Introduction

In this paper, I present an agent based model (ABM) that simulates the effects of gendered citation practices on a cohort of scientists as they progress through their careers by acquiring grants, publishing, and accumulating citations. By *gendered citation practices*, I mean the small differences between the average citation rates of men and women that have been empirically observed in a variety of scientific disciplines including astronomy, psychology, ecology, and a variety of medical research fields. I also investigate how different granting schemes may impact the community and interact with gendered citation practices.

I explore whether proposals focused on giving grants to the most highly cited researchers have disparate impacts on women. The evidence for or against the proposal is not definitive (nor could it be given the constraints of an ABM), but it does suggest that such a proposal would be ill advised in some circumstances. In particular, it seems that the impacts for women are worse in fields where average publication counts are low, especially early in a career and when the observed citation bias against women is high.

To do so, I first will argue that an ABM can be used to isolate the impact gendered citation practices have on publications, grants, and citations over the long term. Many studies show that women face several disadvantages in science (Moss-Racusin et al, 2013; Ley and Hamilton 2013; Whitman et al, 2019; Helmer et al, 2018; Lariviere et al 2013). However, few of these studies investigate if gender citation practices are a cause of disadvantage.[1] An ABM can provide another line of evidence that supports the causal story that women are disadvantaged in science due to gendered citation practices.

I also argue that an ABM is an effective way to test proposals to change grant funding that weigh citations much more heavily than the current system. While (so far) no grant agencies have proposed to make all decisions about grants based solely on citation counts, citations are seen as a potential proxy for measuring impact of past work and may provide information about the impact of future work. If grant agencies wanted to maximize the impact of the work they fund, heavily weighting citations might be a way to do so. The model developed here can test such a proposal and consider if such a proposal should be shelved because it causes disproportionately worse outcomes for women.

[1] Lariviere et al 2013 does discuss gender differences in citation and publication rates and concludes that increasing women's representation and full participation in science is necessary, but there are no simple ways to do so.

The model shows that there is a negative correlation between the variable for level of bias against women and the average grant rate for women compared to men; the stronger the bias, the more negative the outcomes; they receive fewer citations and are less likely to be in the top 25% of most cited researchers. The effect sizes are generally small; the average proportion of women's average citations compared to men's is reduced by about 1%, even when bias is high. I also investigate the potential impact of bias in astronomy using empirical publication and citation rates for the discipline. Women are also less likely to be in the top 25% of most cited researchers and have a lower average proportion of citations.

Agent Based Models

Agent based models are an increasingly common tool for philosophers. Models can be used to provide a possible explanation for observed macro level empirical phenomena by testing whether a set of agents will produce the expected phenomenon given a set of simple rules (Epstein 2006). In contrast to statistical models, ABMs allow the agents to interact with their simulated environment and the environment responds to the individual choices over the course of the model (Bruch and Atwell, 2015).

Philosophers of science have used models to investigate questions such as how information should be shared in a scientific community (Zollman, 2013), how the productivity gap may arise between men and women (Kofi Bright, 2017), and how discrimination can arise in epistemic communities (Rubin and O'Connor, 2018).

Agent based models have also been used to inform policy decisions. Urban planners have used detailed simulations of traffic conditions in a city and simple rules about how people in the simulation move around to show what impact certain transit policies might have on the flow of traffic through the city or how it might change land use (Waddell 2002, 2003). Epidemiological models are used to show that a policy intervention such as mask mandates are effective in controlling the spread of an illness by comparing what happens in the model when agents wear masks to what happens when they do not (Auchincloss and Roux, 2008). Some philosophers have also used the results of their models to make suggestions about public policy, such as Heesen (2017), which considers some proposals about grant allotment and will be discussed in more detail later.

Modeling Gendered Citation Practices

This chapter's ABM has two main aims: to investigate whether gendered citation practices alone make a difference to the long term outcomes of agents of different genders, and to test how gendered citation practices would interact with a proposal to give grants to more highly cited researchers.

There is a great deal of literature about the problems women face in science: a chilly climate (Wylie, 2007; Chisholm et al, 1999), lowered publication output (Larivière et al, 2013), a leaky pipeline leading to lower levels of career attainment and an underrepresentation of women, less access to mentorship, implicit biases (Moss Racusin et al 2014), and (sometimes) lower grant rates[2]. While gendered citation practices are obviously connected to these other issues, some studies indicate that there is bias even after controlling for many of these factors, such as fewer papers published, differences in citations for subdisciplines or topics, and lower average academic rank (Caplar, Tacchella, and Birrer, 2017; Way, 2017).

The productivity of scientists is already a key factor in evaluating researchers for grants and tenure and promotion; the number of publications and prestige of the journal they are published in is seen as a good proxy for the quality of the research. Moher, Goodman, and Ioannidis (2017) look at the tenure and promotion criteria at a variety of US academic biomedicine institutions and find that they are generally focused on productivity measures, including number of articles published, journal impact factor for the venues for those publications, and the citation counts of those publications.

Women are also less likely to be cited than men in a variety of scientific fields. Larivière et al (2013) find that papers with women as the solo author or in prestigious authorship positions--generally the first or last, depending on the discipline--are cited less than papers with men in the same type of authorship role. Their dataset includes over 5 million papers from 2008 to 2012, from nearly every country in the world. Other fields where women have been observed to receive fewer citations than men include astronomy (Caplar, Tacchella, and Birrer 2017), archaeology (Hutson 2002), communication science (Knobloch-Westerwick and Glynn 2013), ecology and evolutionary biology (Kelly and Jennions 2006), epidemiology (Schisterman et al. 2017), gynecological oncology (Hill et al. 2015), international relations (Maliniak, Powers, and Walter 2013), neuroscience (González-Álvarez and Cervera-Crespo 2017), neurosurgery (Khan

[2] Bornmann et al (2007) conducts a meta-analysis of data from a variety of international grant agencies and finds that women receive grants at lower rates than men; Van der Lee and Ellemer (2015) show that women receive fewer grants than men in the Netherland's national research council; Witterman et al (2019) shows that women receive fewer grants from Canada's national research council. But Pohlhaus et al (2010) finds no difference between men and women in NIH grants. Both Van der Lee and Witterman et al find that the gaps are due to lower evaluations of women principal investigators, not due to lower proposal quality, meaning that if women are missing out on grants, it may be because of the compounded effects of gender difference.

et al. 2009), pediatric neurosurgery (Klimo et al. 2014), and psychology (Geraci, Balsis, and Busch 2015).

An ABM can attempt to isolate the impacts of gendered citation practices. An ABM can assume that women and men are, for example, equally likely to publish and equally likely to receive a grant but assume that there is some small difference between the citation rates of men and women and investigate the impact of that difference alone. The difference of men and women scientists' actual careers can be directly measured by empirical investigation, but it is incredibly difficult to determine what caused those differences. An ABM can provide evidence that citation bias alone is responsible for some of the differences in career outcomes between men and women.

Modeling citation-focused grant funding

Citation counts are an increasingly popular way to measure impact of scientific work; many assume that the best papers with the greatest ability to shape future science are the ones that are mostly widely cited. Ioannidis and Khoury (2014) and Heesen (2017) discuss, with varying levels of endorsement, some policy proposals that give grants to researchers that are the most highly cited. In this paper, I will consider what might happen if these trends were taken to their logical conclusion: what if grants were given to researchers solely based on being among the most highly cited?

