Research Policy 48 (2019) 103838

Contents lists available at ScienceDirect

Research Policy

journal homepage: www.elsevier.com/locate/respol

Gender discrepancies in publication productivity of high-performing life science graduate students

Alexandra Graddy-Reed[a,*], Lauren Lanahan[b], Jonathan Eyer[c]

[a] University of Southern California, Price School of Public Policy, 650 Childs Way, Los Angeles, CA, 90089, USA
[b] University of Oregon, Lundquist College of Business, 1208 University St., Eugene, OR, 97403, USA
[c] University of Southern California, Price School of Public Policy, 650 Childs Way, Los Angeles, CA, 90089, USA

ARTICLE INFO

JEL classification:
H5
O3
O38
I25
J24
J16

Keywords:
R&D
Gender
Productivity
Life sciences
Graduate students

ABSTRACT

Despite equal matriculation into life science graduate programs, the gender gap persists for later-stage professional outcomes. To understand this divergence, we examine graduate training and use the competitive NSF Graduate Research Fellowship Program to identify high-quality life science students that are awardees and honorable mentions. We use a differencing research design to estimate the relative difference of the R&D award across gender on publication trajectory. The results of the triple difference estimation show a negative effect for women compared to men from the award. We investigate the driver of this effect by examining trends within gender and find a large, positive effect of the award for men but fail to find such evidence for female awardees. Our results indicate different signaling effects across gender even though the funding is meritocratic.

1. Introduction

"*The formative pre-doctoral years are a critical window, because students' experiences at this juncture shape both their beliefs about their own abilities and subsequent persistence in science.*" – Moss-Racusin et al., 2012: *16475*

In response to the widely documented gender gap in science and engineering (S&E) fields (e.g. Moss-Racusin et al., 2012), policymakers have implemented educational policies – ranging from primary to higher education – to increase female enrollment in S&E training and reduce workforce attrition.[1] Regarding enrollment, these policies appear effective with recent evidence reporting that female matriculation into S&E higher education programs is increasing (e.g. Carrell et al., 2010; Miller and Wai, 2015; Van Arensbergen et al., 2012). In many S&E fields, women now exceed the number of men in both undergraduate and graduate program enrollment (Sugimoto et al., 2015). This varies across academic disciplines with women accounting for half or even the majority in certain fields – including life sciences, psychology, and social sciences.[2]

However, the most recent U.S. National Science Board's S&E Indicator Report notes large discrepancies remain at later career stages. Women hold less than 30 percent of S&E positions in the workforce,[3] and they notably lag in research-related appointments in academia and industry (Moss-Racusin et al., 2012; Van der Lee and Ellemers, 2015; Ceci et al., 2014; Sugimoto et al., 2015). Importantly, this discrepancy appears to be a function of productivity differences in publication activity (Lerchenmueller and Sorenson, 2018).

Given that graduate training serves as a critical bridge between program matriculation and professional placement, there may be differences in graduate training for men and women that account for the diverging trends at later career stages. Prior studies have found evidence of gender separation in graduate training with women serving on smaller research teams (Buffington et al., 2016) and facing implicit bias

* Corresponding author.
E-mail addresses: graddyre@price.usc.edu (A. Graddy-Reed), llanahan@uoregon.edu (L. Lanahan), jeyer@price.usc.edu (J. Eyer).

[1] http://unesdoc.unesco.org/images/0024/002457/245717e.pdf; https://www.nsf.gov/statistics/2017/nsf17310/

[2] National Science Board (NSB): Chapter 2 pg. 2-58 (Appendix Table 2-24). Despite these trends, they consistently lag in other fields that include engineering, computer sciences, and physical sciences.

[3] NSB: Fig. 3-27 (https://www.nsf.gov/statistics/2016/nsb20161/uploads/1/6/chapter-3.pdf)

https://doi.org/10.1016/j.respol.2019.103838
Received 26 November 2018; Received in revised form 18 July 2019; Accepted 30 July 2019

discrimination from senior researchers over their lab assignments (Moss-Racusin et al., 2012). Female graduate students also are not granted the same research collaboration opportunities with advisors as their male counterparts (Pezzoni et al., 2016). Research at this juncture of professional training, however, is scant; more work remains to understand whether and how the gender gap persists at this early career stage.

To address this issue, we document variation in the response to signaling associated with federal funding that leads to divergent trends in research productivity by gender. Specifically, we draw upon the U.S. National Science Foundation's (NSF) prestigious Graduate Research Fellowship Program (GRFP) to identify a sample of life science graduate students with demonstrated research potential. We leverage variation in federal research and development (R&D) funding allocation between groups of high-quality awardees and honorable mentions. Using this variation in funding allocation, we examine productivity trends. We employ a differencing research design to estimate the relative difference of the effect of the award for women on publication productivity.

We document that the award's impact takes effect five years following the GRFP allocation; this is at the time when many students are transitioning from graduate school to their first professional placement. The results of the triple difference estimation show a negative effect for women compared to men from the award. When we investigate the driver of this effect by examining trends within gender, we find a large, positive effect of the award for men but fail to find such evidence for female awardees. The differential effect between the productivity trajectory of female awardees and honorable mentions is nominal in size and statistically insignificant. Further, female awardees underperform both male awardees and male honorable mentions.

We explore a series of mechanisms that may moderate this effect and find preliminary evidence that women engage in different training compared to men that disadvantages them professionally. Graduate advisors collaborate disproportionately less with female awardees. Specifically, highly cited male advisors publish more with their male students than their female students. Further, they also publish more with their female honorable mention students than with their female awardee students. Female advisors producing lower-impact research also publish more with female honorable mentions and male awardees than with female awardees. Given the collaborative nature of this field, this suggests that women do not receive the same boost as male students when receiving the competitive award. Our results indicate that the signal of the award, and the opportunity it confers, varies by gender.

This study provides evidence of differing perceptions of signals along observable characteristics. These impacts are long lasting and have important implications for professional trajectory and the production of science. We advance the growing scholarship on the gender gap by redirecting attention to an earlier career stage during graduate training. Moreover, we show that biases can result in worse outcomes for disadvantaged groups even when policies are meritorious.

2. The differential value of federal R&D funding

In this analysis, we examine the impact of competitive, federal R&D allocation on research production among early career scientists across gender. For all S&E researchers, resource allocation is critical to the rising costs of research production (Stephan, 2012). In addition, securing federal peer-reviewed funding provides a positive signal of the quality of proposed research. This signal may be especially valuable to graduate students as they progress through training. With scientific research becoming increasingly collaborative (e.g. Jones et al., 2008; Wuchty et al., 2007), procurement of an award can provide important network and training opportunities such as placement within a senior faculty member's lab and delegation of tasks.

We argue, however, that the value of federal R&D award varies across gender. Implicit bias has been widely shown to impact student performance and employment outcomes across various dimensions including race and ethnicity (Bertrand et al., 2004; Oreopoulos, 2011) and economic status (Hanna and Linden, 2012). In response, government bodies have created policies to improve equity and combat such perceptions. However, in some cases, the policies themselves exacerbate the inequality they aim to reduce.[4] We argue this bias may also occur with signals from meritocratic policies, such as peer-reviewed R&D funding. For graduate students, this signal variation may contribute to the discrepancies across gender in research production found at later career stages.

2.1. Bias in federal R&D funding – considerations across gender

To assess the presence of signal variation, we examine competitive, federal R&D funding. As discussed, peer-reviewed funding should improve a researcher's productivity both through the provision of financial resources and through a positive quality signal to the larger research community. In this example, male and female graduate students who obtain an award, receive both resources and signals that affect productivity. Alternately, students just shy of the cutoff may receive an honorable mention, which provides a smaller quality signal benefit yet no resource allocation. We assume that the benefit from the financial allocation does not vary across gender. However, we argue the additional quality signals do vary across gender.

We first assess how researcher quality may vary in the absence of the award. If men and women are perceived to have equal baseline quality distributions, we would not expect variation across gender. There is evidence among early-childhood studies to show that intellectual ability – specifically, spatial and mathematical reasoning – is comparable across gender (e.g. Miller and Halpern, 2014). The results from an extended life-course analysis finds compelling evidence that gender discrepancies do not stem from biological differences (Ceci et al., 2014: 75). Based on intellectual capabilities alone, this suggests there would not be a divergence in research productivity across gender.

However, some have directly challenged this assumption. Lawrence Summers, former President of Harvard University, contentiously hypothesized there is "different availability of aptitude at the high end" between male and female scientists.[5] Although many publicly denounced this remark, it underlines the prevalence of both implicit and explicit bias for women in S&E fields that likely accounts for divergence in performance (Barres, 2006). There is compelling evidence to show that environmental factors account for broad discrepancies in performance across gender. Prior studies have found that women face implicit bias discrimination from senior researchers over their lab assignments (Moss-Racusin et al., 2012); female graduate students are not granted the same research collaboration opportunities with advisors as their

[4] For example, recent work on Ban the Box policies provide evidence that minority job applicants receive fewer responses as employers try to guess which applicants have a criminal record (Agan and Starr, 2017; Doleac and Hansen, 2016). Ban the Box refers to policies that prohibit employers from asking applicants if they have a criminal record.

[5] Summers first presented this argument on January 14, 2005 at the Conference on Diversifying the S&E Workforce sponsored by the National Bureau of Economic Research in Cambridge, MA. He goes on to say the following: "So my best guess, to provoke you, of what's behind all of this is that the largest phenomenon, by far, is the general clash between people's legitimate family desires and employers' current desire for high power and high intensity, that in the special case of science and engineering, there are issues of intrinsic aptitude, and particularly of the variability of aptitude, and that those considerations are reinforced by what are in fact lesser factors involving socialization and continuing discrimination." Harvard has since removed the transcript from their website but it was archived and is accessible via the following web address through the Way Back Machine: https://web.archive.org/web/20110823225044/https://www.harvard.edu/president/speeches/summers_2005/nber.php

male counterparts (Pezzoni et al., 2016); and female researchers face bias in recognition of publication contributions with collaborative projects (Lerchenmueller and Sorenson, 2018; Sarsons, 2017). These external factors likely overshadow individual ability and reduce production for women; thus, in the absence of the award, we anticipate men will be more productive than women.

