14 June 2023

Dear Dr Rodgers,

Please find attached a revised version and track changes version of our manuscript “Meta-research: The effect of the COVID-19 pandemic on the gender gap in research productivity within academia” for consideration for publication in *eLife.*

We would like to thank the editors and the reviewers for their constructive and positive comments which have much improved our manuscript. Please find below a detailed description of our responses and alterations in light of these comments, **in bold,** underneath Author action/response, with text from the revised manuscript written in ***bold, blue italics***.

We hope you agree that the revised version is much improved due to the generous feedback!

Yours sincerely,

Kiran Lee (on behalf of all co-authors)

Response to Reviewers comments

Comment: This study provides fundamental evidence substantially advancing our understanding of the gender-specific impacts of the Covid-19 pandemic on academic research productivity. However, there are a number of concerns that need to be addressed.

**Author response**

**We are grateful for this positive feedback on our manuscript and for the constructive comments from the reviewers and the editor.  We have responded to the reviewers’ comments below.**

ESSENTIAL REVISIONS  
Comment: 1) I am concerned about the use of the included studies for a formal meta-analysis. Firstly, it is clear from the authors well-structured description in the methods section that many studies are not measuring the same outcome.  
Many studies use number of submitted or published articles, which may be a fairly comparable metric, but many of the survey studies use very different measures of productivity:  
a. Research time (N = 4), job-loss (N = 1), burnout (N = 1), and number of projects (N = 1)  
b. These are highly different outcomes, and should not be analysed as comparable.

**Author action/response:  
We conducted a sensitivity analysis excluding the survey studies and found little change in the results. We now mention the sensitivity analysis in the main text (L214-219):**

**“*Our results changed little when we conducted a sensitivity analysis that excluded seven effect sizes using four measures of productivity from survey-based studies that are less directly comparable: research time (N = 4), job-loss (N = 1), burnout (N = 1) and the number of projects (N = 1). When excluding these effect sizes, the overall estimate is -0.063, 95% CI = [-0.0892-  -0.0372] , SE = 0.0133, p < 0.001 and the estimate for survey-studies only is -0.239, 95% CI = [-0.3419- -0.1352], SE = 0.0527, p < 0.001*.”**

**We include all the studies in the main  analysis because although they are not directly the same as productivity output in terms of submissions and publications, they are relevant to an academic’s productivity and likely reflect the overall change in the gender bias in academic productivity that our article aims to assess.**

Comment: 2) Secondly, it is not clear how the studies differ in their use of the pandemic as an "intervention".  
a. Some studies will ask respondents to compare their productivity before and after, while some studies will provide actual comparisons of productivity before and after. Both methods will have their share of drawbacks, but I do not think they are comparable

**Author action/response:**

**All but 5 survey studies are about perceived impacts and not actual impacts, so these effects become separated in prediction 1b when we investigate survey and submission and publication studies as moderators. In addition, we recognize that none of our studies measure actual comparisons of individual-level productivity before and after the start of the pandemic. As we mention in the discussion, in the first submission, the findings of the studies comparing publication patterns (L242-243):**

**“*might reflect lower submission and acceptance rates of articles by women compared to their male colleagues or an increased drop-out of women from academia caused by the pandemic*.”**

Comment: b. The different survey studies will also have been fielded at different times, making the "intervention" (the pandemic) very different. Perhaps the loss of research time is most acute in the beginning, before academics have had time to adapt? Or later, because of burn-out?

**Author action/response:**

**We now include this point in the ‘Limitations’ section (L559-564):**

***“We recognise there are differences between article-studies in the length of time sampled before the pandemic (mean = 11 months, standard deviation = 10 months, range = 1-50 months) and during the pandemic (mean = 7 months, standard deviation = 5 months, range = 1-17 months). Survey studies were fielded at different times, (mean = 21/08/2020, standard deviation = 99 days, range = 20/04/2020 – 28/02/21) which potentially affects participants’ feelings of productivity changes.”***

**In addition, we mention in the discussion (L256-258):**

***“*It is likely that the article-studies we could include in our study underestimates the long-term effects of the pandemic, which might span over many years.”**

Comment: 3) Thirdly, the diversity of research design and modelling techniques makes it hard to conduct formal comparisons of effect estimates.  
  