Ioannidis and Khoury (2014) provide a suggestion for what this policy might look like. They argue in favor of giving grants and tenure to researchers who do well in the areas they highlight as important: productivity, quality, replicability, shareability, and translatability. By productivity, they specifically advocate for using a metric such as the top 10% of most highly cited articles to determine who is most productive. They stress the importance of using all of the measures, so would not approve of solely using citation metrics alone. But much of the rest of their proposal is purposefully vague; they want grant agencies and tenure committees to decide for themselves how important each criterion is, and how to evaluate them. The most concrete suggestion they offer is to give grants to those who are most highly cited, and this paper will test this citation-focused implementation of their proposal.

Heesen (2017) discusses the type of scientist who might find such a proposal compelling: a scientist who is convinced that the most highly cited papers are such because they are the most informative papers and deserve all their citations on merit alone. He calls this the competence-based view of academic superstars, where superstars are just those who are very highly cited (4511). Given that granting agencies want to fund work that will be successful and impactful, if the competence-based model is correct, grant agencies should award grants to those who are most highly cited. Heesen ultimately does not endorse the competence-based model because he shows that epistemic luck can explain the emergence of superstars as well as the competence-based model can.

The proposal I test in this paper is: give grants to the most highly cited scientists in the model. I also consider what happens when the cutoff for who constitutes a top performer is higher than the grant rate, and what happens when additional grants are given to those outside of the top performers after each top performer has earned a grant.

There are several reasons to use an ABM to test this proposal. A model can investigate what impacts a policy proposal might have on the careers of scientists without having to experiment with the careers of actual people. Trying to test the impact of changing how grants or tenure is decided is both impractical and unethical. It would require massive changes to how science currently operates, which is unlikely to happen given how many institutions would have to agree to the change.

Further, if the practical and ethical issues of changing the reward structure of science were overcome and the intervention could be implemented in a randomized way, there are still many confounds that would make it hard to interpret any changes caused by changing methods for allocating grant funding; changing the grant structure of science would have an impact on the careers of scientists, as do many other factors that couldn’t be controlled for. While an ABM cannot provide conclusive evidence about the effects of a major policy change, neither could an empirical investigation. However, an ABM can be implemented by one person and provide some evidence within a much shorter timeframe and with many fewer resources than a large experiment would.

Finally, a model can investigate what, if any, disparate impacts there will be on women. Given that gender bias exists in several scientific disciplines, focusing solely on productivity measures may further disadvantage women in science. The model can also test what the specific impacts of gendered citation practices alone will be even if women published at equal rates and were equally likely to receive grants. But if grants are given solely to top performers and women may be less likely to be among the top performers, women’s careers may suffer. The model can investigate how large an impact such a proposal would have when gendered citation practices are the only kinds of bias. If women are less likely to get grants when grants are given to top performers while at the same time they are also less likely to be cited, then it is likely the impact will be even greater when other forms of bias are also in play.

Methods

The model was coded in NetLogo 6.0.4, a programming language designed for agent based models. A description of the model follows, and the full code of the model is in Appendix A.

Initializing the model

In the initial phase, the model first generates scientists, based on a user input for the size of the community. Half are men and half are women; if an odd number is input, there will be one additional man. The choice to have the two populations be roughly the same size is arbitrary. However, using similarly sized populations removes the possibility that differences in population sizes are driving any effects seen in the model.

The starting agents are meant to represent a cohort of researchers at roughly the same career stage, near the beginning of their careers. It is reasonable to think of these agents as scientists starting in postdoc positions, for example, where the papers generated in this initial round represent those published while in graduate school. It can also be used to represent researchers starting a tenure track faculty position, where the initial publications are those published during both graduate school and previous postdocs. The model can be adjusted to represent an accurate size for this community based on graduation rates or available postdoc or tenure track positions for a particular discipline. Large community sizes require large amounts of memory and are only feasible with distributed computing or a supercomputer, so to keep the processing power small and accessible, I only focused on small communities. So, the model could not capture the entire field of chemistry, for example, where 2,810 doctorates were awarded in 2018 (https://ncses.nsf.gov/pubs/nsf20301/data-tables) without using distributed computing. It can capture a smaller subset of chemistry PhDs, however, such as a cohort of PhD graduates in polymer chemistry, where the number of PhDs awarded in 2018 was 132[3].

[3] It also may be that the model can better represent a subdiscipline than a larger discipline. In chemistry, it may be rare for a polymer chemist to cite an organic chemist, for example, but highly likely that they will cite other polymer chemists. The model does not account for any kind of clustering across topics, so it better fits the assumptions of the model to focus on a community where a paper would plausibly cite any other paper, because the topics are close enough. Wang et al (2019), for example, uses citation patterns

Once the appropriate number of scientists has been created, the next step is to generate initial publications for each person. The model uses a log normal distribution to generate the initial number of publications. A log normal distribution is described by the mean and standard deviation of a normal distribution where each value from the normal distribution is then raised to *e*. The user input value for the average number of publications is not the mean used to generate the initial values. Instead, the user input average is used to calculate the mean of the normal distribution according to the formula:

$$mean\ = exp(\mu\ + \frac{\sigma^2}{2})$$

where the mean is the user input average and σ is the standard deviation, which is held constant at .6. Next, the number of initial publications for each agent is generated by drawing a number *x* from a normal distribution with mean μ and a standard deviation of .6. Then, *x* is used as the exponent for *e*, rounded up to the nearest integer, then 1 is subtracted from the value. This ensures that there are only integer values of publications, and that it is possible for an agent to have no publications in the initial phase.

The choice to use a log normal distribution is supported by some empirical results. There is limited data about what the distribution of publications early in a researcher's career looks like. However, I was able to secure access to two data sets, one from sociology and one from computer science that contain information about the number of publications assistant professors have at time of hire.[4] Both also included information about prior position status; the sociology data set noted whether an individual had been ABD, a postdoc, an assistant professor, or some other position prior to appointment, while the computer science data only includes whether each individual had a postdoc immediately prior to appointment. The sociology data covers from 2003 until 2020; the computer science data covers the 1970's to 2011. These analyses, and all subsequent analyses, were done in R. We tested both datasets, focusing on the most recent years of each--2007-2011 for computer science and 2016-2020 for sociology--to determine what distribution fit best. We estimated the parameters of the log normal distribution that most closely fit our data. We tested Poisson, gamma, log normal, negative binomial, and Weibull distributions, because they all have zero as a lower bound and infinity as an upper bound[5]. Most

to sort papers from a corpus into 1,450 subdisciplines; these subdisciplines, as opposed to what they call top-level fields, are better candidates for the model.

[4] The data set from sociology was discussed in Bauldry (2013), and he has continued to add and maintain the data set. The data set from computer science was reported on in Way, Larremore, and Clauset (2016).

[5] There is no hard upper limit to the number of papers someone could write; even the most productive author ever could potentially write one more paper. However, there are of course practical limits to the

people have a low number of publications, but there are always academic superstars with many, many publications; each distribution considered has a long right tail. The log normal distribution provided the best fit for both data sets, including when looking at subsets by year or type of prior position; this was determined visually. The mean number of publications for each data set varied based on what subset was included, but the standard deviation of the normal distribution that describes the logarithmic normal graph was generally between .5 and .75, so I selected .6 to be the standard deviation for the normal distribution used to generate the initial number of publications. This value could easily be changed but including it as a user input would likely reduce the clarity of the model.