Next, we assess how the signal may vary in response to a competitive, federal R&D allocation. Importantly, the funding source serves a critical role in defining the value of this signal (Azoulay et al., 2011). We consider federal programs, which serve as the largest source of R&D within higher education. Within this context, panelists examine both the intellectual merits and broader impacts of the proposal.[6] The former provides a signal of prestige and legitimacy (Stephan, 2012; Suchman, 1995). This is substantiated by the broadly held placement and promotion standards in academia and industry that value grant activity as a metric of research excellence. Theoretically, this activity influences external perceptions of quality for both the project and Principal Investigator (PI) (Merton, 1968; Podolny, 1993). This is particularly salient among graduate students that have higher levels of uncertainty given their early career status.

The latter – broader impacts – explicitly addresses equity issues to encourage a broad representation of the scientific workforce (Fealing et al., 2015). The federal grant may be viewed in part as redistributive where women and under-represented minorities receive the award based on their demographic rather than solely on the merit of the project (DiTomaso et al., 2007a,b; McNeely and Fealing, 2018). For example, Heilman et al (1997) and Garcia et al (1981) find that when redistributive policies are salient, managers and reviewers view minority candidates as less qualified even after controlling for signals of true quality. Quadlin (2018) further shows that hiring preferences differ between male and female candidates, with males prized for competence and commitment while women are evaluated on personality and likability. Contrary to initial expectation, this increases the likelihood that a woman with moderate quality is hired but decreases the likelihood for highly qualified females.

Based on this logic, for men, procurement of R&D funding not only yields access to tangible resources, it provides legitimacy for the line of research and helps the students define their research identity. These likely offer longer-term reputational benefits, which may improve access to broader networks and research opportunities. For women, while they also tangibly receive monetary resources, the federal award may not solely be viewed with the same signal of prestige and legitimacy as conferred for men. Due to the broader impacts criterion, we anticipate the signal of the award for men is greater than the signal for women. This would result in greater productivity for men with the award than female awardees.

2.2. Bias in federal R&D funding – considerations within gender

We also assess the impact of competitive, federal R&D allocation on research production within gender. The logic for men is relatively straightforward. Controlling for research ability, both receipt of the resource allocation and the signal of prestige and legitimacy will likely lead to increased production in contrast to men without the award. We substantiate this with prior scholarship, particularly within the life sciences, that has found evidence of a positive effect of R&D funding on production (Jacob and Lefgren, 2011a,b; Azoulay et al., 2011; Graddy-Reed et al., 2018).

We apply the same logic for women but argue that the differential effect of the award will be less. We attribute this to the fact that all women in S&E face hurdles. These additional barriers will dampen the impact of the award. Evidence of systemic bias spans manuscript review, hiring, and promotion (Ceci and Williams, 2011; Lerchenmueller and Sorenson, 2018). Controlling for research ability, we anticipate that women with the award will produce more than their female counterparts without, given the additional benefits of the monetary allocation and prestige of the award. However, we anticipate this differential among women will be smaller than the differential impact among men. Thus, all else equal, we hypothesize that competitive, federal R&D will increase productivity for men more than for women.

3. Empirical context

The U.S. National Science Foundation's (NSF) Graduate Research Fellowship Program (GRFP) defines the context for this study. This prestigious program has awarded S&E graduate students since 1952 with notable recipients that include Sergey Brin (co-founder of Google), Maxine Singer (recipient of the Vannevar Bush Award), Nina Federoff (recipient of the U.S. National Medal of Science), and Eric Cornell (Physics Nobel Laureate). As of 2018, NSF offers $138,000; this is apportioned as a $34,000 annual stipend and $12,000 annual educational allowance to the recipient's institution for three years. For comparison, NSF offered $69,000 in 1995 and $121,500 in 2005 for the full award, with $14,400 (1995) and $30,000 (2005) for the annual stipend and $8600 (1995) and $10,500 (2005) for the annual educational allowance.[7]

The eligibility requirement uniquely defines this program and makes it salient for this study. Students are eligible to apply as the PI during the first 12 months of their S&E-based graduate training.[8] The solicitation requirements are tailored to this graduate student population by requesting the following: (i) two-page research statement; (ii) three-page personal statement; (iii) three to five reference letters; and (iv) higher education academic transcripts. The two-page research statement stands in contrast to standard NSF programs that require more detailed and developed 15-page project descriptions. Moreover, the additional three components are unique to this program given the early career stage of the graduate student and heightened level of uncertainty concerning their research potential.

This program follows the standard NSF single-blind merit review process. Senior scholars with expertise on the proposal topic review both the intellectual merits and broader impacts.[9] Unique to this program, given the heightened uncertainty due to the early career stage of the applicant base, reviewers identify applicants with "demonstrated potential for significant research achievements in STEM."[10] Senior scholars serving as ad hoc reviewers explicitly distinguish the top 20 percent that are meritorious from the rest.[11] Then among this set, NSF panelists sort the applicants into awardees and honorable mentions (Freeman et al., 2009).

Further, NSF publicizes both GRFP awardees and honorable mentions; the latter group includes competitive applicants just shy of receiving the federal funding. Although NSF generally restricts application data, this program provides the exception.[12] While honorable mentions do not receive the financial benefits of the GRFP award, NSF publicizes this recognition as a signal of merit purporting it as a "significant national academic achievement."[13] With data on the

[6] NSF:https://www.nsf.gov/bfa/dias/policy/merit_review/facts.jsp#1; NIH: https://grants.nih.gov/grants/peer/guidelines_general/Review_Criteria_at_a_glance.pdf

[7] Retrieved September 3, 2018 from GRFP Program Solicitation, NSF 18-573

[8] PIs must be U.S. citizens or permanent residents.

[9] The timeframe for this study is 1995 – 2005. In 1997, NSF added broader impacts as a second merit criterion.

[10] https://www.nsf.gov/pubs/2016/nsf16588/nsf16588.htm

[11] Retrieved December 21, 2016 from: https://www.nsf.gov/od/oia/programs/epscor/GRFP_Webinar.pdf.

[12] While NSF reports applicant data for awardees and honorable mentions for this program, data on the larger set of applicants that did not receive either of these acknowledgements are restricted. Thus, we are unable to assess the range of ability across the full sample of applicants.

population of GRFP award recipients and honorable mentions, not only are we able to draw upon graduate students with comparable research potential across a range of U.S. institutions,[14] this also serves as the basis for the research design.

4. Sample & data

For this study, we focus on the life sciences. This offers a baseline for understanding gender discrepancies where initial enrollment is at parity. Placed within the context of the GRFP, we utilize an individual-level panel dataset of awardees and honorable mentions to examine whether a gender gap persists in research productivity.

We trace graduate students over a standard 16-year timeframe – spanning the five years preceding the GRFP event to the ten years following.[15] To construct this dataset, we selected a sample of GRFP awardees and honorable mentions in the life sciences and then gathered annual metrics on graduate training, professional placement, and research production. Rather than surveying GRFP applicants, which presents potential tradeoffs of recall bias and low response rates (Clarke et al., 2008), we used a series of third-party data sources to triangulate data. We secured complete information for 76 percent of the sample; this exceeds response rates from standard survey designs (Baruch and Holtom, 2008; Anseel et al., 2010). We detail the GRFP sampling and data triangulation efforts in turn.

4.1. GRFP sampling

First, with an interest in tracing these individuals over their standard 16-year timeframe, we selected students that participated in the program between 1995 and 2005. The earlier year denotes the first year that GRFP honorable mention data was publicly available, while the latter allows for a ten-year timeframe to trace these individuals following the GRFP. We then restricted the sample to the life sciences. We selected the three most prominent fields in terms of GRFP activity to control for disciplinary variation within the life sciences division. These include ecology & evolutionary biology (Ecology); biochemistry, biophysics, & structural biology (Biochemistry); and biology, integrated biology, integrated biomedical sciences, & kinesiology (Biology), which comprise 21.32, 20.54, and 12.36 percent of the division's activity, respectively. Appendix Table A1 provides summary statistics by award conferment status for these three fields in reference both to the life science division and the entire GRFP program.

With the intention of drawing upon third party data sources to augment the GRFP dataset, we further restricted the sample by removing graduate students with common last names – specifically, those with a last name that appeared more than five times in the GRFP sample. This approach follows prior studies that are reliant on third party data sources to ease the process of verifying whether the correct individual is being tracked both over time and across sources (Jacob and Lefgren, 2011a). Disambiguating individuals with common names is particularly challenging given that we are tracing a sample within a narrow scope of academic fields.[16] Although, this technique disproportionately removed students with Asian surnames, this as a defensible tradeoff to reduce the presence of false positives.

Finally, we separated the lists of awardees and honorable mentions respectively into the three field-based bins (Ecology, Biochemistry, and Biology) for a total of six bins. Within each bin, we randomly sampled without replacement 150 unique GRFP proposals yielding 900 proposal observations. We defined this number to ensure sufficient statistical power for regression analysis.[17] Of the 900 proposals sampled, 23 students appeared in the dataset twice. This occurred when an individual initially received honorable mention recognition and then an award or honorable mention in the following academic year. In these instances, we recorded the most recent record for the student. Taken together, we identified 877 unique graduate students from the GRFP sample; this defines the full sample.

4.2. Data construction

The NSF GRFP database provides detail on the student's name, year of submission, baccalaureate institution, graduate institution, and field of research. We then tracked the sample of 877 individuals across a series of third-party data sources for additional data on graduate training, professional placement, and publication activity. First, we matched the student in ProQuest, the largest central repository of dissertations and theses. This source provided detail on the student's advisor and graduation. In all, we retrieved data from ProQuest for 697 graduate students (79 percent of the full sample). Second, we conducted systematic online searches to identify the student's first professional placement following graduation. This excludes temporary summer internships. We relied primarily on professional pages on LinkedIn, institutional websites, and personal websites. We retrieved professional placement data for 800 students (91 percent). We relied on the student's name to determine gender; however, we confirmed this assignment based on information available from online searches.

