The leaky pipeline problem, COVID-19 & big data: The impact of the pandemic on the gender gap in research production

Francesco Angeli*
María Agostina Zulli†
Alejandro Danon‡
Ana Carolina Ortega Masagué§

September 2021

Abstract

Gender gaps in research production are well documented in the literature across time and space. Female authors currently account for about 30% of total authors in journal publications, and this gap has been closing over time. The worldwide shock of the COVID-19 pandemic and the lockdown measures to contain it have paralyzed the global economy, and scientific production has not been aliened to this. Furthermore, recent studies claim that the pandemic has not affected all scientists equally due to major shifts in family schedules and routines, imposed by the restrictions. In this paper, we investigate the effect of the pandemic on the gender gap in academic production, as measured by journal publications. Using data from 8.34 million authors in 206 different countries who published articles in journals of several disciplines (STEM -Science, Technology, Engineering and Mathematics- and Economics) from the first quarter of 2015 to the second quarter of 2021, we calculate the percentage of female authors for each discipline, country, and period. Then, we estimate a fixed-effects model to identify the impact of country- and time-varying mobility restrictions on the gender gap in research production. Our results show a negative impact of the lockdown measures on the gender gap in academic production nine months after the mobility measures were implemented. This effect is larger and occurs earlier when considering publications with a single author. We also find heterogeneous results in the magnitude and timing of the effect of the restrictions across countries and disciplines.

Keywords: Gender gap, CODIV-19, Academic publications

JEL Codes: B54, C23, C63, J16, J24

*Faculty of Economics, National University of Tucumán, Argentina, franangeli13@gmail.com
†Faculty of Economics, National University of Tucumán, Argentina, agostinazulli@gmail.com
‡LAPDE-INVECO (Universidad Nacional de Tucumán) y BICE
§INVECO (Universidad Nacional de Tucumán)

1 Introduction

The emergence of COVID-19 has not only altered the health system but also economic, social, and political aspects throughout the world. Among the most common strategies to alleviate the spread of the virus, we find restrictions on circulation, the closure of educational, leisure and work establishments, and the interruption of activities considered non-essential.

The lockdown measures due to COVID-19 have imposed major shifts in family schedules and routines, which has probably reinforced traditional gender roles (Wenham et al. 2020). Scientists have not been excluded from this, and it is plausible that competing demands from home-schooling, family obligations and other caring duties during the pandemic affected the work productivity of women and men differently (Staniscuaski et al. 2020; Collins et al. 2020; Viglione 2020). For instance, a recent survey on 4,535 principal investigators in scientific projects in Europe and the U.S. indicated that academic women and, especially, scientists with young children, have experienced a more substantial decline in research time than academic men and scientists without children (Myers et al. 2020).

A vast literature collects the challenges faced by women in academic careers prior to the pandemic (See Cheng, 2020 Fox et. al, 2017). According to this literature, women published fewer articles, received fewer grants and citations, and were less likely to be granted tenure or promoted than men (Holman et al. 2018; Huang et al. 2020). Generally, these challenges are known as the leaky pipeline problem. This metaphor is used to describe how women become underrepresented minorities in academics, and particularly in STEM fields.

Although there is an ongoing debate about the causes, the main reasons for this leak have been attributed to lack of female mentors, family responsibilities due to feminization of housework and motherhood (Wolfinger et. al, 2017 Cheng, 2020), challenges in reaching hierarchical positions, implicit and subconscious biases (Charlesworth Banaji, 2019), the Matilda effect[1] (Knobloch-Westerwick et. al, 2013), among others.

Given that academic publications are the most common means of disseminating scientific knowledge and the main measure of the productivity of research work (Lameda et al. 2015) and considering they have a direct impact on professional prospects and in the visibility of women in science, the differential impact of COVID-19 on the working life of women scientists

[1]The Matilda effect is a bias against acknowledging the achievements of those women scientists whose work is attributed to their male colleagues. The term was coined in 1993 by science historian Margaret W. Rossiter.

may have important consequences, not only on their careers but also, for the generation of knowledge in general. The underrepresentation of women in science means wasting great talents and has consequences for how we think and solve scientific problems. If science is about building knowledge to understand how the world works, the ability to identify questions and answers from many perspectives will help to build more complete and robust scientific answers (Blickenstaff, 2005). In addition, there is evidence that gender diversity within the workforce is correlated with the creation of value and profits (McKinsey et. al, 2018; Zhang, 2020), and gender diversity in leadership positions increases productivity (Dezso Ross, 2011).

The Covid-19 pandemic has led to unprecedented closures of day-care centres, schools and workplaces. While women were already doing most of the world's unpaid care work, especially those with children[2] prior to the onset of the pandemic (Ceci et al. 2014; Cheng 2020), emerging research suggests that the lockdown response to the pandemic has resulted in a dramatic increase in this burden (Wenham et al. 2020; Alon et. al 2020 Myers et al. 2020). Although academics are used to working at a distance and with flexible schedules, it is plausible that during the pandemic competing demands from home-schooling, family obligations and other caring duties have affected the productivity of women and men differently (Staniscuaski et al. 2020; Collins et al. 2020; Viglione 2020).

Our results support our hypothesis of a decreased participation of women over men in academic papers. The restrictions on mobility have a three-quarters lagged effect on the gender gap in academic production, resulting in a setback of at least a year in terms of gender equality. In addition, this effect is larger and occurs earlier for publications with a single author. We conclude that, although some effects are starting to show after a year of the beginning of the pandemic and the lockdown measures, there might be further effects in the long run over this gap. We also conclude about the differential effects between a group of nineteen countries and all the STEM disciplines considered.

The rest of the work is organized as follows. The next section reviews the previous literature and the evidence found on gender gaps in scientific production. Sections 3 and 4 present the methodology for the construction of the database together with the strategy for identifying the phenomenon. Finally, sections 5 y 6 presents the results and final reflections.

[2] The literature calls it a "motherhood penalty".

2 Literature Review

Among the pioneers of bibliometrics[3] for gender analysis, we find Cole (1979) with a sample of 53 American men and women who received their PhD between 1911 and 1931. His first results showed that men were more likely to publish scientific research in the first five years of their careers, although this probability decreased, becoming more moderate as time passed. It is striking to the authors that this result held for all cohorts despite important changes such as the education of women in the United States. Continuing with this investigation, Cole and Zuckerman (1984) use data collected on 263 couples of men and women who received their PhD between 1960-1970, thus obtaining pairs of scientists with the same educational origins and with degrees obtained in the same year and the same field. Although the scope of their results is limited and is restricted to only 6 scientific fields, their conclusions reveal women productivity of less than 60%, in terms of the number of scientific publications, regarding their male colleagues. These differences appear to increase as the authors advance in their careers and exceed their tenure years.

On the other hand, recent literature is based on new computational methods that allow better data management and non-manual gender and nationality assignment techniques in bibliometric analysis. We can mention a series of publications that seek, through some computational method, to estimate the ratio between the gender of the authors, for many publications.

Concerning the analysis of authorship by gender, West et. al (2013) performs a detailed estimate of the composition by gender of the authors of scientific publications, paying particular attention to positions they occupy in collaborative research. Using the JSTOR repository to obtain academic publications by discipline from 1990 to 2001 and the records of the US Social Security to determine the gender of the authors according to their first name, their results shows that there are still gaps between male and female authors in most fields despite the increase in the percentage of female authors. Likewise, a more detailed analysis shows that new gender disparities appear in the composition of academic authorship where men predominate in the position of first and last author, indicating higher hierarchies, while women are underrepresented among individual authors. These same results are supported for a wider range of countries by Lariviere et al. (2013), their results point out that globally

[3]Bibliometric is a scientific field based on mathematical and statistical methods on all scientific literature and on the authors who produce it to analyse scientific activity. This research is part of a long tradition of studies aware of existing gender inequalities in scientific disciplines, with special attention to the production of academic publications.

women scientists represent less than 30% of authorships worldwide, being the countries with the least scientific production where there is a greater number of female authors.