**Author action/response:**

**Diversity in research designs is common in most fields as there are generally no set of standards on how to conduct a certain type of a study. This is why there is commonly a very large difference in between study heterogeneity in effect sizes. However, a part of the heterogeneity will be explained by certain common features of the studies, and we have tried to capture these in our analysis.**

Comment: None of the included studies are randomised controlled experiments. In fact, the research design (and modelling techniques) differs immensely across included studies. Therefore:  
  
a. We do not know to what degree individual, and thus by extension, individual gender gap estimates can be causally interpreted. This should be more explicitly acknowledged in the article

**Author action/response:**

**We now include the following sentence in the ‘Limitations’ of the ‘Methods’ (L569-571):**

**“*The patterns we describe should be seen as a potential indication that biases exist, but alternative approaches are needed to speculate about potential underlying causes and remedies*.”**

**Our finding that the gender gap has increased in academic productivity, is contextualised around evidence in other studies that the pandemic has enhanced gendered work roles and disadvantaging women. We start the final paragraph of the discussion with (L301-303):**

**“*Our study has several limitations, which are outlined in the Limitations of the methods and summarised here. Causes of the increased gender disparity in research productivity during the pandemic are not identified*.”**

Comment: b. Different estimates may be relatively closer to a causal effect, depending on the research design. By treating all designs equally, the authors gloss over these crucial differences in what we can gleam from each study.

**Author action/response:**

**Unlike medical meta-analyses, our approach does not assume that there is one causal effect that is being estimated by all studies. Rather, we assume that there can be multiple effects that all shape the gender inequality in academic productivity. Our question is to assess whether overall, these effects have changed during the pandemic such that also the gender inequality has either generally increased or decreased. The summary statements we provide reflect the average effect rather than our estimate of the true effect. We now clarify this in the introduction (L91-93):**

**“*We assume that the pandemic might have influenced the multiple aspects that jointly affect gender inequality in research productivity, and our estimate reflects whether on average these effects have increased or decreased gender inequality*.”**

**and in the discussion (L245-247):**

**“*Future studies might therefore detect different effects of the pandemic on gender disparity in productivity, depending on the outcomes they assess and the time-frame over which these changes are analysed*.”**

**and (L310-313):**

**“*Despite the heterogeneity in our sample, with regards to different outcome measures reflecting research productivity and different approaches to estimate effects, we overall find that the majority of studies performed to date indicate that the gender gap in research productivity has increased during the pandemic. This finding suggests that the COVID-19 pandemic has likely influenced many of the processes that contribute to differences in the achieved research productivity of women compared to men*.”**

**The most fundamental difference in research design of studies is between studies that assess productivity by numbers of male and female authors before and during the pandemic, and those that use surveys to assess productivity differences of individuals. We separate these studies in prediction 1b and Figure 2, and then subsequently only use measures of numbers of submissions and publications in downstream analyses investigating effects of research field, size of previous gender gap and authorship position. Finally, while Risk of Bias is commonly a part of meta-analysis in medicine, and sometimes in environmental evidence, it is almost never done in other fields. While we agree that this is an important component of meta-analysis, unfortunately there are no Risk of Bias standards or tools for the kind of studies we have used.**

Comment: c. It is not clear what "estimand" we get from the meta-analytical model. Some studies are cross-sectional, and will give us the average between-subject difference in productivity. Other studies, longitudinal ones, may give us the "difference in differences", i.e. the average difference in within-subject publication productivity trend. These are fundamentally two different answers to the question: Has the pandemic exacerbated the gender gap in research?

**Author action/response:**

**We referred to the difference between cross-sectional and longitudinal studies in our discussion, acknowledging that for cross-sectional studies selective disappearance may occur (L241-243):**

**“*Such a change might reflect lower submission and acceptance rates of articles by women compared to their male colleagues or an increased drop-out of women from academia caused by the pandemic*.”**

**All the included studies investigating productivity before and during the pandemic were cross-sectional. The ‘publishing’ studies aggregate data across individuals, and the surveys were only conducted at a single time point, asking people whether they thought the pandemic had affected their productivity. We also clarify in the ‘Extracting variables’ sub-section on ‘Effect sizes’ that (L254-256):**

**“*Six studies investigated numbers of articles at different time points during the course of the pandemic. From these studies, we calculated 30 effect sizes using numbers of articles across the entire pandemic period*.”**

Comment: d. The guidance from e.g. Cochrane is to use a random effects model to account for variation in what effect is estimated. The authors seem to use a random effects model to account for dependency between effects extracted, but I think they should also be more explicit in leveraging this for accounting for variation in what effect is estimated.

**Author action/response:**

**We are limited in using such an approach, because, as also pointed out by the reviewers, given the diverse factors shaping academic productivity we can only estimate the potential average effect of the pandemic on gender biases in research productivity rather than one true effect.**

Comment: e. Each included paper will have different modelling strategies. E.g. the Myers et al. 2020 paper uses regularised regression (Lasso), while the Andersen et al. (2020) used a hierarchical logit model. These differences may contribute to differences in estimated gender gaps in ways we don't know or lead to overfitting.

