Gender and Attrition in the Changing Organization of Scientific Work

Seokkyun WOO
Korea Advanced Institute of Science and Technology (wsk618@kaist.ac.kr)
You-Na LEE
Georgia Institute of Technology (youna.lee@gatech.edu)

2024.01.16

Abstract:
Despite longstanding concerns about the under-representation of women in science, few studies have approached this issue from the perspective of the changing organization of work in science. Past studies have documented a trend toward increased bureaucratization of scientific work, marked by the growing number of scientists specialized in supporting roles. Using data on publishing careers of scientists from 1951 to 2012 from selected natural and social science fields, we show that these "supporting" career-type scientists have been traditionally associated with women. While we find that the gender difference in career types has converged over the past few decades, this convergence has been largely driven by an increasing share of male scientists taking on supporting roles. We also find that historical gender inequality in career attrition in science is largely attributable to women traditionally occupying "supporting" roles, which suggests that examining work organization is crucial for understanding gender inequality in science. Lastly, using survival analysis, we find that both female "lead" and "supporting" career types face higher attrition rates than their male counterparts. Meanwhile, we find that "lead" career types yield fewer advantages for women compared to men in natural sciences, whereas "supporting" career types are particularly disadvantageous for women in social sciences. Our findings provide science policymakers with insights necessary to tailor support for women scientists by considering the nuances of their production roles in science.

Keywords: Scientific Careers, Organization of Science, Gender

Introduction:

The under-representation of women in science is a longstanding issue (*1-5*). Women not only encounter more structural barriers when entering science (*4, 6-8*), but they are also more likely to leave the field and pursue non-scientific careers (*1, 3, 9*). While many studies have highlighted important factors that are attributed to persistent gender inequality in science, few studies have examined this issue from the perspective of the changing organization of work in science. As science evolved into a more collaborative, team-based endeavor (*10, 11*), scientific work became increasingly bureaucratized, with increasing use of hierarchy, standardization, and division of labor (*12-14*) in the production of knowledge. One important outcome of this trend is the rise of a category of scientists whose main role is to support the lead scientists in their labs and projects (*15-18*). Despite science's increasing reliance on these supporting scientists, studies have shown immense inequalities between lead and supporting scientists (*16, 19*).

Given the increasing division between the lead and supporting roles contributing to the career inequalities in science, we argue that the well-documented gender difference in scientific careers should be examined from the perspective of the changing organization of work in science. **We first empirically examine whether women are more likely to spend their publishing careers as supporting scientists and, if so, whether this has changed over time.** History of science provides much evidence of the confined and often marginalized role of women in science, if acknowledged at all. Well-known historical cases, such as Madame Lavoisier (*5*) and, more recently, Rosalind Franklin (*20*), highlight the struggles faced by women scientists who were perceived by their contemporaries mainly as supporting associates rather than independent scientists in their rights**. We then examine if gender differences in career types (lead vs. supporting scientists) account for historical gender inequalities in career longevity.** While previous studies have pointed out various factors contributing to higher career attrition of women, we focus on the organizational aspect by questioning whether women's historical confinement to a less prestigious production role in science contributed to higher career attrition for women. **In our final analysis, we investigate the potential differential effects of supporting roles on attrition rates between women and men.** Essentially, while our first research question examines whether the "supporting role" is gendered, this question asks whether the "supporting penalty" is gendered. Understanding the differential attrition penalty is crucial as it allows policymakers with insights necessary to tailor support for women scientists by considering the nuances of their career types.

To address these questions, we construct scientists' comprehensive publishing career data from a large-scale bibliometric dataset, SciSciNet (*21*) (SM A.2), specifically focusing on selected fields from natural sciences (Biology and Chemistry) and social sciences (Psychology and Sociology). In total, our constructed data covers 658,049 US-affiliated authors who entered publishing careers from 1951 to 2012. By identifying each career type ("lead" vs. "supporting") (SM A.4) and name-inferred binary gender of scientists (SM A.3), we examine the interplay between career type and gender and how they explain inequality in career longevity. To measure the career type of a scientist, we classify each author into one of two types: "lead" and "supporting" career types (*17, 19*). Lead career types are authors who, at some point in their careers, have produced at least one paper as a first author. Meanwhile, supporting career types are the authors who, at no point in their careers, appear in a publication as a first author. We validated this measure against a subset of the bibliometric dataset, which provides contributorship statements and finds that supporting career types are more likely to perform tasks that are labor-intensive (SM A4.2, Fig. 6D,H,L), which is consistent with the conceptual description of the role of supporting scientists in the literature (*16, 18*).

Results:

Gender Trends in Career Type: A Historical Perspective

We first report the share of scientists among their cohort who spend their careers as lead career types, defined as those who have published at least one first-author paper in their publishing careers (Fig. 1A-D). Our findings show that in natural sciences (chemistry and biology), the "lead" career types have been traditionally associated with male scientists, thus corroborating the long-standing view that women were associated with supporting roles in the production of science (*5, 22*). For instance, during the first decade of the 1950s, only slightly more than half (52.6%) of female scientists in chemistry held the position of lead career types, compared to 88.3% of their male counterparts (Fig. 1B). Similarly, in biology, only 67.4% of women were identified as lead career types, as opposed to 87.1% of men (Fig. 1A). It is worth noting that when Rosalind Franklin was undertaking her pioneering research on uncovering the structure of DNA in the early 1950s, the production of science was heavily gendered, such that women scientists have disproportionately carried out the labor-intensive tasks in natural sciences. Interestingly, our analysis reveals a contrasting trend within the social sciences, where the division of labor between women and men does not reflect the same level of gender disparity as observed in the natural sciences (Fig. 1C,D). While we did notice gender discrepancies within psychology and sociology throughout our observation periods, these differences were not as pronounced as those observed from the natural sciences.

Secondly, we find a striking trend of convergence in career type (lead vs. supporting) between men and women in natural sciences. For example, by the 2010-2012 period, the gender difference in the share of the lead career type significantly narrowed, with 64.6% of women in chemistry identified as lead career types compared to 67.0% of men (Fig. 1B). In biology, the share of lead career type is around 62.3% for women and 63.7% for men (Fig. 1A). While there are still measurable differences between men and women in these fields during this period ($t \approx 3.428$ in chemistry, $t \approx 2.092$ in biology), our findings clearly show a narrowing gender gap over time (Fig. 1A,B). Interestingly, this narrowing trend was not mainly driven by an increase in women occupying lead positions but by a growing proportion of men occupying supporting roles. This suggests that the observed reduction in gender inequality in production roles in the natural sciences has been partly driven by a shift in the career paths of male scientists toward roles traditionally held by women. Our finding from the social sciences contrasts sharply with that of natural sciences, where the proportion of lead career types for women and men is decreasing in a parallel fashion (Fig. 1C,D).

We now examine the publishing careers of scientists as it is a dimension where inequality has been observed and extensively researched. One of the key implications from the literature on the bureaucratic structuring of scientific work has been the notably shorter careers of supporting scientists (*16, 19*). On the other hand, a sociology of science literature has extensively examined gender disparities in scientific career science (*9, 23, 24*). We building on these insights to provide a novel perspective on gender inequality in science. We first report the average publishing career lengths for female and male scientists in chemistry, biology, psychology, and sociology, respectively (Fig. 1E-H). On average, women have shorter publishing career lengths than men in natural sciences (Fig. 1E,F). However, the narrowing gender gap in publishing career lengths in these fields mirrors the narrowing gap in career types in these fields. To further illustrate this connection, we report scatter plots that compare the gender gaps normalized by the proportion of lead career types on the x-axis and the ratio of average career lengths between female and male authors on the y-axis (Fig. 1I-L). For example, a ratio less than 1 implies that women are less

likely to be lead career types (vertical axis) or have shorter career lengths on average compared to men. Our model-free evidence from natural sciences (Fig. 1I,J) shows strong correlations between the variations in the career gaps and career longevity, suggesting that the diminishing gender gap in career types could be a contributing factor to the decrease in gender inequality observed in career longevity within the natural sciences.

Historical Role of Career Type on Gender Inequality in Attrition

To further investigate the role of career types on gender inequality, we use the Cox Proportional Hazard Model to estimate the attrition hazard as a function of gender as well as early career productivity, team size, and citation counts both with and without the inclusion of the career types (lead vs. supporting) variable (SM B.1). These models are estimated from the sample of 63 cohorts from 1951 to 2012 from Natural Sciences (Fig. 2A,C) and Social Sciences (Fig. 2B,D). The estimated coefficients for women, without the inclusion of the career-type variable, are depicted in blue. In contrast, the red dots represent the estimated coefficients for women from the model that includes the career type variable.

In natural sciences, we find that gender inequality in career attrition has been decreasing over time, as shown by the decreasing hazard rate for women (see blue, Fig. 2A). However, after controlling for career types, the decrease in the gender gap is substantially attenuated (see red, Fig. 2A), which suggests that the career type might mediate the relationship between gender and career longevity in natural sciences. To formally examine the mediating effect of the career type, we ran a causal mediation analysis (*25*) (SM B.3) to estimate the proportion of the gender effect explained by its indirect effect via the career type (Fig. 2C,D). The findings from natural sciences (Fig. 2C) suggest that since the latter half of the 20th century, a substantial proportion of the gender effect on career attrition has been mediated through career type. Interestingly, we also find that the proportion of the mediation effect has been decreasing over time in natural sciences (Fig, 2C). In sharp contrast, we do not observe any associations between career type and gender inequality in career attrition in social sciences (Fig. 2B,D). The level of gender inequality in career attrition is lower in these fields, and the historical pattern of inequality remains largely unchanged with the inclusion of the career type variable (Fig. 2B). Moreover, the causal mediation analysis shows that the career type hardly mediates the relationship between gender and career attrition (Fig. 2D).

The Effects of Career Type on Gender-Specific Attrition

Lastly, we employ survival analysis to examine how career type and gender together affect career attrition in science. We use the Cox proportional hazard model, which estimates relative hazard difference (SM B1). While the Cox proportional hazard model has an advantage in not assuming a specific survival function, the model assumes that hazard ratios between different groups are constant over time, which may be a strict assumption given the observed different empirical survival curves for lead and supporting career types in our sample (SM A.7, Fig.8). We thus complement our analysis with the Weibull-based accelerated failure time (AFT) model, which directly estimates the effects of the covariates on time to exit publishing careers (SM B2). To address the potential heterogeneities across individual scientists, we incorporate control variables representing the upper 90th percentile in three key early individual performances: the first 3 years of publication productivity (number of publications), the average number of first five-year citations (c5), and the average size of early collaborative teams. We present the regression results for the natural sciences (biology and chemistry) in Table 1 and the social sciences (psychology and sociology) in Table 2.