Third, we matched the student to the Scopus bibliometric database to trace publication activity. This data source, a subsidiary of Elsevier, houses the largest collection of bibliometric data. We matched 672 students (77 percent) that published over their respective standardized 16-year timeframe. We assume the set of students without records in Scopus did not publish over this timeframe. Among the matched set, we gathered data on 5900 peer-reviewed academic publications. We rely on peer-reviewed journal articles – a standard bibliometric measure – to construct the outcome variable (Adams and Griliches, 1996; Jacob and Lefgren, 2011a). Fourth, given the importance of the student-advisor relationship (Gaule and Piacentini, 2018; Bettinger and Long, 2005), we expanded our data on the student's advisor to include their gender and research productivity. Regarding the latter, we relied on Google Scholar to trace annual citation activity of the student's advisor over their respective graduate student's 16-year timeframe. We confirmed the student's advisor for 778 graduate students (89 percent) with citation data for 427 advisors (49 percent).

Finally, we relied on three additional data sources for detail on the graduate training environment. We matched the student's graduate program to the National Research Council's (NRC) 2010 Survey of Doctoral Programs to gather additional detail on program quality and demographics. This survey reports program-level detail from 2000 to 2006, which overlaps with the second half of the timeframe for this

[13] Page 7: https://www.nsf.gov/pubs/2016/nsf16588/nsf16588.pdf

[14] For the baseline DDD analysis, we rely on data from 108 U.S. institutions from 43 states.

[15] To illustrate, for students that applied for the GRFP in 1995, their 16-year timeframe extends from 1990 – 2005, while someone who applied in 2005 would have a timeframe from 2000 – 2015. Altogether, we draw upon research activity over a 25-year period between 1990 – 2015 for students that received GRFP acknowledgment between 1995 and 2005. Section 3.1 provides more detail.

[16] The 30 most common names in the GRFP database are as follows (we indicate common names of Asian heritage by *italics*): Smith; Lee; Johnson; Brown; Miller; *Chen*; Williams; Anderson; Davis; Jones; *Wang*; Thompson; Wilson; *Chang*; Moore; Green; Thomas; *Kim*; Young; *Liu*; Martin; Jackson; Nelson; *Wong*; Lewis; *Lin*; Roberts; Allen; Evans; and *Yang*. Of these common names, 21.6 percent are of Asian heritage. This level exceeds the national Asian-American population in 2000 (U.S. Census), the middle year of our sample (4.2 percent).

[17] We ran power tests to ensure the sample size was sufficiently large to estimate significance tests. The minimum size for difference in means test with data in wide form is 265 and for regression analysis with data in long form is 6,000. The sample used for the primary analysis yielded 564 observations in wide form and 8,460 observations in long form.

study. The NSF Higher Education Research and Development (HERD) survey provided annual university-division-level data on federal R&D financing. In addition, we matched the student's graduate institution to the National Center for Education Statistics (NCES) Integrated Postsecondary Education Data System (IPEDS) for institutional governance. Altogether, we retrieved data from the three sources as follows: NRC matched to 754, HERD matched to 767, IPEDS matched to 765 graduate students.

We gathered complete data across all sources (except for advisor citations) for 667 graduate students (76 percent of the full sample). This defines the full complete case (CCA) sample. This includes complete information on 360 awardees and 307 honorable mentions of which, 336 are female and 331 are male. Within gender, 52 percent of males are awardees compared to 56 percent of females. This distribution of awards and honorable mentions by gender suggests that the federal program mirrors the matriculation rates of graduate enrollment;[18] in other words, the gender gap is not apparent at this early stage of award assignment.

5. Methods

5.1. Differencing research design

To estimate the effect of gender discrepancies on research performance during early career training we employ a differencing research design that includes both triple differences and difference-in-differences models (Imbens and Wooldridge, 2007; Davidoff et al., 2005). Broadly, this is a quasi-experimental design where we draw upon variation across gender (female verses male), GRFP funding allocation (awardee verses honorable mention), and time (pre- verses post-GRFP acknowledgement).[19] For each graduate student in the sample, we define a standardized 16-year timeframe by the five years prior to the GRFP proposal submission event ($-5 \le t \le -1$), the GRFP year – the sixth year ($t = 0$), and the following 10-year post period ($1 \le t \le 10$). For the outcome measure, we focus on research performance as measured by total publications and new publications. For the latter, we compute annual measures of the three-year moving average. Both measures offer unique, yet complementary insight on publication activity.

Eq. (1) represents the triple difference model (DDD). It is an individual-year-level model where i denotes the student and t denotes the annual period. $Award_i$ is coded one for graduate students that receive the GRFP award; $Post_t$ is coded 1 for the years following the GRFP proposal review[20] ; and $Female_i$ is coded 1 for women. We drop the GRFP year ($t = 0$) to ensure a clear cutoff between the pre and post periods. The triple interaction term – $Award_i * Post_t * Female_i$ – estimates the relative difference of award status across gender on research productivity.

$$Y_{it} = \beta_0 + \beta_1 Award_i + \beta_2 Post_t + \beta_3 Female_i + \beta_4 Award_i{*}Post_t + \beta_5 Award_i{*}Female_i + \beta_6 Post_t{*}Female_i + \beta_7 Award_i{*}Post_t{*}Female_i + Controls + \varepsilon_{it} \tag{1}$$

Eqs. (2a) and (2b) represent a series of stratified difference-in-difference models (DD). We stratify the sample by award designations and gender to tease apart how each sub-group drives the net effect estimated by the triple difference in Eq. (1). Eq. (2a) stratifies by *award* designation to estimate the relative difference across gender on research productivity. In other words, we estimate the impact of gender in the post period for sub-samples of awardees and honorable mentions. Eq. (2b), conversely, stratifies by *gender* to estimate the relative difference across award designation on research productivity. Here, we estimate the impact of the award in the post period for the sub-samples of women and men. We also estimate a differencing model between female awardees and male honorable mentions. As with the DDD model, we drop the GRFP year ($t = 0$) to ensure a clear cutoff between the pre- and post-periods.

$$Y_{it} = \beta_0 + \beta_1 Female_i + \beta_2 Post_t + \beta_3 Female_i{*}Post_t + Controls + \varepsilon_{it} \tag{2a}$$

$$Y_{it} = \beta_0 + \beta_1 Award_i + \beta_2 Post_t + \beta_3 Award_i{*}Post_t + Controls + \varepsilon_{it} \tag{2b}$$

For all models, we include a vector of individual- and institutional-level controls. At the division level, this includes continuous measures for faculty research productivity, the rate of female faculty members, prior GRFP activity, and the average GRE scores for graduate students. We also include a set of indicators: student-advisor gender match; prior publications; a set of dummies for the graduate program size, rank, and academic field; graduate and baccalaureate governance structures (public verses private); and graduate institution flagship and/or land grant institution. Lastly, we include an indicator for the calendar year the student submitted their GRFP proposal to control for annual macroeconomic trends, as well as funding conferment changes within the GRFP.

5.2. Timing – threshold model

In estimating the production function on the outcomes of total and new peer-reviewed, academic publications, it is important to consider dynamic features of the publication process. While the GRFP event takes place at year zero ($t = 0$), producing peer-reviewed articles takes multiple years – from the initial inception of the idea, to data construction and analysis, to journal submission and (very likely) revisions through peer-review, to ultimate publication (Powell, 2016). Rather than assume that the award will affect publication activity in the following year ($t = 1$), we estimate a threshold model to account for *when* the effect begins. Operationally, a threshold model allows coefficient values to be different on either side of a cutoff in a running variable (i.e. time). This approach estimates the value of the running variable at which the coefficient values switch (Tong, 2012). We confirm this with a machine learning approach to estimate the effect year. We randomly assign half of the individuals to a training set and estimate Eq. 1 on each possible year that the GRFP signal could come into effect ($t = 0$ to $t = 9$). We select the best fitting model based on the R-squared value and apply this model to the other half of the data. We then estimate the differencing models – both DDD and stratified DDs – using the threshold year to designate the post period.

5.3. Coarsened exact matching

As a preliminary diagnostic of the differencing research design, we report annual cumulative and new publication activity for the full sample (Appendix Figure A1, Panels A and B, respectively). We illustrate this with four trends lines for the corresponding groups: female awardees, female honorable mentions, male awardees, and male honorable mentions. We label the standardized 16-year timeframe on the x-axis to reflect the five years leading up to the GRFP event ($t = 0$) and the 10 years following.

The trend lines prior to the GRFP are relatively flat with male awardees demonstrating slightly increased levels of publication activity. Female awardees, female honorable mentions, and male honorable mentions have flatter, lower pre-trends. When examining the

[18] In 2000, male enrollment in U.S. Life Sciences graduate programs was 47 percent. Source: NSF, National Center for Science and Engineering Statistics, special tabulations (2014) of the 2013 Survey of Graduate Students and Postdoctorates in Science and Engineering.

