable: Audit scores							
	All top		Algorithms with		Exclude algorithms		Add field specific	
	995 cosine		score ≥ 2.5		with the lowest accuracy		cutoffs	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female	-0.071 (0.190)	0.015 (0.070)	0.067 (0.134)	0.066 (0.050)	-0.014 (0.135)	0.052 (0.052)	-0.040 (0.120)	0.041 (0.048)
Mixed	0.087 (0.129)	0.036 (0.048)	-0.041 (0.102)	-0.036 (0.038)	0.035 (0.099)	0.015 (0.038)	-0.054 (0.089)	-0.021 (0.035)
Undetermined	-0.490 (0.595)	0.177 (0.219)	-0.375 (0.379)	-0.118 (0.142)	-0.456 (0.358)	-0.064 (0.137)	-0.149 (0.348)	-0.006 (0.138)
Observations	671	671	1051	1051	1057	1057	1374	1374
R-sqr	-0.002	-0.003	-0.002	0.001	-0.001	-0.002	-0.002	-0.001
Mean of DV	2.494	0.829	2.515	0.832	2.456	0.818	2.362	0.784

The table presents a regression analysis correlating a paper’s audit score with the gender of the authors on the suggested paper. Papers were rated on a 0-4 scale, with scores of 2 or higher indicating potential citability due to shared fields and topics. In the table, the even columns correspond to the case where the dependent variable is equal to 1 if the score is greater than 2 and 0 otherwise, and the uneven column to the case where the dependent variable is the rating of a pair of articles.

E.3.4 Audit Study: auditors effect and link to other variables

Panel A: Controls for auditors effects and other variables

	Outcome variable: Audit scores							
	Remove Failed		Removed Most		Auditors		Additional	
	Placebo Checks		Lenient Auditor		Fixed effects		Control variables	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female	0.011 (0.124)	0.041 (0.051)	0.075 (0.119)	0.065 (0.051)	0.020 (0.108)	0.044 (0.043)	-0.018 (0.118)	0.037 (0.048)
Mixed	-0.011 (0.091)	-0.023 (0.038)	0.052 (0.088)	-0.001 (0.038)	0.020 (0.079)	-0.014 (0.032)	0.120 (0.090)	0.016 (0.037)
Undetermined	-0.061 (0.341)	0.010 (0.142)	0.014 (0.335)	0.027 (0.144)	0.076 (0.317)	0.073 (0.127)	-0.265 (0.344)	0.023 (0.140)
Observations	1178	1178	1201	1201	1445	1445	1214	1214
R-sqr	-0.003	-0.002	-0.002	-0.001	0.183	0.150	0.143	0.078
Mean of DV	2.282	0.767	2.219	0.755	2.354	0.785	2.379	0.794

Panel B: Link to other variables

	Omission			Audit scores		
	(1)	(2)	(3)	(4)	(5)	(6)
Audit score	-0.070 (0.011)					
Audit score greater than 2		-0.140 (0.028)				
Audit score greater than 3			-0.137 (0.023)			
Algorithm rank based on cosine values				-0.010 (0.002)	-0.003 (0.001)	-0.006 (0.001)
Observations	1214	1214	1214	1214	1214	1214
R-sqr	0.032	0.020	0.029	0.013	0.010	0.022
Mean of DV	0.805	0.805	0.805	2.379	0.794	0.451

The table presents a regression analysis correlating a paper's audit score with the gender of the authors on the suggested paper. Papers were rated on a 0-4 scale, with scores of 2 or higher indicating potential citability due to shared fields and topics. In Panel A: even columns use a binary dependent variable for scores ≥ 2 ; odd columns use direct pair scores. Columns (1)-(2), (3)-(4), (5)-(6), and (7)-(8) pertain to cases excluding failed placebo checks, excluding the most lenient auditor, adding auditors fixed effects, and adding field/style/journal/year controls, respectively. The reference gender category is *All Male*. In Panel B: Column (1) links the audit score to omission via pair ratings. Columns (2) and (3) associate the audit score with omission probability for scores ≥ 2 and ≥ 3 . Columns (4) to (6) relate direct pair scores, scores ≥ 2 and ≥ 3 to average cosine levels.

E.4 Cosine, citation and JEL

E.4.1 Linear relation

Panel A: Citation and cosine

	Outcome variable: Citation					
	Cosine		Cosine		Cosine	
	Top 97 percentile		Top 99 percentile		Top 99 percentile	
					Top 10	
	(1)	(2)	(3)	(4)	(5)	(6)
Algorithm rank based on cosine values	-0.001	-0.001	-0.002	-0.002	-0.003	-0.003
	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)	(0.000)
N	946702	946702	290697	290697	258342	258074
R-sqr	0.022	0.056	0.033	0.080	0.028	0.078
Controls		Y		Y		Y

Panel B: Cosine and JEL code

	Outcome variable: Algorithm rank based on cosine values					
	Cosine		Cosine		Cosine	
	Top 97 percentile		Top 99 percentile		Top 99 percentile	
					Top 10	
	(1)	(2)	(3)	(4)	(5)	(6)
same JEL code	-1.901	-2.495	-2.865	-1.617	-2.964	-0.730
	(0.075)	(0.130)	(0.110)	(0.178)	(0.077)	(0.108)
N	946702	946702	290697	290697	258074	258074
R-sqr	0.001	0.551	0.002	0.554	0.006	0.465
Controls		Y		Y		Y

The table examines the relationship between cosine similarity rankings, citation likelihood, and JEL classification. Higher ranks indicate lower cosine values. Panel A links cosine rank to citation probability, while Panel B relates it to JEL category alignment. Columns (1)-(2), (3)-(4), and (5)-(6) correspond to the top 97th, 99th percentiles, and top 10 ranks in the 99th percentile of cosine similarity, respectively, presented without and then with controls. The set of controls includes the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles; and prior publications. Standard errors are in parentheses.

E.4.2 Variance decomposition of cosine based on same JEL

		$E(y X = x)$		$Var_x E(y X = x)$	$E_x Var(y X = x)$	$Var(y)$	Sum of squares		Sum of squares (more breakdown)	
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
		Same	Not same				Within	Between	Within	Between
		JEL	JEL				groups	groups	groups	groups
TF-IDF		0.49	0.475	6.85E-05	0.067	0.067	430.934	3.601	410.739	23.796
no colocation										
TF-IDF	Colocation (5,0.4)	0.432	0.415	2.4108E-05	0.065	0.065	376.232	5.202	367.5	13.934
with colocation										
Words	Word2Vec	0.873	0.872	5E-07	0.015	0.015	41.562	0.031	35.083	6.51
embeddings										
BERT	Distill-Roberta	0.688	0.683	8.5E-06	0.031	0.031	150.112	0.744	142.321	8.536

The table examines the relationship between cosine similarity rankings and JEL classification, showing a variance decomposition of the cosine based on the same JEL characteristic of a pair of articles. Columns (8) and (9) present the decomposition when breaking down the same JEL category as a union of primary fields.

E.5 Omission and gender: economics (Full Table)

	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.022 (0.005)	0.029 (0.005)	0.026 (0.005)	0.014 (0.005)	-0.003 (0.005)	-0.003 (0.008)
Mixed (Suggested)	-0.031 (0.004)	-0.012 (0.004)	-0.009 (0.003)	0.001 (0.004)	0.000 (0.004)	0.000 (0.006)
Undetermined (Suggested)	0.076 (0.012)	0.068 (0.012)	0.047 (0.012)	0.071 (0.012)	0.053 (0.012)	0.053 (0.016)
Number of authors (Citing)		-0.004 (0.001)	-0.004 (0.001)	-0.004 (0.001)	-0.004 (0.001)	-0.004 (0.001)
Difference in publication years		0.006 (0.000)	0.006 (0.000)	0.006 (0.000)	0.007 (0.000)	0.007 (0.000)
Microeconomics		-0.002 (0.007)	0.012 (0.007)	0.012 (0.007)	0.011 (0.007)	0.011 (0.008)
Macroeconomics		0.035 (0.007)	0.036 (0.007)	0.036 (0.007)	0.035 (0.007)	0.035 (0.009)
International Economics		0.035 (0.008)	0.034 (0.008)	0.033 (0.008)	0.030 (0.008)	0.030 (0.009)
Finance		0.039 (0.007)	0.035 (0.007)	0.036 (0.007)	0.037 (0.007)	0.037 (0.008)
Labour & Education		0.007 (0.008)	0.027 (0.008)	0.026 (0.008)	0.023 (0.008)	0.023 (0.010)
Other (empirical) fields		0.019 (0.008)	0.034 (0.008)	0.033 (0.008)	0.030 (0.008)	0.030 (0.009)
Miscellaneous		0.045 (0.011)	0.039 (0.011)	0.038 (0.011)	0.032 (0.011)	0.032 (0.012)
Same Primary Field		-0.077 (0.003)	-0.079 (0.003)	-0.079 (0.003)	-0.079 (0.003)	-0.079 (0.003)
Number of references (Citing)		-0.006 (0.000)	-0.005 (0.000)	-0.005 (0.000)	-0.005 (0.000)	-0.005 (0.000)
Number of suggested articles		0.015 (0.000)	0.014 (0.000)	0.014 (0.000)	0.014 (0.000)	0.014 (0.000)
Theoretical (Suggested)			-0.037 (0.007)	-0.037 (0.007)	-0.040 (0.007)	-0.040 (0.011)
Empirical (Suggested)			-0.087 (0.008)	-0.083 (0.008)	-0.088 (0.008)	-0.088 (0.012)
Top 5 journals (Suggested)			-0.110	-0.109	-0.102	-0.102

				(0.003)	(0.003)	(0.003)	(0.004)
Affiliation ranking: top 10 (Suggested)				-0.081	-0.073	-0.056	-0.056
				(0.003)	(0.003)	(0.003)	(0.006)
Affiliation ranking: 10-20 (Suggested)				-0.059	-0.053	-0.036	-0.036
				(0.004)	(0.004)	(0.004)	(0.007)
Affiliation ranking: 20-30 (Suggested)				-0.041	-0.035	-0.024	-0.024
				(0.005)	(0.005)	(0.005)	(0.008)
Affiliation ranking: 30-50 (Suggested)				-0.027	-0.020	-0.011	-0.011
				(0.004)	(0.004)	(0.004)	(0.007)
Affiliation ranking: 50-100 (Suggested)				-0.019	-0.015	-0.008	-0.008
				(0.004)	(0.004)	(0.004)	(0.007)
Affiliation ranking: 100-200 (Suggested)				-0.008	-0.004	-0.003	-0.003
				(0.004)	(0.004)	(0.004)	(0.007)
Number of authors (Suggested)					-0.018	-0.012	-0.012
					(0.001)	(0.001)	(0.002)
Publication in top 16: Quintile 2 (Suggested)						-0.027	-0.027
						(0.003)	(0.005)
Publication in top 16: Quintile 3 (Suggested)						-0.048	-0.048
						(0.004)	(0.005)
Publication in top 16: Quintile 4 (Suggested)						-0.067	-0.067
						(0.004)	(0.006)
Publication in top 16: Quintile 5 (Suggested)						-0.072	-0.072
						(0.004)	(0.007)
Observations	151704	151704	151704	151704	151704	151704	151704
R-sqr	0.001	0.047	0.068	0.069	0.071	0.071	
Methodology controls (Citing)		Y	Y	Y	Y	Y	Y
Citing Institution, Journal		Y	Y	Y	Y	Y	Y
Citing Year of publication Fixed effects		Y	Y	Y	Y	Y	Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. Standard errors are reported in parentheses and clustered by the citing papers. Columns (6) replicates columns (5) clustering at the suggested paper level.

E.6 Omission and gender: economics (Logit model)

	Outcome variable: Omission			
	(1)	(2)	(3)	(4)
iomitj				
All Female (Suggested)	0.131 (0.032)	0.157 (0.034)	0.085 (0.034)	-0.038 (0.035)
Mixed (Suggested)	-0.172 (0.019)	-0.049 (0.020)	0.007 (0.021)	-0.003 (0.021)
Undetermined (Suggested)	0.517 (0.099)	0.358 (0.102)	0.571 (0.111)	0.429 (0.105)
Observations	151704	151704	151704	151704
R-sqr	0.001	0.066	0.067	0.070
Basic controls (excluding no. authors of the suggested)		Y	Y	Y
No. authors (suggested)			Y	Y
Prior publications (suggested)				Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The table estimates the baseline equation using a logit model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for the citing paper, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing papers.

E.7 Omission and gender: heterogeneous effects in economics

Panel A: Basic controls

Outcome variable: Omission						
	Full sample	Journal		Affiliation		
		Non Top five	Top five	Top tier	Mid tier	Low tier
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.032 (0.006)	0.022 (0.007)	0.056 (0.012)	0.049 (0.010)	0.003 (0.012)	0.025 (0.009)
Mixed (Suggested)	0.028 (0.004)	0.028 (0.004)	0.027 (0.009)	0.032 (0.007)	0.030 (0.006)	0.012 (0.008)
Team includes women (Citing)	0.010 (0.003)	0.006 (0.004)	0.019 (0.006)	0.012 (0.005)	0.008 (0.005)	0.004 (0.005)
All Female (Suggested) $\times$ Team includes women (Citing)	-0.065 (0.011)	-0.046 (0.013)	-0.101 (0.024)	-0.064 (0.018)	-0.062 (0.023)	-0.058 (0.018)
Mixed (Suggested) $\times$ Team includes women (Citing)	-0.092 (0.008)	-0.082 (0.009)	-0.120 (0.016)	-0.107 (0.012)	-0.085 (0.012)	-0.071 (0.016)
Observations	151704	105152	46552	63565	47274	40865
R-sqr	0.070	0.063	0.083	0.071	0.061	0.052

Panel B: Basic controls + publication records of the suggested papers

Outcome variable: Omission						
	Full sample	Journal		Affiliation		
		Non Top five	Top five	Top tier	Mid tier	Low tier
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.016 (0.006)	0.010 (0.007)	0.023 (0.012)	0.032 (0.010)	-0.012 (0.012)	0.010 (0.009)
Mixed (Suggested)	0.026 (0.004)	0.026 (0.004)	0.025 (0.009)	0.027 (0.007)	0.029 (0.006)	0.016 (0.008)
Team includes women (Citing)	0.009 (0.003)	0.006 (0.004)	0.019 (0.006)	0.012 (0.005)	0.008 (0.005)	0.004 (0.005)
All Female (Suggested) $\times$ Team includes women (Citing)	-0.066 (0.011)	-0.047 (0.013)	-0.102 (0.024)	-0.063 (0.018)	-0.065 (0.023)	-0.058 (0.018)
Mixed (Suggested) $\times$ Team includes women (Citing)	-0.091 (0.008)	-0.081 (0.009)	-0.118 (0.016)	-0.107 (0.012)	-0.083 (0.012)	-0.069 (0.016)
Observations	151704	105152	46552	63565	47274	40865
R-sqr	0.073	0.065	0.087	0.073	0.063	0.057

This table explores the relationship between the omission of the reference and the gender of authors in the suggested paper, with emphasis on the gender of the citing paper and variations across journals and institutions. The dependent variable, omission, is binary, representing if an article does not cite another from the database that belongs to its relevant prior literature, as outlined in Section 3.3. A linear probability model is used for estimations. The reference gender category is *All Male*. *Team includes women (Citing)* notes if the citing article has at least one female author. The set of basic controls includes: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Column (1) shows the entire sample; columns (2) and (3) display results based on the suggested paper's journal rank, and columns (4) to (6) consider the authors' institutional tiers in the suggested paper. Standard errors are clustered by citing articles and given in parentheses.

E.8 Omission and two-sided gender: fields in economics

Panel A: Controls without publication records of the suggested papers

	Outcome variable: Omission							
	Econometrics	Micro-economics	Macro-economics	International Economics	Finance	Labour & Education	Other empirical fields	Miscellaneous
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.040 (0.016)	0.068 (0.019)	-0.017 (0.020)	0.065 (0.013)	0.032 (0.012)	-0.002 (0.018)	0.037 (0.019)	0.062 (0.034)
Mixed (Suggested)	0.048 (0.009)	0.021 (0.013)	-0.009 (0.013)	0.011 (0.013)	0.034 (0.007)	0.033 (0.014)	0.026 (0.016)	-0.057 (0.031)
Team includes women (Citing)	0.022 (0.007)	0.032 (0.010)	-0.018 (0.009)	0.006 (0.009)	0.007 (0.006)	0.015 (0.010)	0.003 (0.011)	0.023 (0.024)
All Female (Suggested) $\times$ Team includes women (Citing)	-0.083 (0.035)	-0.187 (0.049)	-0.022 (0.040)	-0.094 (0.027)	-0.051 (0.022)	-0.007 (0.027)	-0.040 (0.037)	-0.178 (0.074)
Mixed (Suggested) $\times$ Team includes women (Citing)	-0.136 (0.020)	-0.118 (0.029)	-0.066 (0.029)	-0.063 (0.021)	-0.104 (0.013)	-0.066 (0.023)	-0.013 (0.027)	-0.095 (0.058)
Observations	28218	18022	14498	14210	50675	12579	10608	2894
R-sqr	0.074	0.083	0.080	0.099	0.054	0.098	0.092	0.116

Panel B: Controls with publication records of the suggested papers

	Outcome variable: Omission							
	Econometrics	Micro-economics	Macro-economics	International Economics	Finance	Labour & Education	Other empirical fields	Miscellaneous
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.028 (0.016)	0.044 (0.019)	-0.031 (0.020)	0.052 (0.013)	0.014 (0.012)	-0.017 (0.018)	0.024 (0.020)	0.052 (0.035)
Mixed (Suggested)	0.046 (0.009)	0.021 (0.013)	-0.009 (0.013)	0.009 (0.013)	0.032 (0.007)	0.026 (0.014)	0.024 (0.016)	-0.061 (0.031)
Team includes women (Citing)	0.022 (0.007)	0.032 (0.010)	-0.017 (0.009)	0.005 (0.009)	0.007 (0.006)	0.014 (0.010)	0.003 (0.011)	0.023 (0.024)
All Female (Suggested) $\times$ Team includes women (Citing)	-0.084 (0.035)	-0.189 (0.049)	-0.025 (0.040)	-0.094 (0.027)	-0.053 (0.021)	-0.005 (0.027)	-0.039 (0.037)	-0.177 (0.073)
Mixed (Suggested) $\times$ Team includes women (Citing)	-0.133 (0.020)	-0.120 (0.029)	-0.066 (0.029)	-0.063 (0.021)	-0.103 (0.013)	-0.061 (0.023)	-0.012 (0.027)	-0.094 (0.058)
Observations	28218	18022	14498	14210	50675	12579	10608	2894
R-sqr	0.076	0.088	0.084	0.101	0.057	0.102	0.093	0.117

This table shows the relationship between the omission of references and the gender of the authors of the suggested paper, focusing on the heterogeneity by subfields. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variables, the reference category is *All Male*, which represents papers written by only men. The term *Citing* indicates the gender composition of the citing paper. All the models include the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing articles.