We first find that female scientists are positively associated with a greater hazard of leaving a publishing career in both natural and social sciences than men (Table. 1-2(1)). For example, in natural sciences, the hazard rate for female scientists is, on average, 14.1% (=1.141-1) greater than that of male scientists (Table. 1(1)). We find this female penalty consistent even after controlling for early performances (Table. 1(2)) and also for the results from the AFT model, which directly estimates the exiting time (SM B5, Table 3). Meanwhile, we observe a less pronounced gender disparity in attrition within the social sciences; women exhibit a 5.6% higher hazard rate compared to men (Table. 2(1)). The observed gender disparities in career attrition align with our non-parametric estimates (SM A.7, Fig. 9), with the notable exception of the field of sociology. As evidenced by the prior study (*19*), we find that supporting career types face higher attrition risks across both natural and social sciences (Tables. 1-2 (3)). For instance, within the natural sciences, supporting career types experience an 83.7% (=1.837-1) increased hazard rate compared to their lead counterparts (Table 1(3)). Similarly, our finding from the AFT model shows that the expected survival time for supporting career types is only 52.2% of that for those in lead career types (SM B5, Table 3(3)). The supporting penalty was more severe in social sciences, with supporting career types experiencing a 129.4% (=2.294-1) increased hazard rate (Table 2(3)) and exhibiting 41.8% of the expected survival time of the lead career types (SM B5, Table 7(3)).

Next, we examine whether the career-type effect on attrition differs between women and men by including their interaction effect in our regression model (Tables. 1-2(4)). In natural sciences (Table. 1(4)), we find a negative interaction effect, suggesting that the effect of the "supporting" career type on attrition is relatively less severe for female scientists. For example, the attrition hazard for being supporting career types is around 4.6 percent (=1-0.954) less for women than men. In terms of expected survival time (SM B5, Table. 3(4)), we find that women in supporting career types have expected survival time with a factor of 1.067 of male supporting career types, suggesting that the expected career shortening from being supporting types is less severe for women than men. Interestingly, we find an opposite pattern in social sciences, such that women face a 9.2% (=1.092-1) greater hazard rate for being supporting career types than men (Table. 2(4)). Similarly, when considering the expected career duration, female supporting types have a survival time factor of 0.927 compared to male supporting types (SM B5, Table 7(4)). This indicates a more pronounced career shortening for women in supporting types compared to their male counterparts. Thus, within the natural sciences, the "supporting" career type appears to carry a relatively smaller penalty for women than for men. In contrast, the opposite pattern emerges within the social sciences, where the supporting career type tends to have a more detrimental effect on the career prospects of women.

Lastly, we further compare supporting penalties by gender for the entire period (1951 to 2012) with that for the recent period (2000-2012). First, our main findings remain consistent even when we stratify our sample by field and different timeframes (SM B.5, Tables 2-25). However, we can further observe an interesting trend where these supporting penalties for women are also getting smaller over time (from 0.954 to 0.919 for natural sciences and from 1.092 to 1.074 for social sciences both in the Cox model) and also where supporting penalties by gender in social sciences have been becoming closer to those in natural sciences although more slowly. For example, psychology, where scientists often conduct experiments as in natural sciences, shows 1.001 for the interaction term during 2000-2012, which is close to 0.927 in chemistry (SM B5). It is likely that natural sciences have a longer history of supporting positions than the social sciences (*26*) and that social sciences are following this changing nature observed earlier in natural sciences with a time lag. The evidence from running separate regression by cohort yields consistent results

with decreasing trends of interaction effects between gender and supporting career types, initially observed in the natural sciences (Fig. 3A) and subsequently in the social sciences (Fig. 3B).

Discussion

We first find that women were historically more likely to spend their careers as supporting scientists. However, we find this gap has been converging over time. Interestingly, and perhaps not in the desired direction, this convergence has been overwhelmingly driven by a decreasing share of lead career types (i.e., an increasing share of supporting career types) among men. Both male and female scientists are equally likely to take supporting roles now in science, making supporting roles less feminized than before. In terms of career lengths, women have shorter publishing careers than men in natural sciences. However, we find that the gender gap has been decreasing over time, and this gap became much narrower once we take account of their production role as measured by their career types. Thus, our evidence provides strong evidence that the existing gender inequality in scientific careers can partly be explained by the historically gendered division of labor in the production of science. While further research is needed to understand why this pattern is observed in natural sciences but not in social sciences, a potential explanation is the traditionally high level of teamwork associated with natural sciences (*11*), often leading to a more pronounced division of labor (*26, 27*), which is often intertwined with gendered patterns (*28*).

Our survival analyses show that both female scientists and supporting scientists are more likely to leave publishing careers. However, the effect of the "supporting" career type on attrition is relatively less severe for female scientists than for male scientists in natural sciences, while the supporting career tends to have a more detrimental effect on the career prospects of women in social sciences. One possible explanation is that natural sciences have a longer history of supporting positions (*26*), and there have been efforts by research institutions to make supporting scientist careers more formalized and stabilized in the natural sciences (*29*). This standardization of supporting positions with the less feminization of supporting roles may have helped female supporting scientists in natural sciences become less stigmatized with a smaller penalty, while, for men, the supporting role makes them still stigmatized. Social sciences, with a shorter history of supporting positions, show the opposite pattern, with female supporting scientists having a larger penalty than male counterparts.

The results of this study will help us understand gender inequality in science in the context of the evolving landscape of scientific work, which is characterized by extensive teamwork (*10, 11*) and a high division of labor, leading to specialized career paths (*16, 19*). Moreover, the results motivate a recent movement to stabilize supporting scientist careers as a policy to address gender inequality in science (*29*). Stabilizing this position will not only contribute to retaining female scientists traditionally in supporting roles but also contribute to reducing the exit of male scientists increasingly in supporting roles. Thus, policies designed to address gender differences in attrition in scientific publishing need to account for field differences and differential career tracks. In particular, the growth of supporting scientists suggests that understanding the intersection between role and gender may be increasingly important when designing programs such as NSF's ADVANCE programs that are designed to improve the retention of female scientists (*30*).

Data and Methods

Our primary data source is SciSciNet (*21*), a comprehensive, meticulously curated, and open-source dataset comprising over 134 million scientific publications. The dataset is specifically tailored for research

in the science of science domain and provides essential measures to examine scientists' publishing careers. For example, the dataset provides precomputed metrics that would otherwise require resource-intensive data processing and computations. We selected four fields from this dataset: chemistry, biology, psychology, and sociology. We selected these fields for several reasons. Firstly, chemistry and biology serve as representative fields of natural sciences, while psychology and sociology embody social sciences. The bureaucratization of scientific work is most evident in team-based science, where producing scientific knowledge involves formalization, specialization, and division of labor. These features are much more evident in natural sciences. In contrast, social sciences are much more closely aligned with the traditional "craft" scientific production model (*15, 26, 31*), thus offering a contrasting perspective in our research. Secondly, our selection was also influenced by the compatibility of authorship norms (*32*) within these fields with our method for categorizing lead versus supporting career types (see next parapgrah). For instance, our classification may not apply to mathematics and economics, where authorship order follows an alphabetical norm. Similarly, fields with hyper-authorship norms, like high-energy physics, may not effectively capture the specialized career trajectories using our method. Lastly, given that our study examines the historical changes in the production roles of scientists, with the time period going back to the 1950s, fields without sufficiently long histories were not considered.

Following the previous method of classifying scientists by their production roles (*17, 19*), we categorize scientists into **lead career types** as those who have published at least one paper as first authors and **supporting career types** as scientists who have never had positions as first authors in their publishing careers (SM A.4) We provide evidence that this simple measure can effectively capture the essential features of the different production roles in science (SM A.4, Fig.6). We find that supporting career types are more likely to engage in labor-intensive tasks such as performing experiments, while lead career types are much more likely to undertake conceptual or abstract tasks or those related to resource allocation. Meanwhile, it is possible that lead career types have longer publishing careers because of the survival bias. However, our data shows that 90% of the lead career types have transited to this category within 5 years from their first publications (SM A.5, Fig.7), suggesting that our classification is stable across publishing careers.

To construct comprehensive publishing careers of scientists from biology, chemistry, psychology and sociology, we first retrieved all authors who appear in the SciSciNet bibliographic records associated with the four selected fields. To ensure that our sample is comprised of true-field authors, we only include authors who have published over half of their papers in the focal fields. Given the primary argument that the bureaucratization of science is most relevant to the US context (SM A.2), we further refined our sample to include only those authors for whom at least half of their published papers indicate their affiliation as being located in the United States. The binary gender of the authors was inferred using data from SciSciNet. This dataset provides a probabilistic variable ranging from 0 (most male) to 1 (most female). We assigned authors with a probability range of [0, 0.1] as male and [0.9, 1] as female (SM. A.3). The proportion of authors whose gender was identified using this method varied across fields. We were able to infer the binary gender for 80.9% of authors in biology, 79.3% in chemistry, 90.0% in psychology, and 90.6% in sociology (SM A.3, Fig.2). Following previous studies (*9, 19*), we excluded "transient authors" from our study, which is defined as authors who have published fewer than two papers and those who have not produced papers in two or more periods. Our final dataset includes 62 cohorts of authors in four fields whose first publications were from 1951 to 2012. We consider authors to have exited their publishing careers, designating their last publication years as the dropout years if they did have any papers published during the 5 year period from 2017 to 2021. If authors have publication during 2017-2021, they were

right-censored in our study as of 2016. The final dataset includes 658,049 authors, with 186,360 from chemistry, 334,111 from biology, 116,346 from psychology, and 21,232 from sociology.