[19] While a regression discontinuity design (RDD) would be ideal in this model, we are unable to operationalize it. Not only does NSF restrict access to application scores, NSF's merit review scoring relies on ordinal ranking rather than continuous measures. (https://www.nsf.gov/bfa/dias/policy/merit_review/#review)

[20] As we discuss in sections 4.2 and 5.1, we adjust *Post* based on the results from a threshold model to account for *when* the effect begins.

differences in the pre-proposal period, male awardees (0.45 publications) and female honorable mentions (0.30 publications) are both each statistically higher than female awardees (0.21 publications). The difference in distributions (Kolmogorov-Smirnov test), however, are not statistically significant. The diverging trajectory in the pre-period suggests differential pre-trends between the sub-groups. This introduces bias to our research design by violating the assumption of parallel pre-trends. We surmise this divergence is partially attributable to non-random award assignment.[21]

To improve validity, we employ a coarsening exact matching (CEM) procedure to achieve balance along observable factors in the pre-period across these groups (Blackwell et al., 2009; Iacus et al., 2012). Given our central research question, female awardees define the treated group and the remaining sub-groups (female honorable mentions, male awardees, and male honorable mentions) define the control. We rely on the following pre-treatment indicators to coarsen the full sample: any publications prior to GRFP, graduate program research rank, graduate program life science field, average GRE scores for the graduate program, graduate institution flagship, and baccalaureate institution type. This set of variables accounts for individual-, program-, and university-level factors. Prior research has found that these factors affect graduate student productivity and their subsequent research trajectory (Graddy-Reed et al., 2017). We directly match across the set of pretreatment measures between the two groups to achieve balance.[22] This reduces the sample to 707 students – 213 Female Awardees and 494 across the control subgroups. Altogether, this accounts for 81% of the full sample (91 percent of the treated and 78 percent of the control) and allows us to estimate the sample average treatment effect.

5.4. Trend diagnostics for coarsened sample

We report the annual cumulative and new publication activity for the coarsened sample (Fig. 1, Panels A and B, respectively). In contrast to Appendix Figure A1, the pre-trends across all four sub-groups exhibit equivalent trends. In the years following the GRFP event, male awardee productivity increases at a greater rate in contrast to the other groups. Male honorable mentions also deviate from both female post trend lines demonstrating higher publication rates. Both sets of female groups have comparably positive, yet flatter trends following the GRFP event. When considering the pre- and post-trends, the annual average total publication count differs both by gender and award designation, thus substantiating the utility of the differencing research design.

At the cost of an efficiency loss, we improve balance in the pre-period using the coarsening technique. We argue this is a necessary tradeoff and rely on the coarsened sample for analysis. We provide additional detail on the coarsening procedure in Appendix B and report sensitivity checks in Appendix Table B3.

5.5. Descriptive statistics

Table 1 reports the descriptive statistics for the coarsened sample with complete data for 564 PIs.[23] We present the statistics in five columns; column 1 reports for the coarsened sample. Column 2 presents the treatment group of female awardees while columns 3, 4, and 5 present statistics for the control groups of female honorable mentions, male awardees, and male honorable mentions, respectively. For comparison, we present the descriptive statistics for the full sample with complete data (full CCA sample) in Table A2 in the Appendix.

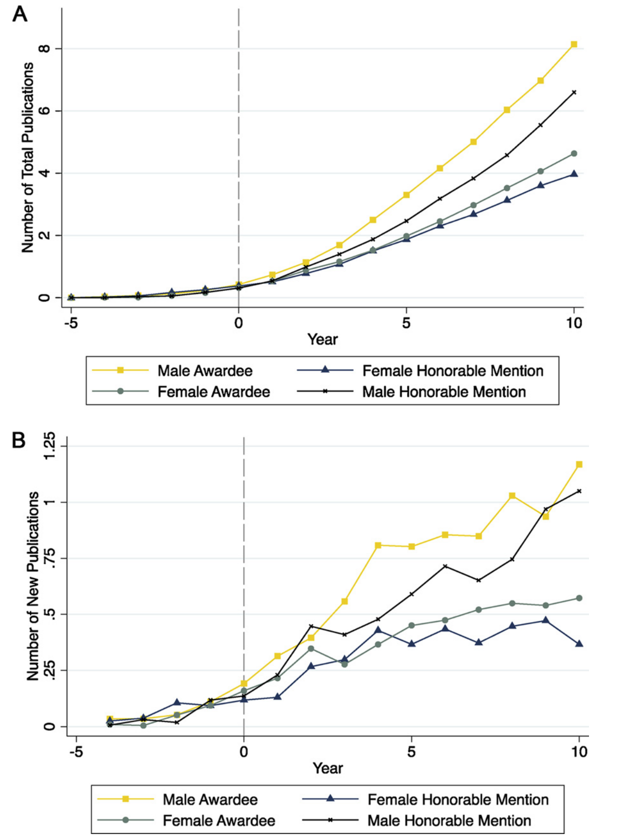

Fig. 1. Coarsened Sample Annual Average Publications by Gender & Award Status (N = 707).
Panel A - Total Publications
Panel B - New Publications

By construction of the coarsening procedure, female awardees exhibit comparable trends with the three other groups for the following: any publications prior to GRFP, graduate program research rank, graduate program life science field, GRE scores for the graduate program, graduate institution flagship, and the baccalaureate institution type. However, it is worth noting that male awardees are less likely to attend a public institution for their baccalaureate training.

When considering post period activity, females are less likely to have any publications – with female honorable mentions exhibiting the lowest rate (79 percent). Moreover, female awardees produce 5.08 publications by 10 years after the GRFP; female honorable mentions produce 4.77 compared to male awardees and honorable mentions with 8.9 and 7.03 publications, respectively. As for new publications – computed as the three-year moving average, we find similar trends with male awardees (0.59) and honorable mentions (0.46) leading female activity (0.33 for awardees and 0.34 for honorable mentions).

In terms of the gender match between the student and their advisor, over three-quarters of the men work with male advisors, while only one quarter of women work with a female advisor. It takes approximately 5.4 years to complete the PhD with comparable rates across each subsample. However, female honorable mentions have the lowest rate of completing a PhD with just 90 percent of the sample compared to 96 percent of female awardees, 95 percent of male awardees, and 98 percent of male honorable mentions. In addition, female awardees and male honorable mentions have the lowest rates of being in an academic position following degree completion (61 and 60 percent respectively compared to 67 percent for female honorable mentions and 70 percent

[21] An NSF panel determines awards via a single-blind review process. With that said, the level of uncertainty with the GRFP panel review is greater than more standard NSF programs given the early career stage of the student-applicant and the abbreviated format of the proposal. The unique programmatic structure alleviates some of this concern.

[22] We construct strata of observations with statistically indistinguishable values between the two groups and then subsequently coarsen the sample to ensure balance.

[23] This accounts for 80 percent of the coarsened sample (707 students); 84 percent of the full CCA sample (667); and 64 percent of the full sample (877).

Table 1
Coarsened CCA Sample Descriptive Statistics.

	Full Sample	Female Awardee	Female Honorable Mention	Male Awardee	Male Honorable Mention
Publication Activity					
Proportion that Ever Publish	0.84	0.83	0.79	0.86	0.85
Proportion with Publications prior to GRFP	0.12	0.11	0.12	0.13	0.10
Total Publications by GRFP (0 - 14)	0.22	0.15	0.30	0.27	0.17
	(0.98)	(0.48)	(1.49)	(1.18)	(0.58)
Three Year Moving Average of Annual	0.43	0.33	0.34	0.59	0.46
Publications (0 - 6.7)	(0.70)	(0.56)	(0.61)	(0.85)	(0.71)
Total Publications 10 years after GRFP (0 - 40)	6.45	5.08	4.77	8.90	7.03
	(6.30)	(4.97)	(5.51)	(7.53)	(6.16)
PI Characteristics					
Baccalaureate Institution: Public	0.45	0.46	0.47	0.35	0.52
Baccalaureate Institution: Liberal Arts College	0.16	0.21	0.16	0.13	0.14
Baccalaureate Institution: R1 Research University	0.76	0.69	0.77	0.82	0.78
Field: Biochemistry	0.36	0.33	0.31	0.36	0.45
Field: Ecology	0.31	0.31	0.34	0.32	0.25
Field: Biology	0.33	0.35	0.35	0.32	0.30
Gender Match with Advisor	0.52	0.24	0.29	0.82	0.79
Completed a PhD	0.95	0.96	0.90	0.95	0.98
Time from Proposal to Degree (0 - 12, N = 558)	5.43	5.62	5.23	5.27	5.50
	(1.62)	(1.53)	(1.78)	(1.42)	(1.76)
Initial Placement in Research (N = 540)	0.82	0.78	0.83	0.86	0.83
Initial Placement in Academy (N = 540)	0.64	0.61	0.67	0.70	0.60
Initial Placement in Post-Doctoral Fellowship	0.54	0.51	0.46	0.61	0.56
Graduate Program Characteristics					
Public Institution	0.52	0.57	0.50	0.44	0.54
Flagship Institution	0.27	0.33	0.29	0.22	0.25
Land Grant Institution	0.27	0.30	0.27	0.24	0.24
Annual Average Publications Per Faculty	2.27	2.27	2.27	2.24	2.32
(0.76 - 3.6)	(0.55)	(0.52)	(0.57)	(0.54)	(0.59)
Proportion of Female Faculty (0.1 - 0.45)	0.24	0.24	0.25	0.24	0.24
	(0.05)	(0.04)	(0.05)	(0.04)	(0.05)
Two-year Prior GRFP Activity (0 - 24.8)	5.96	5.66	5.33	6.48	6.33
	(4.88)	(4.65)	(4.17)	(5.06)	(5.49)
Average GRE Math Score (646 - 784)	731	729	730	733	734
	(28.12)	(27.51)	(27.11)	(29.13)	(28.67)
Program Rank in Top Tercile	0.90	0.90	0.90	0.90	0.90
Observations	564	175	115	146	128

Note: Table presents means with standard deviations in parentheses or proportions.

for male awardees). Male awardees have a higher rate of placing in a post-doctoral position compared to the other three groups and a slightly higher rate of placing in a research position more broadly. In the life sciences, post-doctoral positions commonly precede moving on to tenure-track positions (Sauermann and Roach, 2012; Lerchenmueller and Sorenson, 2018). These trends – particularly post-trend metrics – are roughly comparable for the full CCA sample (N 667) as shown in Appendix Table A2.