E.9 Mathematics: summary statistics

	Full sample	Male	Female	Mixed	Undetermined	P-value (Female-Male)	P-value (Mixed-Male)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total	32,848	25,771	1,005	4,835	1,237		
Number of authors	1.991	1.867	1.269	2.726	2.282	0.000	0.000
Top five publication (%)	27.685	29.390	23.483	21.758	18.755	0.000	0.000
Total authors publications (top 16)	8.663	8.575	4.323	10.578	6.553	0.000	0.000
Total authors publications (top 35)	14.004	14.033	6.7	16.338	10.226	0.000	0.000
Number of references	28.151	28.202	25.732	29.469	23.923	0.000	0.000
Number of references in top 16 journals	2.566	2.505	2.431	2.988	2.280	0.484	0.000
Journal (%)							
Acta Mathematica	0.798	0.916	0.199	0.393	0.404	0.018	0.000
Acta Numerica	0.499	0.559	0.398	0.310	0.081	0.500	0.027
Advances in Mathematics	17.599	18.230	15.423	15.822	13.177	0.023	0.000
American Journal of Mathematics	4.022	4.474	4.179	2.296	1.213	0.657	0.000
Annals of Mathematics	4.107	4.812	2.687	1.572	0.485	0.002	0.000
Annals of Statistics	10.460	10.101	10.846	12.347	10.267	0.442	0.000
Biometrika	8.475	7.291	8.955	13.754	12.126	0.047	0.000
Bulletin of the American Mathematical Society	1.534	1.672	3.682	0.517	0.889	0.000	0.000
Communications on Pure and Applied Mathematics	4.286	4.524	3.881	3.330	3.395	0.334	0.000
Duke Mathematical Journal	5.011	5.619	4.677	2.978	0.566	0.202	0.000
Foundations of Computational Mathematics	1.574	1.595	1.095	1.737	0.889	0.212	0.471
Inventiones Mathematicae	5.891	6.616	4.776	3.144	2.425	0.021	0.000
Journal of Differential Equations	22.893	20.558	30.348	29.059	41.390	0.000	0.000
Journal of the American Mathematical Society	2.073	2.386	1.592	0.931	0.404	0.103	0.000
Mathematical Programming	9.882	9.740	5.970	11.169	10.994	0.000	0.002
Memoirs of the American Mathematical Society	0.895	0.908	1.294	0.641	1.293	0.210	0.066
Affiliation (%)							
Top tier	19.636	19.832	16.219	20.91	13.339	0.005	0.086
Middle Tier	30.139	30.119	21.692	32.947	26.435	0.000	0.000
Low Tier	43.245	42.439	51.94	43.888	50.445	0.000	0.062
Undefined	6.981	7.609	10.149	2.254	9.782	0.004	0.000

The table describes the papers published per journal, and institutions in the database for mathematics from 1985-2019 in the set of top 16 journals. The selected papers exclude proceedings, comments, articles of less than three pages, book reviews, bibliographical items, articles without references and abstracts, editorial material, letters, and corrections.

E.10 Sociology: summary statistics

	Full sample	Male	Female	Mixed	Undetermined	P-value (Female-Male)	P-value (Mixed-Male)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Total	20,591	8,943	5,636	5,796	216		
Number of authors	1.946	1.578	1.506	2.922	2.477	0.000	0.000
Top five publication (%)	38.196	37.035	39.762	38.682	32.407	0.001	0.042
Total authors publications (top 16)	6.966	6.717	5.225	9.086	5.759	0.000	0.000
Total authors publications (top 35)	8.274	8.171	6.066	10.651	6.301	0.000	0.000
Number of references	53.175	52.514	53.427	54.358	42.227	0.081	0.000
Number of references in top 16 journals	6.029	5.357	6.483	6.685	4.384	0.000	0.000
Journals (%)							
American Journal of Sociology	6.357	8.800	3.478	5.452	4.630	0.000	0.000
American Sociological Review	8.232	10.410	5.394	7.885	1.389	0.000	0.001
Annals of Tourism Research	8.528	9.270	5.536	10.041	15.278	0.000	0.959
Annual Review of Sociology	1.486	1.789	1.295	1.208	1.389	0.001	0.003
Gender & Society	4.963	0.962	14.585	1.950	0.463	0.000	0.000
Information, Communication & Society	6.037	5.446	6.902	6.315	0.463	0.307	0.220
Journal of Health and Social Behavior	4.968	3.254	5.181	7.557	0.926	0.000	0.000
Journal of Marriage and Family	14.361	7.648	19.464	19.462	22.222	0.009	0.370
Population and Development Review	6.124	8.632	3.194	4.469	23.148	0.000	0.000
Social Forces	9.106	10.701	6.760	9.058	5.556	0.000	0.128
Social Networks	4.934	6.240	1.366	6.263	8.333	0.027	0.006
Social Problems	4.900	4.819	6.157	3.744	6.481	0.000	0.000
Sociological Methods & Research	3.414	5.323	0.940	2.985	0.463	0.000	0.029
Sociology of Education	3.011	2.885	3.123	3.175	0.926	0.000	0.000
Sociology (Journal of the British Sociological Association)	7.761	8.386	10.131	4.676	2.778	0.000	0.000
Work, Employment and Society	5.818	5.434	6.494	5.763	5.556	0.000	0.000
Affiliation (%)							
Top tier	22.152	19.626	20.972	27.312	18.981	0.045	0.000
Middle Tier	32.050	29.389	31.796	36.663	25.000	0.003	0.000
Low Tier	31.753	32.375	34.226	28.244	35.648	0.014	0.000
Undefined	14.046	18.609	13.006	7.781	20.370	0.000	0.000

The table describes the papers published per journal, and institutions in the database for sociology from 1985-2019 in the set of top 16 journals. The selected papers exclude proceedings, comments, articles of less than three pages, book reviews, bibliographical items, articles without references and abstracts, editorial material, letters, and corrections.

E.11 Algorithms robustness

E.11.1 Economics

	Mean citation	Observations	Basic controls (without no. authors)		Basic controls		Basic controls+ Prior publications	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			All Female	Mixed	All Female	Mixed	All Female	Mixed
BERT (MiniLM-L12 ; 5 most similar)	0.235	137602	0.020 (0.005)	-0.007 (0.004)	0.009 (0.005)	0.004 (0.004)	-0.007 (0.005)	0.003 (0.004)
BERT (Mqnet base ; 5 most similar)	0.247	137496	0.019 (0.005)	-0.005 (0.004)	0.005 (0.006)	0.008 (0.004)	-0.009 (0.006)	0.005 (0.004)
BERT (MiniLM-L6 ; 5 most similar)	0.243	137602	0.025 (0.005)	0.001 (0.004)	0.015 (0.005)	0.010 (0.004)	-0.002 (0.006)	0.008 (0.004)
BERT (Distilbert ; 5 most similar)	0.226	137479	0.018 (0.005)	0.000 (0.004)	0.007 (0.005)	0.010 (0.004)	-0.008 (0.005)	0.008 (0.004)
TF-IDF with colocation (collocation (20,0.1); 5 most similar)	0.146	136025	0.023 (0.004)	0.000 (0.003)	0.015 (0.004)	0.009 (0.003)	0.002 (0.004)	0.007 (0.003)
TF-IDF no colocation (5 most similar)	0.196	136769	0.026 (0.005)	0.006 (0.004)	0.017 (0.005)	0.014 (0.004)	0.002 (0.005)	0.013 (0.004)
BERT (MiniLM-L12 ; 10 most similar)	0.177	275271	0.022 (0.003)	-0.005 (0.003)	0.013 (0.003)	0.004 (0.003)	-0.001 (0.003)	0.002 (0.003)
BERT (Mqnet base ; 10 most similar)	0.187	275059	0.014 (0.004)	-0.003 (0.003)	0.004 (0.004)	0.006 (0.003)	-0.009 (0.004)	0.004 (0.003)
BERT (MiniLM-L6 ; 10 most similar)	0.182	275295	0.025 (0.003)	-0.000 (0.003)	0.017 (0.003)	0.007 (0.003)	0.002 (0.004)	0.005 (0.003)
BERT (Distilbert ; 10 most similar)	0.169	275072	0.017 (0.003)	-0.002 (0.002)	0.009 (0.003)	0.006 (0.003)	-0.005 (0.003)	0.004 (0.003)
TF-IDF with colocation (collocation (20,0.1); 10 most similar)	0.109	265595	0.016 (0.003)	0.003 (0.002)	0.010 (0.003)	0.009 (0.002)	-0.000 (0.003)	0.007 (0.002)
TF-IDF no colocation (10 most similar)	0.14	269604	0.017 (0.003)	0.002 (0.002)	0.010 (0.003)	0.010 (0.002)	-0.003 (0.003)	0.009 (0.002)
Top 3 performing algorithms (5 most similar)	0.163	357836	0.019 (0.003)	-0.001 (0.002)	0.011 (0.003)	0.007 (0.002)	-0.003 (0.003)	0.005 (0.002)
Top 5 performing algorithms (5 most similar)	0.129	549976	0.015 (0.002)	-0.003 (0.002)	0.008 (0.002)	0.004 (0.002)	-0.005 (0.002)	0.002 (0.002)
Exclude algorithms with the lowest accuracy (5 most similar)	0.094	933668	0.012 (0.002)	-0.003 (0.001)	0.006 (0.002)	0.003 (0.001)	-0.003 (0.002)	0.001 (0.001)
BERT (5 most similar)	0.129	549453	0.013 (0.002)	-0.001 (0.002)	0.006 (0.002)	0.005 (0.002)	-0.006 (0.002)	0.003 (0.002)

Word embeddings (5 most similar)	0.106	355797	0.011 (0.003)	-0.009 (0.002)	0.003 (0.003)	-0.002 (0.002)	-0.006 (0.003)	-0.003 (0.002)
TF-IDF with colocation (5 most similar)	0.091	469176	0.016 (0.002)	-0.000 (0.001)	0.010 (0.002)	0.005 (0.002)	0.001 (0.002)	0.004 (0.002)
Top 3 performing algorithms (10 most similar)	0.119	687759	0.016 (0.002)	-0.001 (0.001)	0.010 (0.002)	0.005 (0.001)	-0.001 (0.002)	0.003 (0.001)
Top 5 performing algorithms (10 most similar)	0.091	1055047	0.012 (0.001)	-0.002 (0.001)	0.007 (0.001)	0.002 (0.001)	-0.003 (0.001)	0.001 (0.001)
Exclude algorithms with the lowest accuracy (10 most similar)	0.064	1789389	0.008 (0.001)	-0.002 (0.001)	0.004 (0.001)	0.001 (0.001)	-0.003 (0.001)	0.000 (0.001)
BERT (10 most similar)	0.092	1046502	0.010 (0.001)	-0.002 (0.001)	0.006 (0.001)	0.003 (0.001)	-0.004 (0.001)	0.001 (0.001)
TF-IDF with colocation (10 most similar)	0.064	916866	0.010 (0.001)	0.000 (0.001)	0.006 (0.001)	0.004 (0.001)	-0.001 (0.001)	0.003 (0.001)
Cross-references	0.265	86076	0.014 (0.007)	0.009 (0.005)	0.004 (0.007)	0.018 (0.005)	-0.007 (0.007)	0.017 (0.005)

This table shows the relationship between the omission of references and the gender of the authors of the suggested article for the economics sample, emphasizing the robustness checks associated with the algorithm. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter appears in the former's relevant prior literature, as defined in Section 3.3. The equations are estimated using a linear probability model. Standard errors are reported in parentheses and grouped by citing articles. The rows identify the different alternative methodologies, i.e., the algorithms and how the most similar set is constructed. Column (1) shows the average citation conditional on being in the most similar set of the specific methodology. Column (2) indicates the number of observations. The reference gender category is *All Male*. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles.

E.11.2 Mathematics

	Mean citation	Observations	Basic controls (without no. authors)		Basic controls		Basic controls+ Prior publications	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			All Female	Mixed	All Female	Mixed	All Female	Mixed
Baseline	0.166	134340	0.016 (0.006)	0.015 (0.003)	0.010 (0.006)	0.023 (0.003)	0.000 (0.006)	0.021 (0.003)
Top 10 of each algorithm	0.139	180141	0.009 (0.005)	0.014 (0.002)	0.004 (0.005)	0.019 (0.003)	-0.005 (0.005)	0.017 (0.003)
All top 99 cosine	0.116	220925	0.010 (0.004)	0.012 (0.002)	0.006 (0.004)	0.017 (0.002)	-0.001 (0.004)	0.015 (0.002)
All top 995 cosine	0.169	113025	0.016 (0.006)	0.017 (0.003)	0.011 (0.006)	0.023 (0.003)	0.002 (0.006)	0.021 (0.003)
All top 98 cosine	0.077	428301	0.005 (0.002)	0.007 (0.001)	0.002 (0.003)	0.010 (0.001)	-0.004 (0.003)	0.009 (0.001)
All top 97 cosine	0.06	628919	0.004 (0.002)	0.006 (0.001)	0.002 (0.002)	0.008 (0.001)	-0.003 (0.002)	0.007 (0.001)
Top 3 performing algorithms	0.252	58928	0.018 (0.010)	0.018 (0.005)	0.006 (0.010)	0.030 (0.005)	-0.003 (0.010)	0.028 (0.005)
Top 5 performing algorithms	0.187	95075	0.011 (0.007)	0.014 (0.004)	0.004 (0.007)	0.021 (0.004)	-0.006 (0.007)	0.019 (0.004)
Exclude algorithms with the lowest accuracy	0.121	189495	0.010 (0.004)	0.014 (0.002)	0.006 (0.005)	0.018 (0.002)	-0.002 (0.005)	0.016 (0.002)
BERT	0.17	103950	0.021 (0.006)	0.017 (0.003)	0.013 (0.006)	0.026 (0.004)	0.007 (0.006)	0.024 (0.004)
TF-IDF no colocation	0.286	29599	0.044 (0.014)	0.016 (0.007)	0.031 (0.014)	0.030 (0.008)	0.020 (0.014)	0.026 (0.008)
TF-IDF with colocation	0.205	76004	0.023 (0.008)	0.011 (0.004)	0.013 (0.008)	0.022 (0.004)	0.004 (0.008)	0.019 (0.004)
BERT (MiniLM-L12 ; 5 most similar)	0.150	152288	0.018 (0.005)	0.008 (0.003)	0.010 (0.005)	0.018 (0.003)	0.006 (0.005)	0.016 (0.003)
BERT (Mqnet base ; 5 most similar)	0.148	152280	0.022 (0.005)	0.004 (0.003)	0.015 (0.005)	0.015 (0.003)	0.011 (0.005)	0.013 (0.003)
BERT (MiniLM-L6 ; 5 most similar)	0.149	152308	0.018 (0.005)	0.006 (0.003)	0.011 (0.005)	0.016 (0.003)	0.007 (0.005)	0.015 (0.003)
BERT (Distilbert ; 5 most similar)	0.138	152270	0.013 (0.005)	0.006 (0.003)	0.007 (0.005)	0.015 (0.003)	0.002 (0.005)	0.014 (0.003)
TF-IDF with colocation (collocation (20,0.1); 5 most similar)	0.132	152312	0.021 (0.005)	0.009 (0.003)	0.014 (0.005)	0.019 (0.003)	0.008 (0.005)	0.016 (0.003)
TF-IDF no colocation (5 most similar)	0.144	152297	0.017 (0.005)	0.007 (0.003)	0.010 (0.005)	0.017 (0.003)	0.005 (0.005)	0.015 (0.003)
BERT (MiniLM-L12 ; 10 most similar)	0.100	304315	0.010 (0.003)	0.008 (0.002)	0.004 (0.003)	0.015 (0.002)	0.001 (0.003)	0.014 (0.002)
BERT (Mqnet base ; 10 most similar)	0.099	304264	0.013 (0.003)	0.006 (0.002)	0.007 (0.003)	0.013 (0.002)	0.004 (0.003)	0.012 (0.002)
BERT (MiniLM-L6 ; 10 most similar)	0.099	304341	0.012	0.005	0.007	0.012	0.003	0.011

			(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
BERT (Distilbert ; 10 most similar)	0.092	304236	0.007	0.004	0.002	0.011	-0.003	0.010
			(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
TF-IDF no colocation (10 most similar)	0.096	304428	0.011	0.006	0.006	0.013	0.002	0.011
			(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Top 3 performing algorithms (5 most similar)	0.099	318498	0.011	0.004	0.006	0.012	0.003	0.011
			(0.003)	(0.002)	(0.003)	(0.002)	(0.003)	(0.002)
Top 5 performing algorithms (5 most similar)	0.070	497411	0.006	0.003	0.003	0.007	-0.000	0.006
			(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Exclude algorithms with the lowest accuracy (5 most similar)	0.049	850090	0.004	0.003	0.002	0.006	-0.000	0.005
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BERT (5 most similar)	0.072	492841	0.008	0.004	0.004	0.009	0.001	0.008
			(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Word embeddings (5 most similar)	0.056	335251	0.003	0.003	0.002	0.005	-0.002	0.004
			(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
TF-IDF with colocation (5 most similar)	0.074	445119	0.011	0.005	0.006	0.011	0.002	0.009
			(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Top 3 performing algorithms (10 most similar)	0.064	623836	0.007	0.005	0.003	0.010	0.001	0.009
			(0.002)	(0.001)	(0.002)	(0.001)	(0.002)	(0.001)
Top 5 performing algorithms (10 most similar)	0.044	976957	0.003	0.003	0.001	0.006	-0.001	0.006
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BERT (10 most similar)	0.045	959202	0.004	0.003	0.001	0.007	-0.001	0.006
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Word embeddings (10 most similar)	0.035	666221	-0.001	0.003	-0.001	0.004	-0.004	0.004
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
TF-IDF with colocation (10 most similar)	0.047	872680	0.006	0.003	0.003	0.007	0.000	0.006
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

This table shows the relationship between the omission of references and the gender of the authors of the suggested article for the mathematics sample, emphasizing the robustness checks associated with the algorithm. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter appears in the former’s relevant prior literature, as defined in Section 3.3. The equations are estimated using a linear probability model. Standard errors are reported in parentheses and grouped by citing articles. The rows identify the different alternative methodologies, i.e., the algorithms and how the most similar set is constructed. Column (1) shows the average citation conditional on being in the most similar set of the specific methodology. Column (2) indicates the number of observations. The reference gender category is *All Male*. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles.