As well as new disparities are appearing in authorship composition, indicating differences in hierarchy, Macaluso et. al (2016) from the collection of 85,000 articles published between 2008 and 2013 extracted from PLOS (Public Library of Science) found that gender is a significant variable when determining the probability of carrying out a certain task within the investigation. Even controlling for academic age (years since the first publication), women are more likely to be the ones who perform the experiments, "physical work", and men are more likely to be associated with "conceptual work"[4]. Likewise, their results indicate that the gender of the corresponding author[5] is related to the role played by other scientists within the team.

As we mention, the gender gap across disciplines has been shrinking thanks to the increase in women participation. An annual report from Elsevier (2020), comparing the ratio of female active authors between 1999 and 2003 with the ratio between 2014 and 2018 concludes that for all the countries considered, there is an increase in the number of female researchers compared to 10 years ago. However, Huang et. al (2020) from reconstructing the complete publication history of more than 1.5 million authors, belonging to 13 disciplines, warned that, paradoxically, the increase in the participation of women in science in the last 60 years (between 1955 and 2010) was accompanied by an increase in gender differences both in productivity (number of publications) and impact (total number of citations obtained by the publication within 10 years of publication). In line with these results, Holman et. al (2018), indicates that 75% of the disciplines considered have female participation, within publications, significantly lower than 45% and only 5 disciplines have participation above 55%. The heterogeneity between disciplines is significant, with disciplines related to care and social sciences being the disciplines with a higher proportion of women, evidence already provided by Lariviere et al. (2013). While all disciplines are moving toward parity, the authors predict that some of them will require decades and even centuries in the absence of specific policies.

These previous results support the perception that gender differences in academia are a

[4]Data analysis, writing of the work, contribution with analysis tools or materials and design of the experiment.

[5]This is the individual in charge of communicating with the journal during the submission of the manuscript, the peer review, and the publication process. Generally, they are senior researchers and group leaders.

universal phenomenon that persists in all disciplines, and particularly in STEM fields, across most geographic regions.

Covid-19 pandemic and the mobility restrictions imposed to reduce the spread of the virus has generated new research that proposes to analyse the effects of the pandemic in terms of gender inequalities. Given that most academics have been forced to work from home, the competing demands for home-schooling, child and other care duties might have penalised the scientific productivity of women.

Emerging research shows an overall decrease in the time dedicated to research although there are important differences between disciplines. Scientists working in fields that tend to rely on physical laboratories and time-sensitive experiments—bench sciences such as biochemistry, biological sciences, chemistry, and chemical engineering—reported the largest declines in research time (Myers et. al, 2020).

In terms of differences between genders, female scientists and scientists with young dependents reported that their ability to devote time to their research has been substantially affected, and these effects appear additive: the impact is most pronounced for female scientists with young dependents (Myers et. al, 2020 Deryugina et. al, 2021). The disproportionate productivity slowdown among female scholars since the onset of the pandemic may be explained by the substantial increases in childcare and housework burdens, experienced by both men and women, but women experiencing significantly larger increases than men did.

The disproportional reduction in women's labour productivity in certain disciplines has been studied by Amano-Patiño et al. (2020). The authors explore working papers publications in economic fields, where their analysis suggests that although the relative number of female authors in non-pandemic related research has remained stable concerning recent years (at around 20%), women constitute only 12% of the total number of authors working on COVID-19 research. Moreover, it is primarily senior economists who are contributing to this new area. Mid-career and junior economists record the biggest gap between non-COVID and COVID research, and the gender differences are particularly stark at the mid-career level. This same phenomenon is reported in health science where Andersen et. al (2020) estimated that the proportion of COVID-19 papers with a woman first author was 19% lower than that for papers published in the same journals in 2019. Their findings are consistent with the idea that the research productivity of women, especially early-career women, has been affected more than the research productivity of men.

Faced with this tradition of studies, the impact of the mobility restrictions imposed by the pandemic updates the need for research that combines the management of metadata (big data algorithms) that allow the automatic collection and transformation of a large volume of academic publications, and at the same time, allows to analyse the impact on the production of scientific publications for a large number of disciplines and a large number of countries.

3 Empirical Approach

Given the state of knowledge, in order to perform a multidisciplinary and cross-country analysis, our main contribution is the generation of a unique dataset containing the following variables about authors of academic publications in journals: *Title of the paper, discipline, journal, author´s names, author´s gender, country of residency and position.* Furthermore, our dataset is completely open-source work because the tools we use to have this attribute, and on the other hand the repositories consulted are also open source.

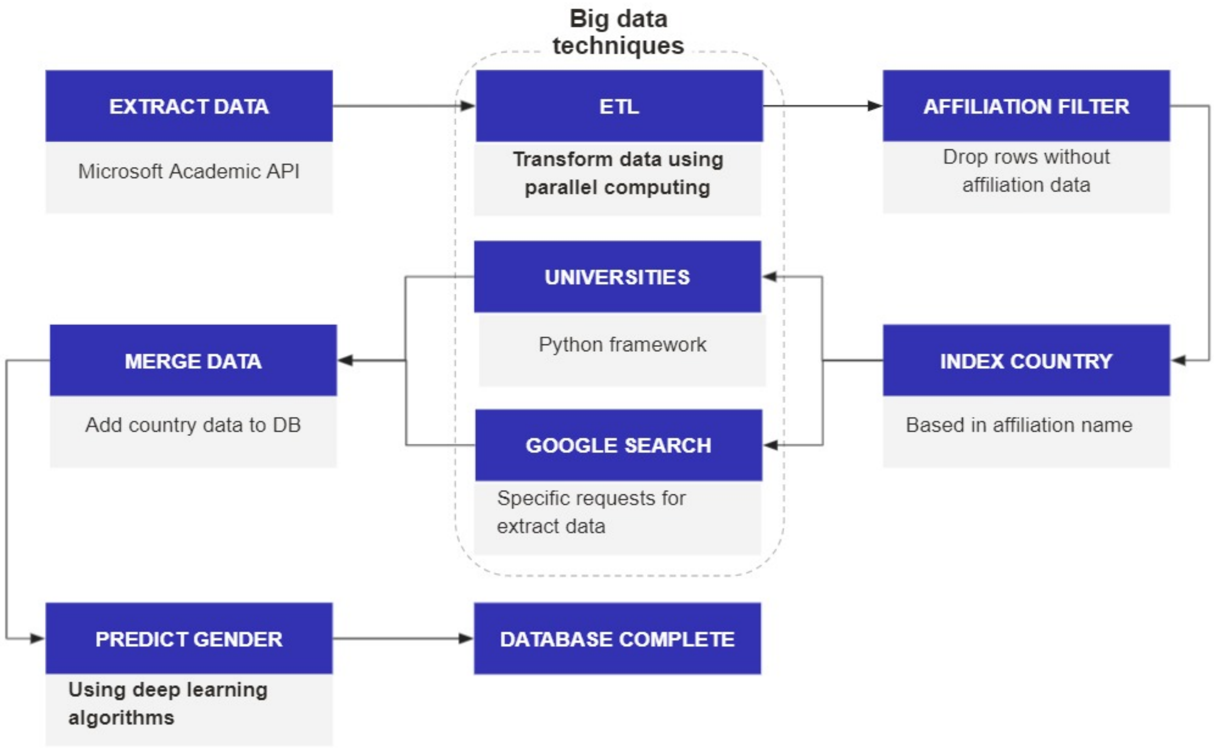
Once we built our dataset, we could set up panel data of the gender gap in authorship between countries and disciplines to estimate the impact of the lockdown measures on the authorship gap between female and male scientists.