We used two modeling frameworks for survival analysis. Firstly, we used the Cox Proportional Hazard Model due to its flexibility in handling survival data without making assumptions about the survival function. Given our discrete measurement of career length in years, we applied the Efron method to adjust for ties. To address the strict assumption about the constant hazards over time, we also employed the Accelerated Failure Time (AFT) model. Both models were estimated using the R survival package. Our main regression is specified as follows.

$$log\, h_i(t) = \alpha(t) + \beta_1 Women_i + \beta_2 Support_i + \beta_3 Women \times Support + Z_i^T \gamma \qquad \text{eq(1)}$$

W modeled the log hazard of exiting publishing careers as a linear function of gender, career types, and other covariates Vector Z, which include the total number of publications, the average number of first five-year citations (c5), the average number of team size from the first 3 years of publishing careers. We used binary variables indicating the upper 90^{th} percentile in these three early performance variables in our regressions. To account for potential unobserved heterogeneity across cohorts, we incorporated cohort fixed effects into our analysis. Detailed information on our estimation methods and descriptive statistics can be found in Supplementary Materials (SM B1-5).

Figures

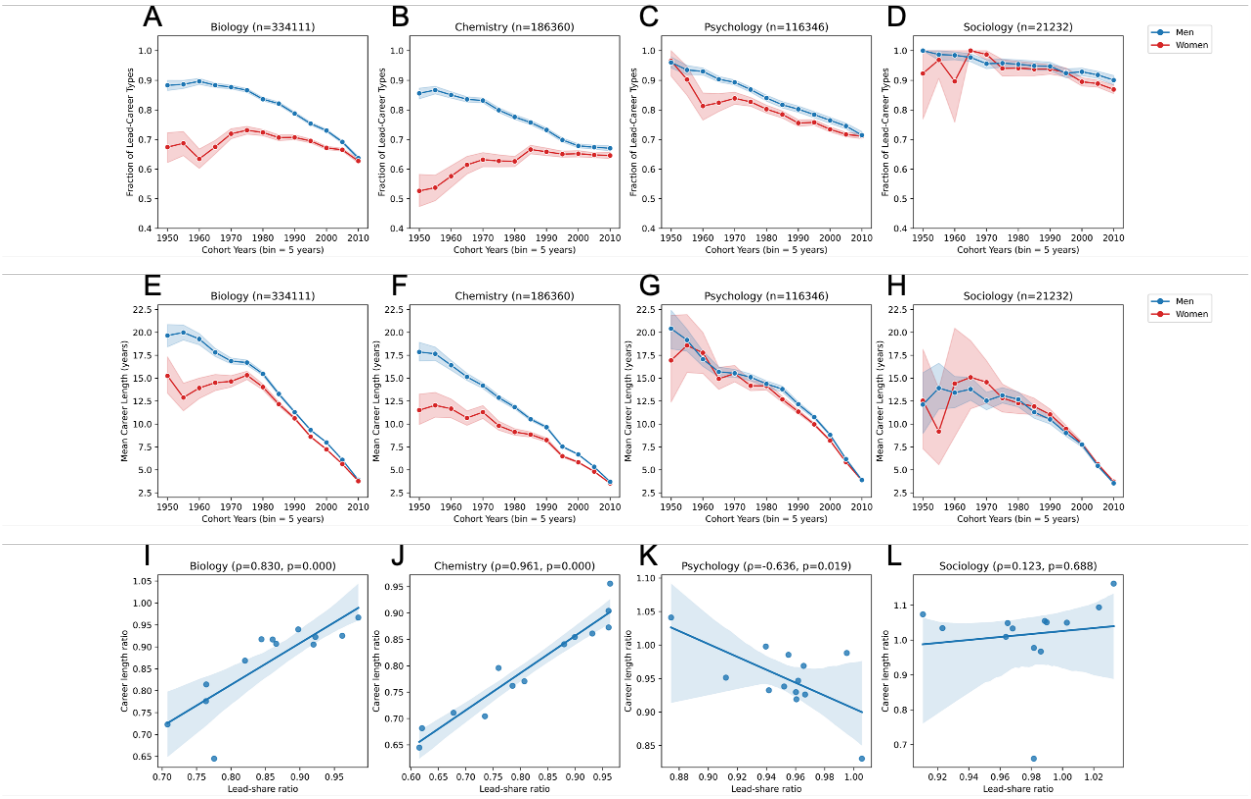

Fig 1. (A-D) Fraction of lead-career types by 5-year cohort, based on binary gender identified from authors' first names. Note that lead-career types are authors who have published at least one lead-author paper during their entire publishing careers. (E-H) Mean publishing career lengths by 5-year cohort, based on binary gender identified from authors' first names. (I-L) Correlation between gender ratio in publishing career length by 5-year cohort and the fraction of lead-career types by 5-year cohort. The vertical axis represents the female-to-male ratio of mean career length, while the horizontal axis represents the female-to-male ratio of the fraction of lead career types.

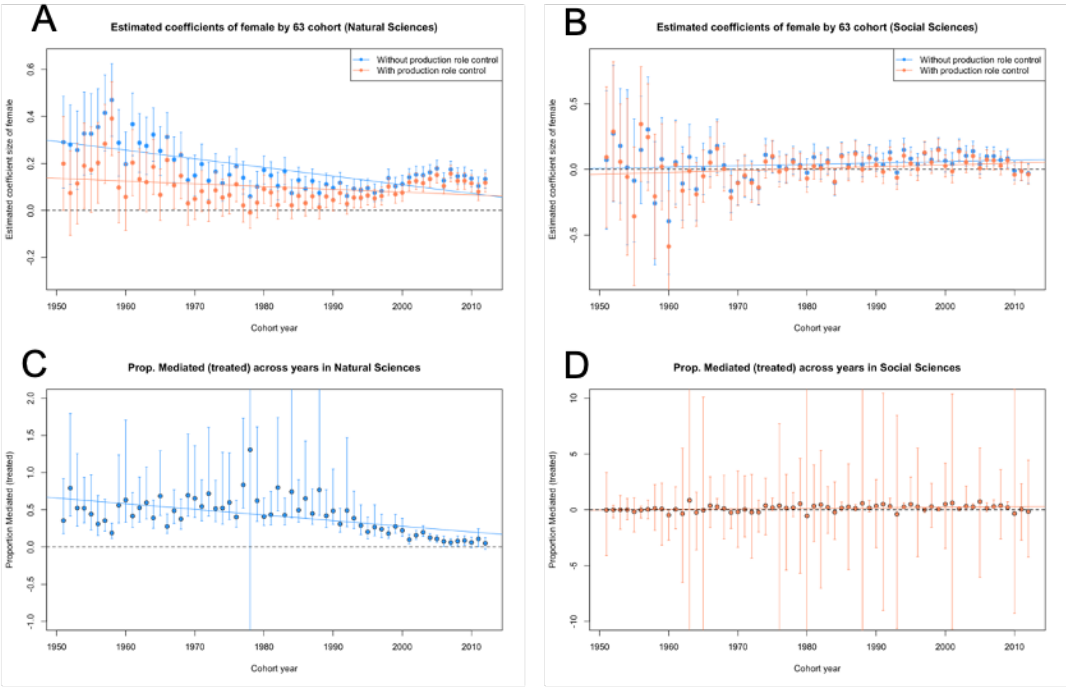

Fig 2. (A-B) Estimated coefficients of women authors by cohorts in natural sciences (A) and social sciences (B). Blue dots represent the coefficients estimated from the Cox Proportional Hazard model, excluding career-type control. The blue line depicts the slope from regressing estimated coefficients without career-types control on cohort years. The red dots represent the estimated coefficients with the career-type control. The red line depicts the slope from regressing estimated coefficients with career-types control on cohort years. Both models include productivity, the number of coauthors, and the average team size from the first 3 years of authors' publishing careers. (C-D) The estimated proportion of the effects of gender on exiting publishing careers mediated by career type across cohorts in natural sciences (C) and social science (D). Both C and D show fitted trend lines indicating the change in the mediated proportion over time, with error bars representing the variability in estimates.

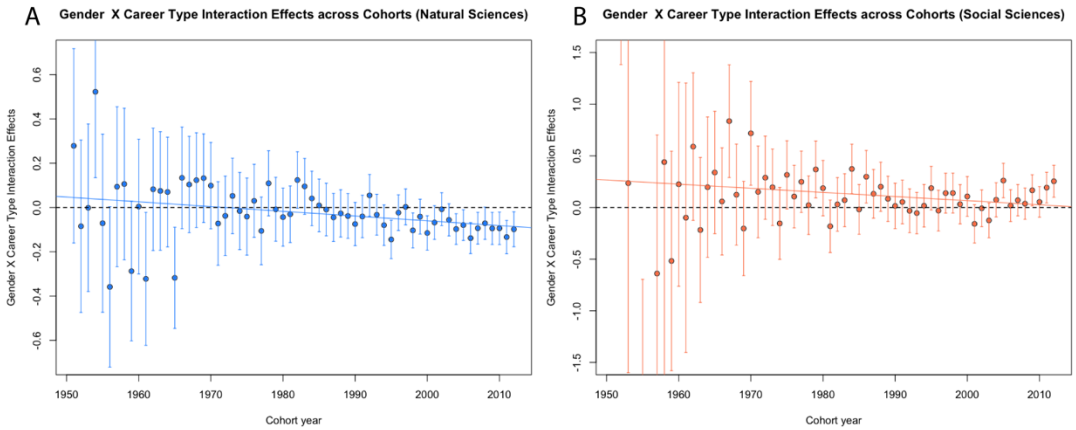

Fig 3. Interaction effects between Gender and Career Types over Cohorts in the natural sciences sample (A) and social sciences sample (B). Interaction effects were estimated using the Cox Proportional Hazard model specified in eq(1) without cohort fixed effects. The vertical axes represent the magnitudes of the estimated interaction effects based on cohort year, which are displayed on the horizontal axes. Vertical lines around the estimated points represent the 95% confidence intervals.