E.11.3 Sociology

	Mean citation	Observations	Basic controls (without no. authors)		Basic controls		Basic controls+ Prior publications	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
			All Female	Mixed	All Female	Mixed	All Female	Mixed
Baseline	0.184	94377	0.017 (0.003)	-0.001 (0.003)	0.017 (0.003)	0.004 (0.004)	0.012 (0.003)	0.005 (0.004)
Top 10 of each algorithm	0.156	144935	0.014 (0.002)	0.001 (0.003)	0.014 (0.002)	0.002 (0.003)	0.009 (0.002)	0.003 (0.003)
All top 99 cosine	0.135	175839	0.013 (0.002)	0.002 (0.002)	0.013 (0.002)	0.003 (0.002)	0.009 (0.002)	0.003 (0.002)
All top 995 cosine	0.184	90185	0.016 (0.003)	0.002 (0.003)	0.016 (0.003)	0.005 (0.004)	0.011 (0.003)	0.006 (0.004)
All top 98 cosine	0.095	340788	0.010 (0.001)	0.003 (0.001)	0.010 (0.001)	0.004 (0.002)	0.007 (0.001)	0.004 (0.002)
All top 97 cosine	0.077	500959	0.008 (0.001)	0.003 (0.001)	0.008 (0.001)	0.003 (0.001)	0.005 (0.001)	0.003 (0.001)
Top 3 performing algorithms	0.224	62293	0.021 (0.004)	0.003 (0.004)	0.021 (0.004)	0.003 (0.005)	0.016 (0.004)	0.002 (0.005)
Top 5 performing algorithms	0.189	90494	0.018 (0.003)	0.007 (0.003)	0.018 (0.003)	0.006 (0.004)	0.013 (0.004)	0.005 (0.004)
Exclude algorithms with the lowest accuracy	0.145	152080	0.014 (0.002)	0.002 (0.002)	0.014 (0.002)	0.003 (0.003)	0.010 (0.002)	0.003 (0.003)
BERT	0.185	91777	0.012 (0.003)	-0.002 (0.003)	0.012 (0.003)	-0.002 (0.004)	0.007 (0.003)	-0.003 (0.004)
TF-IDF no colocation	0.252	27647	0.016 (0.007)	0.003 (0.006)	0.016 (0.007)	0.010 (0.007)	0.012 (0.007)	0.013 (0.007)
TF-IDF with colocation	0.173	59986	0.014 (0.004)	0.008 (0.004)	0.014 (0.004)	0.010 (0.004)	0.010 (0.004)	0.011 (0.004)
BERT (MiniLM-L12 ; 5 most similar)	0.177	92053	0.011 (0.003)	-0.001 (0.003)	0.010 (0.003)	0.007 (0.004)	0.007 (0.003)	0.007 (0.004)
BERT (Mqnet base ; 5 most similar)	0.182	92022	0.013 (0.003)	0.001 (0.003)	0.013 (0.003)	0.005 (0.004)	0.010 (0.003)	0.006 (0.004)
BERT (MiniLM-L6 ; 5 most similar)	0.181	92079	0.014 (0.003)	0.001 (0.003)	0.014 (0.003)	0.006 (0.004)	0.011 (0.003)	0.007 (0.004)
BERT (Distilbert ; 5 most similar)	0.18	92032	0.008 (0.003)	0.004 (0.003)	0.007 (0.003)	0.007 (0.004)	0.004 (0.003)	0.008 (0.004)
TF-IDF with colocation (collocation (20,0.1); 5 most similar)	0.142	91984	0.021 (0.003)	0.012 (0.003)	0.021 (0.003)	0.015 (0.003)	0.017 (0.003)	0.016 (0.003)
TF-IDF no colocation (5 most similar)	0.159	92032	0.023 (0.003)	0.013 (0.003)	0.023 (0.003)	0.017 (0.003)	0.020 (0.003)	0.018 (0.003)
BERT (MiniLM-L12 ; 10 most similar)	0.129	184100	0.008 (0.002)	-0.000 (0.002)	0.008 (0.002)	0.003 (0.002)	0.006 (0.002)	0.004 (0.002)
BERT (Mqnet base ; 10 most similar)	0.134	183973	0.008 (0.002)	0.001 (0.002)	0.008 (0.002)	0.004 (0.002)	0.005 (0.002)	0.004 (0.002)
BERT (MiniLM-L6 ; 10 most similar)	0.132	184144	0.011	0.003	0.011	0.006	0.008	0.006

			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
BERT (Distilbert ; 10 most similar)	0.132	184028	0.006	0.002	0.006	0.004	0.003	0.004
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
TF-IDF with colocation (collocation (20,0.1); 10 most similar)	0.104	183912	0.015	0.008	0.015	0.010	0.012	0.011
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
TF-IDF no colocation (10 most similar)	0.118	184017	0.016	0.007	0.016	0.010	0.013	0.011
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Top 3 performing algorithms (5 most similar)	0.131	198644	0.010	0.000	0.010	0.003	0.007	0.003
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Top 5 performing algorithms (5 most similar)	0.106	292869	0.008	-0.001	0.008	0.001	0.006	0.002
			(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Exclude algorithms with the lowest accuracy (5 most similar)	0.078	486176	0.008	0.001	0.008	0.002	0.006	0.002
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BERT (5 most similar)	0.104	299078	0.007	-0.001	0.007	0.000	0.005	0.001
			(0.001)	(0.002)	(0.001)	(0.002)	(0.001)	(0.002)
Word embeddings (5 most similar)	0.093	175386	0.011	0.003	0.011	0.006	0.009	0.007
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
TF-IDF with colocation (5 most similar)	0.096	240154	0.011	0.006	0.011	0.008	0.008	0.008
			(0.002)	(0.002)	(0.002)	(0.002)	(0.002)	(0.002)
Top 3 performing algorithms (10 most similar)	0.093	384412	0.007	0.001	0.006	0.003	0.004	0.003
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Top 5 performing algorithms (10 most similar)	0.074	564408	0.005	0.001	0.005	0.002	0.004	0.002
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
BERT (10 most similar)	0.073	572708	0.005	0.000	0.005	0.001	0.003	0.002
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)
Word embeddings (10 most similar)	0.066	343405	0.007	0.001	0.007	0.003	0.006	0.004
			(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)

This table shows the relationship between the omission of references and the gender of the authors of the suggested article for the sociology sample, emphasizing the robustness checks associated with the algorithm. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter appears in the former’s relevant prior literature, as defined in Section 3.3. The equations are estimated using a linear probability model. Standard errors are reported in parentheses and grouped by citing articles. The rows identify the different alternative methodologies, i.e., the algorithms and how the most similar set is constructed. Column (1) shows the average citation conditional on being in the most similar set of the specific methodology. Column (2) indicates the number of observations. The reference gender category is *All Male*. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles.

E.12 “Case-controls”

	Outcome variable: Omission					
	Model 1		Model 2		Model 3	
	Economics					
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.078 (0.005)	0.032 (0.005)	0.081 (0.004)	0.033 (0.004)	0.049 (0.006)	0.020 (0.006)
Mixed	0.016 (0.003)	0.009 (0.003)	0.023 (0.002)	0.015 (0.002)	0.009 (0.003)	0.005 (0.003)
Undetermined	0.111 (0.012)	0.060 (0.012)	0.124 (0.010)	0.073 (0.010)	0.065 (0.019)	0.026 (0.019)
Observations	339500	339500	489267	489267	131308	131308
	Mathematics					
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.023 (0.006)	0.008 (0.006)	0.030 (0.010)	0.003 (0.010)	0.018 (0.020)	-0.005 (0.020)
Mixed	0.035 (0.003)	0.032 (0.003)	0.047 (0.005)	0.041 (0.005)	0.064 (0.011)	0.057 (0.011)
Undetermined	0.078 (0.005)	0.057 (0.005)	0.139 (0.009)	0.095 (0.009)	0.087 (0.020)	0.050 (0.020)
Observations	156870	156870	99900	99900	20463	20463
	Sociology					
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.039 (0.003)	0.030 (0.003)	0.047 (0.003)	0.035 (0.003)	0.033 (0.006)	0.021 (0.006)
Mixed	0.004 (0.003)	0.005 (0.003)	0.009 (0.004)	0.012 (0.004)	-0.001 (0.006)	0.002 (0.006)
Undetermined	0.093 (0.011)	0.073 (0.011)	0.121 (0.012)	0.097 (0.012)	0.121 (0.022)	0.102 (0.022)
Observations	150428	150428	154061	154061	45401	45401
Basic controls	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y

This table illustrates the link between omission from references and the gender of suggested paper authors, akin to a case-control analysis. Algorithm predictions are combined with author choices to form a list of potentially citable articles. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. “Model 1” augments author selections with the top 10 similar papers from algorithms. “Model 2” extends this to include the top n similar papers, where n equals the length of the authors’ selections. “Model 3” resembles “Model 2,” but only includes cited papers sharing sub-fields with the citing paper, acknowledging potential sub-field identification limitations (for example, unavailable or self-reported JELs). The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing papers.

E.13 Difference in language?

	Outcome variable: Omission						
	Max IDF	Mean IDF	Median IDF	Basic controls		Basic+ Prior publications	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All Female (Suggested)	-0.027 (0.030)	0.002 (0.010)	0.002 (0.013)	0.014 (0.005)	-0.003 (0.005)	0.014 (0.005)	-0.003 (0.005)
Mixed (Suggested)	-0.004 (0.020)	0.007 (0.007)	0.007 (0.008)	0.001 (0.004)	0.000 (0.004)	0.002 (0.004)	0.000 (0.004)
Undetermined (Suggested)	-0.096 (0.086)	-0.049 (0.029)	-0.047 (0.034)	0.073 (0.012)	0.056 (0.012)	0.071 (0.012)	0.053 (0.012)
Observations	23875	23875	23875	151704	151704	151704	151704
R-sqr	0.039	0.119	0.092	0.069	0.072	0.069	0.072
Basic controls	Y	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)	Y	Y	Y			Y	Y
Control: IDF				Y		Y	
Control: Text readability					Y		Y

This table analyses the difference in vocabulary used by gender taking the example of economics. Columns (1), (2) and (3) are linear regressions where the dependent variables are, respectively, the maximum Inverse document frequency (IDF) of words in a given paper, the mean IDF, and the median IDF by articles. Columns (4) to (7) explores the relationship between the omission from references and the gender of authors in the suggested paper, with emphasis on the gender of the citing paper. The dependent variable, omission, is binary, representing if an article does not cite another from the database that belongs to its relevant prior literature, as outlined in Section 3.3. A linear probability model is used for estimations. The reference gender category is *All Male*. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. In columns (1) to (3), the set of basic controls is one-sided. Standard errors are reported in parentheses.

E.14 Varying rank of most similar papers

Outcome variable: Omission						
	Rank ≤ 2		Rank ≤ 5		Rank ≤ 7	
	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.
Economics						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.018 (0.010)	0.002 (0.010)	0.015 (0.007)	-0.002 (0.007)	0.015 (0.006)	-0.002 (0.006)
Mixed (Suggested)	0.004 (0.007)	0.003 (0.007)	0.006 (0.005)	0.004 (0.005)	0.005 (0.004)	0.004 (0.004)
Undetermined (Suggested)	0.064 (0.023)	0.047 (0.023)	0.066 (0.016)	0.047 (0.016)	0.069 (0.014)	0.051 (0.014)
Observations	47220	47220	97268	97268	122391	122391
Mathematics						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.016 (0.011)	0.007 (0.011)	0.011 (0.007)	0.001 (0.007)	0.013 (0.007)	0.004 (0.007)
Mixed (Suggested)	0.029 (0.006)	0.027 (0.006)	0.025 (0.004)	0.022 (0.004)	0.023 (0.004)	0.021 (0.004)
Undetermined (Suggested)	0.056 (0.009)	0.044 (0.009)	0.049 (0.006)	0.036 (0.006)	0.045 (0.005)	0.032 (0.005)
Observations	47922	47922	91982	91982	112125	112125
Sociology						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.020 (0.006)	0.016 (0.006)	0.018 (0.004)	0.013 (0.004)	0.019 (0.004)	0.014 (0.004)
Mixed (Suggested)	0.018 (0.007)	0.019 (0.007)	0.008 (0.005)	0.009 (0.005)	0.006 (0.004)	0.006 (0.004)
Undetermined (Suggested)	0.038 (0.019)	0.028 (0.019)	0.030 (0.013)	0.020 (0.013)	0.025 (0.011)	0.016 (0.011)
Observations	30533	30533	61433	61433	76516	76516

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper, emphasizing the average relative value of the cosine as proxy by the rank of the suggested paper across the algorithms. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. All the tables include the basic set of controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing papers.

E.15 Granular field controls

Panel A: Economics

	Outcome variable: Omission			
	2-digit JEL (130)		3-digit JEL (760)	
	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.
	(1)	(2)	(3)	(4)
All Female (Suggested)	0.015 (0.005)	-0.001 (0.005)	0.017 (0.005)	0.000 (0.005)
Mixed (Suggested)	0.003 (0.004)	0.001 (0.004)	0.004 (0.004)	0.003 (0.004)
Observations	147676	147676	147676	147676
R-sqr	0.071	0.074	0.073	0.076

Panel B: Mathematics

	Outcome variable: Omission			
	MSC (63)		MSC (570)	
	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.
	(1)	(2)	(3)	(4)
All Female (Suggested)	0.011 (0.006)	0.003 (0.006)	0.011 (0.006)	0.003 (0.006)
Mixed (Suggested)	0.024 (0.004)	0.021 (0.004)	0.024 (0.004)	0.021 (0.004)
Observations	104777	104777	104777	104777
R-sqr	0.089	0.092	0.092	0.095

Panel C: Sociology

	Outcome variable: Omission			
	Optimal clusters (60)		MAG fields (4000)	
	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.
	(1)	(2)	(3)	(4)
All Female (Suggested)	0.016 (0.003)	0.012 (0.003)	0.014 (0.003)	0.010 (0.003)
Mixed (Suggested)	0.005 (0.004)	0.006 (0.004)	0.004 (0.004)	0.005 (0.004)
Observations	89297	89297	89297	89297
R-sqr	0.076	0.080	0.090	0.094

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper, with a more granular field controls. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. MSC is Mathematics Subject Classification. The number in parenthesis near the classification code is the number of distinct categories. Standard errors are reported in parentheses and clustered by the citing papers.

E.16 Splits based on specific categories

	Outcome variable: Omission							
	Less prominent journals		Prominent journals	Fields with few women	Fields with more women			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Economics								
All Female (Suggested)	0.016 (0.007)	0.007 (0.007)	0.050 (0.011)	0.022 (0.011)	0.040 (0.008)	0.022 (0.008)	0.016 (0.010)	0.004 (0.010)
Mixed (Suggested)	0.029 (0.005)	0.027 (0.005)	0.025 (0.007)	0.023 (0.007)	0.031 (0.005)	0.029 (0.005)	0.019 (0.008)	0.017 (0.008)
Team includes women (Citing)	0.002 (0.004)	0.002 (0.004)	0.019 (0.005)	0.019 (0.005)	0.009 (0.004)	0.009 (0.004)	0.010 (0.006)	0.009 (0.006)
All Female (Suggested) × Team includes women (Citing)	-0.039 (0.013)	-0.040 (0.013)	-0.108 (0.022)	-0.108 (0.021)	-0.084 (0.016)	-0.086 (0.016)	-0.036 (0.017)	-0.035 (0.017)
Mixed (Suggested) × Team includes women (Citing)	-0.077 (0.009)	-0.076 (0.009)	-0.112 (0.013)	-0.111 (0.013)	-0.103 (0.010)	-0.101 (0.010)	-0.067 (0.013)	-0.067 (0.013)
Observations	84810	84810	66894	66894	114753	114753	36951	36951
R-sqr	0.053	0.055	0.076	0.080	0.064	0.068	0.096	0.098
Mathematics								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.051 (0.007)	0.045 (0.007)	0.024 (0.012)	0.016 (0.012)	0.050 (0.010)	0.041 (0.010)	0.040 (0.010)	0.032 (0.010)
Mixed (Suggested)	0.047 (0.004)	0.045 (0.004)	0.051 (0.006)	0.049 (0.006)	0.053 (0.006)	0.050 (0.006)	0.055 (0.006)	0.054 (0.006)
Team includes women (Citing)	0.029 (0.004)	0.029 (0.004)	0.020 (0.005)	0.019 (0.005)	0.026 (0.005)	0.025 (0.005)	0.031 (0.005)	0.031 (0.005)
All Female (Suggested) × Team includes women (Citing)	-0.148 (0.017)	-0.148 (0.017)	-0.093 (0.026)	-0.096 (0.026)	-0.111 (0.023)	-0.111 (0.023)	-0.145 (0.022)	-0.146 (0.022)
Mixed (Suggested) × Team includes women (Citing)	-0.114 (0.009)	-0.113 (0.009)	-0.084 (0.013)	-0.083 (0.013)	-0.117 (0.012)	-0.114 (0.012)	-0.116 (0.011)	-0.115 (0.011)
Observations	79168	79168	55172	55172	52488	52488	52289	52289
R-sqr	0.083	0.086	0.105	0.106	0.101	0.105	0.081	0.084
Sociology								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.057 (0.005)	0.053 (0.005)	0.038 (0.013)	0.036 (0.013)	0.072 (0.006)	0.067 (0.006)	0.038 (0.013)	0.032 (0.013)
Mixed (Suggested)	0.028 (0.006)	0.028 (0.006)	0.026 (0.012)	0.028 (0.012)	0.029 (0.006)	0.030 (0.006)	0.040 (0.014)	0.041 (0.013)
Team includes women (Citing)	0.032 (0.005)	0.032 (0.005)	0.026 (0.009)	0.027 (0.009)	0.035 (0.005)	0.035 (0.005)	0.041 (0.010)	0.041 (0.010)
All Female (Suggested) × Team includes women (Citing)	-0.069 (0.007)	-0.068 (0.007)	-0.101 (0.017)	-0.102 (0.017)	-0.080 (0.007)	-0.079 (0.007)	-0.061 (0.015)	-0.060 (0.015)
Mixed (Suggested) × Team includes women (Citing)	-0.045 (0.007)	-0.046 (0.007)	-0.050 (0.014)	-0.051 (0.014)	-0.047 (0.007)	-0.048 (0.007)	-0.053 (0.015)	-0.053 (0.015)
Observations	72202	72202	22175	22175	75681	75681	18696	18696
R-sqr	0.065	0.068	0.120	0.122	0.081	0.084	0.068	0.071
Basic controls	Y	Y	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y		Y

This table illustrates the relationship between citation omissions and the authors' gender in the cited paper, with data segmented by specific criteria. Columns (1) to (4) differentiate journals by their reduction in omissions, ie, publishing in those journals reduces the likelihood of omissions (columns 3 and 4) versus other journals (columns 1 and 2). Columns (5) to (8) categorize papers by subfields where women's share of publications is in the top 25% (columns 7 and 8), with the rest in columns (5) and (6). The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. The set of basic controls includes methodology, journal types, team sizes, publication year difference, references count in the citing article, number of suggested articles, and fixed effects for journals, fields, and publication years of the citing paper and fixed effects for the institutions. Standard errors are reported in parentheses and clustered by the citing papers.