6. Results

6.1. Threshold analysis

First, we report the results from the threshold analysis; this confirms when the effect of the award differs between the pre and post period. The results, presented in Appendix Table A3, indicate this to be five years after the GRFP (t = 5) for the coarsened sample – the time when many students are graduating and transitioning to their first professional placement. We further confirm the treatment effect year with the machine learning estimation and find five years post GRFP (t = 5) best fits the training data for the coarsened sample (Appendix Table A4).[24] We incorporate this result and estimate the differencing models utilizing the threshold year (t = 5) as the year between the pre- and post-periods. As with a standard differencing model, we omit the treatment year (t = 5) from the estimation.

6.2. DDD results

Table 2 reports the DDD results estimating the sample average treatment effect of the GRFP award allocation on publication activity (Eq. 1). The outcome measure for column 1 is total publication levels, and for column 2 it is the three-year moving average of new publications.[25] The parameter of interest is statistically significant for total publications. The average differential effect for women with the award relative to men with the award is 1.17 fewer publications; this differential effect takes place five years following the GRFP (Standard Error (SE) 0.683).

We also estimate a series of robustness assessments for this model. We first estimate on an abbreviated timeframe and include the first five years (pre-GRFP) and last five years after the threshold effect. The treatment effect in this model is a reduction of 1.29 publications for females, on average (SE 0.774). Second, rather than clustering the

[24] In Appendix Tables A3 and A4, we also estimate the threshold model and machine learning approach on the full (un-coarsened) sample. The threshold model yields year 4 to be the effect year in the full sample. In the machine learning approach, years 4 and 5 yield identical R-squared values and have negligible differences in the coefficient.

[25] We also estimated the model using other outcomes that include first author position publications, publications in the top 5% of life science journals, and publications without the student's advisor. Across all three measures, we do not find a statistically significant total effect from the primary DDD model; however, we do find evidence of a heterogeneous effect from the DD models across various stratifications. The results reflect the trends illustrated by total publications in Table 3.

Table 2
OLS DDD Estimation Results.

	Total Publications (1)	New Publications (2)
DDD Treatment Effect Post*Award*Female	**−1.165***	−0.055
	(0.683)	(0.121)
Award*Post	1.308**	0.152
	(0.544)	(0.096)
Award*Female	−0.311	−0.084
	(0.197)	(0.053)
Post*Female	−1.319***	−0.299***
	(0.492)	(0.082)
Award	0.195	0.069*
	(0.122)	(0.038)
Female	−0.029	−0.046
	(0.200)	(0.050)
Post Period	4.524***	0.567***
	(0.363)	(0.062)
Constant	−5.008	−0.727
	(5.850)	(1.242)
Observations	8,460	6,768
Unique PI's	564	564
Adjusted R-Squared	0.365	0.176
Clustered by PI	Yes	Yes
Sample	Coarsened	Coarsened
Controls Included	Yes	Yes
Cutoff	Threshold Year	Threshold Year

Note: Columns estimate DDD on coarsened sample for full timeframe with cutoff year of threshold year (t = 5) omitted for outcome of either total publications or three-year moving average of new publications. All estimations include individual, program, and university-level controls listed in section 5.1. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

standard errors at the PI-level, we cluster by year. The results are robust (1.17 fewer publications, SE 0.096). Lastly, we estimate a PI and year fixed effects model and include the set of time-varying indicators from the DDD model in the regression. The results are robust though smaller in magnitude. Women with the GRFP award produce 0.76 fewer publications (SE 0.187). These results are reported in Table A5 in the Appendix (Panels A and B).

6.3. Stratified DD results

To investigate the driver of the DDD effect, we estimate a series of DD models with stratified sub-samples. Table 3 presents the results for both outcomes – total and new publications – among the stratified groups. Regarding differences within gender, column 1 reports the effect among females across GRFP designation. The interaction is economically small and statistically insignificant (0.14, SE 0.414). We do not find evidence that women with the award produce a differential effect compared to women without the award. This stands in contrast to our expectations. However, as anticipated, we do find a significant positive effect for male awardees compared to male honorable mentions (col. 2: 1.31, SE 0.545).

As for differences across gender, the results in columns 3–5 indicate a negative effect for women compared to men. Female awardees produce 2.48 fewer publications than male awardees (col. 3: SE 0.474) and 0.59 fewer publications than male honorable mentions (col. 5: SE 0.247). Moreover, female honorable mentions produce 1.32 fewer publications (SE 0.494) than men with the same designation.

The results on new publications mimic those of total publications, also showing negative effects for women compared to men across award status. We also run a series of robustness checks for the stratified samples presented in Table A5 of the Appendix. The results are robust when clustering by year and estimating with individual and year fixed effects. Comparing the impact of the award within women, we find evidence of a small, positive effect. Although we do not find an effect with the main model, this lends preliminary evidence to our initial expectations that female awardees will produce more than female honorable mentions. However, the relative impact between female awardees and honorable mentions is smaller than the differential effect for men.

To examine these trends with greater nuance, we run a series of stratified two-period DD estimations. We define the year prior to the GRFP event (t = -1) as the pre-period and consider various post designations from year one to year 10 post GRFP, respectively. Fig. 2 reports the results for the total publications outcome; Fig. 3 reports for new publications.[26] We only report the interaction coefficient in the figures if the treatment effect is statistically significant. For reference, we present the complete set of interaction coefficients in Table A6 in the Appendix.

Regarding the award effects within gender, the positive and increasing trend points to a benefit of the R&D allocation for men – male awardees produce more than male honorable mentions (dark blue line). This difference is significant starting five years after the GRFP, which is consistent with the timing from the baseline, threshold model. Yet, when we estimate the differences between female awardees and honorable mentions, we find no evidence of a differential effect (indicated as black in the legend, results are not presented in the figure due to lack of statistical significance). The coefficients reported in Table A6 (Female: Award*Post) are economically small, especially in contrast to the other results.

Turning to award effects across gender, we find a negative effect for women. The most prominent negative effect lies between female and male awardees with the evidence of a difference starting three years post GRFP (gray line reported in Fig. 2). Ten years following the GRFP, the difference between these two groups is 3.7 publications. We also find a negative effect across gender when we examine only honorable mentions. Without the award, women produce less than men; this effect begins to take effect seven years post GRFP (yellow line). More notable, however, is the fact that we also find evidence that female awardees produce *less* than male honorable mentions; this effect begins six years following the GRFP (light blue). Ten years following the GRFP, the difference between these two groups is 1.2 publications. We find similar trends when we estimate for new publication activity (Fig. 3). However, the positive effect of the award for men loses statistically significance in later years.

7. Exploratory mechanisms

We have reported a series of results that indicate the value of the competitive, federal R&D award is less for women. Female awardees publish at lower rates than male students. Moreover, while male awardees produce more than male honorable mentions, we do not find conclusive results that female awardees produce more than female honorable mentions. In this section, we explore a series of plausible mechanisms that may explain why these differences persist. These include training factors and other funding opportunities that may affect the student during graduate training and longer-term factors pertaining to the student's research orientation.

7.1. Graduate advisor

First, we consider the student's graduate advisor as a source of variation. As introduced in section 1, life sciences research is a collaborative endeavor (Jones et al., 2008; Conti and Liu, 2015). Moreover, graduate training in this field is generally an opportunity to join an

[26] Fig. 2 reports the results from 50 two-period DD regressions (10 post-years across five stratified groups, respectively). Fig. 3 reports the results from 45 two-period DD regressions (9 post-three-year range across five stratified groups).

Table 3
Stratified Panel OLS DD Estimation Results.

	Total Publications					New Publications				
	(1) Female	(2) Male	(3) Awardee	(4) HM	(5) FA/MHM	(6) Female	(7) Male	(8) Awardee	(9) HM	(10) FA/MHM
Award*Post	0.143	1.308**				0.097	0.152			
	(0.414)	(0.545)				(0.074)	(0.097)			
Post*Female			−2.484***	−1.319***				−0.354***	−0.299***	
			(0.474)	(0.494)				(0.090)	(0.082)	
Award*Female					−0.594**					−0.142***
					(0.257)					(0.052)
Award	−0.225	0.219				−0.036	0.074*			
	(0.163)	(0.152)				(0.041)	(0.040)			
Post Period	3.205***	4.524***	5.832***	4.524***		0.268***	0.567***	0.719***	0.567***	
	(0.333)	(0.365)	(0.405)	(0.365)		(0.054)	(0.062)	(0.074)	(0.062)	
Female			−0.431**	0.145				−0.149***	−0.017	
			(0.167)	(0.289)				(0.040)	(0.059)	
Constant	9.516	−19.793**	0.306	−13.121	−0.038	2.564**	−4.564***	0.580	−3.150*	0.427
	(5.779)	(8.910)	(6.886)	(10.545)	(5.599)	(1.042)	(1.754)	(1.562)	(1.855)	(1.235)
Observations	4,350	4,110	4,815	3,645	4,848	3,480	3,288	3,852	2,916	3,939
Adjusted R-Squared	0.338	0.400	0.402	0.347	0.0707	0.151	0.208	0.205	0.177	0.0730
Clustered by PI	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Sample	CEM Female	CEM Male	CEM Awardee	CEM HM	CEM Awardee	CEM Female	CEM Male	CEM Awardee	CEM HM	CEM Awardee
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cutoff: Threshold year omitted	Yes	Yes	Yes	Yes	No	Yes	Yes	Yes	Yes	No

Note: Columns estimate stratified DDs on coarsened sample for full timeframe with cutoff of threshold year (t = 5) omitted on outcome of either total publications or three-year moving average of new publications. All estimations include individual, program, and university-level controls listed in section 5.1. For col. 5 and 10, there is no cutoff year. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

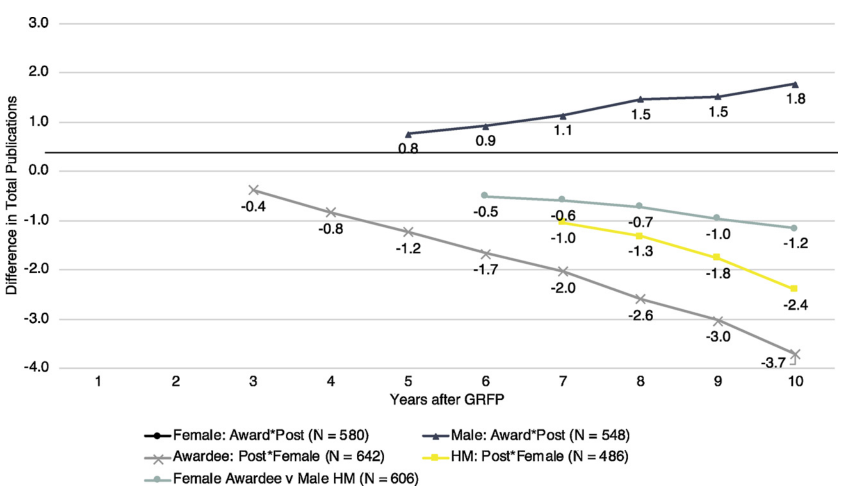

Fig. 2. Annual Stratified Two-Period Difference Estimation Results - Statistically Significant Treatment Effect on Total Publications.
Notes: This figure reports the results from 50 two-period DD regressions (10 post-years across five stratified groups). We report the coefficient on the treatment effect (Post * Treat) for statistically significant coefficients $p < 0.1$. We report the full set of results in Appendix Table A6 Panel A.