When 49 % of variation is explained, simply, by the data employed (survey, publication, or submission, see p. 5 of manuscript), it makes me wonder how much is simply explained by differences in research design or modeling strategy.

**Author action/response:**  
**We include the following in the ‘Limitations’ sections of the ‘Methods’ (L570-574): “*We used raw data to calculate effect sizes using the same modelling techniques wherever possible. This was not possible in 10 studies, where we consequently used effect size as provided in the study. We recognize that their different modelling techniques may have contributed to the estimated effect sizes.”***

Comment: 4) Another concern is the conclusions drawn from the differences in meta-analytic effect drawn from figure 2. To me, the arguments of survey estimates being more "true" to the actual effect, because they are closer to the real-time events (p. 17) could easily be turned on its head  
  
Those estimates are also self-reported, and only a snapshot at one point.  
a. People are bad at judging changes.

**Author action/response:**

**We include the following at the end of the second paragraph of the Discussion (L272-274):**

**“*As with any survey study, survey data are prone to subjectivity and suffer from limited numbers of respondents sampled which may influence the overall estimate*.”**

Comment: b. Productivity may have been severely hampered early on, when surveys were fielded, and then normalised over time

**Author action/response:**

**In our data, surveys were not fielded earlier (or later) compared to when the data collection for the publication studies occurred (both occurred at a median of 6-7 months into the pandemic). In addition, the reports we had found mostly indicated that research projects take longer between conception and final publication than the duration of the pandemic, suggesting that the effect of the pandemic on publication rates might persist for extended periods of time. However, the patterns over time are indeed difficult to predict, and we have therefore amended our statement in the discussion (L256-260):**

**“It is likely that the article-studies we could include in our study underestimate the long-term effects of the pandemic, which might span over many years.*. However, the exact time dynamics are difficult to predict as adjustments and changes in conditions might lead to a normalising in production patterns over time (Clark 2023*).”**

Comment: c. Survey estimates are around 4.2 and 4.9 times larger than estimates from publication data (0.193/0.046) and submission data (0.193/0.039) (p. 7), but also has the smallest sample sizes (I think from looking at 1/SE in figure 2). Couldn't this easily be an effect of higher degrees of noise and lower power in this data ?  
  
**Author action/response:**

**We include a leave-one-out sensitivity analysis for survey-studies only. We did not detect evidence of publication bias that might have substantially shifted the values up, but with the small sample there could be relatively high uncertainty around the mean effect estimate of the survey studies. The results of this are described (L220-223):**

**“​​*Overly large effect sizes did not change the results in a leave-one-out analysis, which repeatedly fitted the overall model as in prediction 1a, but for survey-studies only, leaving out one effect size at a time to see the effect on the overall survey-study estimate (Fig. S2). Leaving out the most influential effect size, the overall estimate was -0.183, 95% CI = [-0.270 -  -0.096] , SE = 0.045, p < 0.001*.”**

**and methods outlined (L534-536):**

**“*We also performed a leave-one-out analysis using the ‘leave1out’ function in ‘metafor’. This was a meta-analysis on survey-studies, leaving out exactly one study at a time to see the effect of individual studies on the overall survey-study estimate.*”**

Comment: d. Often, large-scale international surveys of researcher suffer from very low response rates and often mostly cover few countries (North American and Western Europe). At best, it makes it hard to generalise beyond these areas, and at worst this bias the results.

**Author action/response:**

**We include the following in the ‘Limitations’ section of the ‘Methods’ (L557-559):**

**“*Conclusions from survey studies are also limited to North American and Western European, since 18/23 studies are exclusive to or dominated by respondents from these regions*.”**

Comment: I appreciate a discussion of why survey estimates differ so much from publication- or submission-based studies, but this discussion is currently one-sided. The authors should consider if perhaps survey data suffers from other problems that could inflate estimates.

**Author action/response:**

**We have balanced the discussion for survey data following Comment 4) as in the preceding comments and corresponding author action/responses/**

Comment: 5) The authors seem to conflate gender inequality with gender bias throughout the paper (e.g. on line 55 or 281). In science studies, one would often denote a difference in outcome as a bias only if there is a provable mechanism of discrimination or differential treatment of women in e.g. journal accept rates, grant funding, etc. Instead, an indirect or structural difference could be a disparity or inequality brought on by other factors (e.g. women being disproportionally allotted more teaching load, domestic work, not being promoted to equal degree, etc.).  
  
a. There is a great working paper by Traag and Waltman with a discussion of these, and I suggest the authors use some of their clarification ([doi.org/10.48550/arXiv.2207.13665](http://doi.org/10.48550/arXiv.2207.13665)).