E.17 Citing paper fixed effects

Outcome variable: Omission			
Economics			
	(1)	(2)	(3)
All Female (Suggested)	0.026 (0.005)	0.016 (0.005)	-0.000 (0.006)
Mixed (Suggested)	-0.017 (0.004)	0.006 (0.004)	0.003 (0.004)
Undetermined (Suggested)	0.071 (0.013)	0.057 (0.013)	0.042 (0.013)
Observations	148554	148554	148554
Mathematics			
	(1)	(2)	(3)
All Female (Suggested)	0.020 (0.006)	0.010 (0.006)	0.001 (0.006)
Mixed (Suggested)	0.005 (0.003)	0.021 (0.003)	0.019 (0.003)
Undetermined (Suggested)	0.044 (0.005)	0.044 (0.005)	0.033 (0.005)
Observations	129820	129820	129820
Sociology			
	(1)	(2)	(3)
All Female (Suggested)	0.007 (0.003)	0.014 (0.003)	0.009 (0.003)
Mixed (Suggested)	-0.011 (0.003)	0.005 (0.004)	0.005 (0.004)
Undetermined (Suggested)	0.034 (0.011)	0.029 (0.011)	0.019 (0.011)
Observations	92056	92056	92056
Basic controls		Y	Y
Prior publications (Suggested)			Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper. Contrary to the baseline, the estimation is using a citing paper fixed effect. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are clustered by citing articles and given in parentheses.

E.18 Broader set of fixed effects

Outcome variable: Omission								
Economics								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.036 (0.006)	0.019 (0.006)	0.027 (0.006)	0.012 (0.006)	0.026 (0.006)	0.012 (0.006)	0.026 (0.006)	0.012 (0.006)
Mixed (Suggested)	0.028 (0.004)	0.026 (0.004)	0.026 (0.004)	0.025 (0.004)	0.026 (0.004)	0.024 (0.004)	0.028 (0.004)	0.026 (0.004)
Team includes women (Citing)	0.009 (0.003)	0.009 (0.003)	0.009 (0.003)	0.009 (0.003)	0.010 (0.003)	0.009 (0.003)	0.010 (0.003)	0.010 (0.003)
All Female (Suggested) × Team includes women (Citing)	-0.065 (0.011)	-0.066 (0.011)	-0.066 (0.011)	-0.067 (0.011)	-0.066 (0.011)	-0.067 (0.011)	-0.066 (0.011)	-0.067 (0.011)
Mixed (Suggested) × Team includes women (Citing)	-0.092 (0.008)	-0.091 (0.008)	-0.091 (0.008)	-0.090 (0.008)	-0.091 (0.008)	-0.090 (0.008)	-0.092 (0.008)	-0.091 (0.008)
Observations	151704	151704	151271	151271	151704	151704	151704	151704
R-sqr	0.074	0.076	0.075	0.077	0.075	0.077	0.084	0.086
Mathematics								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.047 (0.006)	0.038 (0.006)	0.040 (0.006)	0.033 (0.006)	0.041 (0.006)	0.034 (0.006)	0.039 (0.006)	0.032 (0.006)
Mixed (Suggested)	0.053 (0.004)	0.051 (0.004)	0.049 (0.004)	0.047 (0.004)	0.048 (0.004)	0.046 (0.004)	0.047 (0.004)	0.045 (0.004)
Team includes women (Citing)	0.026 (0.003)	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)	0.025 (0.003)	0.024 (0.003)
All Female (Suggested) × Team includes women (Citing)	-0.133 (0.014)	-0.133 (0.014)	-0.131 (0.014)	-0.132 (0.014)	-0.132 (0.014)	-0.132 (0.014)	-0.129 (0.014)	-0.130 (0.014)
Mixed (Suggested) × Team includes women (Citing)	-0.107 (0.007)	-0.105 (0.007)	-0.106 (0.007)	-0.105 (0.007)	-0.106 (0.007)	-0.105 (0.007)	-0.107 (0.007)	-0.105 (0.007)
Observations	134340	134340	134340	134340	134340	134340	134340	134340
R-sqr	0.090	0.093	0.091	0.093	0.091	0.093	0.098	0.100
Sociology								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
All Female (Suggested)	0.066 (0.005)	0.061 (0.005)	0.056 (0.006)	0.052 (0.006)	0.052 (0.005)	0.049 (0.005)	0.053 (0.005)	0.050 (0.005)
Mixed (Suggested)	0.032 (0.006)	0.034 (0.006)	0.028 (0.006)	0.029 (0.006)	0.027 (0.006)	0.028 (0.006)	0.030 (0.006)	0.031 (0.006)
Team includes women (Citing)	0.037 (0.004)	0.037 (0.004)	0.032 (0.005)	0.033 (0.005)	0.030 (0.004)	0.030 (0.004)	0.031 (0.004)	0.032 (0.004)
All Female (Suggested) × Team includes women (Citing)	-0.078 (0.006)	-0.078 (0.006)	-0.080 (0.007)	-0.080 (0.007)	-0.074 (0.006)	-0.074 (0.006)	-0.076 (0.006)	-0.076 (0.006)
Mixed (Suggested) × Team includes women (Citing)	-0.048 (0.007)	-0.049 (0.007)	-0.051 (0.007)	-0.052 (0.007)	-0.046 (0.007)	-0.047 (0.007)	-0.048 (0.007)	-0.049 (0.007)
Observations	94377	94377	83041	83041	94377	94377	94377	94377
R-sqr	0.083	0.087	0.099	0.101	0.091	0.093	0.097	0.099
Suggested Field × Year FE	Y	Y						
Time-varying journal impact Factor			Y	Y				
Suggested journals FE					Y	Y		
Suggested journals × Year FE							Y	Y
Prior publications (Suggested)		Y		Y		Y		Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper, with a more granular set of fixed effect controls. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. All the estimations include the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing papers. The impact factor data comes from Marx and Fuegi [2020].

E.19 Prominence: alternative measure (example of economics)

	Outcome variable: Omission			
	Top 16	Top 35	Top 5	Top 16 Last 5 years
	Max prominence			
	(1)	(2)	(3)	(4)
All Female (Suggested)	-0.004 (0.005)	-0.002 (0.005)	0.005 (0.005)	0.001 (0.005)
Mixed (Suggested)	-0.000 (0.004)	0.000 (0.004)	0.000 (0.004)	-0.001 (0.004)
Undetermined (Suggested)	0.051 (0.012)	0.055 (0.012)	0.059 (0.012)	0.055 (0.012)
Observations	151704	151704	151704	151704
R-sqr	0.072	0.071	0.071	0.072
	Average prominence			
	(1)	(2)	(3)	(4)
All Female (Suggested)	-0.003 (0.005)	-0.001 (0.005)	0.004 (0.005)	0.002 (0.005)
Mixed (Suggested)	-0.002 (0.004)	-0.002 (0.004)	-0.001 (0.004)	-0.003 (0.004)
Undetermined (Suggested)	0.050 (0.012)	0.054 (0.012)	0.061 (0.012)	0.055 (0.012)
Observations	151704	151704	151704	151704
R-sqr	0.072	0.071	0.070	0.072
	Total prominence			
	(1)	(2)	(3)	(4)
All Female (Suggested)	-0.005 (0.005)	-0.003 (0.005)	0.004 (0.005)	-0.000 (0.005)
Mixed (Suggested)	-0.001 (0.004)	-0.000 (0.004)	-0.001 (0.004)	-0.002 (0.004)
Undetermined (Suggested)	0.042 (0.012)	0.046 (0.012)	0.055 (0.012)	0.047 (0.012)
Observations	151704	151704	151704	151704
R-sqr	0.072	0.071	0.071	0.072

This table shows the relationship between the omission of references and the gender of the authors of the suggested paper, emphasizing the effect of seniority. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variables, the reference category is *All Male*, which represents papers written by only men. All the columns include the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Seniority is measured by how prolific the authors of the papers are. This includes measures based on the maximum publications, the average publication, the total publications in the top 16 (column 1), top 35 (column 2), top 5 (column 3) and top 16 publications in the last 5 years (column 4). Standard errors are reported in parentheses and clustered by the citing papers.

E.20 Excluding Self-citation

	Outcome variable: Omission						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All Female (Suggested)	0.017 (0.005)	0.022 (0.005)	0.012 (0.005)	-0.004 (0.005)	0.012 (0.006)	-0.003 (0.006)	0.003 (0.008)
Mixed (Suggested)	-0.029 (0.003)	-0.009 (0.003)	0.000 (0.004)	-0.001 (0.003)	0.010 (0.004)	0.008 (0.004)	0.018 (0.008)
Undetermined (Suggested)	0.070 (0.012)	0.043 (0.012)	0.065 (0.011)	0.048 (0.011)	0.072 (0.013)	0.054 (0.012)	0.042 (0.015)
Team includes women (Citing)					-0.004 (0.003)	-0.004 (0.003)	
All Female (Suggested) $\times$ Team includes women (Citing)					-0.003 (0.011)	-0.004 (0.011)	
Mixed (Suggested) $\times$ Team includes women (Citing)					-0.034 (0.008)	-0.033 (0.008)	
Publication in top 16 (Suggested)							-0.020 (0.001)
All Female (Suggested) $\times$ Publication in top 16 (Suggested)							-0.004 (0.004)
Mixed (Suggested) $\times$ Publication in top 16 (Suggested)							-0.007 (0.003)
Observations	147609	147609	147609	147609	147609	147609	147609
R-sqr	0.001	0.063	0.064	0.066	0.064	0.067	0.066
Basic controls (Citing)		Y	Y	Y	Y	Y	Y
Basic controls (suggested)		Y	Y	Y	Y	Y	Y
Number of authors (Suggested)			Y	Y	Y	Y	Y
Prior publications (Suggested)				Y		Y	Y

This table shows the relationship between the omission of references and the gender of the authors of the suggested paper, dropping the instance self-citations (example of economics). Self-citation is defined as any instance where at least one of the authors on the paper cites their own paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are clustered by citing papers and reported in parentheses.

E.21 Multiple citations

Outcome variable: Omission						
Panel A: Same gender substitution						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.113 (0.006)	0.111 (0.006)	0.086 (0.006)	0.036 (0.006)	0.108 (0.006)	0.059 (0.006)
Mixed (Suggested)	0.030 (0.004)	0.055 (0.004)	0.076 (0.004)	0.070 (0.004)	0.108 (0.005)	0.101 (0.005)
Team includes women (Citing)					0.016 (0.004)	0.015 (0.004)
All Female (Suggested) $\times$ Team includes women (Citing)					-0.076 (0.012)	-0.079 (0.012)
Mixed (Suggested) $\times$ Team includes women (Citing)					-0.111 (0.009)	-0.107 (0.009)
Observations	151704	151704	151704	151704	151704	151704
R-sqr	0.003	0.083	0.087	0.108	0.088	0.109
Panel B: Any substitution						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.079 (0.006)	0.079 (0.006)	0.051 (0.006)	-0.007 (0.006)	0.078 (0.007)	0.020 (0.007)
Mixed (Suggested)	-0.035 (0.004)	-0.005 (0.004)	0.019 (0.004)	0.011 (0.004)	0.051 (0.005)	0.042 (0.005)
Team includes women (Citing)					0.014 (0.004)	0.013 (0.004)
All Female (Suggested) $\times$ Team includes women (Citing)					-0.090 (0.013)	-0.093 (0.013)
Mixed (Suggested) $\times$ Team includes women (Citing)					-0.110 (0.009)	-0.106 (0.009)
Observations	151704	151704	151704	151704	151704	151704
R-sqr	0.002	0.087	0.091	0.119	0.093	0.120
Basic controls (Citing)		Y	Y	Y	Y	Y
Basic controls (suggested)		Y	Y	Y	Y	Y
Number of authors (Suggested)			Y	Y	Y	Y
Prior publications (Suggested)				Y		Y

This table analyzes the relationship between the omission of references and the gender of authors (example of economics). It considers situations where another paper by a listed author is cited. I account for those situations by assuming that the suggested paper was also cited. Expressly, Panel A assumes that the suggested paper is also cited if the cited article has the same gender composition as the suggested, and Panel B assumes that the suggested paper is also cited if any paper of at least one of the authors is cited, regardless of the gender composition. The dependent variable “omission” is binary, indicating if an article does not cite another given its relevance as per Section 3.3. The analysis uses a linear probability model with standard errors clustered by citing papers. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles.

E.22 Vary year difference between citing and cited

Outcome variable: Omission						
	> 1		> 3		> 5	
	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.	Basic controls	Basic+ Prior pub.
Economics						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.013 (0.006)	-0.004 (0.006)	0.014 (0.006)	-0.004 (0.006)	0.013 (0.007)	-0.005 (0.007)
Mixed (Suggested)	0.002 (0.004)	0.001 (0.004)	-0.002 (0.004)	-0.004 (0.004)	-0.006 (0.005)	-0.007 (0.005)
Observations	137738	137738	111740	111740	88921	88921
R-sqr	0.075	0.077	0.076	0.079	0.073	0.077
Mathematics						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.006 (0.006)	-0.004 (0.006)	0.001 (0.007)	-0.007 (0.007)	0.004 (0.007)	-0.004 (0.007)
Mixed (Suggested)	0.025 (0.003)	0.022 (0.003)	0.025 (0.004)	0.022 (0.004)	0.023 (0.004)	0.020 (0.004)
Undetermined (Suggested)	0.040 (0.005)	0.027 (0.005)	0.038 (0.005)	0.025 (0.005)	0.037 (0.006)	0.023 (0.006)
Observations	120450	120450	95184	95184	74340	74340
R-sqr	0.086	0.089	0.082	0.085	0.077	0.081
Sociology						
	(1)	(2)	(3)	(4)	(5)	(6)
All Female (Suggested)	0.017 (0.003)	0.013 (0.003)	0.016 (0.004)	0.012 (0.004)	0.015 (0.004)	0.011 (0.004)
Mixed (Suggested)	0.003 (0.004)	0.004 (0.004)	-0.002 (0.004)	-0.001 (0.004)	-0.001 (0.005)	-0.001 (0.005)
Observations	86167	86167	70842	70842	57208	57208
R-sqr	0.081	0.085	0.081	0.084	0.078	0.082

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper, emphasizing the effect of varying the time between the publication year of the citing paper and the cited paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. All the columns include the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are reported in parentheses and clustered by the citing papers.

E.23 Extension to other journals

E.23.1 Top 35 journals

Outcome variable: Omission					
Economics					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.010 (0.002)	0.003 (0.003)	-0.006 (0.003)	0.025 (0.003)	0.016 (0.003)
Mixed (Suggested)	-0.004 (0.002)	0.004 (0.002)	0.002 (0.002)	0.023 (0.002)	0.021 (0.002)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.071 (0.006)	-0.071 (0.006)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.060 (0.004)	-0.060 (0.004)
Observations	444023	444023	444023	440830	440830
R-sqr	0.068	0.068	0.070	0.070	0.072
Mathematics					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.020 (0.001)	0.017 (0.001)	0.015 (0.001)	0.047 (0.002)	0.044 (0.002)
Mixed (Suggested)	0.008 (0.001)	0.011 (0.001)	0.011 (0.001)	0.033 (0.001)	0.032 (0.001)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.078 (0.003)	-0.077 (0.003)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.055 (0.002)	-0.055 (0.002)
Observations	1099815	1099815	1099815	1035339	1035339
R-sqr	0.050	0.050	0.052	0.053	0.054
Sociology					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.012 (0.002)	0.012 (0.002)	0.010 (0.002)	0.044 (0.003)	0.043 (0.003)
Mixed (Suggested)	-0.001 (0.002)	0.003 (0.002)	0.004 (0.002)	0.023 (0.003)	0.023 (0.003)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.054 (0.003)	-0.054 (0.003)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.036 (0.004)	-0.036 (0.004)
Observations	229570	229570	229570	227454	227454
R-sqr	0.063	0.063	0.064	0.064	0.065
Basic controls (excluding no. authors of the suggested)	Y	Y	Y	Y	Y
No. authors (suggested)		Y	Y	Y	Y
Prior publications (suggested)			Y		Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper in a set of 35 high-impact journals in each discipline. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The reference category for gender is papers by male-only author teams. Results are derived from a linear probability model with standard errors, in parentheses, clustered by citing article. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. The mean value of the citation conditional on being in the similar set is 0.163 in economics, 0.098 in mathematics, and 0.126 in sociology.