Tables

Table 1. Cox proportional hazard regressions from Natural Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.141***	1.131***	1.093***	1.111***
	(1.003)	(1.003)	(1.003)	(1.004)
support			1.837***	1.872***
			(1.004)	(1.005)
early_product_top		0.660***	0.726***	0.727***
		(1.006)	(1.006)	(1.006)
early_c5_top		0.893***	0.904***	0.904***
		(1.005)	(1.005)	(1.005)
early_teamsize_top		0.942***	0.807***	0.807***
		(1.006)	(1.006)	(1.006)
Genderfemale:support				0.954***
				(1.007)
Observations	520,471	520,471	520,471	520,471
Log Likelihood	-4,866,630.000	-4,863,680.000	-4,850,208.000	-4,850,186.000
LR Test	7,870.759*** (df = 62)	13,769.370*** (df = 65)	40,713.690*** (df = 66)	40,757.920*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 2. Cox proportional hazard regressions from Social Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.056***	1.043***	1.027***	1.004
	(1.007)	(1.007)	(1.007)	(1.008)
support			2.294***	2.185***
			(1.008)	(1.012)
early_product_top		0.545***	0.609***	0.608***
		(1.014)	(1.014)	(1.014)
early_c5_top		0.795***	0.777***	0.776***
		(1.011)	(1.011)	(1.011)
early_teamsize_top		1.206***	0.941***	0.941***
		(1.011)	(1.012)	(1.012)
Genderfemale:support				1.092***
				(1.015)
Observations	137,578	137,578	137,578	137,578
Log Likelihood	-1,076,422.000	-1,074,905.000	-1,070,069.000	-1,070,052.000
LR Test	1,201.837*** (df = 62)	4,236.148*** (df = 65)	13,908.570*** (df = 66)	13,942.450*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

References

1. H. Etzkowitz, C. Kemelgor, B. Uzzi, *Athena unbound: The advancement of women in science and technology*. (Cambridge University Press, 2000).
2. J. S. Long, M. F. Fox, Scientific careers: Universalism and particularism. *Annual review of sociology*, 45-71 (1995).
3. A. E. Preston, *Leaving science*. (Russell Sage Foundation, 2004).
4. H. Zuckerman, J. R. Cole, Women in American science. *Minerva*, 82-102 (1975).
5. J. R. Cimpian, T. H. Kim, Z. T. McDermott, Understanding persistent gender gaps in STEM. *Science* **368**, 1317-1319 (2020).
6. S.-J. Leslie, A. Cimpian, M. Meyer, E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262-265 (2015).
7. E. Reuben, P. Sapienza, L. Zingales, How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences* **111**, 4403-4408 (2014).
8. J. Huang, A. J. Gates, R. Sinatra, A.-L. Barabási, Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, (2020).
9. J. D. Adams, G. C. Black, J. R. Clemmons, P. E. Stephan, Scientific teams and institutional collaborations: Evidence from US universities, 1981–1999. *Research policy* **34**, 259-285 (2005).
10. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036-1039 (2007).
11. M. Weber. (CA: University of California Press, 1978).
12. D. S. Pugh, D. J. Hickson, C. R. Hinings, C. Turner, Dimensions of organization structure. *Administrative science quarterly*, 65-105 (1968).
13. J. P. Walsh, Y.-N. Lee, The bureaucratization of science. *Research Policy* **44**, 1584-1600 (2015).
14. W. O. Hagstrom, Traditional and modern forms of scientific teamwork. *Administrative Science Quarterly*, 241-263 (1964).
15. E. J. Hackett, Science as a vocation in the 1990s: The changing organizational culture of academic science. *The journal of higher education* **61**, 241-279 (1990).
16. Y.-N. Lee, J. P. Walsh, Rethinking science as a vocation: One hundred years of bureaucratization of academic science. *Science, technology, & human values* **47**, 1057-1085 (2022).
17. S. Milojević, F. Radicchi, J. P. Walsh, Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences* **115**, 12616-12623 (2018).
18. C. R. Sugimoto, V. Larivière, *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*. (Harvard University Press, 2023).
19. B. Maddox, *Rosalind Franklin: The dark lady of DNA*. (HarperCollins New York, 2002).
20. Z. Lin, Y. Yin, L. Liu, D. Wang, SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* **10**, 315 (2023).
21. F. Xu, L. Wu, J. Evans, Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences* **119**, e2200927119 (2022).
22. T. K. Kelly, W. P. Butz, S. Carroll, D. M. Adamson, G. Bloom, "The US scientific and technical workforce: improving data for decisionmaking," (RAND CORP SANTA MONICA CA, 2004).
23. S. Milojević, F. Radicchi, J. P. Walsh, Reply to Hanlon: Transitions in science careers. *Proceedings of the National Academy of Sciences* **116**, 17625-17626 (2019).
24. H. Zuckerman, R. K. Merton, Patterns of evaluation in science: Institutionalisation, structure and functions of the referee system. *Minerva*, 66-100 (1971).
25. L. Roberts, S. Rockman, A. Hui, Historiographies of science and labor: From past perspectives to future possibilities. *History of Science* **61**, 448-474 (2023).
26. J. S. Long, M. F. Fox, Scientific careers: Universalism and particularism. *Annual review of sociology* **21**, 45-71 (1995).

27. M. F. Fox, Women and scientific careers. *Handbook of science and technology studies*, 205-223 (1995).
28. K. Imai, L. Keele, D. Tingley, A general approach to causal mediation analysis. *Psychological Methods* **15**, 309-334 (2010).
29. L. L. Hargens, Patterns of scientific research. *Washington, DC: American Sociological Association*, (1975).
30. J. P. Walsh, Y.-N. Lee, in *ResearchProfessional News*. (2022).
31. D. Peterson, All that is solid: Bench-building at the frontiers of two experimental sciences. *American Sociological Review* **80**, 1201-1225 (2015).
32. S. Jabbehdari, J. P. Walsh, Authorship norms and project structures in science. *Science, Technology, & Human Values* **42**, 872-900 (2017).

Supplemental Materials

Gender and Attrition in Research Careers

Seokkyun Woo

You-NA Lee

Appendix A: Dataset and Descriptive Results

A1. Introduction

In this paper, we empirically examine the extent to which the career type of scientists is gendered, how the gendering of career type explains existing inequality in publishing careers in science, and how career type differentially affects (by gender) the publishing careers of scientists. To address these empirical questions, we need to construct comprehensive publishing careers of scientists in our selected fields. We do this by obtaining the following information about each scientist: the time in which scientists entered and left their publishing careers, taking into account any right censoring, their binary gender, career type (lead vs. supporting career types), and various measures of their early-career academic activities. All of this information is sourced from bibliographic records available through SciSciNet(*21*) in the fields of biology, chemistry, psychology, and sociology.

A2. Dataset

Our main bibliographic dataset comes from SciSciNet (*21*), a large-scale open data lake that incorporates open-source bibliometric datasets covering over 134 million scientific publications. The advantage of this dataset for our research is not only that it is open source but also that it provides crucial information for the operationalization of our key variables, including disambiguated authors and full names, which can be used to infer binary gender information. In SciSciNet, there are a total of 311 fields, consisting of 19 top-level fields and 292 sub-level fields. Of 117,633,905 bibliographical records with field information, 99.38% are assigned with at least one top-level field. Furthermore, 99.35% of them are assigned a single top-level field. So, most records are assigned with single top-level fields. In our paper, we concentrate on extracting comprehensive bibliographic records associated with the domains of Biology, Chemistry, Psychology, and Sociology, which fall within these 19 top-level fields.

To construct the publishing careers of scientists in these four fields, our approach involves initially identifying all authors from bibliographic records within each of these fields. Subsequently, since some of these authors may also publish papers beyond the selected fields, we collect all publication records authored by these individuals that fall outside of the selected fields. While our analysis is operationalized at the field-author level, including authors who have published at least one paper in a focal field may not be desirable. We use authors who have published at least half of their publications in the focal field as focal-field authors. Previous studies documented the growth of transient authors, authors who have only published a single paper in their careers. To address the inflated count of supporting types, we exclude transient authors in similar manners defined by previous studies (*9, 19*), as individuals who have published

only one paper and have a career duration of fewer than two years. Additionally, we excluded authors who have published an excessive number of papers, specifically more than 20 in a single year.

Meanwhile, given that different countries have different science institutions, in particular ways in which they provide permanent positions to scientists, we exclusively focused on US scientists. This is particularly important because, in the US, basic science research is mostly led by universities, and the role of non-university public research institutions is not as pronounced as in Europe (cite). Thus, we consider an author to be a US author if more than 50% of the papers are published with US-affiliated organizations. Finally, we isolated our sample authors to those authors who started to publish from 1951 to 2012, which included a total of 62 cohorts. After excluding authors whose binary gender information cannot be inferred (see next section), our final career dataset is comprised of 334,111 authors from the field of biology, 186,360 from chemistry, 116,346 from psychology, and 21,232 from sociology.

A3. Gender Disambiguation

In order to proxy a binary gender of an author, we relied on SciSciNet's gender identification score, which built upon the work of Van Buskirk, Clauset and Larremore (*33*). This gender score falls within the range of (0,1), where a score of 0 indicates a strong male association and a score of 1 suggests a strong female association. Our data shows that distribution of these scores is predominantly clustered around values close to 0 and close to 1 (Fig. 1). In our paper, we adopt the following criteria for determining an author's gender: An author is categorized as a woman if their gender score is less than or equal to 0.1. Conversely, an author is classified as a man if their gender score is greater than or equal to 0.9.

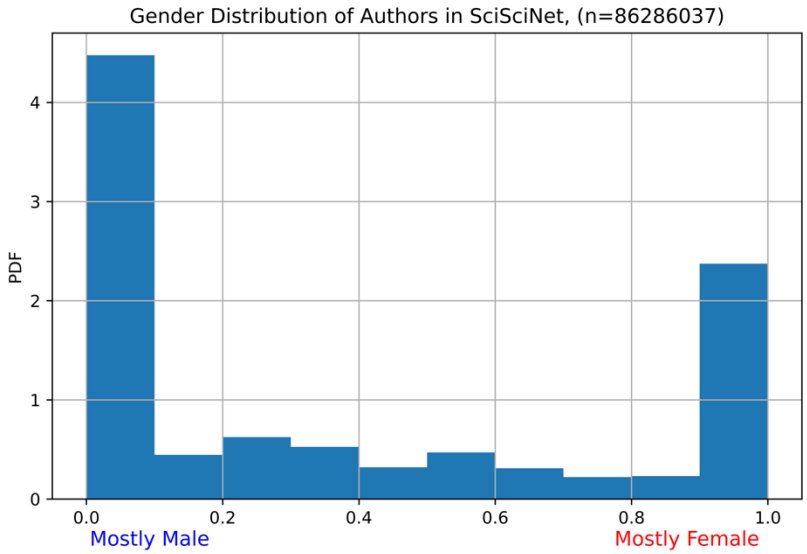

Fig 1. Gender distribution of authors in SciSciNet (n= 86286037). The score is close to 0 if an author's name is more male, while it is close to 1 if the name is more female.