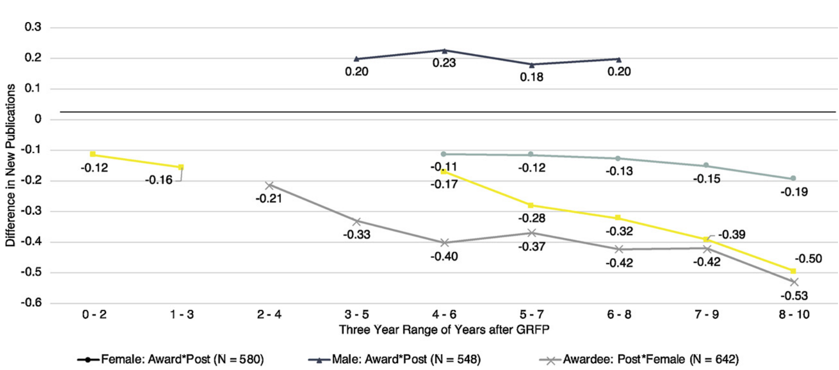

Fig. 3. Annual Stratified Two-Period Difference Estimation Results - Statistically Significant Treatment Effect on Three-Year Moving Average of New Publications.
Notes: This figure reports the results from 45 two-period DD regressions (9 post-three-year range across five stratified groups). We report the coefficient on the treatment effect (Post * Treat) for statistically significant coefficients $p < 0.1$. We report the full set of results in Appendix Table A6 Panel B.

advisor's lab and research team. Although the federal funds support student research during graduate school, the intention of this program is to enhance their training experience, of which the advisor plays a critical role. Yet, there is evidence that high-quality advisors mentor more men than women and that the mentoring approach varies based on the gender of the advisee (Pezzoni et al., 2016; Carrell et al., 2010; Gaule and Piacentini, 2018; Sheltzer and Smith, 2014). One potential explanation is that advisors view quality as more uncertain for female students and rely on paid research assistance-ships to discern quality. The GRFP fellowship removes this opportunity and the subsequent quality discovery. This may have perverse consequences on how female awardees are mentored.

We examine the matched pairs of students and advisors by both the gender pairing and the advisor's citation impact.[27] We use this information to examine the advisor's joint co-authoring behavior with the student. For the second measure, we compute the advisor's cumulative forward citation count at year 10 (corresponding to 10 years after the GRFP),[28] and then bifurcate the sample placing advisors above and below the median level, respectively.[29] The median citation count for advisors is 799.5 at year 10 with an average of 50 citations a year, compared to 480.5 in year 5 with an average of 43.7 a year. We assume that advisors with higher research impact – those with citation levels above the median – prioritize research during graduate training. This should translate to higher levels of research activity for the student given the prevalence of collaborations in this field.

Due to the limited citation data and multiple dimensions, we estimate a series of t-tests to illustrate discrepancies. Table 4 presents these exploratory results. Panels A and B report the results on joint publication activity for male (col. 1–12) and female (col. 13–24) advisors, respectively. We report the results for advisors with above median citations in columns 1–6 and 13–18 and below median citations in columns 7–12 and 19 – 24. T-tests reveal that male advisors producing high-impact research are less likely to publish with female awardees compared to the other types of students. Of note, these male advisors publish more with female honorable mentions than female awardees (col. 5). This finding goes against our initial expectations and suggests that female awardees incur a penalty when receiving the award. However, male advisors producing lower-impact research do not reveal such differences.

Regarding female advisors, those producing high-impact research also publish more with male students than female; this holds across award status. There is a divergence, however, for female advisors producing lower-impact research. They publish more with male awardees than female awardees, but more with female honorable mentions than male. However, similar to high-impact male advisors, they publish more with female honorable mentions than female awardees and more with male awardees than male honorable mentions. Again, this is contrary to our expectation and suggests female awardees incur a penalty.[30]

7.2. Alternative pre-candidacy federal funding activity

The GRFP is the largest U.S. federal program that provides support for S&E graduate students – specifically pre-candidacy. However, if a graduate student secured another pre-doctoral R&D award, this may confound the results. With a focus on the life sciences, we turned to the U.S. National Institutes of Health pre-doctoral training programs to assess whether this sample of individuals – either awardee or honorable mention – additionally secured one of the competitive NIH grants.[31] Among the sample of students that ever publish, only one student (a male) secured a pre-candidacy training grant. We are not concerned that NIH funding is confounding the results.

7.3. Student's research orientation

One might be concerned that women are generally less likely to remain active in research following the GRFP award due to gender-specific factors (e.g., childbirth). If females leave academic research, this could explain the large negative effect across gender. To explore this possibility, first, we examine the extensive margin for the interaction of award status and gender on having *any publications*. As a diagnostic, we report the annual likelihood of having any publications for the coarsened sample in Appendix Figure A2. The trends closely overlap among male awardees, female awardees, and male honorable mentions; of note, these groups exhibit a cumulative distribution with a notable increase between the year prior to the GRFP and five years after before leveling off at around 80 percent. Female honorable mentions, however, diverge at the GRFP year (t = 0) and exhibit a flatter trend with a lower likelihood of ever publishing.

To probe this further we look beyond graduate training and consider professional placement following degree conferment. While female awardees are more likely to ever publish than female honorable mentions, it is plausible that there is a selection effect across gender *after* graduate training. The discrepancies in research productivity may result from women pursuing non-research professional positions. To account for this possible selection effect, we draw upon *post*-treatment data – specifically, the student's initial professional placement after graduate training. Although professional placement is likely correlated to award conferment, we use this metric for exploratory purposes to stratify the sample between those that continue a career that prioritizes research productivity verses those that do not.

Among the set of students that place into research-active, academic, or post-doctoral positions, the negative effect *increases* in magnitude from the baseline DDD model. Women that both receive the federal funding in graduate training and professionally place in a research position following graduate training publish approximately 1.7 (SE 0.796) to 2.4 (SE 0.941) fewer publications, on average, than male awardees in similar positions. Recall, we estimate a coefficient of 1.2 fewer publications (SE 0.683) for the primary model (Table 2). This larger negative result likely captures environmental barriers and biases that impact research production for women (e.g. Moss-Racusin et al., 2012; Pezzoni et al., 2016; Lerchenmueller and Sorenson, 2018; Sarsons, 2017). Table 5 presents these results. When we consider heterogeneity estimating the DD model across various stratifications, again, we find similar trends as presented in Figs. 2 and 3. Male awardees exhibit a positive significant effect (compared to male honorable mentions) and female awardees exhibiting a negative effect (compared to male awardees), while there is no statistically significant difference within women.[32]

8. Discussion

We introduced this study by highlighting recent evidence of improved gender equality in the matriculation of women into S&E graduate programs. However, there remains compelling evidence to show pronounced gender discrepancies at later career stages. We focus at the intersection of these milestones to understand whether and how the gender gap forms in graduate training. Drawing upon a sample of life science graduate students with demonstrated research potential – as determined by an NSF review panel – we find evidence that the gender

[27] It is important to mention potentials for sorting differences across fields – of our three sub-fields, ecology has a slightly higher concentration of female advisors: 24.6% compared to 20.5% in biochemistry and 20.6% in biology.

[28] The advisors may be at different points in their career; however, the student's training year remains constant.

[29] We also estimated impacts calculating the advisor's citation count at year 5, when the GRFP impact begins. Fifty advisors change status between above or below the median from these two time periods. The results of the t-tests are consistent with five estimations switching from statistically insignificant to significant and one switching from significant to insignificant.

[30] We also estimate the DDD model on the outcome of joint publications, stratified by the quality threshold of the advisor. The differential effect of the award is 1.4 fewer publications for female awardees with lower-impact advisors. Full results are available upon request.

[31] This includes the following NIH training programs: F30, F31, F99, R36, R90, T32, and T90.

[32] Stratified DD results available upon request.

Table 4
Student-Advisor Joint Publications by Advisor Productivity and Gender.