Every publication has exactly one author. It is common in most scientific fields that multi-author collaborations are the norm, so this assumption of the model may seem inaccurate. However, the model only includes a small cohort of scientists entering the field at roughly the same time. There are many other members of the community than just this small set of agents, and those others could be acting as co-authors and collaborators for each of these papers. Given that each agent represents someone working as a postdoc or early faculty member in a lab, it is reasonable to interpret these agents as belonging to distinct labs and therefore being unlikely to collaborate with other agents in the cohort. This assumption may be less realistic as the model progresses; as each agent has been in the community longer, they are more likely to interact and collaborate with a larger network of researchers, including those who incidentally entered the community at roughly the same time. However, the current version of the model does not account for this. It would require additional empirical investigation to what extent such collaborations do take place. If they do, then a future direction for the model could be to integrate such interactions.

Once every agent has an appropriate number of publications generated, citations from each publication are generated. The number of citations per paper is governed by a user input for the average proportion of papers cited within the community. The average citations may vary based on the field considered, and the model is meant to be general enough to investigate a variety of scientific fields. To find the average number of citations per publication this represents, the user input proportion is multiplied by the total number of publications. For each

number of papers one author can produce. Ioannidis et al (2018) identify Akihisa Inoue as the researcher with the most publications; he had published 2,566 papers between 1976 and 2016. (Ioannidis fails to note that he is likely to surpass Akihisa Inoue as the most productive; he has over 2,000 publications on his google scholar profile, starting in 1994, meaning he has an average of 80 publications per year to Inoue's mere 64). While an impossible value for the number of publications pulled from a log normal distribution is always possible, it is a very unlikely event, given how numbers are generated from a distribution.

publication, a number is drawn from a log normal distribution with the natural log of the average number of citations as its mean and a standard deviation of 1. If the value drawn from the distribution is less than 1, or greater than the total number of publications that can be cited, a new number from the same distribution is generated.

Log normal is used again since citations are also bound at zero but with no upper limit and have a long right tail. There is no empirical evidence or even investigation, to my knowledge, of what the shape of citation counts for early career researchers are. Most empirical work on citation counts looks at a whole field at a time, with researchers of all career stages, and finds that citations obey a power law. Using log normal to generate citations over time does lead to a distribution that appears to be a power law. Additionally, I was able to collect and analyze a small dataset about citations of early career researchers in astronomy, which provides some empirical support that log normal is a reasonable distribution.

Commented [1]: Nice! Can you cite the datasets you refer to throughout this chapter?

Each publication has an equal chance to be cited in this set up phase. A publication cannot cite itself, but it can cite a publication written by the same author[6]. The number of citations generated by the log normal function for each person represents finding papers to cite and compiling a bibliography for their paper. The number of bibliographic entries in a paper in most scientific fields is quite large, but again, the model is meant to represent a small slice of the community, so it is reasonable to use input values for the average number of citations that are not representative of how many entries are in a bibliography.

Running the Model

Once the community is generated, a new set of functions is used to run the model. There are two modes that the user can turn on or off that govern how the model runs. First, one can control whether the model considers the gender of the author of a publication when deciding to make a new citation; if this mode is in effect, there is also an input that adjusts how strong the bias towards authors who are men is. Second, one can control whether the model gives grants to those agents who are highly cited. Both modes work independently of each other, so there are four conditions for the model to run under: bias off and grants for top performers off; bias on and grants for top performers off; bias off and grants for top performers

[6] The model, if anything, likely undercounts the extent to which self-citation occurs; King et al (2017) find that approximately 10% of citations in a paper are to the author's own previous publications.

on; and bias on and grants for top performers on. I will first discuss the most basic case, where both sliders are off.

In each step of the model, three functions are called that represent the work of scientists. First, each agent has an opportunity to receive a grant. Next, each agent has an opportunity to publish. Finally, new publications generate new citations.

The first function determines which agents get a grant in this round of the model. The rate at which grants are given is a user input, expressed as a decimal. The user input determines how many grants are given, then distributes the grants randomly. This way of distributing grants is essentially a lottery; any agent can receive a grant, and previous grants do not increase the likelihood of future grant success.

Next, the agents who have received a grant in this round are able to publish. Every grant results immediately in one publication. The model also does not allow an agent to publish without a grant in that turn of the model. Both assumptions represent a simplification of the scientific process, since of course many grants do result in multiple publications which may not be published immediately and since scientists with other resources at hand (e.g., startup funds, shared instrumentation, graduate students, etc.) are able to publish without grant funding.

Finally, new publications generate new citations. The number of average citations at any given step is equal to the current number of publications times the user provided value for the average proportions of papers cited. This means that each new paper will cite more and more of the publications in the community, meaning that the average number of citations per paper and per author will increase exponentially instead of linearly.[7] This makes sense, given that as the cohort of researchers age, they will be more likely to be familiar with and build on (or critique) each other's work.

In this phase, unlike the initial phase, each publication does not have an equal chance of being cited. First, a random publication is chosen. Then, one of that publication's links is chosen. Then, one end of that link is chosen; it could either be the original publication or the existing paper cited by the new paper. Roughly 50% of the time, the original publication will be chosen, resulting in an equal chance for every paper to be selected. However, the other 50% of the time the model will tend to pick publications with more citations; a paper with more links connecting it to other papers is more likely to 'win' this lottery. The choice of the publication connected to will be driven by the degree distribution of the network--the distribution of the number of links each publication has. The distribution of new citations in the model will not be

[7] This accords with empirical citation patterns across science (Barabási and Albert 1999).

uniform as a result. The process can be thought of as the author of the new publication doing a literature search and finding a relevant paper, then choosing to cite either the original paper, or a paper in its bibliography.

In the case where gender bias is introduced into the model, the grant and publish functions are identical to the case without bias; it is only in generating citations that the gender of the author matters. The evidence for whether the rate of success in receiving grants is affected by gender is mixed. One reason women may receive grants less frequently than men is that women publish less, on average than men. I am interested in investigating one possible mechanism, gendered citation practices, for citation disparities and attempting to show what would happen if only this mechanism were operating. This tells us how much impact this mechanism alone has on citation disparities.

In the biased method of generating citations, the function starts out in the same way as the unbiased case. It chooses a publication at random, then chooses one of its citations at random. Next, however, the function checks what gender the two authors at either end of the link are. If they are the same, it will function just like the unbiased version; it will choose one end at random. If they are different, then the amount of bias is considered. The amount of bias can vary from 1% to 10%, where women are that percentage less likely to receive a citation when the authors selected are of different genders. So, when there is bias in the model, authors of either gender will choose to cite the male author over the female one that percentage of the time[8].

The other variation in the model changes how grants are allotted. When the option for grants for top performers is turned off, grants will be equally distributed across all agents. But when the user turns on grants for top performers, grants will instead mostly go to the individuals with the highest citation counts. How the model decides who gets grants under this condition varies based on user inputs. There are two relevant user inputs: the grant rate and the cutoff for who counts as a top performer. When the grant rate and cutoff are the same, all grants are given to the top performers. When the grant rate is lower than the cutoff--such as when the grant rate is 10% but the cutoff is 25%--the model will select 10% of the total number of authors

[8] This would likely reduce the total number of citations for women by less than the amount of bias would indicate. The author only discriminates when the choice is to choose a man over a woman to cite. However, given that there are only author pairs with different genders approximately 33% of the time (at least in early stages of the model), this means a 10% chance to cite a man over a woman is only realized one third of the time, so the ‘true’ bias level is closer to 3.3%. This value may evolve over the course of the model, however, if the number of heterogeneous author pairs changes over time. But if the parameter for 10% bias does not accurately represent 10% fewer citations for women, this may make a substantial impact to how the results of the model are interpreted.

to receive grants, but the only eligible authors are those in the top 25%. When the grant rate is higher than the top performer cutoff--for example, the grant rate is 50% and the cutoff is 30%--the top 30% of performers will all receive grants, and another 20% of the total number of authors will receive grants, although only those not in the top 30% will be eligible for the additional grants (no one can receive more than one grant in this iteration of the model).