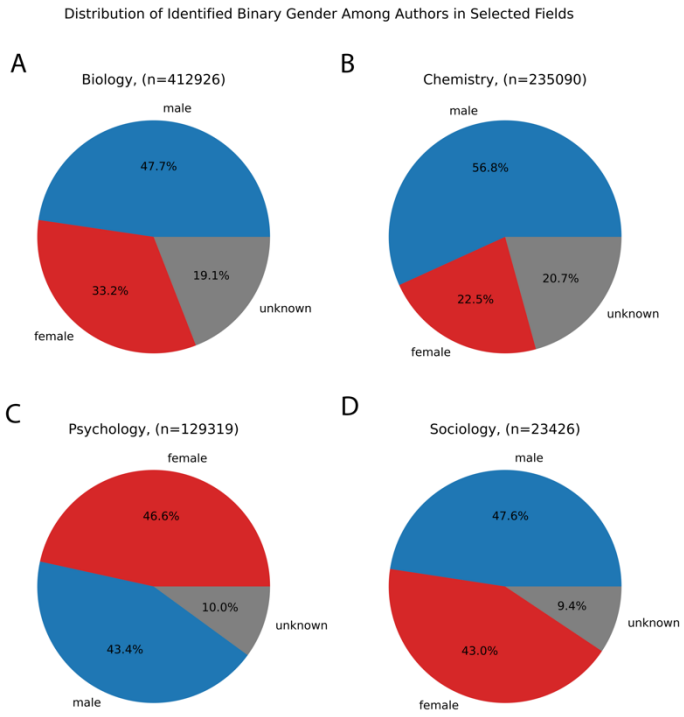

Fig 2. Distribution of identified binary gender among authors in four selected fields. An author is categorized as a "female" if their gender score is less than or equal to 0.1. Conversely, an author is classified as a "male" if their gender score is greater than or equal to 0.9. We code the remaining authors as "unknown."

Fig 2. illustrates the fraction of identified genders from the cohorts of authors whose names start to appear in their respective fields from 1951 to 2012. The fraction of authors that were not identified ranges from 9.4% in sociology to 20.7% in chemistry. We code these as missing and drop them from our final data. We compared our identified gender composition with the reported statistics from the NSF's Survey of Earned Doctorates (SED) from 2003 to 2013 (*34*). Our identified gender ratios, in terms of direction and trend, were mostly consistent with the SED statistics (Fig. 3). The observed discrepancy in the magnitude may be due to a distinction between those who publish in these fields and those who have PhDs in these fields, highlighting that our population of interest is those who publish in journals covered by SciSciNet. These results suggest that the population based on the degree of PhD may have a different gender mix than the publishing researcher population (*35*).

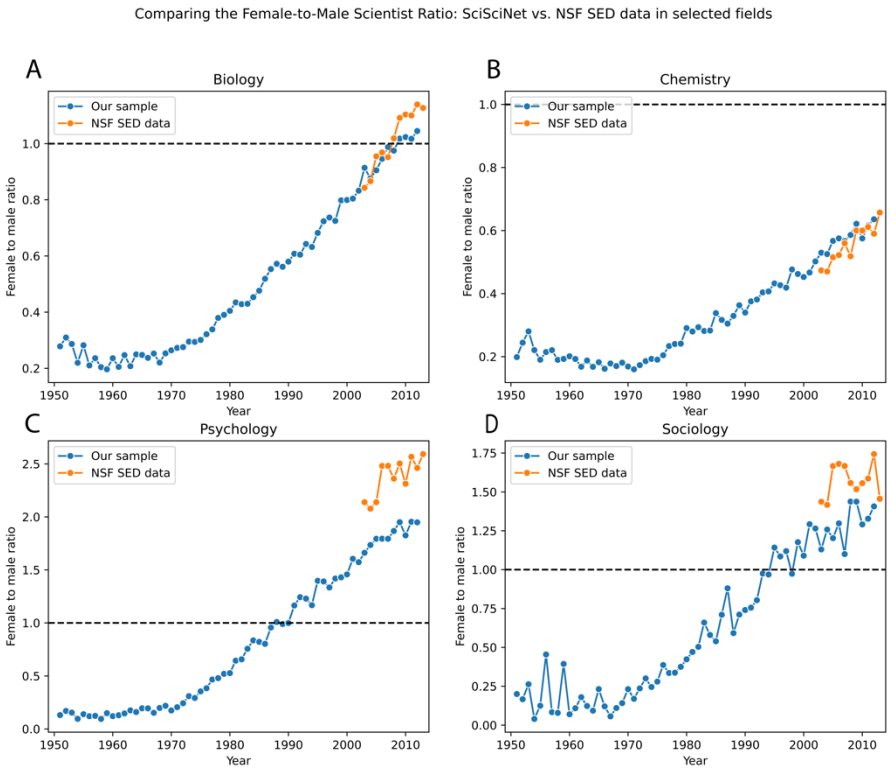

Fig 3. Comparison of Female-to-Male Scientist Ratios: Our Data vs. NSF Survey of Earned Doctorates (SED). The y-axis represents the female-to-male scientist ratio, while the x-axis depicts the doctoral receiving year for NSF SED (in orange) and the years in which scientists start their publishing careers (in blue), which comes from our final dataset. A y-axis value greater than 1 indicates the overrepresentation of women. Please note that SED data provides the number of doctoral recipients by sex from 2003 to 2013.

A4. Career Types

A4.1 Career Types Classification

Prior studies have discussed and documented the increasing role differentiation in science (*14-16, 19, 26*). To measure the presence of this divergence in the scientific workforce in terms of the roles they play in knowledge production, we classify each author into one of two types: lead career type and supporting career type (*17, 19*). Lead career types are authors who, at any point in their careers, have produced at least one paper as first authors. In our data, an author of a single-authored paper is categorized as a lead career type. For multiple-authored papers, we define lead career type as the first author. Supporting career types are the authors who, at any point in their careers, have not appeared in a publication as a first author (Fig. 4).

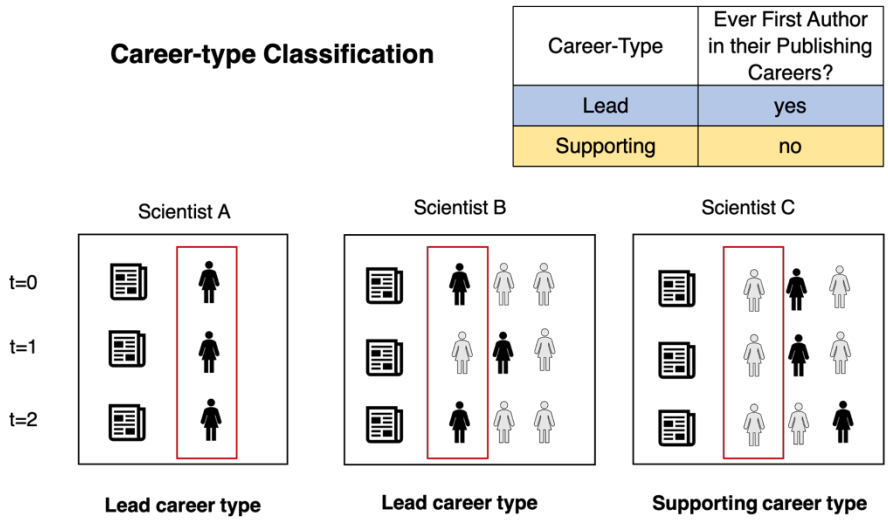

Fig 4. Classification of career type. This figure demonstrates the coding scheme used to classify scientists into lead vs. supporting career types based on the authorship position of scientists' publication history. Scientists are classified as lead career types if they have ever published a first-authored paper in their publishing careers. Conversely, scientists are classified as supporting career types if they have never had a first-author position in their publishing careers.

As suggested by the existing literature on the organization of science, the work of supporting career types is exemplified by labor-intensive tasks, such as performing experiments and generating data, while lead career types focus on abstract and conceptual tasks, such as research design and formulating research questions. We attempt to validate these ideal types by exploiting information on authorship contribution roles (*17, 36, 37*).

A4.2 Career Types and Task Division

To obtain task division information from the contribution statements, we used a dataset provided by, Lin, Frey and Wu (*38*). This dataset (herein, the LFW dataset) provides information from 57,887 records about which authors participated in one of the four standardized tasks in paper production: "conceived," "wrote," "analyzed," and "performed." The LFW dataset covers four journal sources: Science, Nature, PNAS, and PLOS ONE. By matching authors' MAGIDs (Microsoft Academic Graph IDs) from these records, we were able to assign career types to 30,829 authors in biology, 3,755 authors in chemistry, 1,288 authors in psychology, and 5 authors in sociology. The limited coverage in sociology may be attributed to the tendency of sociologists to publish less frequently in general audience journals. Due to the small sample size, we excluded sociology from the validation exercise.

From this merged LFW dataset, we report the fraction of tasks performed across authors and fields by career type (Fig. 6A-L). Our findings suggest a striking difference between lead and supporting career types in terms of task specializations. For example, in biology, more than half of the lead career types have performed "conceived" roles from the list of their publications (Fig. 6A). This contrasts sharply with only 16.7% of supporting career types performing this task (Fig. 6A). This pattern is consistent across fields, with chemistry showing 49.3% vs. 23.6% (Fig. 6E) and psychology with 62.6% vs. 20.8% (Fig. 6I) for lead vs. supporting career types, respectively. Moreover, this division of labor is similarly evident in

the task involving writing drafts (Fig. 6B,F,J). Meanwhile, the difference between the career types is less pronounced when it comes to analyzing data (Fig. 6C,G,K). In psychology, the differences between lead vs. supporting career types are not statistically significant (Fig. 6K). Looking at the tasks involving bodily labor, such as "performance" (i.e., performing experiments and producing data), a clear division of labor emerges. Supporting career types are overwhelmingly engaged in this task. For example, in biology, 73.2% of supporting career types have undertaken this task from the list of their published papers, compared to 60.6% of the lead career types (Fig. 6D). A similar pattern is observed in chemistry (Fig. 6H) and psychology (Fig. 6L). Thus, by linking our dataset to the contribution statements, we demonstrate that our operationalization of lead vs. supporting career types using authorship positions from their publication history is consistent with our priors about the role of supporting scientists in the increasingly bureaucratized production of science (*15, 16, 18*).