**Action/comments by authors:**

**We have revisited where we express “bias” and now avoid use of the word unless in context of mechanisms at play.**

Comment: b. I think this is important, not because I want to downplay potential bias, but because it matters for possible interventions. What we should do to ameliorate unfair differential productivity differs according to the causal mechanism at work.

**Action/comments by authors:**

**We have added a sentence to our ‘Conclusions’ (L317-321):**

**“*Our study cannot speak to the potential mechanisms that might have led to changes in gender inequality during the pandemic and is therefore limited in deriving suggestions for potential interventions to potentially ameliorate unfair differential productivity besides indicating that such inequalities appear pervasive.”***

Comment: 6) How was "before " and "during" pandemic defined? - this is not specified in the inclusion criteria, nor in how the data was extracted.  
Lines 456-457: unclear how you define "before the pandemic" and form where you get the data on the numbers. E.g., my first impression was that you analysed yourself all publications for a given discipline as your baseline. Please clarify.

**Author action/response:**

**We specify in the ‘Methods’ for inclusion criteria for the full-text screen that time periods for before and after the pandemic are chosen according to the authors’ discretion (L388-391):**

**“*Full texts were then screened to identify articles that had included four qualifiers: 1) both genders, 2) some quantifiable metric of academic productivity measured, and 3) compared a time period before the pandemic with a time period during the pandemic (time periods chosen according to the authors’ discretion), 4) primary data.”***

**We remind readers of this in the ‘Methods’ (L433-435):**

**“*Effect size: We extracted values needed to calculate 130 effect sizes from 55 articles investigating the effect of the pandemic on academic research productivity of both genders, comparing the productivity before and during the pandemic, using time periods chosen according to authors’ discretion”***

***and (L487-490):***

***“Previous gender disparity: For the article-studies with available data (N=99 effect sizes), we recorded the number of female and male authors before the pandemic, as defined by the time-period sampled in the respective study and use this ratio as a proxy for size of the previous gender disparity in that population sampled*.”**

Comment: 7) Screening  
# Study screening section: the inclusion criteria are too vague to be reproducible. It is also incomplete? Was publication status used to exclude evidence? How about publication language and full text availability?

**Author action/response:**

**We include the following in the ‘Study screening’ section of the ‘Methods’ (L375-376):**

**“*We included any text returned by the database, of any publication status, including grey literature in any language, though all texts had abstracts with an English version*.”**

**We include in the ‘Iterating the search’ of the ‘Methods’ (L407-408):**

**“*Two studies with full texts in Spanish were translated to English by Google translate, of which one was included*.”**

**We include in the ‘Iterating the search’ section of the ‘Methods’ (L391):**

**“*All full-texts were available*”.**

**We include the following in the ‘Extracting variables’ sub-section of the ‘Methods’ (L498-503):**

**“*We also extracted data for the following variables, to enable description of the datasets: date range before the pandemic; date range during the pandemic; geographic region; data availability for gender and geographic region interaction effect; data availability for gender and career stage/age interaction effect; data availability for gender and parent status interaction effect; gender assignment accuracy threshold for article-studies and gender inference method used (Supplementary file 2, see also ‘Limitations’).”***

Comment: Was there a limitation on the geographical scope or targeted career stages for the included studies and was this information coded in any way?

**Author action/response:**

**We acknowledge in the ‘Limitations’ section of the ‘Methods’ (L552-564):**

**“*Most studies explored academic populations worldwide (N = 99 effect sizes), or from Western (N = 28 effect sizes) regions, but not the Global South (N = 3 effect sizes) limiting investigation of interaction effects between geographic regions. Although 22/130 effect sizes from 8/55 studies held data subdivided between geographic regions, we did not extract separate effect sizes as they differed in the scale of geographic region sampled, which limited our ability to make geographic comparisons. Conclusions from survey studies are also limited to North American and Western European, since 18/23 studies are exclusive to, or have the majority of respondents from these regions. Additionally, we recognise there are differences between article-studies in the length of time considered as before the pandemic (mean = 11 months, standard deviation = 10 months, range = 1-50 months) and during the pandemic (mean = 6.6 months, standard deviation = 5 months, range = 1-17 months). Survey studies were fielded at different times, (mean = 21/08/2020, standard deviation = 99 days, range = 20/04/2020 – 28/02/21) which potentially affects participants’ beliefs of productivity changes*”**

Comment: # Line 360: why not all articles were double-screened to avoid errors? It would not add much work.

**Author action/response: We include the following the ‘Study screening’ section of the ‘Methods’ (L383-387):**

**“*To reduce workload, we did not double-initial-screen 160 texts because of 97% agreement rates achieved.. We acknowledge this may miss up to 13% of studies ​​(Waffenschmidt et al., 2019)“***

Comment: # Line 368 and 382: no details provided how full-text articles were screened, e.g., in terms of the used software, also it is not specified if they were double-screened.