Panel A: Male Advisor

Above Median Citations at Year 10

	(1)		(2)		(3)		(4)		(5)		(6)	
	By Student Gender		By Awardee Gender		By HM Gender		F-A vs. M-HM		By Female Award Status		By Male Award Status	
Joint Publications	Male (1008)	Female (1136)	Male (512)	Female (640)	Male (496)	Female (496)	Male HM (496)	Female A (640)	HM (496)	Awardee (640)	HM (496)	Awardee (512)
Mean	1.34	0.72	1.49	0.55	1.19	0.95	1.19	0.55	0.95	0.55	1.19	1.49
Std. Err.	0.082	0.051	0.12	0.05	0.11	0.095	0.11	0.052	0.095	0.052	0.11	0.12
T-Statistic	6.56		7.59		1.66		5.76		3.93		−1.85	
One-Side P-Value	**0.00**		**0.00**		**0.049**		**0.00**		**0.00**		**0.032**	

Below Median Citations at Year 10

	(7)		(8)		(9)		(10)		(11)		(12)	
	By Student Gender		By Awardee Gender		By HM Gender		F-A vs. M-HM		By Female Award Status		By Male Award Status	
Joint Publications	Male (816)	Female (1184)	Male (464)	Female (832)	Male (352)	Female (352)	Male HM (352)	Female A (832)	HM (352)	Awardee (832)	HM (352)	Awardee (464)
Mean	0.56	0.53	0.65	0.58	0.45	0.43	0.45	0.58	0.43	0.58	0.45	0.65
Std. Err.	0.056	0.038	0.084	0.044	0.066	0.076	0.066	0.044	0.076	0.044	0.066	0.084
T-Statistic	0.47		0.89		0.14		−1.61		−1.71		−1.84	
One-Side P-Value	0.32		0.19		0.44		**0.054**		**0.044**		**0.033**	

Panel B: Female Advisor

Above Median Citations at Year 10

	(13)		(14)		(15)		(16)		(17)		(18)	
	By Student Gender		By Awardee Gender		By HM Gender		F-A vs. M-HM		By Female Award Status		By Male Award Status	
Joint Publications	Male (304)	Female (288)	Male (112)	Female (176)	Male (192)	Female (112)	Male HM (192)	Female A (176)	HM (112)	Awardee (176)	HM (192)	Awardee (112)
Mean	0.944	0.55	1.25	0.63	0.77	0.43	0.77	0.63	0.43	0.63	0.77	1.25
Std. Err.	0.11	0.07	0.17	0.10	0.14	0.079	0.14	0.1	0.079	0.10	0.14	0.17
T-Statistic	3.01		3.33		1.77		0.77		−1.39		−2.19	
One-Side P-Value	**0.0013**		**0.0005**		**0.039**		0.22		**0.083**		**0.015**	

Below Median Citations at Year 10

	(19)		(20)		(21)		(22)		(23)		(24)	
	By Student Gender		By Awardee Gender		By HM Gender		F-A vs. M-HM		By Female Award Status		By Male Award Status	
Joint Publications	Male (272)	Female (384)	Male (112)	Female (272)	Male (160)	Female (112)	Male HM (160)	Female A (272)	HM (112)	Awardee (272)	HM (160)	Awardee (112)
Mean	0.47	0.43	0.82	0.27	0.22	0.82	0.22	0.27	0.82	0.27	0.22	0.82
Std. Err.	0.080	0.055	0.18	0.05	0.048	0.14	0.048	0.048	0.14	0.048	0.048	0.18
T-Statistic	0.37		4.055		−4.61		−0.73		4.70		−3.81	
One-Side P-Value	0.36		**0.00**		**0.00**		0.23		**0.00**		**0.0001**	

Note: Each column presents a *t*-test by the specified dimension on joint publications between the advisor and student. The sample is stratified by the advisor's gender and citation impact. P-values in bold indicate statistical significance. We estimate the advisor's productivity based on their citation levels as reported in Google Scholar at year 10 (10 years post-GRFP for their corresponding student). Based on the distribution, we identify the set above and below the median. At year 10, the median citation count for advisors was 799.5 with an average of 50 citations a year. In addition, the cutoff was also assigned at year 5 as a robustness check (5 years post-GRFP, the threshold cutoff). For this time period, the median citation count was 480.5 with an average of 43.7 citations a year. Fifty advisors changed between above and below the median across these two assignments. The results are very similar between the two assignments with only six *t*-tests switching between weak statistical significance and statistically insignificant.

gap exists in graduate training.

The results of the triple difference estimation show a negative effect for women compared to men from the award. When we investigate the driver of this effect, we fail to find a statistically significant impact on productivity within women compared to a large, positive effect within men of the award. This takes effect five years following the GRFP allocation – at the time when many students are transitioning from graduate school to their first professional placement. While most would expect conferment of a competitive, meritorious NSF award to improve research productivity, we fail to find that the award carries an effect for women. Moreover, female awardees underperform both male awardees and male honorable mentions.

We explore a series of mechanisms that may account for these differences. Notably, we find preliminary evidence that male students

Table 5
OLS DDD Estimation Results by First-Placement.

	Total Publications		
	(1)	(2)	(3)
DDD Treatment Effect Post*Award*Female	**−1.696****	**−2.049****	**−2.412****
	(0.796)	(0.965)	(0.941)
Award*Post	1.843***	1.696**	2.708***
	(0.615)	(0.725)	(0.702)
Award*Female	−0.335	−0.337	−0.398
	(0.235)	(0.296)	(0.313)
Post*Female	−1.148**	−0.579	−0.849
	(0.581)	(0.712)	(0.675)
Award	0.198	0.078	0.318*
	(0.139)	(0.161)	(0.181)
Female	0.066	0.266	0.160
	(0.261)	(0.325)	(0.346)
Post Period	4.813***	4.971***	4.800***
	(0.423)	(0.502)	(0.473)
Constant	−8.571	−12.776*	−1.687
	(7.328)	(7.620)	(6.594)
Observations	6,675	4,665	4,545
Unique PI's	445	311	303
Adjusted R-Squared	0.396	0.435	0.448
Clustered by PI	Yes	Yes	Yes
First Placement Sample	Research	Academic Research	Post-Doctoral
Controls	Yes	Yes	Yes
Cutoff	Threshold Year	Threshold Year	Threshold Year

Note: Columns estimate DDD on coarsened sample for full timeframe with cutoff of threshold year (t = 5) omitted for sub-samples labeled. Outcome is total publications. All estimations include individual, program, and university-level controls listed in section 5.1. Standard errors in parentheses. *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$.

benefit from collaboration during their training. Specifically, we find that both female and male advisors co-author less with female awardees. While the amount of funding is constant, this implies that the award bestows different training opportunity based on the student's gender with women at a disadvantage. One explanation for this effect is that the advisors rely more on paid research assistance-ships to discern quality of female students and the award separates women from the natural mentoring that would occur if they were financially dependent on their advisor. Future research should expand upon this finding.

Given the lack of a definitive mechanism from this study, it is important to discuss the lingering concern of quality. One could argue, in line with Lawrence Summers' statement, that NSF is giving awards to less qualified females, thus providing higher quality male students with honorable mention status to increase diversity. Despite the features of our research design to control for quality, we would assume that if NSF is capable of assigning the award to high quality males it should also be capable of assigning award status within the group of females to those of the highest quality. Thus, these higher qualified female awardees *should* outperform the female honorable mentions given the positive signal and additional resource allocation the award confers for men. But they do not.

Similarly, while we cannot eliminate the hypothesis that the male treatment effect is due to underlying quality, such a theory would be inconsistent with the general assumption that the grants provide signaling and resource impacts. Moreover, quality differences at this early stage are generally small. Thus, the presence of a treatment effect for men with the absence of one for women suggests a gender bias.

We test this concern empirically using the introduction of the *broader impacts* criterion in 1997. Prior to this, *intellectual merits* was the sole review criterion. Descriptive statistics comparing individuals who applied in 1995 and 1996 versus those who applied between 1997 and 2005 show that the rate of females applying (and female awardees, more specifically) in the early sample are not statistically different as compared to those that proposed from 1997 to 2005. In the early sample, 50 percent were female, while among the broader impacts sample 51 percent were female. Among awardees, 55 percent were female in the early sample and 53 percent were female in the broader impacts sample. This suggests that the reviewers maintained consistent review of proposals when the sole emphasis was on *intellectual merits*. We also ran the primary model omitting these early years to examine the policy change; the results are consistent (Appendix Table A7). The additional emphasis on broader impacts has not produced a notable downward bias.

In sum, we do not find conclusive evidence of a difference in productivity between female awardees and honorable mentions. So, either there is no quality difference between female awardees and honorable mentions *and* the award then offers no positive signal for female students; or if the female awardees are of higher quality than the female honorable mentions, they actually suffer a *negative* signal from the award. Moreover, we find preliminary evidence of a penalty for female awardees when considering joint publication patterns with advisors. Thus, even if there were underlying quality differences between these high-quality male and female students, the value of the award still exhibits a differential effect across gender.

9. Conclusion

The NSF GRFP provides competitive R&D funding for early career graduate students with demonstrated research potential. Although we find evidence that men and women receive the award at comparable rates that reflect trends of graduate program matriculation, the program does not produce the same outcomes across gender. It appears that the signal conferred by the grant and training following the award vary with women at a disadvantage.

These results indicate that more work remains to ensure that graduate students are supported appropriately. The NSF Postdoctoral program, for example, requires a Postdoctoral Researcher Mentoring Plan as part of the application. The mentor is required to articulate professional training components that include research, teaching, and even career counseling. This additional component could help to ensure equitable training and reduce the potential of the Pygmalion effect whereby student performance varies with mentor expectations (Rosenthal and Jacobson, 1968).

More broadly, this study contributes to the literature on the perverse consequences of policies. Previous research has documented the divergent signaling effects for disadvantaged groups in response to policies aimed at reducing the inequality. However, we find evidence of divergent signals in response to policies that are meritorious. While the GRFP aims to provide high-quality students with resources and a positive signal, women do not receive the same positive effect felt by men. These effects are persistent and carry important implications for professional trajectory and the production of science at large.

Declaration of Competing Interest

None.