3.1 Dataset

To build a model that captures the effects of the lockdown on the author gap, firstly, we had to build a database with certain features of authors publishing research papers in academic journals such as country of residency, discipline, journal, the title of the paper, names of the authors, gender of the authors, author position, etc.

In order to build this database, we had to go through several instances, detailed by the following workflow.

Figure 1: Workflow of the construction of the database



Note: (**API**) Application Programming Interface is a connection between computers or between computer programs.

Before starting to explain the stages of this workflow, it should be noticed this part of the work is closely related to branches of computer science such as algorithm optimization, machine learning techniques and big data[6] solutions.

For this first stage, it is important to extract as much information as possible from papers that meet the main conditions of having been published in the period 2010 to 2021 and that they belong to STEM and economic fields.

With this motivation, and with the current state of knowledge, different recognized data sources such as Jstor, PubMed, PLOS, Scopus (Elsevier), Google Scholar, Sci-Hub and Microsoft Academic, among others, were considered. Microsoft Academic is used due to the fact it contains the authors' affiliation data (since this will index the author´s residence), something that was only found in a low proportion in the others (for instance, in JSTOR, approximately 3% of all papers contained this information). Another no less important

[6]Placed in a Big Data context, metadata provides us with information about each dataset based on a database schema. The use of these allows us to manage an architecture in a scalable way. On the other hand, microdata marks specific information indicating something about a specific element.

advantage is that the database provides a free connection via API (this feature is not available in the others).

3.1.1 Extract data

We created an account on the Microsoft[7] site to generate a "subscription-key", then we understood which were the available parameters and filtered those that needed to be extracted within each query. The attributes chosen were: "Publication Type", "Publication type name", "Title", "Year published", "Date Published", "Citation count", "Publisher", "Numeric position in author list", "Author id", "Author name", "Author affiliation name", "Field Name", "Journal name". On the other hand, data was extracted only from "Papers", and the subjects chosen were "economy", "internal medicine"[8], "biochemistry", "computer science", "engineering", "environmental science", "geography", "geology", "mathematics", "molecular biology", "physics", "materials science", "statistics", "data science", and "organic chemistry".

On the other hand, it was considered that the API has a request limit of fifty thousand papers. After this, automation was generated in the requests where each one was set to return fractions that did not exceed this limit (a total of 302 requests were made).

3.1.2 ETL

The data returned by each request has a .Json format, inside it there is metadata containing information about each publication. Approximately the weight of the set of .Json files is 20 Gigabytes.

This data must be unstructured at the microdata level in order to generate a useful dataset. For this purpose, an ETL (Extract Transform Load) was built which allowed us to obtain this resulting dataset. Since the volume of data was so large, hardware was required to be able to execute them, optimizing to the maximum the ETL Transform process and reducing the time complexity[9].

[7] See the documentation on https://docs.microsoft.com/en-us/academic-services/project-academic-knowledge/reference-interpret-method

[8] Internal medicine and material science were finally removed because they presented large imbalances over time.

[9] In computer science, the time complexity is the computational complexity that describes the amount of computer time it takes to run an algorithm. Time complexity is commonly estimated by counting the number of elementary operations performed by the algorithm, supposing that each elementary operation takes a fixed amount of time to perform. Thus, the amount of time taken and the number of elementary operations performed by the algorithm are taken to differ by at most a constant factor.

Something necessary when handling this volume of data is the use of parallel computing. This is a form of computation in which many instructions are executed simultaneously. It operates on the principle that large problems can be broken down into smaller ones. For this, different frameworks can be chosen, such as PySpark, Koalas, Dask, etc. In the case of this project, we used Dask since we already knew about this library.

Once this framework was implemented in the ETL, the result was a dataset of 26 million authors, where we found as main variables: "paper title", "author's name and surname", "affiliation", "field of study", "number of citations", "journal name", "publication dates", among others.

3.1.3 Affiliation Filter

This is an intermediate but necessary step. Since there are many papers in which the authors did not have any type of affiliation (indicated in the article), we drop these observations since this is a necessary condition for the analysis to be cross-country.

3.1.4 Index Country

3.1.4.1 Universities Method: We used a framework called Universities which facilitates a connection via API to a database, which contains information from many universities around the world. Among the attributes that each university contains in this database is the country to which it belongs. Then, a list was generated which contains the name of the affiliations of the whole dataset without repetitions. Having thus been reduced to a list of 17,000 universities worldwide. Iterating over the list of universities mentioned above, the Universities method is applied to try to find a match and then extract the "Country" field. Approximately 50% of the universities in the list were matched to a country.

3.1.4.2 Google Search Method: A slightly more rudimentary method was used for the remaining 50%. We simulated a request to the Google search engine (Google Search), with the name of each university adding the word address as a search attribute. This request returns many parameters but among them, we find a long address containing the name of the country to which it belongs. One of the stoppers is that google bans the IP from where these queries are made every 1000 requests. But to fix this we used an iterator combined with a proxy alternator, where after 999 queries were made, it alternated to another proxy IP.

As a result, we obtained a list of dictionaries for example: ["University of Cambridge":

"The Old Schools, Trinity Ln, Cambridge CB2 1TN, UK"].

Taking advantage of PyCountry framework, calls were made to the framework for each element of each list of addresses in order to generate the corresponding match and thus obtain the value of the country as a return. Employing this methodology, the missing data could be collected.

After this, the two lists of countries found by both methods were unified and included as the "Country" column in the dataset, merging them to the column of affiliations of each author.

3.1.5 Merge data

In this instance, we had to generate a dictionary with the names of the universities as **"Key"** and the name of the countries as **"Values"** of the dictionary. With this we were able to create a ***Country*** column in the dataset, where the Affiliation column is compared against the dictionary keys, and where there is a match, the value is placed in the Country value.

3.1.6 Predict Gender

Different algorithms were selected to be used to make the predictions. After comparing several options, two were chosen as the most outstanding in terms of performance: Sexmachine (Gender - Guesser) and Genderize.

The first one is trained with 40 thousand names, then based on this a query can be made and the algorithm returns results as male, female, mostly male, mostly female, or ambiguous. On the other hand, it offers an argument where you can tell which country the name belongs to improve the probability of the prediction. This algorithm contains information from several countries, including countries such as China and India.

The second incorporates large databases where it also crosses information from users with their profiles in social networks. Genderize also has a neural network that is training itself every second with new datasets to improve its predictions. Also, after performing queries, it returns a parameter of what was the probability with which the prediction was made.

We go through the whole list of unique names (230 thousand names approx.) using the first algorithm, and then the names that gave an ambiguous result (usually Asian names), were passed through Genderize. It was decided to do this in this sequence since the use of Genderize involves a monetary cost.

It can be said that the list was fully predicted with the use of both algorithms. Then in the exploratory analysis, the desired probability cohort for dropping observations will be established. In this case, the cohort selected was based on an accuracy of at least 90%.