**Author action/response: We include in the ‘Study screening’ section of the ‘Methods’ (L388-391):**

**“*Full texts were then screened to identify articles that had included four qualifiers: 1) both genders, 2) some quantifiable metric of academic productivity measured, and 3) compared a time period before the pandemic with a time period during the pandemic (time periods chosen according to the authors’ discretion), 4) primary data. Full-text screening was conducted by recording which, if any of the four qualifiers were missing for each text in an excel spreadsheet and only performed by K.L.”***

Comment: # Table 1 header: first and only mention that articles were screened from 2020 only - this is a first mention on having such a limit on publication searches - this should be in the search strategy description.

**Author action/response: We include in the ‘Search process’ section of the ‘Methods’ (L336):**

**“*All searches were filtered for texts from 2020 onwards*.”**

Comment: # The authors should describe whether some quality weighting was used or quality assessment played a part in the inclusion criteria.

**Author action/response: We include in the ‘Study screening’ section of the ‘Methods’ (L391-392): “*Studies were not assigned quality weightings as there is no common standard to do this*.”**

Comment: 8) Data and meta-data  
# Extracting variables section: please refer the reader to a proper meta-data SI file describing in detail all extracted variables, at a level that enables replication of the data extraction and fully informed data reuse (the "meta-data" file archived on Zenodo does not provide such information in sufficient detail and some of the variables in the provided raw data file are quite cryptic).

**Author actions/response:**

**We include a datafile (Supplementary file 2, “Variables” sheet) describing all variables extracted. We include the following in the ‘Extracting variables’ section of ‘Methods’ (L498-503):**

**“*We also extracted data for the following variables to enable description of our datasets: date range before the pandemic; date range during the pandemic; geographic region; data availability for gender and geographic region interaction effect; data availability for gender and career stage/age interaction effect; data availability for gender and parent status interaction effect; gender assignment accuracy threshold for article-studies and gender inference method used*. *Please see* *Supplementary file 2, “Variables” sheet, for descriptions of all the variables*”**

Comment: # A data file is needed including a list of articles at the full-text screening stage, with individual reasons for exclusion.

**Author actions/response:**

**We include a datafile (Supplementary file 1) describing individual reasons for excluding articles.**

**We included in the ‘Iterating the search’ section of the ‘Methods’ (L403-407):**

**“*For these 169, we screened the full texts, excluding 120 studies (53 studies missed one qualifier, 38 studies missed two qualifiers, 27 studies miss three qualifiers, 1 study missed four qualifiers) and keeping 50 (including the 25 identified in the original search) that all contained the necessary information to calculate the effect sizes (Supplementary file 1)*”**

Comment: 9) Lines 324-327: Only 17 out of 21 eligible articles were included in the extracted data set - what happened with the remaining 4? Why were they not included in the dataset? From the data files available on Zenodo I can see these were not published as journals articles, however inclusion criteria do not state that only grey literature should be excluded.

**Author action/response:**

**There are five, not four studies, because the original submission listed 14/17 scoped articles indexed in Scopus picked up by the search, when this should have been 13/17 articles. This has since been corrected.**

**We have included the five articles that were scoped but not picked up in the databases and re-run the analyses using these data and state this in the Iterating the search section of the ‘Methods’ (L408-410):**

**“*Five studies found in the scoping search but not returned in database searches were also included, resulting in 55 total studies used to extract variables for analysis*.”**

Comment: 10) Line 339-352: The search string used had only 14/17 sensitivity (82%). However, the next step of search strategy refined did not improve the sensitivity and just removed some of the irrelevant studies to reduce the screening effort. Given the limited sample size, and limited scope of the searches, it is of concern that a significant proportion of potentially available evidence has been missed (at least 20%). The main reason for using "test set" (benchmarking) articles is to improve search sensitivity, with search precision being less important. Since it is hard to get 100% sensitivity in database searches, additional search strategies should be used to locate potentially missed articles, for example by collecting and screening articles that were cited by or cited already included articles (after the main database searching and screening is completed). Also, grey literature should have been searched.

**Author action/response:**

**Minimising work load at the expense of potentially missing some available evidence was a trade-off we were willing to make given the need to quickly produce results in the context of a potentially pressing issue. For this, we chose to use litsearchR to improve efficiency of finding texts and to not perform forwards and backwards searches. Additionally, given the research topic is so new, we will consistently be finding new articles if we performed backwards and forwards searches and so we chose to not perform these. We did not search specific grey literature databases, but did not filter against grey literature during screening. We welcome future research to incorporate our scripts and methods to monitor this new research topic including more research as and when it becomes available.**

Comment: 11) Lines 350-251: the recommended 10% hit rate should be calculated from pilot screening round not from how many "test set" articles have been captured by the search string.