Results

Parameter Sweep

A parameter sweep is necessary to determine if the results of the model are robust across a number of parameter values. It is done by varying the inputs of the model across a set of plausible values to determine what the parameter space of the model is.

Choosing parameter sweep values

I ran the parameter sweeps in four subsets of conditions for the model: bias off and grants for top performers off; bias on and grants for top performers off; bias off and grants for top performers on; and bias on and grants for top performers on. This helps to avoid duplication of runs; for example, when both bias and grants for top performers are off, there would be 90 identical runs where the program iterates through the nine combinations for those two values. As each run gets longer and more complex, this duplication of runs can add significant time to the parameter sweep and increases the likelihood of errors related to inadequate memory.

The values selected are meant to represent plausible values for the inputs. For community size, 50, 100 and 250 are meant to represent a small, medium, and large size community. The average number of publication values of 3, 7, and 10 may be low for most scientific disciplines, but larger values, especially for community size of 250, cause memory errors. The average proportion of citations within the community is difficult to estimate, but even for the smallest community of 50 with the lowest number of average publications of 3, a proportion of .01 means that each paper will have approximately two citations from within the community. For a community size of 250, lower proportions of average citations are used, since using the same proportion of .01 would represent substantially more citation per papers than it does in smaller community sizes, so the proportions are smaller to keep better in line with a similar absolute number of initial citations per paper. Even at the lowest value of three initial publications and with an average citation proportion of .001, each paper will have an average of one citation (and will quickly accumulate more, since the proportion is rescaled to the number of papers after each step of the model). The values for grant rate are given to represent a high, medium, and low value, although 50% may be an unrealistically high grant rate for many funding agencies. The values chosen for grants given to top performers ensures that each case where the grants to top performers is greater than, less than, or equal to the grant rate will occur. The bias level is chosen to represent a high, medium, and low value for bias.

Parameter sweep	
Community Size	50, 100, 250
Average publications	3, 7, 10, [25, 50 for community size 50]
Average proportion of citations	[.01, .005 for community sizes of 50 and 100] [.005, .001 for community size of 250]
Grant rate	10%, 25%, 50%
Bias level	1%, 5%, 10%
Percent of top performers to receive a grant	10%, 25%
Gender bias	True, false
Grants given to top performers	True, false

Table 1

Models of proportional averages of citation, publications, and grants

We fit three linear models for each of the possible outcomes for the proportional average difference between men and women: grants, publications, and citations. For each outcome, the average value for women was divided by the sum of the means of men and women. This ensures that the differences are scaled in terms of the outcome variable, making it possible to compare the differences between grant rates for different average numbers of publications. If men and women have a difference of two publications when everyone has three initial publications on average, it would not be equivalent to a difference of two publications when there is initially an average of 50 publications. Dividing by the sum of the means scales the differences so that they can be more reasonably compared.

Each model used the same set of predictors. Gender bias and bias level were treated as one variable; where when gender bias was false, it was treated as being 0. Grants given to top performers and the percent of top performers to receive a grant were also treated as the same variable, where when giving grants to top performers was false, it was treated as grants were given to the top 100% (i.e., anyone in the community was equally likely to receive a grant). We created a categorical variable representing the various combinations of community size and average proportion of citations used. This was necessary because different values for the proportion were used across the different community sizes; although .005 was used for all three, it was the high value for a community of 250 and the low value for communities of 100 and 50. Publication count and grant rate were also included as predictors. The interaction of each other predictor with gender bias were also included in each model. All predictors were treated as categorical variables.

For the first model, the outcome variable was the average proportional difference between men and women on citations across a single run. First, we investigated what, if any, parameters can be relaxed. If a parameter does not increase the predictive power of a model, it can be relaxed. One way to test this is to generate new reduced models to compare the predictive power. In the reduced model, the parameter value is dropped from the model. If the reduced model yields equal predictive power to the original model, then the parameter can be relaxed. Dropping community size and the grant rate given to top performers (including whether it was on, since the model treated grants for top performers off the same as having the input set to 100%) did not reduce the predictive power of the model. So, these inputs, community size and grant rate for a top percentage of performers, did not add to the predictive power of the model. This is only true for communities between 50 and 250 and grants given to the top 10-100% of performers, since those parameter values were the ones tested.

Once the model parameters that do not add predictive power are removed, the model can be fit again, and the results interpreted. The summary table for the model on average proportion of citations for women is shown in Table 2. Bias level is a significant main effect. When there is bias of five or ten points, women on average have significantly fewer citations; women’s proportion of citation was lowered by 1% when bias is set to five versus the case where there is no bias, and 3% lower when bias is set to ten.

There are also a number of significant interactions. In particular, as the average number of publications increases, the negative impact on women’s average citations decreases; when bias is set to ten, but there are an average of 50 initial publications, women on average perform the same as men. This may be explained by the fact that there is no bias in the set up phase, regardless of what the bias level is set to. The model runs for 50 turns, so when the average number of publications at the start is large (such as 25 or 50), the citations generated at the start of the model without bias protect against any effect the bias level might otherwise have. The bias level also interacts with grant rate. Higher grant rates are worse for women, with the largest effect when bias is the highest, too; at a grant rate of 50% and a bias level of 10%, women have 2% fewer citations than when there is no bias.

	Estimate	P-value	
(Intercept)	0.50022	< 2e-16	***
Average publications: 7	0.00152	0.500622	
Average publications: 10	-0.00086	0.704221	
Average publications: 25	0.002587	0.452918	
Average publications: 50	-0.00304	0.377031	
Bias level: 1	-0.00017	0.954138	
Bias level: 5	-0.01022	0.000568	***
Bias level: 10	-0.03027	< 2e-16	***
Grant rate: 0.25	-0.0022	0.288728	
Grant rate: 0.5	0.001184	0.568487	
Average publications: 7 x Bias level: 1	-0.00471	0.161047	
Average publications: 10 x Bias level: 1	-0.00066	0.844636	
Average publications: 25 x Bias level: 1	-0.00031	0.949765	
Average publications: 50 x Bias level: 1	0.009659	0.050167	.
Average publications: 7 x Bias level: 5	0.001658	0.622135	
Average publications: 10 x Bias level: 5	0.007036	0.036464	*
Average publications: 25 x Bias level: 5	0.008579	0.08196	.
Average publications: 50 x Bias level: 5	0.018292	0.000209	***
Average publications: 7 x Bias level: 10	0.013692	4.72E-05	***
Average publications: 10 x Bias level: 10	0.020002	2.83E-09	***

Average publications: 25 x Bias level: 10	0.023999	1.15E-06	***
Average publications: 50 x Bias level: 10	0.033971	6.03E-12	***
Bias level: 1 x Grant rate: 0.25	-0.0001	0.973574	
Bias level: 5 x Grant rate: 0.25	-0.00688	0.024898	*
Bias level: 10 x Grant rate: 0.25	-0.00927	0.00251	**
Bias level: 1 x Grant rate: 0.5	-0.00482	0.116201	
Bias level: 5 x Grant rate: 0.5	-0.01107	0.000305	***
Bias level: 10 x Grant rate: 0.5	-0.02083	1.14E-11	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 2

The next model predicts the proportion of publications for women. The model was selected for as in the previous example, and community size, average publications and grant rate did not add to the predictive power of the model. The summary table is shown in Table 3. Bias level of ten was the only significant main effect, and it lowered the proportion of papers authored by women by .9% compared to when there is no bias. The only significant interaction was between bias level 10 and grants to top performers off; this raised the proportion of papers by women by roughly .9% when compared to the case where publications went only to the top 10% of performers. However, there was not a significant effect for the difference between grants to the top 10% and 25% of performers at the highest bias level. This effect does indicate that women have fewer publications when grants are given to top performers and gender bias is high compared to when grants are given equally.