3.1.7 Database Complete

Our final data set is formed with 2.16 million papers from 28.900 different journals belonging to twelve disciplines within STEM and economics fields.

The period considered for our research covers from the first quarter of 2015 to the second quarter of 2021 included. The broad period of analysis will be useful to capture the pre-trends of the gender gap and learn about the particularities between countries and disciplines.

From the two million papers we obtained 8.34 million authors who were affiliated with 16.895 different institutions. After we estimated the residence country of the author from the affiliation information, we obtained authors residing in 206[10] countries. Taking into account the country we predicted the gender of each one from their first name. The distribution between gender is 2.31 million female scientists and 6.03 million male scientists.

We construct a panel data of the gender gap of authorship quarterly by country and discipline from the first trimester of 2015 to the second trimester of 2021 included. Since not all countries had enough authors to estimate accurately the female gap in authorships, to construct our panel data and make our estimations we kept all countries with a participation of at least 1% of our total dataset. This let us with just nineteen countries.

[10]Some countries are very small territories that are not recognized as countries for the UN.

3.2 Identification strategy

To estimate the impact of the lockdown measures on the gender gap in the authorship of academic publications, we perform to estimate the following model:

$$log(\frac{r_{ijt}}{1 - r_{ijt}}) = \alpha_{ij} + \theta_k S_{jt-k} + \delta_{ijt} + \varepsilon_{ijt} \tag{1}$$

Where r_{ijt}[11] refers to the gender gap in authorship define as the ratio between female authors and the sum of female and male authors, α_{ij} refers to the fixed effect for country and discipline, δ_{ijt} refers to a set of variables to control for time effects and finally, S_{jt-k} refers to the average stringency index in the country j, for lags k = 0,1,2,3,4.

In order to estimate the impact of the restrictions to mobility on the academic production of publications, we use the stringency index from the University of Oxford to acknowledge the different degrees of severity of the lockdown measures around the world. The index is a daily composite measure based on nine response indicators: school closures, workplace closures, and travel bans rescaled to a value from 0 to 100 (100 = strictest). Our hypothesis suggests that more severe lockdown restrictions will place a stronger burden on families and, inside the families, on women, on the house and care work, reducing their time to do research and publishing papers.

We consider four lags for the stringency index as we have warned that publishing academic research has a large process, we can divide at least into three parts: the production process of the investigation, most affected by the lockdown measure and the increase in housework, then we have the revision and approval process of the research and finally the publication stage when it is uploaded[12] in the journal. The length of this process also de-

[11]

$$r_{ijt} = \frac{female_{ijt}}{female_{ijt} + male_{ijt}} = \frac{f_{ijt}}{f_{ijt} + m_{ijt}} \tag{2}$$

$$\frac{r_{ijt}}{1 - r_{ijt}} = \frac{\frac{f_{ijt}}{f_{ijt}+m_{ijt}}}{1 - \frac{f_{ijt}}{f_{ijt}+m_{ijt}}} = \frac{f_{ijt}}{m_{ijt}} \tag{3}$$

$$log(\frac{r_{ijt}}{1 - r_{ijt}}) = log(\frac{female_{ijt}}{male_{ijt}}) \tag{4}$$

[12]Microsoft Academics tends to use the online date or the "first seen" date as publication date.

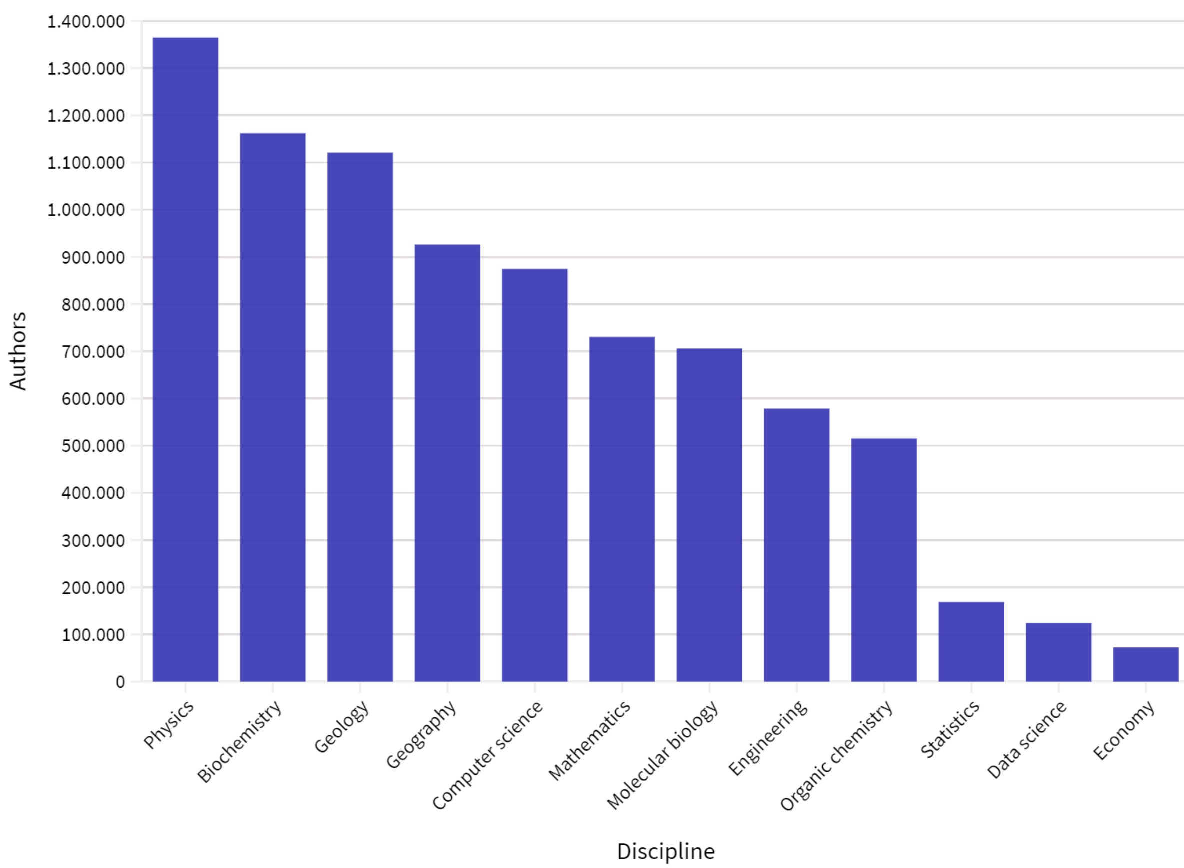
pends on the discipline and the theme of the investigation itself. Then, we also estimate a separate regression model for each discipline.

Considering that lockdown measures have been established in function of the epidemiological situation of the country, we test for the presence of correlation between the fixed effects and the variable S to decide whether OLS is enough to obtain consistent estimates or if we should obtain the within estimators.

4 Summary Statistics

Concerning the distribution of authors by disciplines, our dataset is made up of eleven disciplines within the STEM plus economics, where our five most present disciplines represent 65%[13] of our total authors: physics accounts for 16,4% of the total authors, biochemistry for 13,9%, geology for 13,4%, geography represents 11,1% of the total authors and computer science a 10,5%. The remaining 35% of authors belong to mathematics (8,8%), molecular biology (8,5%), engineering (6,9%), organic chemistry (6,2%), statistics (2,0%), data science (1,5%) and economics (0,9%).

Figure 2: Quantity of authors by discipline, from 2015 to 2021.