**Author action/response:**

**We have deleted the statement about “above the 10% hit rate recommended by Foo et. al”.**

Comment: 12) Lines 411-413: regression models per se are not effect sizes, please specify which estimates from these models were used as effect sizes and how they were combined with other effect sizes.  
  
**Author action/response:**

**We include in the ‘Extracting variables’ section of the ‘Methods’ (L437-440):**

**“*For 10 effect sizes which did not have data available to calculate our own effect sizes, we used already calculated percentage changes in the gender gap in academic productivity as predicted from lasso regression (N = 2), Somers’ delta (N = 2), ordered logistic regression (N = 1) and mixed-effect models (N = 5).*”**

Comment: 13) Analyses section: no description of any sensitivity analyses (e.g., doing analyses leaving out outliers or leaving-one-study-out approach).  
  
**Author action/response:**

**We include the following sensitivity analysis leaving studies in the ‘Are our results robust?’ section of the ‘Results’ (L213-223):**

**“*Our results show little change in a sensitivity analysis that excludes seven effect sizes using four measures of productivity from survey-based studies that are less directly comparable: research time (N = 4), job-loss (N = 1), burnout (N = 1) and the number of projects (N = 1). When excluding these effect sizes, the overall estimate is -0.063, 95% CI = [-0.0892-  -0.0372] , SE = 0.0133, p < 0.001 and the estimate for survey-studies only is -0.239, 95% CI = [-0.3419- -0.1352], SE = 0.0527, p < 0.001*.”**

**We now include a leave-one-out analysis for the overall effect of the pandemic on the gender gap and record the results in the ‘Are our results robust?’ section of the ‘Results’ (L220-223):**

**“*Overly large effect sizes did not change the results in a leave-one-out analysis, which repeatedly fitted the overall model as in prediction 1a, but for survey-studies only, leaving out one effect size at a time to see the effect on the overall survey-study estimate (Fig. S2). Leaving out the most influential effect size, the overall estimate was -0.183, 95% CI = [-0.270 - -0.096] , SE = 0.045, p < 0.001.*” and display this as a forest plot as Figure S2.**

**These analyses are written in ‘Analyses’ section of the ‘Methods’ (L530-536):**

**“*We applied a sensitivity analysis testing prediction 1a (overall pandemic effect on gender gap) and prediction 1b (method of research productivity influence on pandemic effect on gender gap) excluding seven effect sizes using four measures of productivity from survey-based studies that are less directly comparable: research time (N = 4), job-loss (N = 1), burnout (N = 1) and the number of projects (N = 1). We also performed a leave-one-out analysis using the ‘leave1out’ function in ‘metafor’. This performed a meta-analysis on survey-studies, leaving out exactly one study at a time to see the effect of individual studies on the overall survey-study estimate.”***

Comment: 14) ‘Limitations’ section: there are many potential limitations of this study that are not mentioned here. For example, limitations of the search conducted to find the evidence - not searching grey literature, no other additional searches, not searching in languages other than English. This potentially resulted in biased evidence as well as limited the sample size. Also, some statistical analyses are less reliable due to small numbers of effect sizes within some of the levels of categorical variables tested.

**Author action/response:**

**We have considerably expanded our ‘Limitations’ section in the ‘Methods’, including all the examples given by the reviewer (L538-571):**

**“*Our focus is on comparing the effect of the pandemic on women relative to men. We recognize that gender extends beyond this comparison, and that biases are even more likely to target individuals whose identities are less represented and often ignored. These biases also reflect in a lack of studies of the full diversity of gender. While several of the surveys we include had the option for respondents to identify beyond the binary women/men, none of these studies report on these individuals, presumably because of the respective small samples. In addition, all studies using numbers of submissions or publications (38 out of 55) to measure research productivity used automatic approaches that are more likely to mis-gender individuals as they inferred binary gender based on first names. While these approaches seemingly offer the potential to identify trends in larger samples, they themselves introduce and reinforce biases in relation to gender that are hard to assess, intersecting with biases in ethnicity as these approaches are often restricted to names common in English speaking countries (Mihaljević et al., 2019). We do not include grey literature databases in our searches, which may bias our samples to studies with positive effects. We did not perform forwards or backwards searches, meaning we may have missed some relevant studies. However, we expect the literature on the topic to grow, and hope that further work will build on our study and add these new effect sizes to our dataset. Most studies explored academic populations worldwide (N = 99 effect sizes), or from Western (N = 28 effect sizes),) regions, but not the Global South (N = 3 effect sizes),) limiting investigation of interaction effects between geographic regions. Although 22/130 effect sizes from 8/55 studies held data subdivided between geographic regions, we did not extract separate effect sizes as they differed in the scale of geographic region sampled, which limited our ability to make geographic comparisons. Conclusions from survey studies are also limited to North American and Western European, since 18/23 studies are exclusive to or have the majority of respondents from these regions. We recognise there are differences between article-studies in the length of time considered as before the pandemic (mean = 11 months, standard deviation = 10 months, range = 1-50 months) and during the pandemic (mean = 7 months, standard deviation = 5 months, range = 1-17 months). Survey studies were fielded at different times, (mean = 21/08/2020, standard deviation = 99 days, range = 20/04/2020 – 28/02/21) which potentially affects participants’ beliefs of productivity changes. Investigating research field and authorship position effects is limited by the unequal and sometimes small sample sizes of variables that are compared. We use 10 studies that provide effect sizes without raw data available to calculate our own effect size and recognise their differing modelling techniques may contribute to the estimated effect sizes. The patterns we describe should be seen as a potential indication that biases exist, but alternative approaches are needed to speculate about potential underlying causes and remedies*.”**