	Estimate	P-value	
(Intercept)	0.50102	<2e-16	***
Grants for top performer proportion: 0.25	-0.000106	0.964961	
Grants for top performer proportion: 1	-0.001212	0.637173	
Bias Level: 1	6.43E-05	0.980037	
Bias Level: 5	-0.002726	0.288754	

Bias Level: 10	-0.009516	0.000214	***
Grants for top performer proportion: 0.25 x Bias Level: 1	-0.003896	0.283653	
Grants for top performer proportion: 1 x Bias Level: 1	0.000328	0.930159	
Grants for top performer proportion: 0.25 x Bias Level: 5	-0.002534	0.485654	
Grants for top performer proportion: 1 x Bias Level: 5	0.004061	0.277705	
Grants for top performer proportion: 0.25 x Bias Level: 10	-0.001262	0.728332	
Grants for top performer proportion: 1 x Bias Level: 10	0.009656	0.00987	**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 3

Finally, the last model focuses on the proportion of grants received by women. The results are very similar to those of the model for publications. Community size, average publications, and grant rate also do not add to the power of the model. The summary table is provided in Table 4. The highest level of bias is the only significant main effect, although the medium bias level of five bias points, is close to significance (p ≈ .07). High bias reduces the average proportion of grants for women by 1.6%. The only significant interaction is between ten points of bias and grants given to top performers off, where women fare better than when grants are given to the top 10% of performers. The negative effect of bias on grant rates is nearly erased when grants are given equally.

	Estimate	P-value	
(Intercept)	0.502283	<2e-16	***
Bias Level: 1	-0.000334	0.93477	
Bias Level: 5	-0.007435	0.0688	.
Bias Level: 10	-0.016291	6.72E-05	***
Grants for top performer proportion: 0.25 x Bias Level: 1	-0.001588	0.7834	
Grants for top performer proportion: 1 x Bias Level: 1	-0.000314	0.95786	
Grants for top performer proportion: 0.25 x Bias Level: 5	0.000677	0.90673	
Grants for top performer proportion: 1 x Bias Level: 5	0.007796	0.19	

Grants for top performer proportion: 0.25 x Bias Level: 10	0.0018	0.75532	
Grants for top performer proportion: 1 x Bias Level: 10	0.016714	0.00496	**

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 4

Model of percentage of women in top 25%

The previous models only used the averages for all men and all women in the model. However, given that the proposal being tested is directed towards the top performers, it is important to consider whether giving grants to only top performers would reduce the occurrence of women in the set of top performers. Each run of the parameter value also reported how many men and how many women were in the top 25% of performers; the top 25% was always reported, even when grants were given randomly, and when grants were given to the top 25% and top 10%.

To normalize for differences in community size, the number of women and the number of men reported in the top 25% for each run was converted into a percentage of the top performers who were women. An ANOVA analysis was performed on the data, and the summary table is shown in Table 5. The analysis shows that when there is the highest level of bias, women are significantly less likely to be in the top 25% of performers; the percentage of women was reduced by two percentage points for the highest level of bias. Figure 1 shows the differences in the proportion of women across each level of bias; there is a clear downward shift from each box to the next.

Average publications were also a significant main effect. However, the effect size was very small; changing the number of average publications lowered the proportion of women by less than .1 percentage points. Additionally, there are significant interactions between the level of bias and grant rate; when bias and grant rates increase, the percent of women in the top performers decreases. The effect is largest when bias is high and the grant rate is high; when bias is high and the grant rate is 50%, the proportion of women in the top 25% is reduced by 6 percentage points, relative to no bias and a 10% grant rate. Since the grant rate drives publication rates, this makes sense; when the grant rate is higher and bias is higher, there are more opportunities for new publications to fail to cite women.

	Estimate	P-value	
(Intercept)	4.89E-01	<2e-16	***
Bias Level: 10	-2.28E-02	2.73E-05	***
Bias Level: 1	8.27E-04	0.87879	
Bias Level: 5	-7.29E-03	0.17896	
Grant Rate: 0.25	-2.96E-03	0.58528	
Grant Rate: 0.5	4.31E-03	0.42732	
Average publications	6.45E-04	2.98E-14	***
Average citation proportion: 0.005	-1.74E-03	0.61074	
Average citation proportion: 0.01	-6.04E-03	0.09125	.
Grants for top performer proportion: 0.1	4.85E-03	0.0738	.
Grants for top performer proportion: 0.25	1.60E-02	3.77E-09	***
Bias Level: 10 x Grant Rate: 0.25	-2.02E-02	0.00861	**
Bias Level: 1 x Grant Rate: 0.25	-6.66E-03	0.3851	
Bias Level: 5 x Grant Rate: 0.25	-1.47E-02	0.05478	.
Bias Level: 10 x Grant Rate: 0.5	-6.30E-02	2.39E-16	***
Bias Level: 1 x Grant Rate: 0.5	-1.45E-02	0.05796	.
Bias Level: 5 x Grant Rate: 0.5	-3.37E-02	1.13E-05	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 5

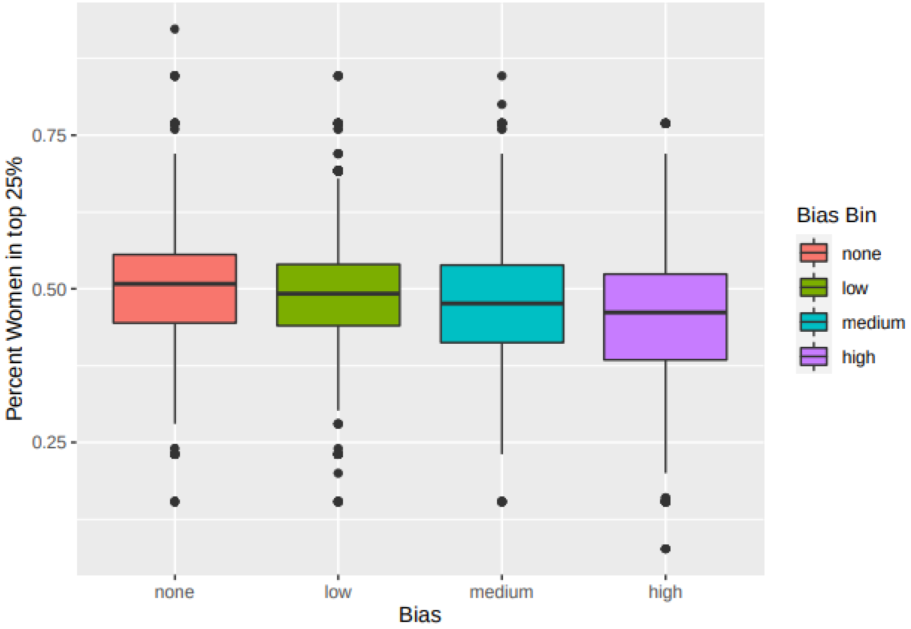

Figure 1

Experiment: Astronomy

The model can be used to test what might happen to a community if a policy to give grants to top performers was instituted. The model can use input values about a specific scientific discipline and provide a prediction about what might occur if that discipline implemented the policy.