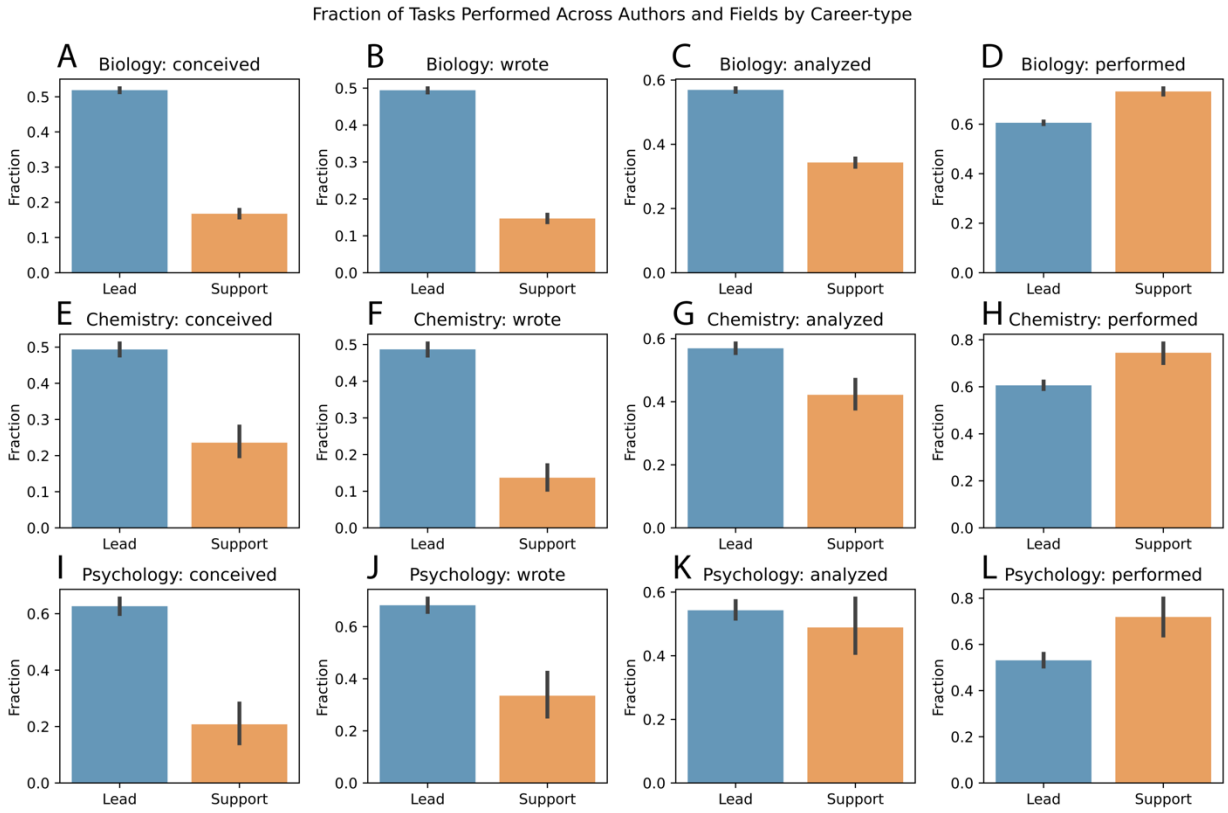

Fig 6. Fraction of tasks performed across authors by career types in Biology (subplots A-D), Chemistry (subplots E-H), and Psychology (subplots I-L). Each subplot illustrates the proportion of lead or supporting career types performing tasks based on available publication records with contribution statement information. The first row represents 30,829 authors from biology, the second row includes 3,755 authors in chemistry, and the last row is from 1,288 authors from psychology.

A5. Career Type Transition

In our paper, our goal is to investigate differential publishing career longevity by gender and career types. As previously explained, we categorize authors as either lead or supporting careers based on their publication history using authorship positions. However, it is possible that among lead career types, some

may have transitioned to this role only later in their careers. This could exacerbate measurement errors with a bias towards the increasing longevity of lead-career types as those with longer careers have more chances to eventually become lead scientists. To address this concern, we examine the distribution of career transition times among lead-career authors to examine whether such cases are prevalent in our sample data. Specifically, we present a distribution of time-to-lead for eventual lead-career types (Fig. 7).

The majority of lead scientists have transited to lead-career type early in their publishing careers (Fig. 7). For both men and women lead-career types, 90% achieve this status, varying by field: 5 years for Biology, 4 years for Chemistry and Psychology, and 2 years for Sociology. Additionally, our analysis indicates that gender does not significantly influence this career progression in natural sciences. However, in the social sciences (psychology and sociology), women scientists take an additional year to reach a lead-career status compared to their male counterparts, with 90% of women achieving this status in one year longer than men. Our findings suggest that the distinction between lead and supporting career types is a more deterministic feature of scientists' career paths. While our data does show that some scientists assume lead roles later in their careers, the majority make these transitions early on. These findings further support the evidence that the production of science is becoming increasingly bureaucratized, with a clear bifurcation of career paths into lead and supporting roles (*14, 16, 19*).

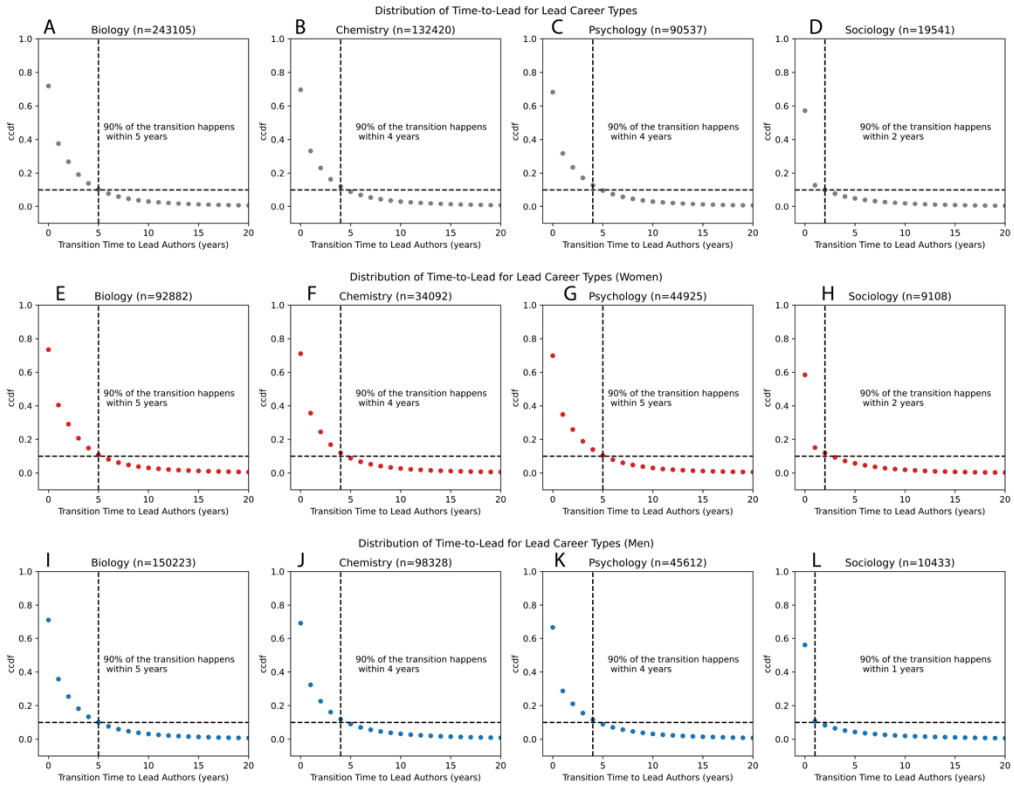

Fig 7. This figure displays the distribution of time-to-lead career types across all authors in the sample (subplots A), women scientists (subplots E-H), and men scientists (subplots I-L). The horizontal axes represent the transition time to reach lead status in years, while the vertical axes depict the complementary cumulative distribution function (CCDF).

A6. Career Longevity

One difficulty with debates on the careers of scientists is the lack of or the continually shifting definition of the STEM workforce (*39, 40*). While a strict definition could only include those who currently are in academia, a broader definition could define the STEM workforce by the extent to which a person's current occupation is related to STEM or the extent to which a person's field of degree is related to STEM. Indeed, a person who stopped publishing in peer-reviewed journals and pursues a non-academic career can still contribute to science in many ways (*41*). Therefore, who should be considered a scientist or an active scientist is a non-trivial question. In our case, we explicitly focus on those who have contributed to the production of science by publishing papers in peer-reviewed journals, which is the major platform by which scientists communicate their findings (*42*). Using this criterion, we identify authors and their careers by the incidence and duration of their peer-reviewed publication activities. In our definition, each scientist starts their career when he or she first appears on our field-level data. If the authors' names appear in the last five years of our dataset, which is during 2017-2021, we consider their careers as right truncated. If their names do not appear during that period, we consider that they have exited at the point of their last publications. For authors whose names only appear once in our data, we conceive of them as transient authors and subsequently drop them from our sample of scientists.

One advantage of our method is that we tackle inequalities in role and career by gender within the same cohort of authors, thereby avoiding survival bias and other forms of bias created by the changing demography of the scientific workforce. We identify cohorts of authors who first start to appear in their respective fields from 1951 to 2012. We stop at the 2012 cohort to ensure that we capture sufficient career lengths in the analysis. Note that our bibliographic records from 2017 to 2021 are used to identify active status for all authors in our data. For example, an author whose name first appears in 2012 can have a maximum career length of 4 years such that depending on the appearance of his or her name in the last three years of our dataset (2017-2021), his or her career can either be right-truncated or ends by the 4th year. Focusing on authors whose primary institutions are in the US (see section), our final data includes full career information for 375,525 authors from natural sciences and 110,078 from social sciences. At the field level, we have 243,105 authors from biology, 132,420 authors from chemistry, 90,537 authors from psychology, and 19,541 authors from sociology (See Table).