E.23.2 More than 90 journals

Outcome variable: Omission					
Economics					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.007 (0.002)	0.002 (0.002)	-0.002 (0.002)	0.024 (0.002)	0.021 (0.002)
Mixed (Suggested)	-0.013 (0.001)	-0.004 (0.001)	-0.005 (0.001)	0.018 (0.001)	0.016 (0.001)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.067 (0.004)	-0.067 (0.004)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.063 (0.003)	-0.063 (0.003)
Observations	863081	863081	863081	849113	849113
R-sqr	0.073	0.073	0.074	0.075	0.075
Mathematics					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.024 (0.002)	0.020 (0.002)	0.016 (0.002)	0.057 (0.002)	0.054 (0.002)
Mixed (Suggested)	0.007 (0.001)	0.013 (0.001)	0.012 (0.001)	0.044 (0.001)	0.043 (0.001)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.111 (0.004)	-0.110 (0.004)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.086 (0.002)	-0.086 (0.002)
Observations	1134916	1134916	1134916	1090169	1090169
R-sqr	0.083	0.084	0.085	0.087	0.088
Sociology					
	(1)	(2)	(3)	(4)	(5)
All Female (Suggested)	0.008 (0.001)	0.007 (0.001)	0.007 (0.001)	0.031 (0.001)	0.031 (0.001)
Mixed (Suggested)	-0.009 (0.001)	0.000 (0.001)	0.001 (0.001)	0.018 (0.001)	0.019 (0.001)
All Female (Suggested) $\times$ Team includes women (Citing)				-0.045 (0.002)	-0.045 (0.002)
Mixed (Suggested) $\times$ Team includes women (Citing)				-0.035 (0.002)	-0.036 (0.002)
Observations	740050	740050	740050	729140	729140
R-sqr	0.101	0.102	0.102	0.103	0.104
Basic controls (excluding no. authors of the suggested)	Y	Y	Y	Y	Y
No. authors (suggested)		Y	Y	Y	Y
Prior publications (suggested)			Y		Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper in a set of more than 90 journals in each discipline. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The reference category for gender is papers by male-only author teams. Results are derived from a linear probability model with standard errors, in parentheses, clustered by citing article. The set of basic controls includes methodological variables, journal type of the suggested paper, team sizes, publication year difference, main field alignment, references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. The mean value of the citation conditional on being in the similar set is 0.167 in economics, 0.159 in mathematics, and 0.094 in sociology. Due to computational constraints and the algorithm's robustness, the top five most similar articles from the best performing algorithm have been used in the estimation.

E.24 Working papers information

	Outcome variable: Omission			
	(1)	(2)	(3)	(4)
All Female	0.025 (0.005)	0.023 (0.005)	0.013 (0.005)	-0.011 (0.005)
Mixed	-0.001 (0.003)	-0.006 (0.003)	0.002 (0.003)	-0.000 (0.003)
Undetermined	0.074 (0.012)	0.050 (0.012)	0.068 (0.012)	0.040 (0.012)
Observations	183806	183806	183806	183806
R-sqr	0.000	0.058	0.059	0.065
Basic controls (suggested)		Y	Y	Y
Number of authors (Suggested)			Y	Y
Prior publications (Suggested)				Y

This table shows the relationship between the omission of references and the gender of the authors of the suggested paper, including the earliest publication date available from the working paper version of those papers. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are clustered by citing papers and reported in parentheses.

E.25 Position and Gender of the most prolific author

E.25.1 Economics

	Outcome variable: Omission					
	Position			Most prominent		
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.014 (0.005)	-0.003 (0.005)	0.016 (0.006)			
Mixed female first	0.004 (0.005)	0.003 (0.005)	0.035 (0.006)			
Mixed male first	-0.000 (0.005)	-0.002 (0.004)	0.020 (0.005)			
Team includes women (Citing)			0.009 (0.003)			0.009 (0.003)
All Female $\times$ Team includes women (Citing)			-0.066 (0.011)			
Mixed female first $\times$ Team includes women (Citing)			-0.109 (0.011)			
Mixed male first $\times$ Team includes women (Citing)			-0.078 (0.010)			
All Female				0.013 (0.005)	-0.003 (0.005)	0.016 (0.006)
Mixed female senior				-0.009 (0.007)	-0.014 (0.007)	0.016 (0.008)
Mixed Male senior				0.000 (0.004)	0.002 (0.004)	0.026 (0.005)
Other mixed				0.045 (0.011)	0.026 (0.011)	0.062 (0.012)
All Female $\times$ Team includes women (Citing)						-0.066 (0.011)
Mixed female senior $\times$ Team includes women (Citing)						-0.100 (0.015)
Mixed Male senior $\times$ Team includes women (Citing)						-0.084 (0.009)
Other mixed $\times$ Team includes women (Citing)						-0.127 (0.026)
Observations	151704	151704	151704	151704	151704	151704
R-sqr	0.069	0.071	0.073	0.069	0.072	0.073
Basic controls	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y

This table presents the link between reference omission and author gender in suggested papers, focusing on the gender of the author at the first position (Columns (1) to (3)) and the gender of the most prominent author (Columns (4) to (6)). The binary dependent variable 'omission' indicates if an article does not cite another from the database that's part of its relevant prior literature, as defined in Section 3.3. A linear probability model is used for estimation. The baseline gender category is *All Male*, denoting papers by male-only teams. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors, clustered by citing papers, are in parentheses.

E.25.2 Mathematics

	Outcome variable: Omission					
	Position			Most prominent		
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.010 (0.006)	0.001 (0.006)	0.037 (0.006)			
Mixed female first	0.026 (0.004)	0.024 (0.004)	0.054 (0.005)			
Mixed male first	0.021 (0.004)	0.018 (0.004)	0.046 (0.005)			
Team includes women (Citing)			0.026 (0.003)			0.026 (0.003)
All Female $\times$ Team includes women (Citing)			-0.133 (0.014)			
Mixed female first $\times$ Team includes women (Citing)			-0.111 (0.010)			
Mixed male first $\times$ Team includes women (Citing)			-0.102 (0.009)			
All Female				0.010 (0.006)	0.001 (0.006)	0.036 (0.006)
Mixed female senior				0.017 (0.006)	0.013 (0.006)	0.041 (0.007)
Mixed Male senior				0.025 (0.004)	0.026 (0.004)	0.056 (0.004)
Other mixed				0.025 (0.008)	-0.002 (0.008)	0.022 (0.010)
All Female $\times$ Team includes women (Citing)						-0.133 (0.014)
Mixed female senior $\times$ Team includes women (Citing)						-0.094 (0.014)
Mixed Male senior $\times$ Team includes women (Citing)						-0.111 (0.008)
Other mixed $\times$ Team includes women (Citing)						-0.099 (0.020)
Observations	134340	134340	134340	134340	134340	134340
R-sqr	0.082	0.085	0.088	0.082	0.085	0.088
Basic controls	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y

This table presents the link between reference omission and author gender in suggested papers, focusing on the gender of the author at the first position (Columns (1) to (3)) and the gender of the most prominent author (Columns (4) to (6)). The binary dependent variable 'omission' indicates if an article does not cite another from the database that's part of its relevant prior literature, as defined in Section 3.3. A linear probability model is used for estimation. The baseline gender category is *All Male*, denoting papers by male-only teams. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors, clustered by citing papers, are in parentheses.

E.25.3 Sociology

	Outcome variable: Omission					
	Position			Most prominent		
	(1)	(2)	(3)	(4)	(5)	(6)
All Female	0.017 (0.003)	0.013 (0.003)	0.061 (0.005)			
Mixed female first	0.010 (0.004)	0.010 (0.004)	0.048 (0.007)			
Mixed male first	-0.002 (0.005)	-0.001 (0.005)	0.016 (0.007)			
Team includes women (Citing)			0.036 (0.004)			0.036 (0.004)
All Female $\times$ Team includes women (Citing)			-0.076 (0.006)			
Mixed female first $\times$ Team includes women (Citing)			-0.065 (0.008)			
Mixed male first $\times$ Team includes women (Citing)			-0.033 (0.008)			
All Female				0.017 (0.003)	0.012 (0.003)	0.061 (0.005)
Mixed female senior				-0.016 (0.006)	-0.011 (0.006)	0.023 (0.010)
Mixed Male senior				-0.001 (0.004)	0.006 (0.004)	0.028 (0.007)
Other mixed				0.050 (0.006)	0.023 (0.006)	0.053 (0.010)
All Female $\times$ Team includes women (Citing)						-0.076 (0.006)
Mixed female senior $\times$ Team includes women (Citing)						-0.059 (0.012)
Mixed Male senior $\times$ Team includes women (Citing)						-0.040 (0.008)
Other mixed $\times$ Team includes women (Citing)						-0.053 (0.012)
Observations	94377	94377	94377	94377	94377	94377
R-sqr	0.076	0.079	0.081	0.076	0.079	0.081
Basic controls	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y

This table presents the link between reference omission and author gender in suggested papers, focusing on the gender of the author at the first position (Columns (1) to (3)) and the gender of the most prominent author (Columns (4) to (6)). The binary dependent variable 'omission' indicates if an article does not cite another from the database that's part of its relevant prior literature, as defined in Section 3.3. A linear probability model is used for estimation. The baseline gender category is *All Male*, denoting papers by male-only teams. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors, clustered by citing papers, are in parentheses.

E.26 Using only title and only abstract to find the similar set (Example of economics)

Outcome variable: Omission					
Only titles					
	(1)	(2)	(3)	(4)	(5)
All Female	0.005 (0.004)	0.009 (0.004)	0.009 (0.004)	-0.004 (0.004)	-0.008 (0.004)
Mixed	-0.027 (0.003)	-0.005 (0.003)	-0.002 (0.003)	-0.001 (0.003)	0.004 (0.003)
Undetermined	0.031 (0.010)	0.041 (0.010)	0.028 (0.010)	0.019 (0.010)	0.034 (0.010)
Observations	138267	138267	138267	138267	138267
R-sqr	0.001	0.040	0.052	0.055	0.056
Only abstracts					
	(1)	(2)	(3)	(4)	(5)
All Female	0.014 (0.005)	0.020 (0.005)	0.017 (0.005)	-0.000 (0.005)	-0.004 (0.005)
Mixed	-0.042 (0.004)	-0.015 (0.004)	-0.010 (0.004)	-0.008 (0.004)	-0.003 (0.004)
Undetermined	0.030 (0.012)	0.029 (0.012)	0.012 (0.012)	0.002 (0.012)	0.012 (0.012)
Observations	137567	137567	137567	137567	137567
R-sqr	0.001	0.040	0.057	0.060	0.060
Basic controls (Citing)		Y	Y	Y	Y
Basic controls (suggested)			Y	Y	Y
Prior publications (Suggested)				Y	Y
Number of authors (Suggested)					Y

This table shows the relationship between the omission from references and the gender of the authors of the suggested paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. For the gender variable, the reference category is *All Male*, which represents papers written by only men. Panel A show the results using only titles and panel B shows the results using only abstracts. For computational reasons and given that the results are not too sensitive to the algorithmic procedure, the top five most similar articles from the best performing algorithm have been used in the estimation. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors, clustered by citing papers, are in parentheses.

E.27 Link between gender and the other regressors (Case of economics)

Panel A: Pearson correlation

	All Female vs. All Male		Mixed vs. All Male		Share of female	
	Pearson correlation	P-Value	Pearson correlation	P-Value	Pearson correlation	P-Value
Publication in top 16	-0.192	0.000	0.039	0.000	-0.127	0.000
Publication in top 35	-0.197	0.000	0.050	0.000	-0.126	0.000
Publication in top 5	-0.130	0.000	-0.024	0.000	-0.108	0.000
Publication in top 35 (5 years before)	-0.145	0.000	0.032	0.000	-0.096	0.000
Avg publication in top 16	-0.167	0.000	-0.034	0.000	-0.142	0.000
Avg publication in top 35	-0.170	0.000	-0.028	0.000	-0.141	0.000
Avg publication in top 5	-0.110	0.000	-0.072	0.000	-0.118	0.000
Avg publication in top 35 (5 years before)	-0.113	0.000	-0.035	0.000	-0.103	0.000
Total publication in top 16	-0.205	0.000	0.067	0.000	-0.124	0.000
Total publication in top 35	-0.212	0.000	0.083	0.000	-0.123	0.000
Total publication in top 5	-0.135	0.000	-0.012	0.000	-0.107	0.000
Total publication in top 35 (5 years before)	-0.146	0.000	0.048	0.000	-0.088	0.000
Theoretical	-0.072	0.000	-0.148	0.000	-0.127	0.000
Empirical	0.108	0.000	0.178	0.000	0.166	0.000
Structural	-0.031	0.000	-0.017	0.003	-0.033	0.000
Top 5 journals	-0.007	0.272	-0.032	0.000	-0.024	0.000
Number of authors	-0.171	0.000	0.340	0.000	0.032	0.000

Panel B: Test of independence

	Gender categorial variable		
	Chi2	P-Value	Cramer's V
Publication in top 16	7200	0.000	0.126
Publication in top 35	8900	0.000	0.140
Publication in top 5	3100	0.000	0.083
Publication in top 35 (5 years before)	3900	0.000	0.093
Avg publication in top 16	15000	0.000	0.183
Avg publication in top 35	18000	0.000	0.201
Avg publication in top 5	9300	0.000	0.143
Avg publication in top 35 (5 years before)	9400	0.000	0.143
Total publication in top 16	8800	0.000	0.139
Total publication in top 35	10000	0.000	0.149
Total publication in top 5	3400	0.000	0.086
Total publication in top 35 (5 years before)	4200	0.000	0.096
Top 5 journals	57.177	0.000	0.043
Affiliation	336.044	0.000	0.060
Number of authors	4800	0.000	0.229
Field	649.901	0.000	0.083
Journal	801.86	0.000	0.093

This table shows the correlation and the test of independence between the variable characterizing the gender of the authors on an article and the other control variables.

E.28 Gender perception

	All Female sample			All Male sample		
	Proba $\geq$ 0.8	Proba $<$ 0.8	p-value	Proba $\geq$ 0.8	Proba $<$ 0.8	p-value
Number of authors	1.218	1.368	0.000	1.866	1.886	0.398
Top 5 journals	0.316	0.258	0.107	0.325	0.258	0.000
Affiliation: Top tier	0.368	0.353	0.690	0.412	0.333	0.000
Econometrics	0.095	0.111	0.489	0.158	0.259	0.000
Microeconomics	0.111	0.111	0.990	0.181	0.177	0.723
Macroeconomics	0.089	0.063	0.232	0.098	0.075	0.007
International Economics	0.132	0.132	0.978	0.079	0.066	0.090
Finance	0.188	0.274	0.006	0.255	0.266	0.416
Labour & Education	0.232	0.132	0.002	0.094	0.061	0.000
Other (empirical) fields	0.123	0.121	0.944	0.108	0.080	0.001
Theoretical	0.519	0.490	0.342	0.644	0.568	0.000
Empirical	0.416	0.416	0.992	0.252	0.217	0.000
Structural	0.065	0.094	0.114	0.104	0.215	0.000
At least one non white author	0.206	0.684	0.000	0.284	0.709	0.000
Publication in top 16	1.729	1.974	0.312	6.341	6.489	0.568
Publication in top 35	2.735	2.911	0.620	9.049	9.152	0.767
Publication in top 5	0.759	0.621	0.302	2.714	2.326	0.008

This table shows the balance check comparing papers with (fe)maleness below 80 percent to those with (fe)maleness above 80 percent.

E.29 Effect of connection between the citing and the suggested papers

	Outcome variable: Omission									
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
All Female	0.020 (0.005)	-0.006 (0.005)	0.019 (0.005)	-0.005 (0.005)	0.016 (0.005)	-0.005 (0.005)	0.017 (0.005)	-0.007 (0.005)	0.014 (0.005)	-0.006 (0.005)
Mixed	-0.008 (0.003)	-0.000 (0.004)	-0.009 (0.003)	-0.002 (0.004)	-0.008 (0.004)	-0.001 (0.004)	-0.009 (0.003)	-0.002 (0.004)	-0.008 (0.004)	-0.001 (0.004)
Same affiliation	-0.048 (0.007)	-0.043 (0.007)								
All Female $\times$ Same affiliation	0.059 (0.032)	0.055 (0.032)								
Mixed $\times$ Same affiliation	-0.012 (0.017)	-0.017 (0.017)								
Co-authors			-0.162 (0.007)	-0.152 (0.007)						
All Female $\times$ Co-authors			0.021 (0.041)	0.020 (0.041)						
Mixed $\times$ Co-authors			0.025 (0.018)	0.022 (0.018)						
Co-authors and Co-authors of Co-authors					-0.096 (0.004)	-0.087 (0.004)				
All Female $\times$ Co-authors and Co-authors of Co-authors					0.010 (0.022)	0.008 (0.022)				
Mixed $\times$ Co-authors and Co-authors of Co-authors					0.002 (0.010)	-0.001 (0.010)				
Same affiliation— Co-authors							-0.115 (0.005)	-0.108 (0.005)		
All Female $\times$ Same affiliation— Co-authors							0.055 (0.027)	0.051 (0.027)		
Mixed $\times$ Same affiliation— Co-authors							0.019 (0.013)	0.015 (0.013)		
All levels of connection									-0.090 (0.004)	-0.081 (0.004)
All Female $\times$ All levels of connection									0.024 (0.019)	0.021 (0.019)
Mixed $\times$ All levels of connection									0.001 (0.009)	-0.002 (0.009)
Observations	147609	147609	147609	147609	147609	147609	147609	147609	147609	147609
R-sqr	0.063	0.067	0.069	0.071	0.069	0.071	0.067	0.070	0.069	0.071
Basic controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
Prior publications (Suggested)		Y		Y		Y		Y		Y

This table shows the relationship between the omission and the gender of the authors of the cited or omitted paper emphasizing the effect of being in the same affiliation or being in a group of coauthors. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. This regression excludes the self-citation. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are clustered by papers and reported in parentheses.

E.30 Validation of innovativeness index: Link with citations

E.30.1 All top 16 journals

	Outcome variable: Omission					
	(1)	(2)	(3)	(4)	(5)	(6)
Innovativeness 0-3	0.272 (0.014)	0.259 (0.014)				
All Female	-0.045 (0.025)	0.054 (0.025)	-0.074 (0.027)	0.046 (0.027)	-0.064 (0.029)	0.067 (0.028)
Mixed	-0.039 (0.015)	-0.015 (0.015)	-0.036 (0.017)	-0.006 (0.016)	-0.028 (0.018)	0.006 (0.017)
Undetermined	-0.116 (0.067)	0.024 (0.065)	-0.191 (0.074)	-0.024 (0.072)	-0.199 (0.078)	-0.017 (0.075)
All Female $\times$ Innovativeness 0-3	-0.040 (0.042)	-0.046 (0.042)				
Mixed $\times$ Innovativeness 0-3	0.041 (0.024)	0.053 (0.024)				
Undetermined $\times$ Innovativeness 0-3	-0.026 (0.116)	-0.016 (0.104)				
Innovativeness 0-5			0.568 (0.023)	0.540 (0.023)		
All Female $\times$ Innovativeness 0-5			-0.080 (0.043)	-0.094 (0.043)		
Mixed $\times$ Innovativeness 0-5			0.030 (0.022)	0.046 (0.022)		
Undetermined $\times$ Innovativeness 0-5			-0.152 (0.101)	-0.084 (0.090)		
Innovativeness 0-7					0.788 (0.029)	0.743 (0.028)
All Female $\times$ Innovativeness 0-7					-0.061 (0.041)	-0.080 (0.041)
Mixed $\times$ Innovativeness 0-7					0.020 (0.021)	0.038 (0.021)
Undetermined $\times$ Innovativeness 0-7					-0.141 (0.091)	-0.055 (0.081)
Observations	27593	27593	27593	27593	27593	27593
R-sqr	0.258	0.291	0.310	0.349	0.340	0.381
Publications controls		Y		Y		Y

This table shows the relationship between the number of citations and the innovativeness index. Citation is measured using the asinh function where $asinh(x) = \ln(x + \sqrt{x^2 + 1})$. Innovativeness is built based on the description in section D. In particular, the logarithm of the innovativeness index is used. For the gender variables, the reference category is *All Male*, which represents papers written by only men. All the model accounts for factors like journal, field, institution, publication year, number of authors. Standard errors are reported in parentheses.