I selected astronomy for several reasons. First, there are roughly 150 graduates from PhD programs every year, and approximately 60-75 faculty positions advertised per year (Trump et al 2020). The small size makes it easier to collect information about what tenure track assistant professors look like at the time of hire, and astronomy has a jobs wiki that records both searches underway and, in some cases, the applicant who received the job. Additionally, the small size is well within the community size where the model runs efficiently, so a large number of replications can be done. Second, while there are several possible funding sources in astronomy, the two most prestigious for US-based researchers are NSF and NASA, both of which publicly report their grant success rates. As of 2014, NSF funded 17% of grant applications, while NASA funded approximately 20% (Astronomy and Astrophysics Advisory Committee, 2016). Finally, Caplar, Tacchella, and Birrer (2017) shows that papers authored by women in the top five astronomy journals receive 10% fewer citations than predicted, given other characteristics of their papers; astronomy is a field where gendered citation practices seem to occur. Therefore, it is likely good estimates are available for all the parameters needed for the model.

Commented [2]: This section is great. I like the rationale. Is Astronomy a field that tends to involve research across labs (because they need to share a few very expensive telescopes)? If so, this is not a blow down problem for your paper. But, in the discussion, you might want to mention that this assumption may not be the best fit for Astronomy in particular; but, you don't expect the results to change under different conditions, etc.

Choosing model values for the experiment

First, I determined the appropriate community size for the model. There were approximately 70 tenure track positions listed on the astronomy jobs wiki in the 2019-2020 academic year at universities in the US and Canada[9]. Positions at universities outside of the US and Canada were excluded in the count, as were research institutes. The astronomy jobs wiki is an unofficial listing of posted jobs, so it may contain inaccuracies, but the wiki is clearly used by many and frequently updated. There are notes about when first round interviews went out, updates on additional application materials requested, lists of campus visit candidates, and sometimes who the job was offered to and accepted by. Job applicants have a strong interest in having a centralized resource for job listings and information, so despite the informal nature of the list, it is likely the best list of jobs and hired candidates available[10]. Given that the model is meant to represent researchers when starting an academic position such as a postdoc or assistant professor, 70 is a reasonable value for the community size parameter.

The values for the grant rate can easily be taken from the average success rates for NSF and NASA grants, 17% and 20% respectively. Given the closeness of the values, I choose to only test the NASA grant rate of 20% to reduce the number of runs necessary, since the grant rate does not seem to have a large impact on the outcomes of the model. I choose to test what would happen if grants were given to the top 10%, 20%, and 30%; this ensures that there is a

[9] The wiki also includes information about postdocs. Given the nature of astronomy, many of these postdocs are based internationally at various telescopes, but many candidates will be from the US and Canada, making it difficult to determine which institutions to include or exclude.

[10] Some of these searchers were ultimately cancelled due to Covid-19, so there were not 70 tenure track hires. However, it is unclear how many of the job searches were cancelled, and it is unclear what impact Covid-19 may have had on who was hired and when. While this was not a typical academic job cycle, it is not possible to collect data on any previous year’s hires, at least with the expectation that the data collected will represent what their bibliometric profile looked like at time of hire.

case where the grant rate is higher, lower, and (for the runs using the 20% grant rate from NASA), equal to the percentage of top performers who receive grants.

Caplar, Tacchella, and Birrer (2017) provides good estimates for the values of bias points to be tested. They found that in the top five astronomy journals, men are cited more than women. The authors used a random forest algorithm to model the number of citations based on "the seniority of the first author, the number of references, the total number of authors, the year of publication, the journal of publication, the field of study and the geographical region of the first author's institution" (2, 2017) and trained it on a set of papers authored by men. Their model predicted the numbers of citations expected based on the non-gendered characteristics of a second set of papers, which were written by women, which the authors then compared with the actual numbers of citations made to these papers. They found that women receive 10% fewer citations than expected. The model predicts that, given the features of the papers written by women, women should receive 4% more citations than men, but instead men receive 6% more citations. Given that the bias level functions as the percentage by which women are less likely to be cited, this makes it reasonable to set the bias level at 6%, to represent the 6% more citations men receive, and with 10%, to represent the 10% more citations they predict women should receive. Additionally, I tested with just one bias point to see what would happen with the minimum amount of gender bias.

For the values for average number of publications, I collected a small dataset based on information from the astronomy jobs wiki. 24 out of the 70 listings noted who received the position, so I compiled a list of names and new institutional affiliations. Then, I looked for Google Scholar profiles for each of these researchers. Only 18 researchers had Google Scholar profiles. For each of these 18 profiles, I collected the number of total citations as calculated by Google Scholar, as well as information about how many peer reviewed papers they had written, how many citations each paper had, and when their first peer reviewed academic publication was published. Most Google Scholar profiles included a number of conference presentations as well as papers posted on arXiv, and these entries were not counted in the peer reviewed paper totals, even if they had citations. I also tried to exclude any papers that seemed to be obviously included in error; one profile had a few medical research papers from the 80s that could not be the work of the same researcher who wrote papers in astronomy, for example.

It is also unclear how representative this sample may be of assistant professor hires for 2019-2020, given that is a convenience sample of researchers whose names were listed on the jobs wiki website as receiving a position (many of whom likely contributed their own names to the website) and who also had a Google Scholar profile; these profiles may systematically differ from those who were hired but not listed on the jobs wiki, or from those who were listed but did not have Google Scholar profiles. This data was collected in November 2020, so it includes publications from 2020 that were not listed on the CVs of applicants when they compiled their application materials in the fall of 2019. However, despite its limitations, this data should represent a snapshot of some assistant professors in astronomy early on in their professorial careers.

The average number of publications for this group was 35, with a standard deviation of 33.37; the minimum number of publications was 7 and the maximum was 139. Several distributions were fit to the data including gamma, Poisson, and log normal distributions, and the log normal gave the best fit[11]; the parameters for the log normal function are a mean of 3.31 and a standard deviation of .696. I used the formula to calculate the mean of the log normal function:

$$mean = exp(\mu + \frac{\sigma^2}{2}) = 34.3$$

Given the closeness of the values for the log normal distribution and the actual data, I used 35 for the model parameter. The model is programmed to use .6 as its standard deviation and it is not an input that can be changed without changing the code, but I left it at .6, since the values are close, and if anything, there is likely less spread to the actual number of initial publications, given that .696 comes from a small sample with a large range.

The final parameter for the model is the average proportion of citations. I collected data about total citations for each individual in the astronomy data set as well as information about citations for each paper. The average number of total citations on peer reviewed papers per individual was 2,655 with a standard deviation of 3,412; the minimum was 141, the maximum was 12,940, and the median was 1468. The average citations per paper was 75.5 with a standard deviation of 195; the minimum was 0, the maximum was 2032, and the median was 29.[12]

The large spread of citations can be partially explained by the fact that researchers have had different length careers. I collected information about the first year each individual published a peer reviewed article in astronomy, since papers from five years ago will generally have more citations than papers from one year ago. Using this information, I was able to calculate an

[11] Log normal was the best fit, but Weibull, gamma, and negative binomial were also good fits; Poisson was a very poor fit. This dataset is very small, so it is easier to fit a curve to it than a large dataset but given that log normal was also the best fit for the data from computer science and sociology, log normal is most likely the best way to describe the shape here as well.