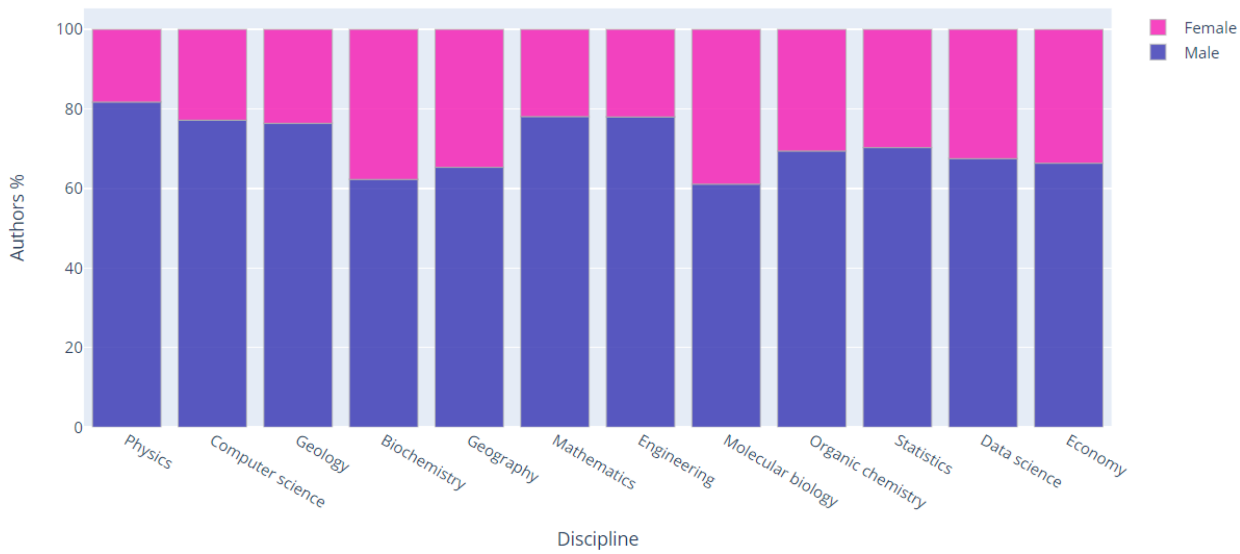


Note: Each author is considered more than once if it has more than one paper published. **Source:** Own elaboration from our dataset.

[13]See table 7 in the appendix.

In terms of gender, on average for our period of study, the disciplines with the lower percentage of female's participation in academic research we find physics with less than 20% of female authors, mathematics, and engineering with 22%, these results are consistent with the previous evidence[14]. As we can see from figure 2, with exception of molecular biology, biochemistry, and geography, where the participation of female scientists is between 35% and 38%, all the other disciplines are still far away from parity.

Figure 3: Average composition of authorships by gender and discipline, from 2015 to 2021.



Source: Own elaboration from our dataset.

Our data shows the smallest gender gaps in authorship in countries from South America, like Argentina and Uruguay, also in Asia like Indonesia and in some countries from Africa. Our data is consistent with the previous evidence[15] of a greater number of female authors in countries with less scientific production.

Female scientists account for less than 30%[16] all around the world within the STEM and economics fields. Prior to the pandemic, we can see a slight upward trend that even continues and seems to increase slightly during 2020. The interannual variation shows that during the first year of the pandemic, female participation increased in the four trimesters of 2020 compared to 2019. However, data shows an important decrease[17] during the first quarter of 2021 with a slight recovery in the second quarter of 2021. There is a 12% drop in female participation compared to

[14] See Elsevier annual report from 2017 to 2020.

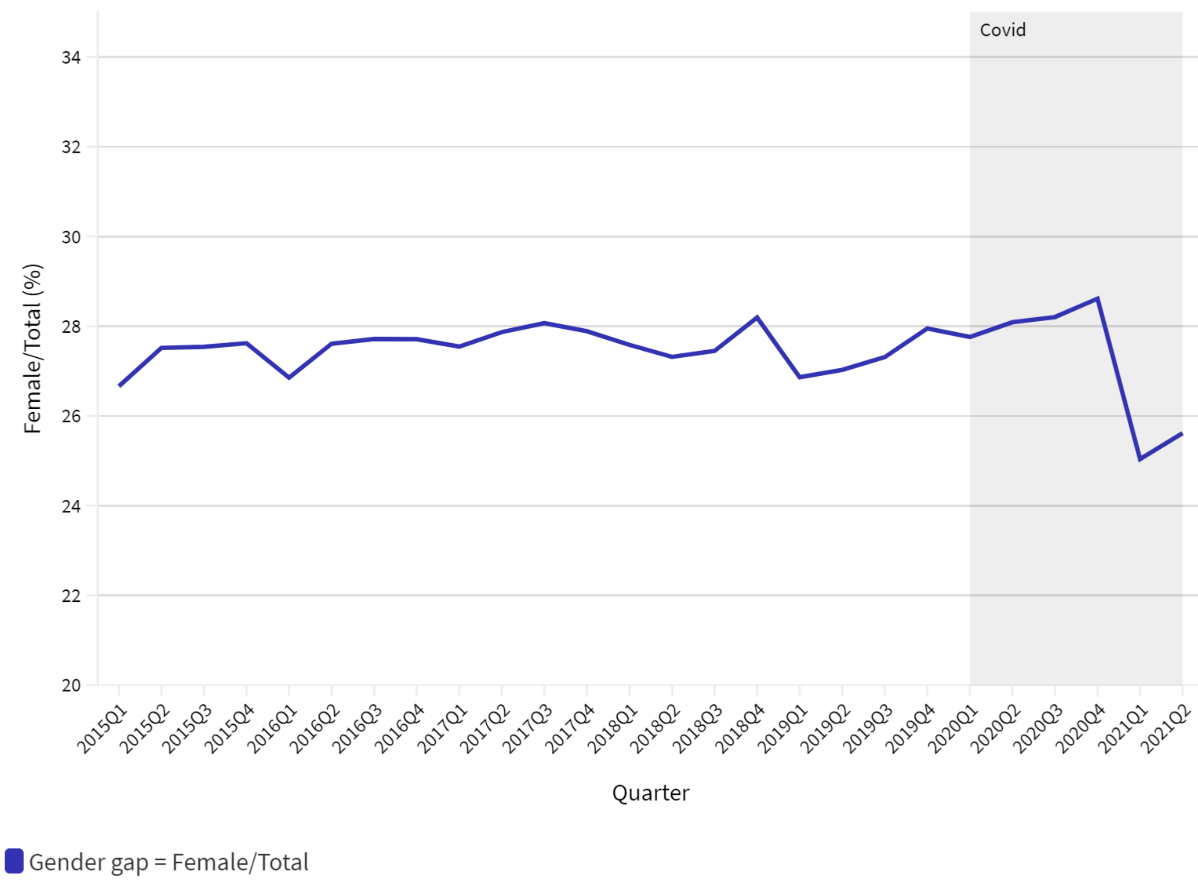
[15] See Elsevier (2020) Lariviere et al. (2013)

[16] This result is consistent with the participation of female scientists found by Lariviere et al. (2013).

[17] Also the chow test for breakpoints rejects the null hypothesis of no breaks at 2021Q1. See the test in the appendix.

the previous quarter, the fourth quarter of 2020, and a 10% decrease compared to the first quarter of 2020.

Figure 4: Evolution of the gender gap in authorship, 2015Q1- 2021Q2.