**We also close the ‘Discussion’ with a summary of limitations (L301-307):**

**“*Our study has several limitations, which are outlined in the Limitations section of the methods and summarised here. Causes of the increased gender disparity in research productivity during the pandemic are not identified. Gender is investigated only as a binary variable, mostly using first-name prediction tools, with no investigation into non-binary or transgender. Geographic regions are not investigated, limiting generalisability as our samples are not equally representative of all geographic regions. All studies are to some extent unique by combined differences in sampling method and analyses performed*.”**

Comment: 15) Table S4 item 20d: Figures 2, 3, 5 show meta-regression results testing secondary hypotheses. Those are not sensitivity analyses.

**Author action/response:**

**We changed the entry for 20d sensitivity analyses in Table S4 to refer to the places where we added the results from the sensitivity analyses in the ‘Are our results robust?’ section of the methods: “Page 15, Figure S2” (L213-224).**

MINOR COMMENTS  
Comment: a) First paragraph of the "Results" sections should provide a better overview of the dataset because Methods sections comes only after Discussion section. For the same reason, each subsection of the Results should have a brief sentence introducing what sort of data (full/subset) was used to obtain a given result. Without this information reading this section is very confusing.

**Author action/response:**

**We now added this information to the ‘Results’ (L106-111):**

***“Our systematic literature review identified 55 studies that met the inclusion criteria (for details on the procedure please see the Methods ). All of the identified studies only compared women to men (see Limitations). We extracted and calculated 130 effect sizes from these studies and performed a meta-analysis and meta-regression to test our three hypotheses and related predictions. Out of 130 effect sizes, 23 are based on survey responses (survey-studies), and 107 are article-studies that measure the number of submitted or published articles (article-studies).”***

Comment: b) Discussion lines 270-272: "financial structures or childcare arrangements" are usually institution- or individual-level and as such cannot explain field-level differences.

**Author action/response:**

**This has been deleted and replaced with (L323-325):**

**“*For example, tenure-clock extensions designed to accommodate pandemic disruption may inadvertently exacerbate the gender gap by extending the period that advantaged individuals can outperform*”**

Comment: c) Lines 305-307: unclear sentence, please rewrite.”

**Author action/response:**

**Changed to (L323-325):**

**“*For example, tenure-clock extensions designed to accommodate pandemic disruption may inadvertently exacerbate the gap by extending the period that advantaged individuals can outperform*.”**

Comment: d) Lines 311-312: unclear sentence, please rewrite.

**Author action/response:**

**The sentence was removed, and replaced with a more general ending (L326-329):**

**“*We recommend future studies to investigate potential mechanisms of the increased gender gap in academic productivity, to continue monitoring the gendered disparity of academic productivity, and to investigate any long-term implications that can arise from reduced productivity of women*.”**

Comment: e) Table S3: define timeframes for comparison and intervention.

**Author action/response**:

**We have included the following in Table S3 under ‘3) comparison’:**

**“*There is no exact start date to the global pandemic. Include all studies mentioning the pandemic as the exact dates defining the duration of the pandemic will be chosen at the discretion of the authors*.”**

**and in ‘2) intervention’:**

**“*There is no exact start date to the global pandemic. Include all studies mentioning the pandemic as the exact dates defining the duration of time before the pandemic will be chosen at the discretion of the authors*.”**

Comment: f) Table S4 item 11 (page 21) does not mention any assessment of within-study risk of bias (this is not a question about how the studies were screened; it is about assessing how valid their methodology was).