Acknowledgements

This research was funded by the National Science Foundation's Science of Science and Innovation Policy Program (NSF 1548288 & 1661157). We thank Ina Ganguli, Kaye Husbands-Fealing, Julia Lane, Maryann Feldman, Dyanne Vaught, Chris Hayter, Daniel Armanios, Peter Younkin, Ouafaa Hmaddi, and, Jeff Schroeder for their comments on earlier versions of this paper. We also thank our editor and two anonymous reviewers for their constructive comment. This paper benefited from feedback at the 2017 Association for Public Policy and Management and the 2018 Academy of Management annual meetings.

We are greatly indebted to our research assistants for their contributions with building the database. Specifically, we would like to thank Georgiy Sichinava, Soro Soukpafolo, Elsa Carolina Mantilla Garcia, Nicole M. V. Ross, Zachary Trudo, and Emily Plews. We would also like to thank Austin McLean for his assistance with the Proquest data and John McDonald for his assistance with the Scopus data. All errors are our own. Alex Graddy-Reed and Lauren Lanahan contributed equally and are considered joint first-authors. Alex Graddy-Reed is the corresponding author.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.respol.2019.103838.

References

Adams, J., Griliches, Z., 1996. Measuring science: an exploration. Proc. Natl. Acad. Sci. 93 (23), 12664–12670.
Agan, A., Starr, S., 2017. Ban the box, criminal records, and racial discrimination: a field experiment. Q. J. Econ. 133 (1), 191–235.
Anseel, F., Lievens, F., Schollaert, E., Choragwicka, B., 2010. Response rates in organizational science, 1995–2008: a meta-analytic review and guidelines for survey researchers. J. Bus. Psychol. 25 (3), 335–349.
Azoulay, P., Graff Zivin, J.S., Manso, G., 2011. Incentives and creativity: evidence from the academic life sciences. Rand J. Econ. 42 (3), 527–554.
Barres, B.A., 2006. Does gender matter? Nature 442 (7099), 133.
Baruch, Y., Holtom, B.C., 2008. Survey response rate levels and trends in organizational research. Hum. Relat. 61 (8), 1139–1160.
Bertrand, M., Duflo, E., Mullainathan, S., 2004. How much should we trust differences-in-differences estimates? Q. J. Econ. 119 (1), 249–275.
Bettinger, E.P., Long, B.T., 2005. Do faculty serve as role models? The impact of instructor gender on female students. Am. Econ. Rev. 95 (2), 152–157.
Blackwell, M., Iacus, S., King, G., Porro, G., 2009. CEM: coarsened exact matching in Stata. Stata J. 9, 524–546.
Buffington, C., Cerf, B., Jones, C., Weinberg, B.A., 2016. STEM training and early career outcomes of female and male graduate students: evidence from UMETRICS data linked to the 2010 census. Am. Econ. Rev. 106 (5), 333–338.
Carrell, S.E., Page, M.E., West, J.E., 2010. Sex and science: how professor gender perpetuates the gender gap. Q. J. Econ. 125 (3), 1101–1144.
Ceci, S.J., Williams, W.M., 2011. Understanding current causes of women's underrepresentation in science. Proc. Natl. Acad. Sci. 108 (8), 3157–3162.
Ceci, S.J., Ginther, D.K., Kahn, S., Williams, W.M., 2014. Women in academic science: a changing landscape. Psychol. Sci. Public Interest 15 (3), 75–141.
Clarke, P.M., Fiebig, D.G., Gerdtham, U.G., 2008. Optimal recall length in survey design. J. Health Econ. 27 (5), 1275–1284.
Conti, A., Liu, C.C., 2015. Bringing the lab back in: personnel composition and scientific output at the MIT Department of Biology. Res. Policy 44 (9), 1633–1644.
Davidoff, A., Blumberg, L., Nichols, L., 2005. State health insurance market reforms and access to insurance for high-risk employees. J. Health Econ. 24 (4), 725–750.
DiTomaso, N., Post, C., Parks-Yancy, R., 2007a. Workforce diversity and inequality: power, status, and numbers. Annu. Rev. Sociol. 33, 473–501.
DiTomaso, N., Post, C., Smith, D.R., Farris, G.F., Cordero, R., 2007b. Effects of structural position on allocation and evaluation decisions for scientists and engineers in industrial R&D. Adm. Sci. Q. 52 (2), 175–207.
Doleac, J.L., Hansen, B., 2016. Does "ban the box" help or hurt low-skilled workers? Statistical discrimination and employment outcomes when criminal histories are hidden (No. w22469). National Bureau of Economic Research.
Fealing, K.H., Lai, Y., Myers Jr, S.L., 2015. Pathways vs. Pipelines to broadening participation in the STEM workforce. J. Women Minor. Sci. Eng. 21 (4).
Freeman, R.B., Chang, T., Chiang, H., 2009. Supporting" the Best and brightest" in science and engineering: NSF graduate research fellowships. Science and Engineering Careers in the United States: An Analysis of Markets and Employment. University of Chicago Press, pp. 19–57.
Garcia, L.T., Erskine, N., Hawn, K., Casmay, S.R., 1981. The effect of affirmative action on attributions about minority group members. J. Pers. 49 (4), 427–437.
Gaule, P., Piacentini, M., 2018. An advisor like me? Advisor gender and post-graduate careers in science. Res. Policy 47 (4), 805–813.
Graddy-Reed, A., Lanahan, L., Ross, N.M., 2018. The effect of R&D investment on graduate student productivity: evidence from the life sciences. J. Policy Anal. Manage. 37 (4), 809–834.
Graddy-Reed, A., Lanahan, L., Ross, N.M., 2017. Influences of academic institutional factors on R&D funding for graduate students. Sci. Public Policy 44 (6), 834–854.
Hanna, R.N., Linden, L.L., 2012. Discrimination in grading. Am. Econ. J. Econ. Policy 4 (4), 146–168.
Heilman, M.E., Block, C.J., Stathatos, P., 1997. The affirmative action stigma of incompetence: effects of performance information ambiguity. Acad. Manage. J. 40 (3), 603–625.
Iacus, S.M., King, G., Porro, G., 2012. Causal inference without balance checking: coarsened exact matching. Political Anal. 20 (1), 1–24.
Imbens, G., Wooldridge, J., 2007. What's new in econometrics: difference-in-Differences estimation. Lecture Notes 10.
Jacob, B.A., Lefgren, L., 2011a. The impact of research grant funding on scientific productivity. J. Public Econ. 95 (9-10), 1168–1177.
Jacob, B.A., Lefgren, L., 2011b. The impact of NIH postdoctoral training grants on scientific productivity. Res. Policy 40 (6), 864–874.
Jones, B.F., Wuchty, S., Uzzi, B., 2008. Multi-university research teams: Shifting impact, geography, and stratification in science. science 322 (5905), 1259–1262.
Lerchenmueller, M.J., Sorenson, O., 2018. The gender gap in early career transitions in the life sciences. Res. Policy 47 (6), 1007–1017.
McNeely, C.L., Fealing, K.H., 2018. Moving the needle, raising consciousness: the science and practice of broadening participation. Am. Behav. Sci. 62 (5), 551–562.
Merton, R.K., 1968. The Matthew effect in science: the reward and communication systems of science are considered. Science 159 (3810), 56–63.
Miller, D.I., Wai, J., 2015. The bachelor's to Ph. D. STEM pipeline no longer leaks more women than men: a 30-year analysis. Front. Psychol. 6, 37.
Miller, D.I., Halpern, D.F., 2014. The new science of cognitive sex differences. Trends Cogn. Sci. (Regul. Ed.) 18 (1), 37–45.
Moss-Racusin, C.A., Dovidio, J.F., Brescoll, V.L., Graham, M.J., Handelsman, J., 2012. Science faculty's subtle gender biases favor male students. Proc. Natl. Acad. Sci. 109 (41), 16474–16479.
Oreopoulos, P., 2011. Why do skilled immigrants struggle in the labor market? A field experiment with thirteen thousand resumes. Am. Econ. J. Econ. Policy 3 (4), 148–171.
Pezzoni, M., Mairesse, J., Stephan, P., Lane, J., 2016. Gender and the publication output of graduate students: a case study. PLoS One 11 (1), e0145146.
Podolny, J.M., 1993. A status-based model of market competition. Am. J. Sociol. 98 (4), 829–872.
Powell, K., 2016. Does it take too long to publish research? Nature News 530 (7589), 148.
Quadlin, N., 2018. The mark of a woman's record: gender and academic performance in hiring. Am. Sociol. Rev. 83 (2), 331–360.
Rosenthal, R., Jacobson, L., 1968. Pygmalion in the classroom. Urban Rev. 3 (1), 16–20.
Sarsons, H., 2017. Recognition for group work: gender differences in academia. Am. Econ. Rev. 107 (5), 141–145.
Sauermann, H., Roach, M., 2012. Science PhD career preferences: levels, changes, and advisor encouragement. PLoS One 7 (5), e36307.
Sheltzer, J.M., Smith, J.C., 2014. Elite male faculty in the life sciences employ fewer women. Proc. Natl. Acad. Sci. 111 (28), 10107–10112.
Stephan, P.E., 2012. How Economics Shapes Science Vol. 1 Harvard University Press, Cambridge, MA.
Suchman, M.C., 1995. Managing legitimacy: strategic and institutional approaches. Acad. Manage. Rev. 20 (3), 571–610.
Sugimoto, C.R., Ni, C., West, J.D., Lariviére, V., 2015. The academic advantage: gender disparities in patenting. PLoS One 10 (5), e0128000.
Tong, H., 2012. Threshold Models in non-Linear Time Series Analysis Vol. 21 Springer Science & Business Media.
van Arensbergen, P., van der Weijden, I., Van den Besselaar, P., 2012v. Gender differences in scientific productivity: a persisting phenomenon? Scientometrics 93 (3), 857–868.
Van der Lee, R., Ellemers, N., 2015. Gender contributes to personal research funding success in the Netherlands. Proc. Natl. Acad. Sci. 112 (40), 12349–12353.
Wuchty, S., Jones, B.F., Uzzi, B., 2007. The increasing dominance of teams in production of knowledge. Science 316 (5827), 1036–1039.