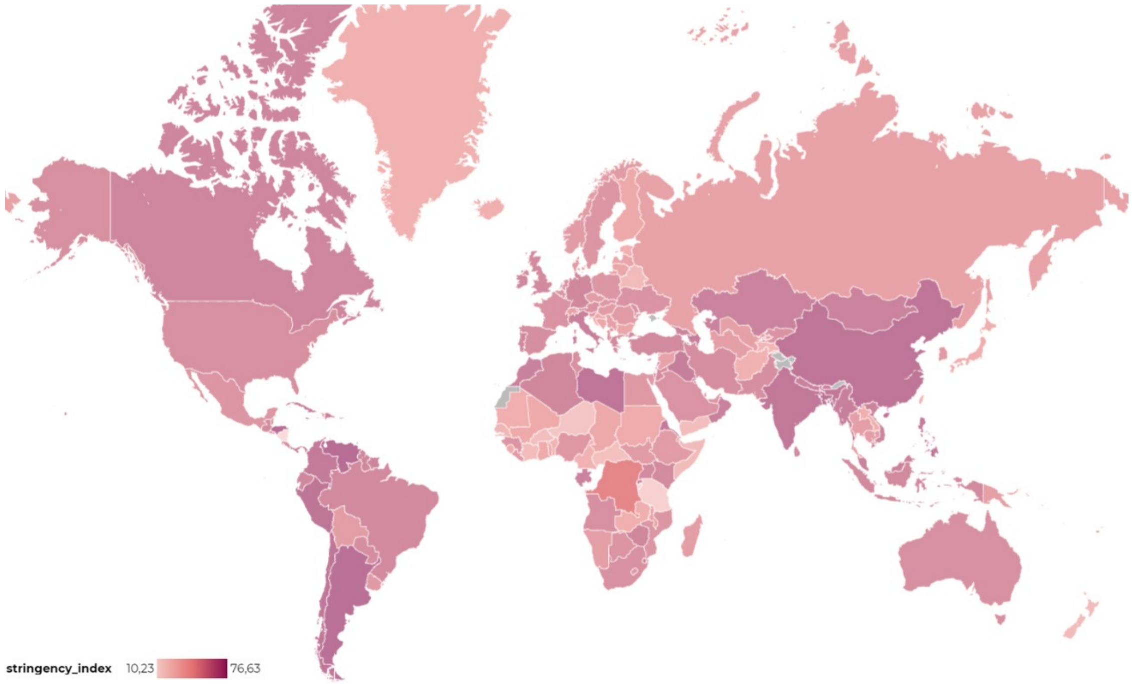


Source: Own elaboration from our dataset.

Although the impact of the pandemic seems to have a lagged effect over these disparities it is expected since, as mentioned earlier, the moment a paper is published in a journal is different from the period of research and production of the publication, the moment where lockdown measures have its worst impact.

The stringency of the lockdown measures is heterogeneous between time and countries. Countries from South America and some Asian countries show an average stringency index higher than the rest of the countries. On the other hand, countries with the lowest average strictness during the pandemic period are countries like Russia, the United States, and some countries from Europe.

Figure 5: Average stringency index by country, period 2020 Q1 to 2021 Q2.



Note: The index is a composite measure based on nine response indicators including school closures, workplace closures, and travel bans, rescaled to a value from 0 to 100 (100 = strictest). **Source:** own elaboration from Our World in Data.

5 Results

We first present the results from estimations relying on our panel data of the gender gap in authorship by country and disciplines, from the first quarter of 2015 to the second quarter of 2021. Our results from **table 1** show that, although all our stringency index variables show the expected negative effect on the gender-gap, the only effect statistically significant is the stringency index from three quarters lagged, in other words, the lockdown measures from nine months before the publishing had a negative effect on women's participation. In particular, the estimated coefficient for this lag is 0.0004, so that an increase of 10 units in the stringency index would decrease the participation of women by 0.4%. These results remain if we consider the overall period and time effects or if just consider the pandemic period 2020 and 2021.

Table 1: FE estimates on the gender-gap

	Dependent variable:				
	log(Female/Total)/(1-Female/total))				
	(1)	(2)	(3)	(4)	(5)
Stringency(t)	0.0001				
	(0.0002)				
Stringency(t-1)		-0.00005			
		(0.0001)			
Stringency(t-2)			-0.0001		
			(0.0001)		
Stringency(t-3)				-0.0004***	
				(0.0001)	
Stringency(t-4)					-0.0003
					(0.0002)
FE TIME	Yes	Yes	Yes	Yes	Yes
FE COUNTRY	Yes	Yes	Yes	Yes	Yes
FE DISCIPLINE	Yes	Yes	Yes	Yes	Yes
Observations	6,095	5,855	5,615	5,375	5,135
R^2	0.045	0.044	0.044	0.041	0.041
Adjusted R^2	0.005	0.002	-0.0004	-0.005	-0.008
F Statistic	27.777***	26.088***	24.696***	22.162***	23.208***
	(df = 10; 5845)	(df = 10; 5605)	(df = 10; 5365)	(df = 10; 5125)	(df = 9; 4886)
Note:	*p<0.1; **p<0.05; ***p<0.01				
	Robust standard errors clustered by group in parentheses				

5.1 Impact on the gender gap by discipline

Disciplines such as medicine, economics, biology, among others, have seen an important increase in the demand for knowledge creation in the urge to understand and act on the expansion of the virus. Furthermore, time in research and publication may vary between them. For these reasons, we enquire about the impact of the restrictions on each discipline as shown in **table 2**. As before, we report our results of the lagged restrictions only.

Table 2: FE estimates on the gender gap by discipline

Discipline	Stringency(t-1)	Stringency(t-2)	Stringency(t-3)	Stringency(t-4)	Observations
Statistics	0.0002	-0.00043	-0.0022**	-0.00261	516
Data Science	0.00018	-0.00031	-0.0018*	0.00191**	514
Computer Science	0.00016	-0.00037***	-0.0009***	nd	400
Organic Chemistry	0.00000	-0.00018	-0.0008*	0.00103**	520
Mathematics	0.00007	-0.00006	-0.0004**	-0.00028	520
Economy	-0.00096	-0.00011	-0.0004*	-0.00212**	505
Geology	-0.00009	0.00020	0.0001	-0.00006	520
Physics	-0.00076***	-0.00027	0.0001	-0.0006***	520
Biochemistry	-0.00011	-0.00002	0.0002	-0.0003***	520
Geography	0.00029	0.00035**	0.0002	0.00000	520
Molecular Biology	0.0002	-0.0002	0.0002	-0.00007	520
Engineering	0.00026	0.00005	0.0003	-0.00040	520

Note: Each line shows the estimated coefficient for the country on the lagged index. FE model ran with robust standard errors. *p<0.1; **p<0.05; ***p<0.01.

As shown in **table 2**, for most of the disciplines we find a negative statistically significant effect of the restrictions to mobility on the ratio, except for geography where the effect seems to be positive and statistically significant for lag two, and molecular biology, engineering, and geology where we do not find statistically significant effects.

Regarding the disciplines where we do find a negative statistically significant effect, in most of them, we can say that an increase in 10 units in the stringency index three quarters after the publication in the journal reduces the ratio of female to male by 0,4% for mathematics and economics, a 0,8% for organic chemistry, a 0,9% computer science, a 1,8% for data science and a 2,2% for statistics. Additionally, we find a negative statistically significant effect on physics on lag one and four; in economics, we also find a negative effect on the ratio for a year after the publication and in computer science on lag two also.