E.30.2 All top 5 journals

	Outcome variable: Omission					
	(1)	(2)	(3)	(4)	(5)	(6)
Innovativeness 0-3	0.294 (0.026)	0.278 (0.025)				
All Female	-0.031 (0.051)	0.074 (0.051)	-0.108 (0.056)	0.023 (0.055)	-0.126 (0.057)	0.016 (0.056)
Mixed	-0.039 (0.031)	-0.012 (0.031)	-0.035 (0.033)	-0.003 (0.032)	-0.014 (0.034)	0.022 (0.033)
Undetermined	0.031 (0.201)	0.088 (0.184)	-0.084 (0.229)	-0.017 (0.214)	-0.110 (0.244)	-0.050 (0.230)
All Female $\times$ Innovativeness 0-3	-0.191 (0.085)	-0.186 (0.082)				
Mixed $\times$ Innovativeness 0-3	-0.019 (0.050)	-0.013 (0.051)				
Undetermined $\times$ Innovativeness 0-3	-0.533 (0.482)	-0.354 (0.403)				
Innovativeness 0-5			0.628 (0.040)	0.587 (0.038)		
All Female $\times$ Innovativeness 0-5			-0.174 (0.101)	-0.188 (0.098)		
Mixed $\times$ Innovativeness 0-5			-0.013 (0.046)	0.006 (0.046)		
Undetermined $\times$ Innovativeness 0-5			-0.232 (0.314)	-0.091 (0.275)		
Innovativeness 0-7					0.849 (0.050)	0.791 (0.048)
All Female $\times$ Innovativeness 0-7					-0.124 (0.100)	-0.154 (0.096)
Mixed $\times$ Innovativeness 0-7					-0.024 (0.043)	-0.002 (0.043)
Undetermined $\times$ Innovativeness 0-7					-0.226 (0.325)	-0.098 (0.292)
Observations	8523	8523	8523	8523	8523	8523
R-sqr	0.175	0.209	0.219	0.261	0.243	0.288
Publications controls		Y		Y		Y

This table shows the relationship between the number of citations and the innovativeness index. Citation is measured using the asinh function where $asinh(x) = \ln(x + \sqrt{x^2 + 1})$. Innovativeness is built based on the description in section D. In particular, the logarithm of the innovativeness index is used. For the gender variables, the reference category is *All Male*, which represents papers written by only men. All the model accounts for factors like journal, field, institution, publication year, number of authors. Standard errors are reported in parentheses.

F Additional Figures

F.1 Example: construction of cosine similarity and omission indicator

Raw Data Data Cleaning $\Rightarrow$ Tokens (Set of words)

	paper p	paper p'	paper p''
trade_credit	6	7	5
retailer	5	1	1
supplier	6	4	1
purchase	3	3	0
bank	1	1	0
cash_constrain	1	1	0
creditor	1	1	0
finance	1	2	1
prediction	1	1	0
free_rider	3	0	0
competition	3	0	1
product_substitutability	2	0	0
...	...	...	...
Total	...	...	...

(a) Step 1

w : word ; p : paper, N : total number of papers in the database

Term Frequency TFIDF (w,p) $\times$ Inverse Document Frequency

$$TF(w,p) = \frac{\text{Number of } w \text{ in } p}{\text{Total Number of words in } p}$$

$$IDF(w) = \frac{\text{Number of papers with } w}{N}$$

	paper p	paper p'	paper p''
trade_credit	$\frac{6}{Total_p} * IDF(\text{trade_credit})$	$\frac{7}{Total_p} * IDF(\text{trade_credit})$	$\frac{5}{Total_p} * IDF(\text{trade_credit})$
retailer	$\frac{5}{Total_p} * IDF(\text{retailer})$	$\frac{1}{Total_p} * IDF(\text{retailer})$	$\frac{1}{Total_p} * IDF(\text{retailer})$
...	...	...	...

(b) Step 2

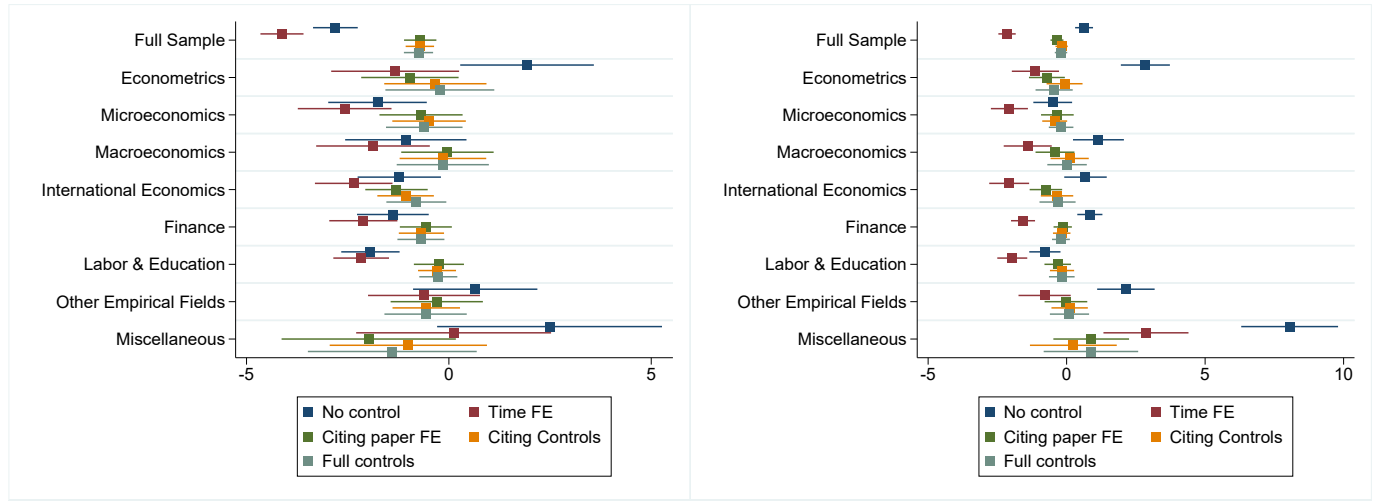
	p: Citing paper (2019)	p': Female paper Omitted (2010)	p'': Male paper cited(2004)
trade_credit	6	7	5
retailer	5	1	1
supplier	6	4	1
purchase	3	3	0
bank	1	1	0
cash_constrain	1	1	0
creditor	1	1	0
finance	1	2	1
prediction	1	1	0
free_rider	3	0	0
competition	3	0	1
product_substitutability	2	0	0
...	...	...	...
Total	...	...	...

(c) Step 3

Figure:
Example

This figure illustrates the construction of the cosine similarity and the omission indicator. Panel (a) shows a part of the document-term matrix for a sample of three papers. Panel (b) presents the construction of the TF-IDF. Panel (c) presents the construction of the omission indicator.

F.2 Cosine, Gender and Fields (Economics)



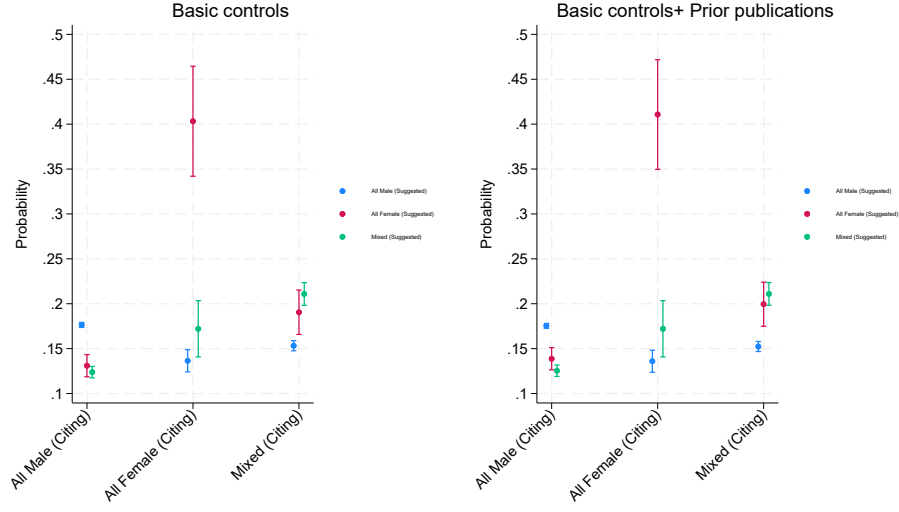
(a) All Female vs All Male

(b) Mixed Gender vs All Male

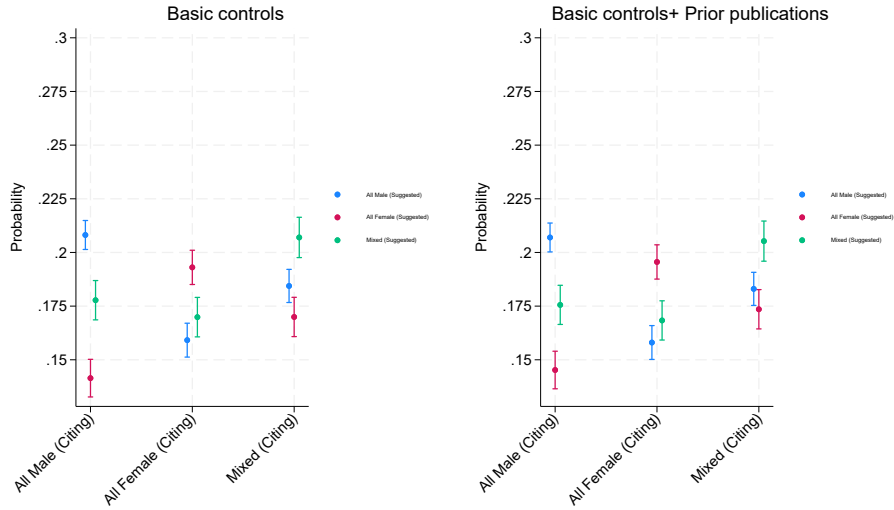
Figure:
Cosine similarity, field and gender

The figure displays the coefficients of linear regressions. In these regressions, the dependent variable is the ranking derived from the cosine similarity, while the variable of interest is gender, with the reference category being all-male authors. A negative coefficient indicates gender-based differences in cosine values, implying that the considered category tends to have higher cosine values than the reference category (all-male authors). Conversely, a positive coefficient means the opposite, and a null coefficient suggests no significant difference. Panel (a) referred to the comparison between all-female authors and all-male authors while Panel (b) referred to the comparison between mixed-gender teams and all-male authors.

F.3 Citation in the most similar set and the gender of the citing: mathematics and sociology



(a) Mathematics



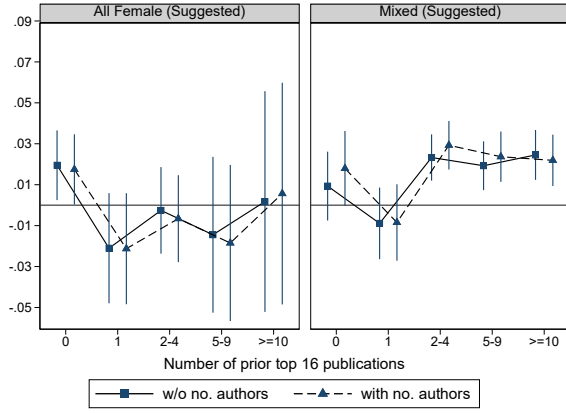
(b) Sociology

Figure:

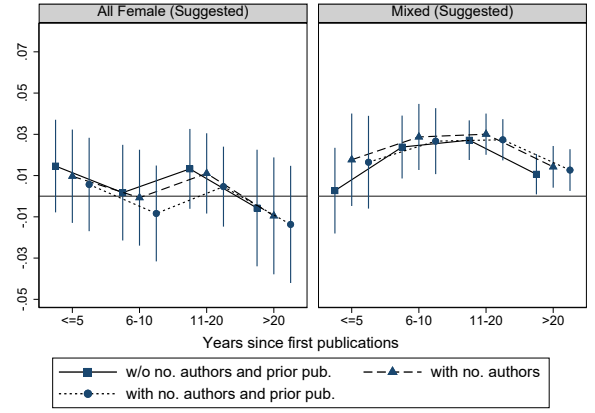
Citation in the most similar set and the gender of the citing

This figure illustrates the relationship between the citation in the most similar set and the gender of the authors of the suggested paper, with emphasis on the gender of the citing paper in mathematics and sociology. The dependent variable, citation conditioning on being in the most similar set is $1 - omission$, where omission is binary, representing if an article does not reference another from the database that belongs to its relevant prior literature, as outlined in Section 3.3. A linear probability model is used for estimations. The reference gender category is *All Male*. *Team includes women (Citing)* notes if the citing article has at least one female author. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. Standard errors are clustered by the citing papers.

F.4 Omission, gender and seniority: mathematics and sociology

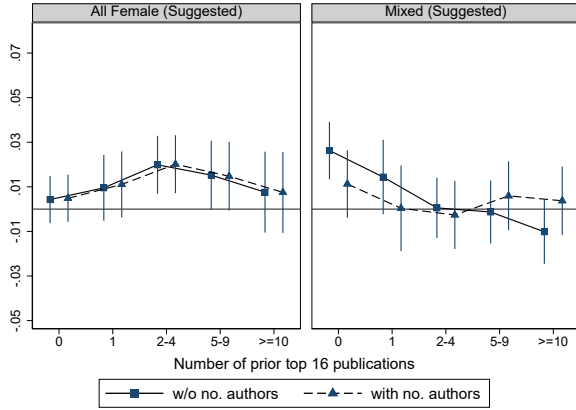


(a)

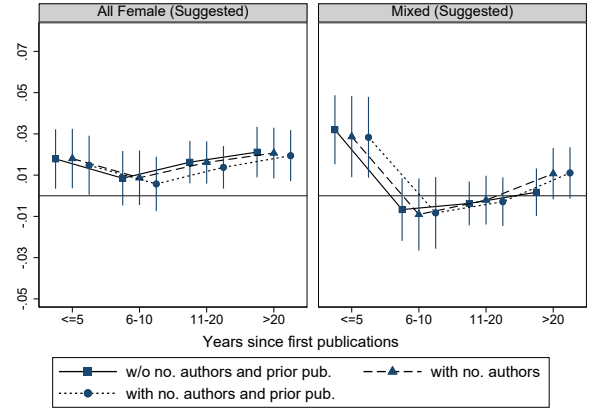


(b)

Part A: Mathematics



(c)



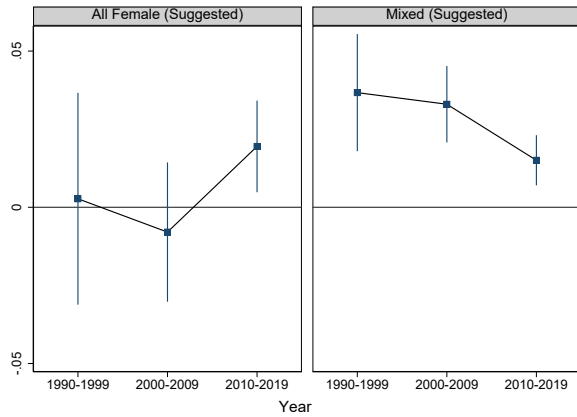
(d)

Part B: Sociology

Figure:
Omission and seniority

The figures show the relationship between the omission of references and the gender of the authors of the suggested paper, emphasizing the effect of seniority. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. For the publications variables, the values on the horizontal axis correspond to the quartile of the distribution. All the figures are estimated using a linear probability model and include controls for methodological variables, journal type of the suggested paper, team size, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. For the gender variables, the reference category is *All Male*, which represents papers written by only men. Standard errors are clustered by the citing papers.