A7. Non-parametric estimations of career longevity by career types and gender

In this section, we present non-parametric Kaplan-Meier survival curves comparing lead and supporting career types (Fig. 8). The data is derived from the comprehensive cohort of individuals who entered publishing careers between 1951 and 2012 (Fig. 8A-D). To provide a more contemporary perspective, we also include Kaplan-Meier curves for a subset of scientists who entered publishing careers from 2001 to 2012 (Fig. 8E-H). Just looking visually, there is a stark difference between lead and supporting roles, with the supporting career types showing significantly shorter career longevity when compared to their lead counterparts. For instance, in the field of biology, the half-life of scientists' publishing careers, which estimates the time it takes for half of the initial cohort to exit their publishing career, is 9 years for lead career types and 4 years for supporting career types. This difference between the two distributions is statistically significant based on non-parametric tests involving the log-rank test and the Wilcoxon test (Fig. 8A). Moreover, this pattern is consistent across different fields (Fig. 8B-D) and for recent cohorts (Fig. 8E-H).

We also present Kaplan-Meier curves representing the publishing careers of male and female scientists (Fig. 9). In the fields of biology, chemistry, and psychology, male scientists have longer half-lives compared to their female counterparts. The differences in their distributions of career longevity are statistically significant, as evidenced by both the log-rank test and the Wilcoxon test (Fig. 9A-C, E-G). However, an opposite pattern is observed in sociology, where women exhibit longer half-lives than men. This difference in distribution is statistically significant, confirmed by the Wilcoxon test for the entire sample (Fig. 9D) and the Log-rank test for the recent cohort sample (Fig 9G).

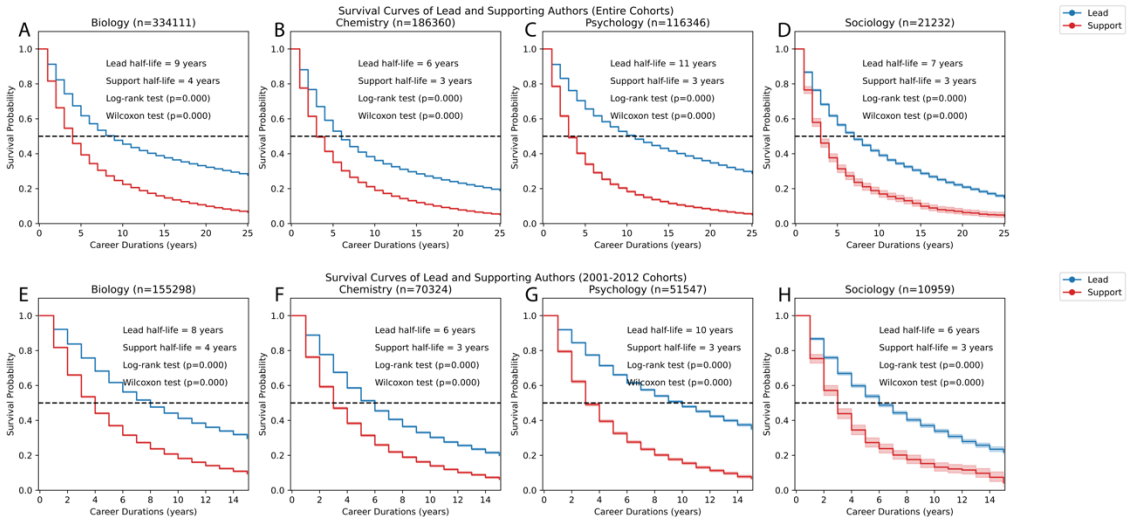

Fig 8. Kaplan-Meier Estimated Survival Curves for Lead and Supporting Career Types for cohorts from the entire period (A-D) and for the post-2000 cohorts. Horizontal axes depict career durations in years, while the vertical axes represent survival probability given career durations. Blue curves indicate the survival curves for the lead career types while supporting career types are depicted in red.

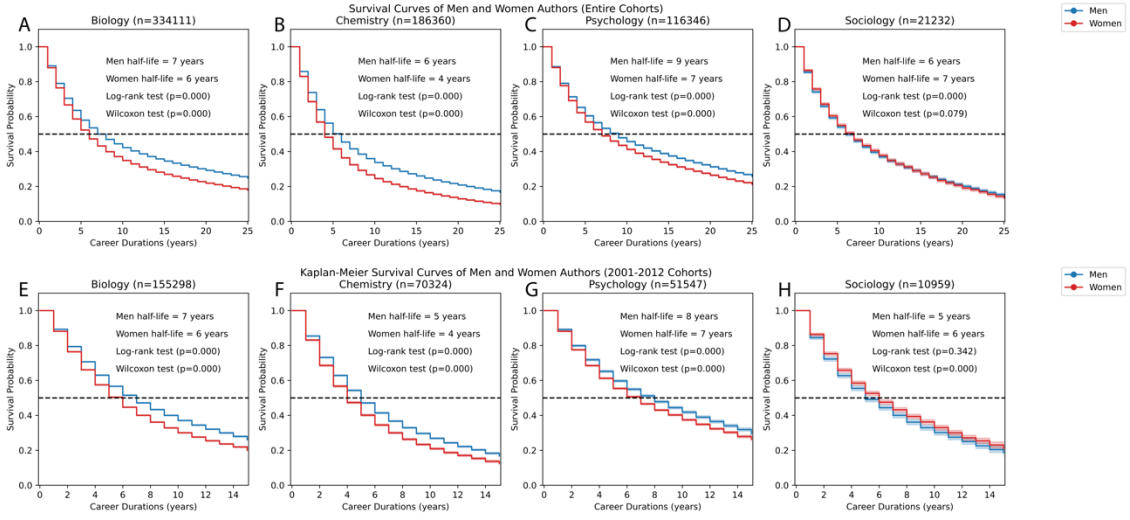

Fig 9. Kaplan-Meier Estimated Survival Curves for Women and Men authors for cohorts from the entire period (A-D) and for the post-2000 cohorts. Horizontal axes depict career durations in years, while vertical axes represent survival probability given career durations. Blue curves indicate the survival curves for the lead career types while supporting career types are depicted in red.

Appendix B: Regression Estimations

The regression table and figure in the main manuscript are based on estimates from a multivariate analysis using the Cox proportional hazard model (*43*) and the Accelerated Failure Model (AFT) based on Weibull distribution. In this section, we briefly describe our models and specifications.

B1. Cox Proportional hazard model

The proportional hazard model is known to be the most flexible method to estimate survival data, with its advantage being avoiding making any assumptions about the function of time in the model (*44*). Because of the discrete nature of our measurement of career length, which was measured in years, we used the Efron method to correct for ties.

We model the scientist's career attrition using the continuous hazard function shown: $h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$, with *h(t)* as the instantaneous hazard rate at which a scientist exits his or her field at time *T* between interval $[t, t + \delta]$ given that the scientist was at risk of leaving science at time *t*. We model this hazard function with the Cox proportional hazard model as follows: $h(t|X) = h_0(t)exp(X\beta)$, where $h(t|X)$ is the hazard at time t for a scientist with a given set of covariates X. $h_0(t)$ is the time-dependent baseline hazard, representing the hazard for a scientist with the baseline values of the covariates. Taking the logarithm of this function, we get the following log-hazard function with our model specification.

$$log\, h_i(t) = \alpha(t) + \beta_1 Women_i + \beta_2 Support_i + \beta_3 Women \times Support + Z_i^T \gamma \qquad \text{eq(1)}$$

The proportional hazard model is estimated with the partial likelihood method, which discards the time-dependent variable $\alpha(t)$, which is assumed to be the same for all scientists, allowing us only to parameterize the effects of time-independent variables. In this specification, β_1 captures the difference in log-hazard between women and men for lead career types, while β_2 captures the difference between supporting and lead career types for men. B_3 captures the interaction effect between gender and career type, which can be interpreted as the difference in the supporting career type effects on log hazards between women and men.

For example, $\beta_1 > 0$ would suggest that the log-hazard of exiting publishing career for women is greater than that of men. Exponentiating this coefficient, e^{β_1}, would give the hazard ratio, indicating the factor by which the exit hazard for women is multiplied relative to men. Lastly, γ is the vector of parameters for our control covariates vector Z. Our control variables include three measures of the scientists' early academic performance: scientists' number of publications, average c5 citations (citations from the first 5 years of publications), and the average team size during the first three years of their careers. We also include a set of binary variables that correspond to the year in which the author starts to publish, which allows us to control for any cohort-specific effects on log hazards.

B2. Accelerated Failure Time (AFT) model

It is important to note that the proportional hazard model assumes that hazards attributed to covariates are constant in time (*45*), an assumption that could be violated when considering the different empirical

survival functions observed between lead and supporting scientists (see Fig. 9). To address this issue, we consider an alternative approach using an accelerated failure time model (AFT). This model directly estimates the survival time, T_i, rather than the hazard ratios.

We use Weibull distribution in the AFT model due to its flexibility in accommodating increasing, constant, or decreasing hazard rates depending on its shape parameter ρ. The survival function of the Weibull distribution is given by $S(t) = exp(-(\lambda t)^{\rho})$, where, ρ is a shape parameter and λ is a scale parameter. The general form of the AFT model is: $log(T_i) = \alpha + X_i\beta + \sigma \mathrm{W}_i$, where $log(T_i)$ is the natural logarithm of the survival time for the i[th] scientist, α is the intercept, X_i is the vector of covariates for the i[th] scientist, β is the vector of coefficients indicating the effect of the covariates on the log-transformed survival time, σ is the scale parameter for the error term W_i. The model specification is identical to those used for the cox proportional hazard model (see eq(1)). To integrate the AFT model with the Weibull distribution, we can parameterize the scale parameter λ of the Weibull distribution as a function of the linear predictor such that $\lambda = exp\left(-\frac{\alpha + X\beta}{\rho}\right)$. This function is then used to substitute λ into the Weibull probability density function, $f(t; \lambda, \rho) = \rho\lambda(\lambda t)^{\rho-1}e^{(-\lambda t)^{\rho}}$, for maximum likelihood estimation.

In the context of the AFT model, the coefficient vectors β represent the covariates' effect on the log of survival time. For example, β_1 from eq(1) would indicate the difference in the log of the expected survival time between women and men. If $\beta_1 > 0$, for binary covariate like gender, it indicates that the expected log survival time for women is greater than that for men. Exponentiating the coefficient, e^{β_1}, would provide the multiplicative effect on the survival time, i.e., for women compared to men. All regression estimates were performed using the "survival" package in R (*46*) for both the Cox proportional hazards model and the AFT (Accelerated Failure Time) model with a Weibull distribution.