5.2 Impact on the gender gap in authorship by country

Since countries are very heterogeneous in terms of care systems and gender policies, we can expect different effects on the gender gap in authorships. In consequence, we estimate our model for each country. In **table 3** we report the results of the estimated coefficient for each country. Since data support a lagged effect of the pandemic on the gender gap, we report our results of the lagged restrictions only.

Table 3: FE estimates on the gender gap by country

Country	Stringency(t-1)	Stringency(t-2)	Stringency(t-3)	Stringency(t-4)	Observations
Russia	0.0001	0.0003	-0.0030	-0.0017	261
Iran	-0.0005	-0.0010	-0.0029**	0.0005	264
Switzerland	0.0013	-0.0007	-0.0028*	0.0013	268
Belgium	-0.0001	-0.0006	-0.0024***	0.0020***	270
Japan	0.0001	-0.0008	-0.0019***	-0.0023**	270
Poland	-0.0010	-0.0002	-0.0011*	0.0013*	269
Italy	-0.0003	-0.0006**	-0.0011***	0.0011*	270
United Kingdom	0.0000	0.0001	-0.0002	0.0006 ***	270
India	0.0003**	0.0002	-0.0001	-0.0001	270
France	0.0006*	0.0002	-0.0001	0.0001	270
China	-0.0002	-0.0002	0.0001	0.0041	270
Canada	0.0005*	0.0000	0.0002	-0.0013***	270
Germany	0.0005	0.0003	0.0003**	0.0001	270
United States	0.0001	0.0002**	0.0005**	0.0002**	270
South Korea	0.0000	0.0007	0.0010	-0.0018	268
Spain	-0.0011**	-0.0005	0.0010	-0.0017***	270
Taiwan	-0.0022	-0.0037***	0.0010	-0.0388	265
Brazil	0.0009**	0.0005	0.0012	0.0007*	270
Australia	0.0003	0.0005***	0.0014***	0.0004	270

Note: Each line shows the estimated coefficient for the country on the lagged index. FE model ran with robust standard errors. *p<0.1; **p<0.05; ***p<0.01.

The coefficient estimates of the parameter Ø, presented in **table 2**, are negative for most of the countries considered, with exceptions of countries like the United States, Australia and Germany. We found big heterogeneity between countries and between the time when restrictions impacted the gender gap. For instance, in half of the countries considered we do find statistically significant effects. This means that in those cases, we can reject the hypothesis that restrictions to mobility had no impact on the female scientist´s ratio. As shown in **table 2**, most of these significant effects are on our lagged index from three quarters before. Some exceptions are Spain with statistically significant effects of lag one and four and Taiwan on lag two.

On the other hand, in Russia, China and South Korea we do not find statistically significant effects of the lockdown measures on the gender gap in publications.

5.3 Impact on the gender gap on single authors

It is plausible that during isolation measures and increases in the burden of housework, single authors researchers had it more difficult to complete or start new investigations. Presuming that authors are substitute inputs in the production of papers, we expect a bigger negative effect in terms of gender among single authors.

We now turn to the results obtained from the same fully extended model with a new panel data constructed with only single authors publications for the nineteen countries considered, for the 2015-2021 period.

Table 4: FE estimates on the gender gap on single authors

	Dependent variable:				
	log(Female/Total)/(1-Female/total))				
	(1)	(2)	(3)	(4)	(5)
Stringency(t)	-0.001*				
	(0.0005)				
Stringency(t-1)		-0.001**			
		(0.0003)			
Stringency(t-2)			-0.001**		
			(0.0005)		
Stringency(t-3)				-0.001	
				(0.001)	
Stringency(t-4)					-0.001
					(0.001)
FE TIME	Yes	Yes	Yes	Yes	Yes
FE COUNTRY	Yes	Yes	Yes	Yes	Yes
FE DISCIPLINE	Yes	Yes	Yes	Yes	Yes
Observations	5,093	4,453	4,614	4,37	4,14
R^2	0.028	0.026	0.026	0.024	0.023
Adjusted R^2	-0.022	-0.026	-0.029	-0.034	-0.038
F Statistic	14.077***	12.349***	11.567***	10.056***	10.97***
	(df = 10; 4843)	(df = 10; 4604)	(df = 10; 4367)	(df = 10; 4131)	(df = 9; 3897)

Note: *p<0.1; **p<0.05; ***p<0.01
Robust standard errors clustered by group in parentheses

Results in **table 4** show that when considering the gender gap among single authors publications only, the estimated coefficients remain with a negative sign as expected but the effects seem to be larger than when considering all publications. Furthermore, the estimated coefficient is negative and statistically significant for period t, lag one and lag two. This means that the effects of the restrictions to mobility are shown sooner since the substitution between authors is impossible in the case of single authors investigations.

In this respect, the effects of the restrictions to mobility have a greater impact on female participation in publishing. Our estimations show that an increase in one standard deviation of the stringency index reduces the ratio of female to male single authors by 1% or by 0,03 percentage points.

5.4 Robustness

To test the robustness of our work and to corroborate whether our model is capturing the impact of the pandemic over the gender gap in publishing, we perform a fake test. The test consisted in performing the same model but faking the period when covid happened.

Table 5: FE estimates on the fake experiment

	Dependent variable:				
	log(Female/Total)/(1-Female/total))				
	(1)	(2)	(3)	(4)	(5)
Stringency(t)	0.00003				
	(0.0001)				
Stringency(t-1)		-0.00004			
		(0.0001)			
Stringency(t-2)			-0.0001		
			(0.0002)		
Stringency(t-3)				-0.0001	
				(0.0001)	
Stringency(t-4)					-0.0004
					(0.0001)
FE TIME	Yes	Yes	Yes	Yes	Yes
FE COUNTRY	Yes	Yes	Yes	Yes	Yes
FE DISCIPLINE	Yes	Yes	Yes	Yes	Yes
Observations	6,133	6,133	6,133	6,133	6,133
R^2	0.045	0.045	0.045	0.045	0.045
Adjusted R^2	0.005	0.005	0.005	0.005	0.005
F Statistic	27.677***	27.684***	27.709***	27.714***	27.685***
	(df = 10; 4843)	(df = 10; 4604)	(df = 10; 4367)	(df = 10; 4131)	(df = 9; 3897)
Note:	*p<0.1; **p<0.05; ***p<0.01				

As shown in Table 5, when simulating that covid appeared in the first quarter of 2017 until the second quarter of 2018, our results are not statistically significant for any of the lags considered. In this way, the evidence supports our previous results of a negative effect on the gender gap in authorship caused by the restrictions to mobility and possibly more accentuated gender roles.

6 Conclusions

The rapid spread of the COVID-19 pandemic and subsequent countermeasures, such as school closures, the shift to working from home, and social distancing have disrupted the family routines and distribution of time between women and men. Due to competing demands from home-schooling, family obligations and other caring duties, the productivity of women and men may have been affected differently.

Prior to the pandemic, the underrepresentation of women in academia was well-documented. After more than one year since the onset of the pandemic and with the third wave of infections arriving in most countries in Europe and North America, it is relevant to study the differential impact the responses to the pandemic have in terms of gender. As recent evidence shows, the lockdown measures have exacerbated the disparities between female scientists and male scientists thanks to the greater burden of house and care work that falls mainly on women, all over the world.

Our investigations aimed to estimate the impact of the restrictions to mobility on the female gap in academic research focusing on the academic publications in journals. Since they directly impact professional prospects and the visibility of women in science, fewer publications could have an important impact on professional prospects for female scientists and the creation of knowledge in general.