F.5 Evolution over time: Mathematics and sociology

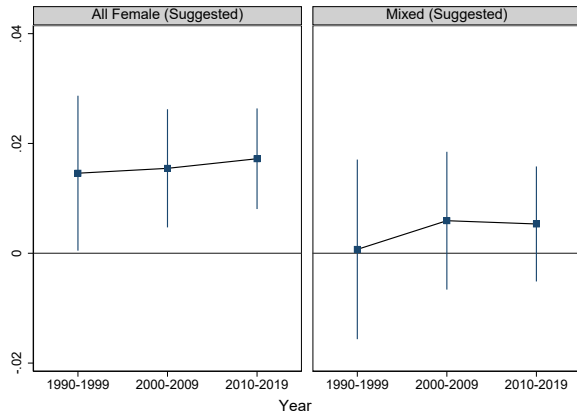


(a) Basic controls

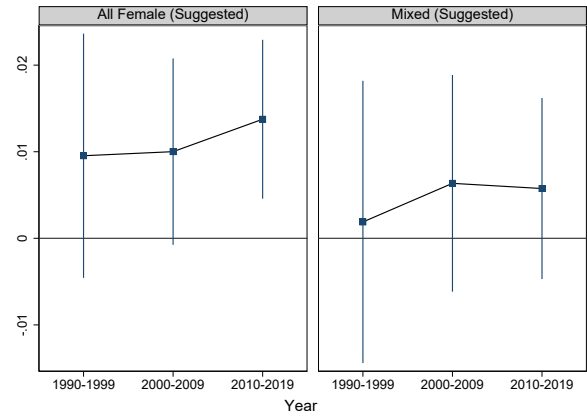


(b) Basic controls+ prior publications

Part A: Mathematics



(c) Basic controls



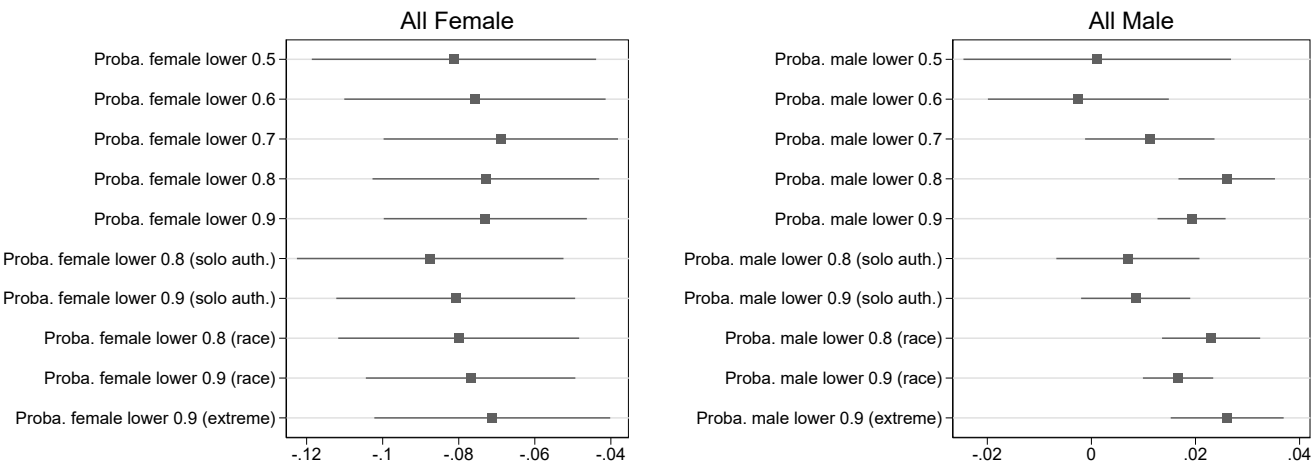
(d) Basic controls+ prior publications

Part B: Sociology

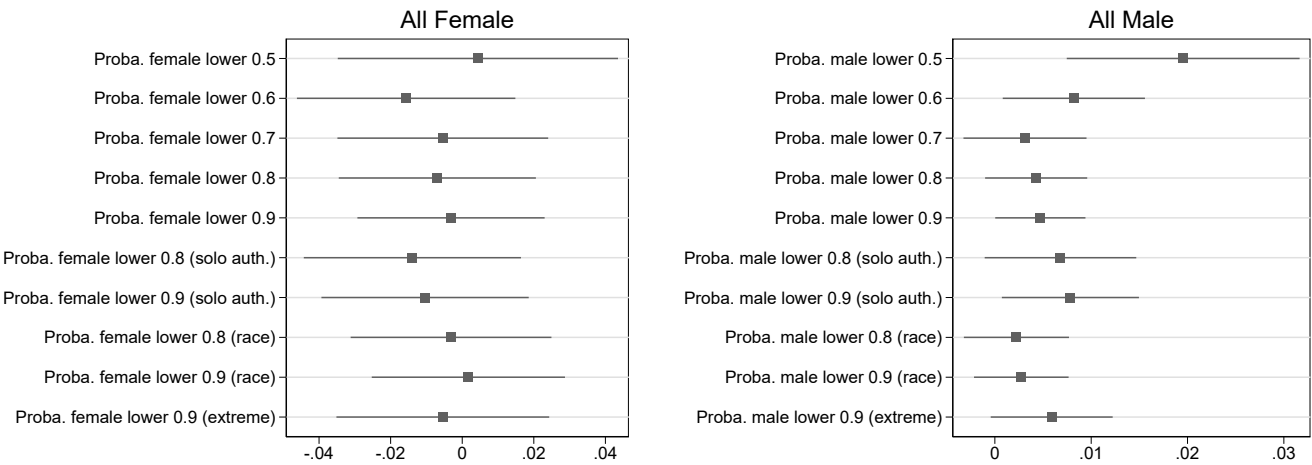
Figure:
Gender and omission: Evolution over time

This figure illustrates the relationship between the omission of references and the gender of the authors of the suggested paper, emphasizing the effect over time. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former, as defined in Section 3.3. The equations are estimated using a linear probability model. The set of basic controls includes methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles. For the gender variable, the reference category is *All Male*, which represents papers written by only men. Standard errors are clustered by the citing papers.

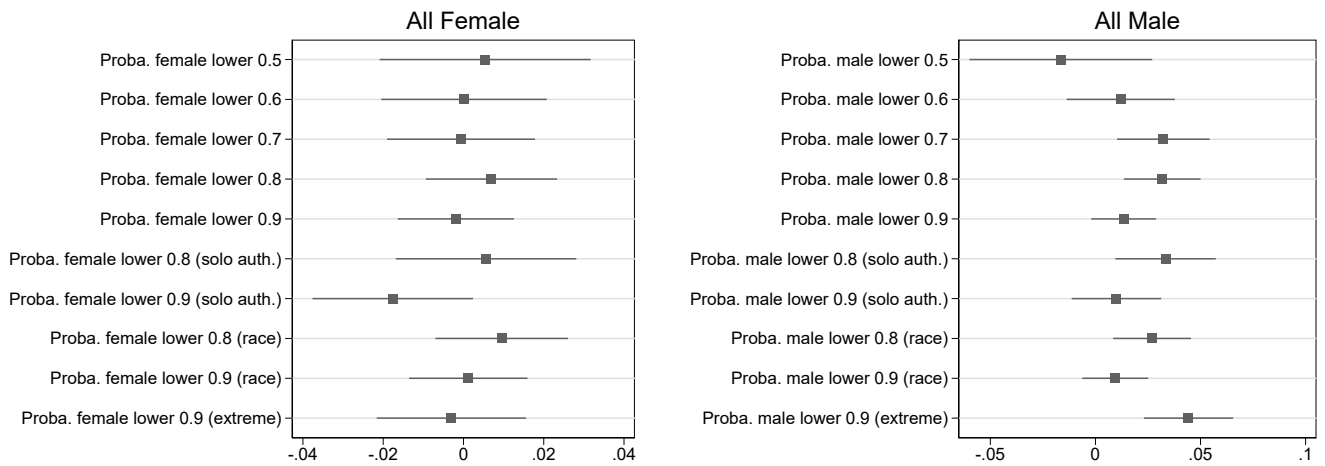
F.6 Gender perception analysis



a) Economics



b) Mathematics

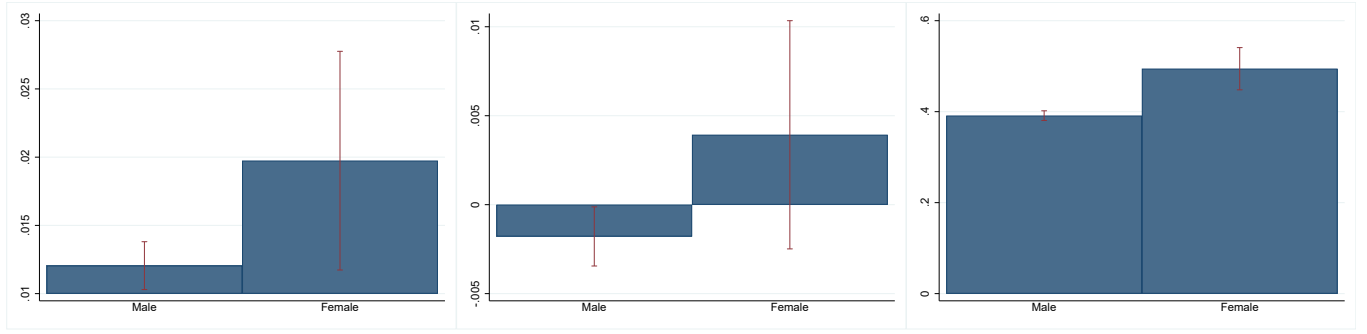


c) Sociology

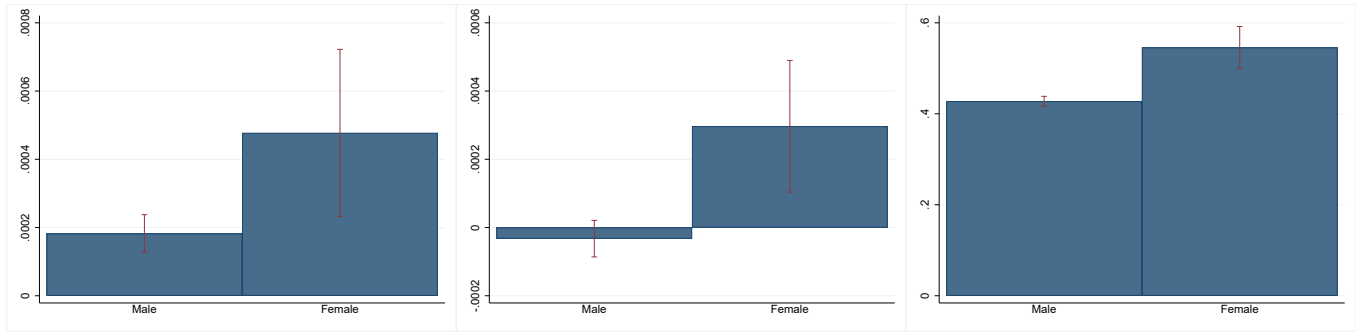
Figure:
Gender perception and omission

This figure explores the link between omission of references and perceived author gender in the suggested paper. The dependent variable, omission, is binary and indicates whether an article does not cite another article in the database given that the latter is in the relevant prior literature of the former (Section 3.3). Linear probability models are employed with the set of basic controls: methodological variables, journal type of the suggested paper, team size for both citing and suggested articles, publication year difference, main field alignment, bibliographic references count in the citing article, number of suggested articles, and fixed effects for institutions, journals, fields, and publication years of the citing paper and fixed effects for the institutions of suggested articles; and prior publications of suggested articles. Each plot focuses on a specific gender category and varies the degree of gender perception by: varying gender perception dummy variable cutoffs, focusing on solo-authors, including race controls, and analyzing extreme cases (probability above 0.9 vs. below 0.7). Self-citations are excluded to increase the chance that the hypothesis of unknown true gender is valid. The race control variable identifies articles with non-white names based on the “ethnicolr” race algorithm (<https://github.com/appeler/ethnicolr>).

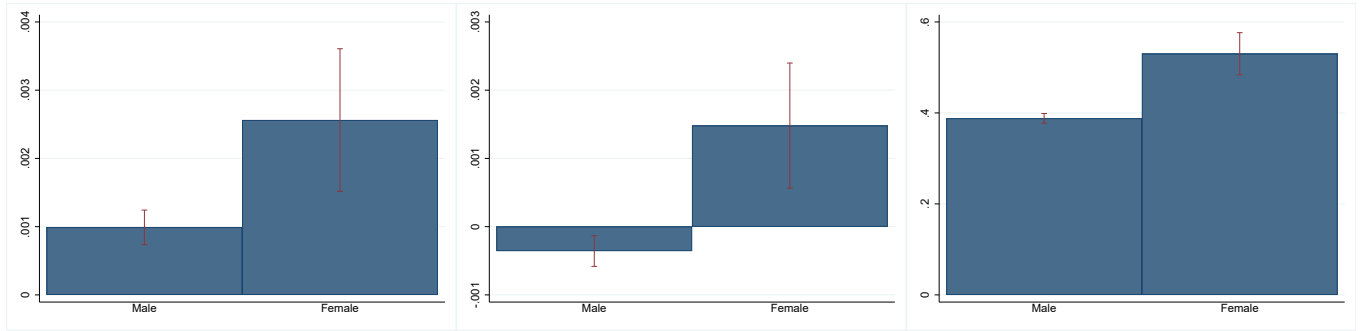
F.7 Innovativeness index and gender



Part A: TF-IDF no collocation



Part B: Word embeddings



Part C: BERT

Figure:
Innovativeness

The figures compare the innovativeness index by gender in economics, differentiating between cited articles and those recommended but uncited and for alternative time windows to construct the innovativeness index. After adjusting for variables like journal, field, institution, publication year, author count, and publication frequency, I present plots across various dimensions: Graph (a) shows average innovativeness scores for cited articles by gender; Graph (b) examines the innovativeness of suggested but uncited articles; and Graph (c) depicts the share of uncited articles with innovativeness scores matching or exceeding the median of those cited. The innovativeness is based on the 0 – 5-window.

F.8 Cosine on abstract, introduction and full papers

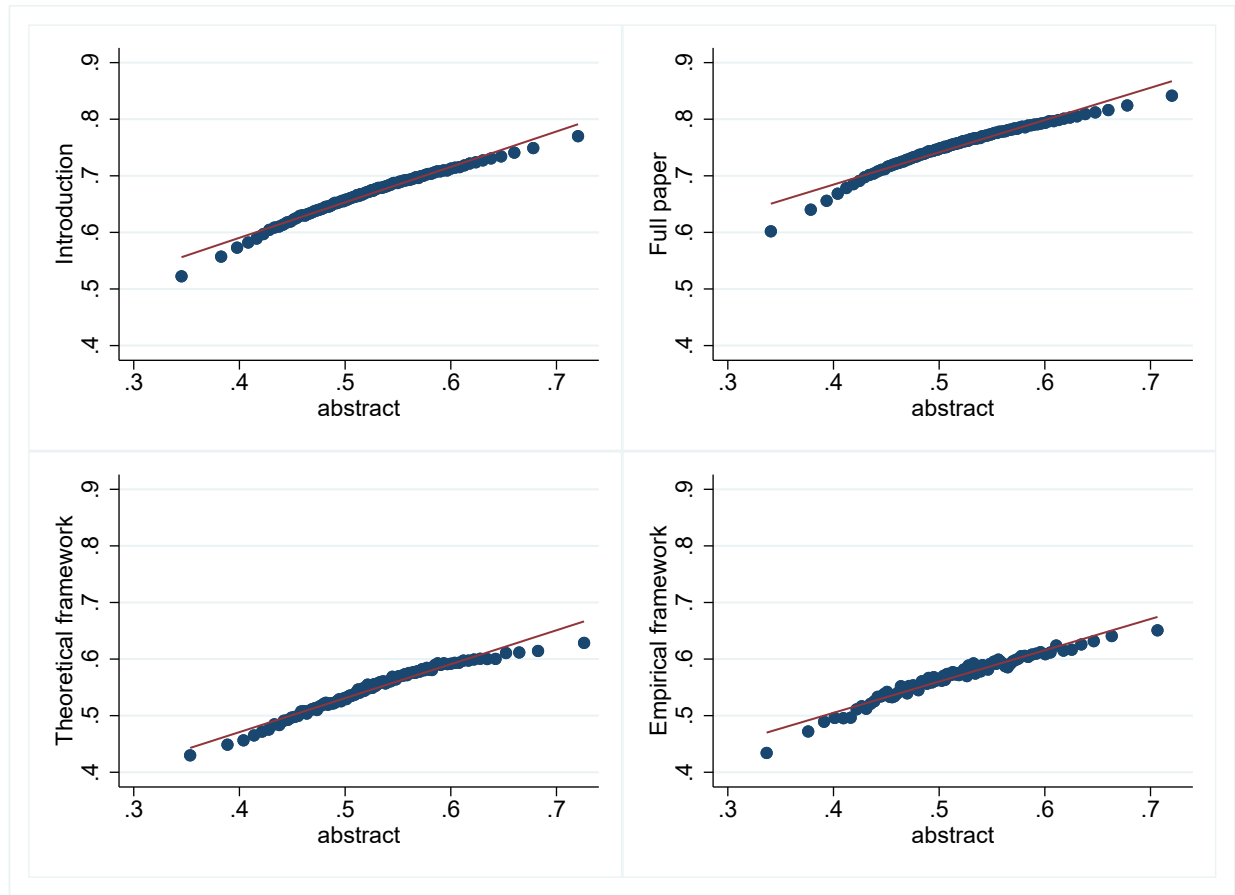


Figure:
Link cosine from abstract, introduction, and full paper

Sample includes around 10,000 downloaded papers from 1992-2019. The papers were downloaded via web crawling. Panel (a) plots the binscatter of the pairwise cosine based on papers' abstracts against the pairwise cosine based on the introduction papers up to 500 most similar papers. Panel (b) does the same for the abstract and the entire paper, panel (c) for the abstract and the theoretical framework and panel (d) for the abstract and the empirical framework. The coefficients have the following respective slopes: 0.63, 0.57, 0.60, and 0.55.