**Author action/response**:

**Risk of Bias assessment is not common in the meta-analysis in fields outside of the medicine and environmental sciences. Consequently, there is no common tool or method to assess the bias of studies. While we recognize this is suboptimal, we also do not feel positioned to come up with a system to evaluate risk of bias in our case. Note that some of the authors of this paper are working on the development of a common risk of bias assessment that can be applicable across the fields.**

Comment: g) I have a hard time understanding the connection between the numbers presented on page 9 (line 155-164) and the numbers presented in figure 3.  
a. On the page, the text shows social sciences as having the greater gender gap (- 0.079) but the figure suggests an effect size of -0.092. Should these be the same?  
b. In the figure, I also seems that multidisciplinary sciences showed the highest gap, but this differs in the text.

**Author action/response:**

**Yes, there was a mistake when transcribing the numbers, and we have now corrected this (L155-161):**

*“****When considering research fields individually, social sciences showed the greatest increases in the academic productivity gender gap during the pandemic (-0.084, 95% CI = -0.143 to -0.024, SE = 0.030, p = 0.006), followed by medicine (-0.066, 95% CI = -0.102 to -0.029, SE = 0.019, p < 0.001). The pandemic showed little effect in multidisciplinary fields (-0.050, 95% CI = -0.125 to 0.024, SE = 0.038, p = 0.188), biological sciences (-0.003, 95% CI = -0.057 to 0.050, SE = 0.027, p = 0.902) or Technology, Engineering, Mathematics, Chemistry and Physics fields (-0.003, 95% CI = -0.053 to 0.048, SE = 0.026, p = 0.916).”***

Comment: h) Publication bias  
a. The authors argue that small studies with large effect sizes do not skew their meta-analytical estimate (p. 15, line 214-218), but how this this the case with such a big difference in slope between article-output studies and survey-studies (-0.026 vs. -0.190). I think this should be more clearly spelled out.

**Author action/response:**

**We have highlighted our reasons for separating between article and survey studies (L525-529):**

**“*We tested for publication bias by performing a multilevel regression model (Nakagawa et al., 2022) which investigates whether small studies have large effect sizes, including research productivity measure as a moderator because of differences in sample size between article-studies and survey studies*.”**

**We have expanded on the Publication bias section of the Results (L206-212):**

**“*The multilevel meta-regression, including article-studies and survey studies as a moderator because of their differences in sample size, showed no evidence of publication bias (Article-studies: slope= -0.025, 95% CI = [-0.059- 0.009], SE = 0.017,  p = 0.148; Survey-response studies: slope= -0.157, 95% CI = [-0.244-  -0.071], SE = 0.044,  p <0.001). This model correlates standard error with effect size and negative slopes suggest small studies do not have large effect sizes. meta-regression. A visual inspection of the funnel plots (Fig. 6) similarly did not indicate any suggestion of publication bias*.”**

Comment: i) Discussion  
a. In the discussion the authors report an overall estimated gender gap of 7 % and 4 % for article-output studies. The authors should report this number for survey-studies as well in order to show how these differ.

**Author action/response:**

**We have included in the discussion (L245-247):**

**“​​*When measured by surveys, productivity reduction was 17% higher than in men.*”**

Comment: b. Please ignore this if I am wrong, but is it 7 % or 7 percentage points of change? Could you show your calculation just as you do with a back-of-the-envelope calculation on line 237-240?

**Author action/response:**

**The percentage changes in the discussion refer to the effect sizes calculated and stated in the results and reflects a gender inequality during the pandemic that is 7% larger than the inequality that existed before the pandemic. Given that this effect size estimate reflects very different measures (e.g. for some surveys a shift in the average score on the Likert scale by 7%, whereas for publications a change in the proportion of authors who are women), we cannot provide a back-of-the envelope calculation. We have changed the sentence in the discussion to clarify that the 7% reflects a relative, not an absolute scale change (L233-235):**

***“Overall, the studies summarised in our meta-analysis suggest that gender gaps in research productivity within academia increased on average by 7% relative to the gender gaps that existed before the pandemic”.***

Comment: j) QM metric  
- Please introduce/briefly explain the QM metric this first time it is mentioned in the article

**Author action/response:**

**We have included the following in the ‘Analyses’ section of the ‘Methods’ (L511-513):**

**“*Models that include moderators use an omnibus test of parameters reported as a QM metric, which tests the null hypothesis that all moderator effect sizes are equal and is significant when at least two moderators are different.*”**

- Please add a short description to the analysis section about how you tested for prediction 3a

**Author action/response:**

**We have included the following in the ‘Analyses’ section of the ‘Methods’ (L524-525): “*To test prediction 3a, we tested in a model authorship position as a moderator on effect size for publication studies*.”**