To accomplish our objective, we had to build our database from academic publications in journals using machine learning and big data solutions to extract information about scientists. After predicting the gender and the country of residency for more than eight million authors we could build a multidisciplinary and cross-country panel data that let us estimate with a fixed-effect model the impact of the lockdown measures on the gender gap in authorships. Our new dataset is a contribution with regard to recent previous studies that have worked with either specific disciplines or working papers for a much smaller number of publications.

Among our conclusions, we emphasize the importance of generating data from new algorithmic and statistical techniques that allow us to address a recent phenomenon from a non-traditional source of information.

Regarding the results, our data show an important decrease in female scientists among the authors publishing during the pandemic period, which translates into an increase in the gender gap in authorship. As expected, the impact of these lockdowns is starting to show during 2021. This is supported by our estimations where we find that the increase in the stringency index nine months after the publication, which accounts for the severity of the lockdown measures, in one sigma decreases the ratio of female to male by 0,02 percentage points by quarter, resulting in a setback of at least a year in terms of gender equality. Overall, these effects seem to be larger when considering just the gender gap in the publication of single authors, where the increase in one sigma reduces the ratio by 0,03 percentage points. In addition, since the substitution between authors is impossible in the case of single authors investigations the effects of the restrictions to mobility are shown sooner than when considering all publications.

However, when considering the estimations for each country and each discipline we do find heterogeneous results. For instance, although the estimated coefficients are negative for most of the countries considered, lockdown measures in countries like the United States, Australia, and Germany positively affect the gender gap. In addition, concerning the disciplines, in the majority of them, the impact is negative and the increase in the index reduces the ratio between 0,4% and 2,2%.

In conclusion, although it might be too soon to study and capture the impact of the responses to the pandemic in terms of gender within academic research output, our results support the hypothesis that productivity between female and male scientists has been affected differently, increasing the gender gap in academia. As expected, the effect on journal publications is starting to show in 2021 and for deeper analysis of professional implications on women and the creation of knowledge in general we need a longer period.

References

Alon, T., Doepke, M., Olmstead-Rumsey, J., Tertilt, M. (2020). The impact of COVID-19 on gender equality (No. w26947). National Bureau of economic research.

Amano-Patiño, N., Faraglia, E., Giannitsarou, C., Hasna, Z. (2020). The Unequal Effects of Covid-19 on Economists' Research Productivity.

Andersen, J. P., Nielsen, M. W., Simone, N. L., Lewiss, R. E., Jagsi, R. (2020). Meta-Research: COVID-19 medical papers have fewer women first authors than expected. Elife, 9, e58807.

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

Charlesworth, T. E., Banaji, M. R. (2019). Gender in science, technology, engineering, and mathematics: Issues, causes, solutions. Journal of Neuroscience, 39(37), 7228-7243.

Cheng, S. D. (2020). Careers Versus Children: How Childcare Affects the Academic Tenure-Track Gender Gap.

Cole, J. R. (1979). Fair science. Free Press.

Cole, J. R., Zuckerman, H. (1984). The productivity puzzle. Advances in Motivation and Achievement. Women in Science. JAI Press, Greenwich, CT.

Dezsö, C. L., Ross, D. G. (2012). Does female representation in top management improve firm performance? A panel data investigation. Strategic management journal, 33(9), 1072-1089.

Elsevier. (2017). Gender in the global research landscape. Elsevier Report. Available: https://www.elsevier. com/research-intelligence/resource-library/gender-report.

Elsevier. (2020). Gender report 2020: The Researcher Journey Through a Gender Lens. Elsevier Report. Available: https://www.elsevier.com/research-intelligence/resource-library/gender-report

Fox, M. F., Whittington, K., Linkova, M. (2017). Gender,(in) equity, and the scientific workforce. Handbook of science and technology studies, 701-731.

González Ramos, A. M., Navarrete Cortes, J., Cabrera Moreno, E. (2015). Dancers in the dark: Scientific careers according to a gender-blind model of promotion. Interdisciplinary Science Reviews, 40(2), 182-203.

Holman, L., Stuart-Fox, D., Hauser, C. E. (2018). The gender gap in science: How long until women are equally represented?. PLoS biology, 16(4), e2004956.

Howe-Walsh, L., Turnbull, S. (2016). Barriers to women leaders in academia: tales from science and technology. Studies in Higher Education, 41(3), 415-428.

Huang, J., Gates, A. J., Sinatra, R., Barabási, A. L. (2020). Historical comparison of gender inequality in scientific careers across countries and disciplines. Proceedings of the National Academy of Sciences, 117(9), 4609-4616.

Hunt, V., Prince, S., Dixon-Fyle, S., Yee, L. (2018). Delivering through diversity. McKinsey Company, 231.

Knobloch-Westerwick, S., Glynn, C. J., Huge, M. (2013). The Matilda effect in science communication: an experiment on gender bias in publication quality perceptions and collaboration interest. Science communication, 35(5), 603-625.

Lameda, C., Garcia, L. S., Vargas, R. U., Garcia, C. Z. (2015). Importancia de publicar artículos científicos desde las perspectivas individual, de las organizaciones y la sociedad. Revista Digital de Investigación y Postgrado, 5(4), 1.

Landivar, L. C., Ruppanner, L., Scarborough, W. J., Collins, C. (2020). Early signs indicate that COVID-19 is exacerbating gender inequality in the labor force. Socius, 6, 2378023120947997.

Larivière V, Ni C, Gingras Y, Cronin B, Sugimoto CR. (2013). Bibliometrics: global gender disparities in science. Nature; 504: 211–213. PMID: 24350369

Macaluso B, Larivière V, Sugimoto T, Sugimoto CR. (2016). Is science built on the shoulders of women? A study of gender differences in contributorship. Acad Med; 91: 1136–1142. https://doi.org/10. 1097/ACM.0000000000001261 PMID: 27276004

Microsoft Academic (Data set) https://docs.microsoft.com/en-us/academic-services/project-academic-knowledge/reference-interpret-method

Myers, K. R., Tham, W. Y., Yin, Y., Cohodes, N., Thursby, J. G., Thursby, M. C., ... Wang, D. (2020). Unequal effects of the COVID-19 pandemic on scientists. Nature human behaviour, 4(9), 880-883.

Nicholas H. Wolfinger, Mary Ann Mason Marc Goulden (2008) Problems in the Pipeline: Gender, Marriage, and Fertility in the Ivory Tower, The Journal of Higher Education, 79:4, 388-405, DOI: 10.1080/00221546.2008.11772108

Olga Shurchkov (23 de abril de 2020). Is COVID-19 turning back the clock on gender equality in academia?. https://medium.com/@olga.shurchkov/is-covid-19-turning-back-the-clock-on-gender-equality-in-academia-70c00d6b8ba1.

Staniscuaski, F., Reichert, F., Werneck, F. P., de Oliveira, L., Mello-Carpes, P. B., Soletti, R. C., ... Kmetzsch, L. (2020). Impact of COVID-19 on academic mothers. Science, 368(6492), 724-724.

Viglione, G. (2020). Are women publishing less during the pandemic? Here's what the data say.

Wenham, C., Smith, J., Morgan, R. (2020). COVID-19: the gendered impacts of the outbreak. The lancet, 395(10227), 846-848.

West JD, Jacquet J, King MM, Correll SJ, Bergstrom CT. (2013). The role of gender in scholarly authorship. PLoS ONE. Public Library of Science; 8: e66212. https://doi.org/10.1371/journal.pone.0066212 PMID: 23894278

Zhang, L. (2020). An institutional approach to gender diversity and firm performance. Organization Science, 31(2), 439-457.