G Extended Sample

G.1 List of Journals in Economics

Journal	Total downloaded	With abstract and references	With Abstract and references/total
AMERICAN ECONOMIC JOURNAL-APPLIED ECONOMICS	424	422	1.0
AMERICAN ECONOMIC JOURNAL-ECONOMIC POLICY	412	408	1.0
AMERICAN ECONOMIC JOURNAL-MACROECONOMICS	339	336	1.0
AMERICAN ECONOMIC JOURNAL-MICROECONOMICS	398	395	1.0
AMERICAN ECONOMIC REVIEW	5379	3000	0.6
AMERICAN JOURNAL OF AGRICULTURAL ECONOMICS	3338	2274	0.7
APPLIED ECONOMICS	7833	7476	1.0
APPLIED ECONOMICS LETTERS	6148	6118	1.0
BROOKINGS PAPERS ON ECONOMIC ACTIVITY	459	251	0.5
CANADIAN JOURNAL OF ECONOMICS	1471	1399	1.0
ECONOMETRIC THEORY	1401	1203	0.9
ECONOMETRICA	1677	1533	0.9
ECONOMIC DEVELOPMENT AND CULTURAL CHANGE	922	383	0.4
ECONOMIC INQUIRY	1891	1866	1.0
ECONOMIC JOURNAL	2461	2143	0.9
ECONOMIC POLICY	332	325	1.0
ECONOMIC RECORD	1021	985	1.0
ECONOMIC THEORY	1855	1854	1.0
ECONOMICA	1012	987	1.0
ECONOMICS LETTERS	9072	8994	1.0
ECONOMIST	236	171	0.7
ENERGY JOURNAL	1090	1081	1.0
ENERGY POLICY	10664	10492	1.0
EUROPEAN ECONOMIC REVIEW	2933	2662	0.9
EXPLORATIONS IN ECONOMIC HISTORY	712	686	1.0
FEDERAL RESERVE BANK OF ST LOUIS REVIEW	457	365	0.8
GAMES AND ECONOMIC BEHAVIOR	2731	2568	0.9
IMF STAFF PAPERS	320	296	0.9
INDUSTRIAL & LABOR RELATIONS REVIEW	758	697	0.9
INTERNATIONAL ECONOMIC REVIEW	1572	1520	1.0
INTERNATIONAL JOURNAL OF GAME THEORY	1155	1119	1.0
INTERNATIONAL JOURNAL OF INDUSTRIAL ORGANIZATION	1787	1678	0.9
INTERNATIONAL LABOUR REVIEW	719	630	0.9
JOURNAL OF APPLIED ECONOMETRICS	1337	1316	1.0
JOURNAL OF BANKING & FINANCE	4606	4580	1.0
JOURNAL OF BUSINESS	548	533	1.0

JOURNAL OF BUSINESS & ECONOMIC STATISTICS	1261	1228	1.0
JOURNAL OF DEVELOPMENT ECONOMICS	2135	2074	1.0
JOURNAL OF ECONOMETRICS	3193	3160	1.0
JOURNAL OF ECONOMIC BEHAVIOR & ORGANIZATION	3552	3525	1.0
JOURNAL OF ECONOMIC DYNAMICS & CONTROL	2960	2891	1.0
JOURNAL OF ECONOMIC GROWTH	260	260	1.0
JOURNAL OF ECONOMIC HISTORY	831	716	0.9
JOURNAL OF ECONOMIC ISSUES	1939	1238	0.6
JOURNAL OF ECONOMIC LITERATURE	3987	518	0.1
JOURNAL OF ECONOMIC PERSPECTIVES	1642	1200	0.7
JOURNAL OF ECONOMIC THEORY	2868	2847	1.0
JOURNAL OF ECONOMICS & MANAGEMENT STRATEGY	883	868	1.0
JOURNAL OF ENVIRONMENTAL ECONOMICS AND MANAGEMENT	1573	1362	0.9
JOURNAL OF FINANCE	2144	2104	1.0
JOURNAL OF FINANCIAL ECONOMETRICS	304	270	0.9
JOURNAL OF FINANCIAL ECONOMICS	2306	2304	1.0
JOURNAL OF HEALTH ECONOMICS	1893	1740	0.9
JOURNAL OF HUMAN RESOURCES	1009	990	1.0
JOURNAL OF INDUSTRIAL ECONOMICS	811	774	1.0
JOURNAL OF INTERNATIONAL ECONOMICS	1858	1848	1.0
JOURNAL OF INTERNATIONAL MONEY AND FINANCE	2193	2114	1.0
JOURNAL OF LABOR ECONOMICS	898	891	1.0
JOURNAL OF LAW & ECONOMICS	783	665	0.8
JOURNAL OF LAW ECONOMICS & ORGANIZATION	738	603	0.8
JOURNAL OF LEGAL STUDIES	662	547	0.8
JOURNAL OF MATHEMATICAL ECONOMICS	1623	1612	1.0
JOURNAL OF MONETARY ECONOMICS	1649	1622	1.0
JOURNAL OF MONEY CREDIT AND BANKING	2105	1731	0.8
JOURNAL OF POLITICAL ECONOMY	1283	1223	1.0
JOURNAL OF PUBLIC ECONOMICS	2816	2789	1.0
JOURNAL OF REGIONAL SCIENCE	979	965	1.0
JOURNAL OF TAXATION	2008	217	0.1
JOURNAL OF THE EUROPEAN ECONOMIC ASSOCIATION	908	885	1.0
JOURNAL OF URBAN ECONOMICS	1472	1350	0.9
KYKLOS	780	704	0.9
LABOUR ECONOMICS	1243	1242	1.0
MACROECONOMIC DYNAMICS	1198	1148	1.0
MANAGEMENT SCIENCE	4393	4309	1.0
NATIONAL TAX JOURNAL	1028	887	0.9
NBER MACROECONOMICS ANNUAL	118	30	0.3

OXFORD	1145	1056	0.9
BULLETIN OF ECONOMICS AND STATISTICS			
OXFORD	1233	1095	0.9
ECONOMIC PAPERS-NEW SERIES			
QUARTERLY	1206	1177	1.0
JOURNAL OF ECONOMICS			
RAND	1182	1159	1.0
JOURNAL OF ECONOMICS			
REVIEW	903	882	1.0
OF ECONOMIC DYNAMICS			
REVIEW	1307	1284	1.0
OF ECONOMIC STUDIES			
REVIEW	2250	2188	1.0
OF ECONOMICS AND STATISTICS			
REVIEW	1783	1780	1.0
OF FINANCIAL STUDIES			
REVIEW	1011	950	0.9
OF INCOME AND WEALTH			
SCANDINAVIAN	1013	950	0.9
JOURNAL OF ECONOMICS			
SOCIAL	1671	1639	1.0
CHOICE AND WELFARE			
SOUTHERN	1835	1725	0.9
ECONOMIC JOURNAL			
THEORY	1218	1162	1.0
AND DECISION			
WORLD	868	720	0.8
BANK ECONOMIC REVIEW			
WORLD	4826	4734	1.0
DEVELOPMENT			
Total	169207	152698	0.9

G.2 List of journals in Mathematics

Journals	Number of papers	Journals	Number of papers
ABSTR APPL ANAL	4,379	J NONLINEAR FUNCT AN	221
ACTA MATH-DJURSHOLM	81	J NONLINEAR MATH PHY	1,199
ANN APPL STAT	1,092	J NONLINEAR SCI	848
ANN MATH	1,033	J NONLINEAR SCI APP	50
ANN PROBAB	2,419	J NONLINEAR SCI APPL	1,115
ANN STAT	3,019	J NONLINEAR VAR ANAL	81
APPL COMPUT HARMON A	1,088	J R STAT SOC A STAT	664
B AM MATH SOC	417	J R STAT SOC B	593
B MATH BIOL	2,308	J R STAT SOC C-APPL	805
B MATH SCI	125	J ROY STAT SOC A STA	364
BIOMETRICS	3,794	J ROY STAT SOC B	407
BIOMETRIKA	2,277	J ROY STAT SOC B MET	319
BIOSTATISTICS	886	J ROY STAT SOC C-APP	294
BULL MATH ANAL APPL	105	J ROY STAT SOC D-STA	369
COMM APP MATH COM SC	84	LECT NOTES MATH	7
COMMUN NONLINEAR SCI	4,710	LOND MATH S	9
COMMUN PUR APPL MATH	1,192	MATH BIOSCI	2,617
CONSTR APPROX	960	MATH BIOSCI ENG	1,350
FIXED POINT THEOR-RO	523	MATH MED BIOL	323
FIXED POINT THEORY	50	MATH MOD METH APPL S	1,934
FIXED POINT THEORY A	1,614	MEM AM MATH SOC	734
FOUND COMPUT MATH	510	MULTISCALE MODEL SIM	945
GEOM TOPOL	1,097	NONLINEAR ANAL-HYBRI	791
IMA J APPL MATH	1,268	NONLINEAR ANAL-MODEL	519
IMA J MANAG MATH	289	NONLINEAR ANAL-REAL	2,859
IMA J MATH CONTROL I	762	NONLINEAR ANAL-THEOR	8,359
IMA J NUMER ANAL	1,329	NUMER MATH	2,362
INFORM FUSION	821	PROBAB THEORY REL	2,018
INFORMATICA-LITHUAN	667	PUBL MATH	23
INT J BIOMATH	840	PUBL MATH-PARIS	114
INVENT MATH	1,401	SIAM J APPL MATH	2,824
INVERSE PROBL IMAG	633	SIAM J CONTROL OPTIM	3,347
J ALGEBR COMB	1,272	SIAM J MATH ANAL	3,149
J ALGEBR STUD	30	SIAM J NUMER ANAL	3,541
J ALGEBRAIC GEOM	478	SIAM J SCI COMPUT	4,138
J AM MATH SOC	34	SIAM REV	743
J AM STAT ASSOC	3,690	STAT SCI	776
J DIFFER EQUATIONS	6,766	STAT SCI INT RES	21
J DIFFER GEOM	1,047	STAT SCI INTERDISC R	72
J EUR MATH SOC	916	STATA J	700
J FOURIER ANAL APPL	1,023	STRUCT EQU MODELING	923
J MATH BIOL	2,138	STUD APPL MATH	972
J MATH PURE APPL	1,422	TECHNOMETRICS	1,072
J NONLINEAR CONVEX A	1,303		

G.3 List of journals in Sociology

Journal	Number of papers	Journal	Number of papers	Journal	Number of papers
ACTA SOCIO	528	CURR SOCIO	632	J HEALTH SOC BEHAV	793
ADV CHILD DEV BEHAV	9	DEBATES SOCIO	46	J HIST SOCIO	467
AFR SOCIO REV	53	DEVIANC SOC	299	J IMMIGR REFUG STUD	107
AGER	62	DEVIANT BEHAV	1,183	J INTERCULT STUD	220
AGR HUM VALUES	771	DISCOURSE SOC	722	J LAW SOC	531
AGRIC HUMAN VALUES	7	DISTINKTION	43	J LEISURE RES	662
AM J CULT SOCIO	80	DRUS ISTRAZ	1,106	J MARRIAGE FAM	2,226
AM J ECON SOCIO	959	DU BOIS REV	185	J MATH SOCIO	359
AM J SOCIO	858	EAST EUR COUNTRYS	7	J POLIT MIL SOCIO	225
AM SOCIO REV	1,150	EAST EUR COUNTRYSIDE	95	J PROF ORGAN	57
ANDULI	52	ECON SOC	614	J SCI STUD RELIG	1,054
ANN TOURISM RES	1,520	ECON SOC REV	423	J SOCIO	647
ANNU REV LAW SOC SCI	278	ETHNIC RACIAL STUD	2,053	J SOCIO SOC WELFARE	210
ANNU REV SOCIO	373	ETHNOGRAPHY	307	J SPORT SOC ISSUES	385
ANTHROZOOS	700	ETHOL SOCIOBIOL	143	J WORLD SYST RES	86
ARCH EUR SOCIO	266	EUR J SOC THEORY	327	KOLNER Z SOZIO	891
ARMED FORCES SOC	869	EUR SOC	485	LANG SOC	534
AUST NZ J SOCIO	85	EUR SOCIO REV	1,004	LAT STUD	107
BERL J SOZIO	574	FILOS-SOCIO	434	LAW SOC REV	642
BIODEMOGR SOC BIOL	149	FOOD CULT SOC	315	LECT NOTES ECON MATH	15
BMS-B SOCIO METHOD	4	GENDER SOC	926	LEISURE SCI	724
BODY SOC	283	GLOBAL NETW	450	MCS-MASC SOC CHANG	61
BRIT J SOCIO	908	HEALTH SOCIO REV	353	MEDIA CULT SOC	1,091
BRIT J SOCIO EDUC	1,099	HIST SOCIO	69	MEN MASC	390
CAH INT SOCIO	76	HUM ECOL	1,065	Methaodos	11
CAMBIO	99	HUM ECOL REV	166	METHAODOS	120
CAN GRAD J SOCIO CR	9	HUM STUD	515	MOBILIZATION	304
CAN J SOCIO	413	IFIP TRANS A	8	MULTICULT PERSPECT	114
CAN REV SOC ANTHROP	304	INFORM COMMUN SOC	898	MUNDO AGRAR	106
CAN REV SOCIO	225	INNOVATION-ABINGDON	334	NATIONS NATL	452
CHINESE SOC ANTHROP	88	INT J COMP SOCIO	396	NETH J SOC SCI	80
CHINESE SOCIO REV	133	INT J INTERCULT REL	1,243	OSTERR Z SOZIO	179
CITY COMMUNITY	310	INT J JPN SOCIO	47	PAPERS	120
COMMUNITY WORK FAM	185	INT J SOCIO	90	PLAY CULTURE	29
COMP SOCIO	155	INT J SOCIO EDUC	66	POETICS	569
COMP STUD SOC HIST	263	INT J SOCIO LAW	79	POL SOCIO REV	408
CONTEMP-REV SOCIO U	92	INT J SOCIO SOC POL	272	POLIT SOC	368
CONTRIB INDIAN SOC	298	INT POLIT SOCIO	266	POPUL DEV REV	605
CONVERGENCIA	311	INT REV SOCIO	139	Prax Sociol	9
CORNELL HOSP Q	420	INT REV SOCIO SPORT	524	PRAX SOCIO	63
CORNELL HOTEL REST A	19	INT SOCIO	687	QUAL RES	465
CORVINUS J SOCIO PO	70	INT STUD SOCIO EDUC	94	QUAL SOCIO	269
COSMOP CIV SOC	67	ISTANB J SOCIO STUD	90	RACE CLASS	321
CRIME MEDIA CULT	215	ISTANB UNIV SOSYOLOJ	20	RACE SOC PROBL	134
CRIT SOCIO	499	J CLASS SOCIO	109	RATION SOC	456
CUEST SOCIO	79	J CONSUM CULT	300	RECH SOCIOGRAPHIQUES	17
CULT SOCIO	61	J CONTEMP ETHNOGR	629	RES SOC STRAT MOBIL	330
CULT SOCIO-LONDON	236	J CULT ECON-UK	204	REV COLOMB SOCIO	115
CULTURALES	73	J ECON CULT SOC	33	REV ESP INVESTIG SOC	344
CURR PERSPECT SOC TH	86	J ECON SOCIO	144		

Journal	Number of papers	Journal	Number of papers
REV ESP SOCIO	144	SOCIO RELIG	508
Rev Estud Cotid	11	SOCIO RES ONLINE	1,127
REV ESTUD COTID	17	SOCIO REV	1,108
REV FR SOCIO	607	SOCIO RURALIS	705
REV INT ORGAN-ESP	66	SOCIO SCI	168
REV INT ORGANIZ	3	SOCIO SOC RES	30
REV INT SOCIO	389	SOCIO SPECTRUM	680
REV RELIG RES	662	SOCIO SPORT J	571
RUDN J SOCIO	252	SOCIO TEC	53
RURAL SOC	73	SOCIO THEOR	387
RURAL SOCIO	716	SOCIO THEOR METHOD	256
S AFR REV SOCIO	94	SOCIO TRAV	617
SCAND J HOSP TOUR	347	SOCIOLOGIA	328
SECU NONRELIGION	36	SOCIOLOGIA-BRATISLAV	287
SEXUALITIES	553	SOCIOLOGICA	56
SOC ALTERN	133	SOCIOLOGIJA	148
SOC ANIM	481	SOCIOLOGUS	92
SOC BIOL	237	SOCIOLOGY	1,387
SOC COMPASS	777	SOTSIO ISSLED+	1,595
SOC FORCES	1,611	SOZ WELT	468
SOC INDIC RES	3,778	SPORT SOC	609
SOC JUSTICE RES	238	STUD SOCJOL	15
SOC MENT HEALTH	140	STUD SOCJOLOGICZNE	119
SOC MOVEMENT STUD	358	STUD SYM IN	23
SOC NATUR RESOUR	1,729	STUD SYMB INTERAC	2
SOC NETWORKS	902	STUD SYMB INTERACT	199
SOC PROBL	745	SYMB INTERACT	626
SOC SCI QUART	2,090	SZOCIO SZLE	15
SOC SCI RES	1,751	TEACH SOCIO	583
SOCIETAMUTAMENTOPOLI	113	TEMPO SOC	328
SOCIETIES	276	THEOR SOC	346
SOCIETY	396	THEORIA	9
SOCIO-ECON REV	268	THEORIA-NY	94
SOCIO ANAL	31	THESIS ELEV	216
SOCIO CAS	827	TURK J SOCIO	71
SOCIO COMPASS	413	URBAN LEAGUE REV	36
SOCIO EDUC	472	VIRAJES	96
SOCIO FOCUS	19	WORK EMPLOY SOC	991
SOCIO FORSKNIN	359	WORK OCCUPATION	449
SOCIO FORUM	878	YOUNG	264
SOCIO HEALTH ILL	1,521	YOUTH SOC	766
SOCIO INQ	685	Z SOZIO	644
SOCIO METHOD RES	582	OTHERS	80
SOCIO METHODOL	243	SOCIO NAUK TEHNOL	96
SOCIO OBOZR	192	SOCIO PERSPECT	884
SOCIO PROST	78	SOCIO QUART	857
SOCIO RACE ETHNIC	46		

References

- Simran Arora, Avner May, Jian Zhang, and Christopher Ré. Contextual embeddings: When are they worth it? *arXiv preprint arXiv:2005.09117*, 2020.
- Gary S Becker. *The economics of discrimination*. University of Chicago press, 2010.
- J Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019.
- Lutz Bornmann, Alexander Butz, and Klaus Wohlrabe. What are the top five journals in economics? a new meta-ranking. *Applied Economics*, 50(6):659–675, 2018.

- David Card, Stefano DellaVigna, Patricia Funk, and Nagore Iriberry. Gender gaps at the academies. *Proceedings of the National Academy of Sciences*, 120(4):e2212421120, 2023.
- Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180*, 2020.
- Morris H DeGroot. *Optimal statistical decisions*. John Wiley & Sons, 2005.
- Kristie M Engemann, Howard J Wall, et al. A journal ranking for the ambitious economist. *Federal Reserve Bank of St. Louis Review*, 91(3):127–139, 2009.
- Nikhil Garg, Hannah Li, and Faidra Monachou. Dropping standardized testing for admissions: Differential variance and access. *arXiv preprint arXiv:2010.04396*, 2020.
- Erin Hengel. Publishing while female: Are women held to higher standards? evidence from peer review. *The Economic Journal*, 132(648):2951–2991, 2022.
- Pantelis Kalaitzidakis, Theofanis P Mamuneas, and Thanasis Stengos. Rankings of academic journals and institutions in economics. *Journal of the european economic association*, 1(6):1346–1366, 2003.
- Pantelis Kalaitzidakis, Theofanis P Mamuneas, and Thanasis Stengos. An updated ranking of academic journals in economics. *Canadian Journal of Economics/Revue canadienne d’économique*, 44(4):1525–1538, 2011.
- Bryan Kelly, Dimitris Papanikolaou, Amit Seru, and Matt Taddy. Measuring technological innovation over the long run. *American Economic Review: Insights*, 3(3):303–320, 2021.
- Yolanda K Kodrzycki and Pingkang Yu. New approaches to ranking economics journals. *The BE Journal of Economic Analysis & Policy*, 5(1), 2006.
- David N Laband and Michael J Piette. The relative impacts of economics journals: 1970-1990. *Journal of economic Literature*, 32(2):640–666, 1994.
- Matt Marx and Aaron Fuegi. Reliance on science: Worldwide front-page patent citations to scientific articles. *Strategic Management Journal*, 41(9):1572–1594, 2020.
- Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.
- Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.
- Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.
- Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. MpNet: Masked and permuted pre-training for language understanding. *Advances in Neural Information Processing Systems*, 33: 16857–16867, 2020.
- Jevin D West, Jennifer Jacquet, Molly M King, Shelley J Correll, and Carl T Bergstrom. The role of gender in scholarly authorship. *PloS one*, 8(7):e66212, 2013.
- zbMATH. zbmeth, mathematics. <https://zbmath.org/>, 1985-2019.
- Dan Zeltzer. Gender homophily in referral networks: Consequences for the medicare physician earnings gap. *American Economic Journal: Applied Economics*, 12(2):169–97, 2020.