[12] The data sets from sociology and computer science that were discussed in the methods section did not have any information about the number of citations per paper at the time of hire (or at any other time).

average number of citations per paper per year: 19.5, with a median of 12.9, a minimum of 5.5 and a maximum of 104.3.

However, this data does not immediately help when determining what is a reasonable value for the *proportion* of papers in the model (i.e., written by other early career researchers) that are cited by another researcher in the model. The Astrophysics Data Service (ADS) has 13 million database records including peer reviewed and non-peer reviewed publications and preprints. If every article has 1,000 citations, it would only cite a vanishingly small proportion of the total articles. However, those 13 million entries cover publications since essentially the start of the discipline. In a rapidly changing and developing field, it is likely that more recent articles get the bulk of citations. An analysis of citation practices in astronomy from 1981 shows that most articles get the bulk of their citations in the first five years after publication (Abt, 1981); that timeline may be even more compressed at this point. So, it is likely that early career authors are citing one another, given that they are publishing in the same time frame and citations tend towards more recent articles[13].

But when experimenting with values for the proportion of articles, it became clear that values such as .01 were likely far too high. When exploring the possible parameters for astronomy, I found that using a .01 proportion of papers within the model cited lead to authors having an average of roughly 2,000 citations in total at the start of the model. While this is not so far off from the average number of citations per individual calculated above, it does not make sense given the assumptions of the model: the model only represents the proportion of papers cited from within the model, not the entire astronomical community. So, the average number of citations per person should be much lower than the observed number, given the assumption of the model[14]. As plausible estimates, I choose .005 and .001 since these result in average initial citations at more reasonable values.

Again, the experiments were run in four groups: bias off and grants for top performers off; bias on and grants for top performers off; bias off and grants for top performers on; and bias on and grants for top performers on. Each iteration of the model was run 1000 times, resulting in 32,000 runs.

[13] The model does not explicitly account for age in the calculations for likelihood to cite. If anything, it is more likely to cite older papers, since they will have a higher degree of connection.

[14] How much lower is, of course, an open question for which it would be very difficult to collect data on. I also want to avoid putting too much confidence in the value of average citations per person based on such a small sample that may be unrepresentative.

Test Parameters: Astronomy	
Community Size	70
Average publications	35
Average proportion of citations	.005, .001
Grant rate	20%
Bias points	1, 6, 10
Percent of top performers to receive a grant	10%, 20%, 30%
Gender bias	True, false
Grants given to top performers	True, false

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 6

Results of the astronomy experiment

Although there were supposed to be 32,0000 runs of the experiment, five are missing[15]. Thus, the analysis was performed on 31,995 runs. We used the same methods as described above: an ANOVA was performed using all the parameters that varied, then compared to a series of reduced models to determine, which, if any parameters could be safely dropped from the model. The outcomes were again the average proportion of citations, publications, and grants for women and the percent of women in the top 25% of all scientists by total citation

[15] The five that are missing are all from the same combination of parameters and occurred sequentially. Either there was some issue internal to what happened in each run, or the issue was external to the simulation. There were no errors recorded to immediately explain what happened to these five runs; NetLogo does record errors when the problem is inadequate memory, for example, although it may not exhaustively report errors. When a simulation gets stuck in an infinite loop, NetLogo will not report it as an error, but in my experience, it also does not terminate the program at any point. This set of simulations ran to their (seemingly) successful endpoint without having to be terminated or otherwise interfered with. Given that the five runs occurred sequentially in a group of 16,200 runs, it seems unlikely that five sequential runs ran into the same error and that all were an error that Netlogo fails to report. I do not know what other kind of errors external to the simulation might be likely culprits for these five runs failing, but I do not believe it's a problem internal to the simulation. While the five are all from the same combination of parameters, the difference between 1,000 results and 995 results to analyze in terms of the power of the model is negligible, so the runs were not repeated and added to the analysis.

counts. The summary table for the ANOVA performed on the reduced model is reported for each outcome.

For the average proportion of grants and publications, there was no significant impact for any of the parameters, including bias. Even at the highest level of bias, it did not have a significant effect on the difference between men and women.

For the average proportion of citations, bias level did have a significant effect at the medium and high level. As the bias level increased, the average proportion of citations going to women decreased. However, the effect is quite small; even at the highest level of bias, women receive only .5% less of the proportion of citations. The summary table is shown in table 7.

	Estimate	P-value	
	Estimate	P-value	
(Intercept)	0.49938	<2e-16	***
Bias Level: 1	-0.000329	0.601	
Bias Level: 6	-0.003582	1.26E-08	***
Bias Level: 10	-0.005395	<2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 7

For the proportion of women in the top 25% of researchers by total citations, bias level did have a significant effect at the highest two levels. The summary table for this model is provided in Table 8. Again, the effect size is fairly small; women still represent 49% of the top 25% of performers, even at the highest level of bias.

	Estimate	P-value	
(Intercept)	0.498938	<2e-16	***
Bias Level: 1	-0.001028	0.523	
Bias Level: 6	-0.00825	3.01E-07	***
Bias Level: 10	-0.011362	1.76E-12	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 8

The model so far has assumed that there are equal numbers of men and women. But in astronomy, more men than women who graduate with PhDs each year in astronomy. Women represented 40% of the PhD graduates in 2014 and 40% of recently hired faculty in astronomy in 2016 (Porter and Ivie, 2019). Women (appeared to) represent 10 out of the 24 researchers identified as tenure track assistant professors starting in 2020 on the astro job wiki. A lower percentage of women could amplify the effects of bias against women, as shown, for example in Rubin and O'Connor (2018), where biased strategies against minority groups are more likely to develop as the size of the minority group decreases.

I added a parameter that allowed changing the percent of women in the model, and ran 8,600 additional simulations, where women represented 40% of the researchers in the model. Since the average proportion of citations did not affect the outcomes, it did not vary. Instead, only bias level and grants for top performers varied in the same way as described in the table above.

When adding these additional observations to the dataset of 31,995 runs where women represented 50% of the researchers, the results were almost identical. Bias level was still the only significant main effect for average proportion of citations; publications and grant average proportions were unaffected by any of the parameters. The summary chart for the average proportion of citations is shown in Table 9.

	Estimate	P-value	
(Intercept)	0.499732	<2e-16	***
Bias Level: 1	-0.000348	0.616	
Bias Level: 6	-0.002858	3.96E-05	***
Bias Level: 10	-0.00597	<2e-16	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 9

The percentage of women was a predictor of the composition of the top 25% of most cited researchers, but essentially having 10% fewer women in the model simply reduced their percentage in the top 25% by 10%. The summary table for the model is provided in table 10. Lowering the percentage of women in the model did not seem to make a substantial difference to the results for astronomy, although the percentage of women in a community may begin to make a difference as the percentage of women decreases.

	Estimate	P-value	
(Intercept)	0.497484	<2e-16	***
Percent Women: 40	-0.096837	<2e-16	***
Bias Level: 1	-0.000557	0.75688	
Bias Level: 6	-0.005526	0.00212	**
Bias Level: 10	-0.010546	4.52E-09	***

Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Table 10

Discussion

The two main questions this model intended to answer were: does bias alone, independent of any differences in productivity or grant attainment by gender, negatively affect the careers of women scientists in the model? And will productivity based grant proposals have a disparate impact on women if implemented? Based on the model output, the answer to the first question is yes, and the answer to the second question is maybe, under some circumstances.

In both the parameter sweep and experiment, bias consistently lowered the average proportion of citations for women and lowered the percentage of women in the top 25% of performers, as shown in tables 2 and 5. This shows that women are affected negatively by gender bias in their discipline in terms of citations and being highly cited relative to male peers.

In the parameter sweep, bias also impacted the average proportion of both grants and publications, as shown in tables 3 and 4. Women were less likely to receive grants, and since grants lead directly to publications, also lowered the rates at which women published relative to men. Awarding grants to top performers further reduced the rates of grants and publications for women. The base parameterization of the model awarded grants to the top 10% of performers. When grants were given equally to all researchers, there was a positive effect on grant and publication proportions for women. The significant interactions between the highest level of bias and grants given equally to every researcher show that women were disadvantaged when grants were given to only the top 10%. This interaction was only seen at the highest level of bias, but lower levels of bias did not differ enough to constitute a main effect, so it is not surprising that there was no significant interaction at lower levels of bias, when there was no significant main effect, either.

Additionally, the difference between giving grants to the top 10% and top 25% at the highest level of bias did not have a significant interaction, meaning that women are disadvantaged at either threshold.[16] So, in the parameter sweep, this does support that women may be disadvantaged by productivity based grant funding, and provides some evidence against implementing a new grant funding system.

However, when using values taken to represent astronomy, the negative effect of bias for women and any interaction with giving highly cited researchers grants in terms of grants and publications disappears, while the effect of bias on citations and proportion of top performers persists. So, the evidence against implementing productivity based grants may be weaker than the parameter sweep indicates.

There may be other explanations for why this is the case. The initial conditions of the experiment seem to be driving the results. In particular, the relatively high number of initial publications coupled with a relatively low grant rate may have a somewhat protective effect on the outcomes for women. There is no bias in how citations are allotted when the model is generated. The rate at which women and men occur in the top 25% at the start of the model will be roughly equal. When the grant rate is only 20% and the model runs for 50 turns, each author will add an average of 10 publications over the course of the model, which is lower than what one would expect over an entire academic career.[17] So a large proportion of the total publications are generated in the initial stages of the model, and there are fewer publications generated later, when bias can occur. Many women who started the model in the top 25% will persist, even if there is a high level of bias, simply due to the size of their starting citations, and the relatively small effect of any additional biased citations.

As a result, the evidence from the model does not clearly indicate that women will be disadvantaged by citation-based grant funding. But the difference between the results of the

[16] Erosheva et al (2020) note that even small differences--differences of 0.3 points on a scale of 9--in the evaluation of candidates as measured by impact scores can affect grant success when grant rates are low, around 10% or 20% for the NIH grants they consider. While the NIH funding process is substantially more complex than the one implemented in the model, they show that what amount to small differences between candidates can result in disparate grant outcomes.

[17] This probably means that 50 rounds of the model are not adequate to represent what happens at say, retirement, but it may be enough to represent a career at time of tenure or promotion to full professor. Some of the possible ways to address this issue will be discussed in the next section.

parameter sweep and astronomy experiment do show some potentially important implications for such a policy.

First, the amount of citation bias makes a difference. In fields where women and men are cited at similar rates, even if there are small but significant differences, these disparities may not result in statistically significant differences over longer time scales and may not have a negative impact on women's careers. Lower productivity and lower grant attainment would also impact citation values in the real world, but the impact of gendered citation practices alone is not significant at low levels. If women and men have similar publications and grant success rates and there are only small differences in citation rates in a field, emphasizing citation metrics in grant decisions is less likely to disadvantage women substantially.

Additionally, having many early publications with no citation bias has a seemingly protective effect on future grant success when citations are used to determine grants. In astronomy, the bulk of publications were generated at the start of the model, and the function for citation bias was only used in subsequent rounds. The initial publications are meant to represent those written as postdocs or graduate students. In astronomy, publications with a large number of authors are the norm. The model treated every paper as single authored (assuming any co-authors would not be represented within the model). Depending on the authorship norms of astronomy, it may be that graduate students and postdocs are in a less prestigious authorship position, where their names are likely not generally used to determine whether to cite a paper or not[18]. So, it may be that in their early publications, any gender differences in citations are unlikely to appear. Presumably, as the researchers in the model age, they are more likely to move into prominent authorship positions, where gendered citation practices are more likely to be in effect. In fields with similar authorship norms of large, multi-author publications, this protective effect of many early publications somewhere in the authorship list may line up with assumptions of the model and lessen the impact of any later gender bias on grant success when citations are used.

Commented [3]: How do you think your work here extrapolotes to the case of biomedicine, where Ioannidis and Khoury made their suggestion?

Additionally, astronomers have a high degree of collaboration across domestic and international universities. Adams et al (2007) shows that almost 65% of astronomy papers published in major journals in 1999 included a research team with multiple affiliations within the US and 24.5% of papers included at least one internationally affiliated author. If these large, multi-authored papers often contain multiple researchers from the same cohort, spread out

[18] The model is agnostic as to why a researcher chooses to cite a paper written by a man instead of one written by a woman. The topic, publication venue, and relevance to the current paper all surely make a difference to the choice to cite a given paper in the 'real world'. But if the cause is explicit or implicit bias, it seems most plausible that it is directed at the researchers in the more prominent authorship positions.

across multiple domestic labs, that would violate the assumptions of the model. However, it is unclear if relaxing that assumption would substantially change the outcome and testing it would require a more complex model.

Conclusion

Future Directions for the model

This model is simple, and there are a number of ways to extend and improve the model.

The model assumes that all scientists and all papers are created equal. There are no differences between each scientist in the model beyond having been assigned a gender. However, scientists do vary in the quality of their work. Some papers are better and more informative than others and deserve to be read and cited more frequently. Some grant proposals are stronger than others and deserve to be funded over weaker ones. While there is no reason to think women or men are systematically better at science than one another, there is good reason to think that talent, access to resources, and luck are not distributed equally to all members of the community.

Deciding how scientific quality or importance is distributed across scientists, papers, or grant applications is, of course, a challenging empirical proposition. However, building any model requires making assumptions about agents, and there are certainly a variety of assumptions about how scientists differ that can be defended or backed with at least some empirical support. For example, NIH has data about the scores for each grant submitted for a particular grant type. With this data, it would be possible to find the shape of the distribution of grant scores, and to build into the model the process of assigning grant scores to each scientist based on this distribution. Additionally, a future version of the model could use the relationship between past and future grants to give more grants to past recipients.

The model currently does not allow more than one publication resulting from a single grant. This likely is not an accurate assumption for most fields; grants generally lead to many publications, especially if they are multiyear grants. Scientists also may publish papers that are not specifically tied to any particular grant, which this version of the model also does not represent. There are many possibilities for how to address this concern but allowing scientists to publish more frequently for each grant or providing a function to publish without earning a grant in each cycle would increase the accuracy of the model.

The model also only represents a small slice of the community. There are no papers that are not tied to a scientist in the model. Papers without authors in the community being modeled could be added to the model in the initial stage and in subsequent steps of the model that could also cite and be cited by the researchers. This would also make it easier to estimate what the average proportion of citations should be, since it would be more closely tied to the average number of citations per paper for a field. Further, the only members are those who enter the community at the same time; there are no pre-existing members, no one is added later, and no one leaves the model. Adding more scientists to the model would increase its complexity quite substantially but may also provide more accurate results.

While the model could be extended in any number of these ways, it has answered the primary questions of this paper. Citation bias lowers the proportion of citations, grants, and publications women earn over the course of the model. The model does not definitively show that women would or would not be disproportionately affected by citation-focused grant funding but does provide some evidence about the conditions under which it is more likely to negatively affect women.