B3. Mediation Analysis

We performed a mediation analysis to examine how gender inequality is realized via the type of career paths that scientists take in the production of science. According to Imai, Keele, Tingley and Yamamoto (*47*), the conventional mediation analysis may potentially lead to biased estimations when the mediator variable is not randomly assigned. Moreover, the conventional mediation analysis makes it intractable to work with non-linear models. We adopted the average causal mediation effect (ACME) framework (*48*) to address the non-random assignment of the mediator variable (Fig.2 C,D in the main manuscript).

The average casual mediation analysis requires sequential ignorability assumptions (*47*), which involve two key features. Firstly, it assumes that the treatment (gender) is independent of the mediator (career type) and outcome (hazard rate), conditional on observed covariates (early performances). Secondly, it assumes that the mediator (career type) is independent of the outcome (hazard rate) conditional on the treatment and covariates, suggesting that there are no unobserved confounders influencing both the mediator and outcome. We posit that the first part of the sequential ignorability assumption is plausible in our case, given that our measure of gender may highly correlate with a biological characteristic. The second feature assumes the independence of career type from the hazard rate, conditional on observed covariates, including gender. While we show that the career transition into lead career types occur early in their careers (fig), we cannot rule out potential unobserved factors that may affect the career transition and the outcome variable. This is a limitation of our study that relies on observational data.

To estimate the average casual mediation effect (ACME), we first fit a linear regression model, the mediator model, which explains the variation in the mediator variable (career type) based on our covariates, including gender and three early performance measures (total number of publications, average number of the first five-year citations (c5), average number of team size from the first 3 years of publishing careers). We then performed the Cox proportional hazard model with the specification shown in eq(1) to examine how the mediator (career type) and other covariates, including gender, jointly influence the hazard rate. These two models are used to calculate the ACME (represents the treatment effect mediated through the mediator) and the total effect (total effect of the treatment on the outcome). The proportion of the gender effect on career longevity, mediated by career type, is calculated as the ratio of ACME to the total effect. This mediation analysis is conducted for each cohort, ranging from 1951 to 2012, and the results are depicted in Figures 2C and 2D in our main manuscript.

B4. Descriptive Statistics

Table. 1 Descriptive Statistics by Fields

	N	mean	std	min	25%	50%	75%	max
Natural Sciences (Biology and Chemistry)								
Career lengths	520,471	9.09	9.75	1	3	5	12	65
Droped out	520,471	0.75	0.43	0	1	1	1	1
Women	520,471	0.37	0.48	0	0	0	1	1
Lead	520,471	0.72	0.45	0	0	1	1	1
Early pubs	520,471	2.95	2.18	1	2	2	4	52
Avg. early c5	520,471	22.54	55.95	0	5.67	12.53	25	17310.5
Avg. early teamsize	520,471	5.97	24.2	1	3	4.5	6.5	5363
Cohort	520,471	1994.85	13.76	1951	1987	1998	2006	2012
Social Sciences (Psychology and Sociology)								
Career lengths	137,578	9.71	9.95	1	3	6	13	65
Droped out	137,578	0.71	0.45	0	0	1	1	1
Women	137,578	0.51	0.5	0	0	1	1	1
Lead	137,578	0.8	0.4	0	1	1	1	1
Early pubs	137,578	2.48	1.9	1	1	2	3	35
Avg. early c5	137,578	9.83	16.66	0	1.6	5	12.33	1538.5
Avg. early teamsize	137,578	3.34	6.45	1	1.85	2.75	4	996.9
Cohort	137,578	1994.83	14.27	1951	1985	1999	2007	2012
Biology								
Career lengths	334,111	9.37	9.69	1	3	6	12	65
Droped out	334,111	0.72	0.45	0	0	1	1	1
Women	334,111	0.41	0.49	0	0	0	1	1
Lead	334,111	0.73	0.45	0	0	1	1	1
Early pubs	334,111	2.94	2.16	1	2	2	4	52
Avg. early c5	334,111	27.18	67.89	0	7	15.14	29.67	17310.5
Avg. early teamsize	334,111	6.58	29.69	1	3.33	4.89	7	5363

Cohort	334,111	1996.12	13	1951	1989	1999	2006	2012
Chemistry								
Career lengths	186,360	8.58	9.83	1	2	5	11	65
Droped out	186,360	0.83	0.38	0	1	1	1	1
Women	186,360	0.28	0.45	0	0	0	1	1
Lead	186,360	0.71	0.45	0	0	1	1	1
Early pubs	186,360	2.97	2.22	1	2	2	4	37
Avg. early c5	186,360	14.22	19.31	0	4.25	9.18	17.6	1361.5
Avg. early teamsize	186,360	4.89	7.37	1	3	4	5.67	798.5
Cohort	186,360	1992.58	14.76	1951	1982	1996	2005	2012
Psychology								
Career lengths	116,346	9.97	10.2	1	3	6	13	65
Droped out	116,346	0.7	0.46	0	0	1	1	1
Women	116,346	0.52	0.5	0	0	1	1	1
Lead	116,346	0.78	0.42	0	1	1	1	1
Early pubs	116,346	2.6	1.98	1	1	2	3	35
Avg. early c5	116,346	10.99	17.64	0	2.2	6	14	1538.5
Avg. early teamsize	116,346	3.61	6.67	1	2	3	4.27	996.9
Cohort	116,346	1994.47	14.35	1951	1984	1998	2007	2012
Sociology								
Career lengths	21,232	8.28	8.29	1	2	5	11	65
Droped out	21,232	0.75	0.43	0	1	1	1	1
Women	21,232	0.47	0.5	0	0	0	1	1
Lead	21,232	0.92	0.27	0	1	1	1	1
Early pubs	21,232	1.85	1.18	1	1	2	2	17
Avg. early c5	21,232	3.45	6.74	0	0	1.3	4	201
Avg. early teamsize	21,232	1.83	4.81	1	1	1.25	2	670
Cohort	21,232	1996.84	13.68	1951	1989	2001	2008	2012

B5. Regression results

B5.1 Natural Sciences

Table 2. Cox proportional hazard regressions from Natural Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.141***	1.131***	1.093***	1.111***
	(1.003)	(1.003)	(1.003)	(1.004)
support			1.837***	1.872***
			(1.004)	(1.005)
early_product_top		0.660***	0.726***	0.727***
		(1.006)	(1.006)	(1.006)
early_c5_top		0.893***	0.904***	0.904***
		(1.005)	(1.005)	(1.005)
early_teamsize_top		0.942***	0.807***	0.807***
		(1.006)	(1.006)	(1.006)
Genderfemale:support				0.954***
				(1.007)
Observations	520,471	520,471	520,471	520,471
Log Likelihood	-4,866,630.000	-4,863,680.000	-4,850,208.000	-4,850,186.000
LR Test	7,870.759*** (df = 62)	13,769.370*** (df = 65)	40,713.690*** (df = 66)	40,757.920*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 3. Weibull regressions from Natural Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.865***	0.873***	0.909***	0.888***
	(1.004)	(1.004)	(1.003)	(1.004)
support			0.522***	0.508***
			(1.004)	(1.005)
early_product_top		1.568***	1.401***	1.399***
		(1.006)	(1.006)	(1.006)
early_c5_top		1.135***	1.116***	1.116***
		(1.006)	(1.005)	(1.005)
early_teamsize_top		1.066***	1.251***	1.251***
		(1.006)	(1.006)	(1.006)
Genderfemale:support				1.067***
				(1.007)
Constant	18.864***	17.869***	19.512***	19.616***
	(1.042)	(1.042)	(1.041)	(1.041)
Observations	520,471	520,471	520,471	520,471
Log Likelihood	-1,352,473.000	-1,349,242.000	-1,334,335.000	-1,334,295.000
chi^2	25,989.000*** (df = 62)	32,450.840*** (df = 65)	62,263.220*** (df = 66)	62,344.710*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 4. Cox proportional hazard regressions from Natural Sciences (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.155***	1.146***	1.127***	1.165***
	(1.005)	(1.005)	(1.005)	(1.007)
support			1.934***	2.012***
			(1.005)	(1.007)
early_product_top		0.602***	0.682***	0.683***
		(1.011)	(1.011)	(1.011)
early_c5_top		0.896***	0.903***	0.904***
		(1.009)	(1.009)	(1.009)
early_teamsize_top		0.970***	0.790***	0.790***
		(1.009)	(1.009)	(1.009)
Genderfemale:support				0.919***
				(1.010)
Observations	240,652	240,652	240,652	240,652
Log Likelihood	-1,866,385.000	-1,864,986.000	-1,857,668.000	-1,857,634.000
LR Test	1,473.721*** (df = 13)	4,272.657*** (df = 16)	18,908.380*** (df = 17)	18,976.120*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 5. Weibull regressions from Natural Sciences (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.882***	0.889***	0.904***	0.879***
	(1.004)	(1.004)	(1.004)	(1.005)
support			0.580***	0.560***
			(1.004)	(1.006)
early_product_top		1.528***	1.366***	1.364***
		(1.009)	(1.009)	(1.009)
early_c5_top		1.098***	1.088***	1.087***
		(1.007)	(1.007)	(1.007)
early_teamsize_top		1.025***	1.211***	1.211***
		(1.007)	(1.007)	(1.007)
Genderfemale:support				1.075***
				(1.008)
Constant	10.694***	10.139***	11.647***	11.793***
	(1.008)	(1.008)	(1.008)	(1.008)
Observations	240,652	240,652	240,652	240,652
Log Likelihood	-495,655.700	-494,192.500	-486,393.300	-486,355.200
chi^2	5,726.972*** (df = 13)	8,653.429*** (df = 16)	24,251.750*** (df = 17)	24,328.040*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

B5.2 Social Sciences

Table 6. Cox proportional hazard regressions from Social Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.056***	1.043***	1.027***	1.004
	(1.007)	(1.007)	(1.007)	(1.008)
support			2.294***	2.185***
			(1.008)	(1.012)
early_product_top		0.545***	0.609***	0.608***
		(1.014)	(1.014)	(1.014)
early_c5_top		0.795***	0.777***	0.776***
		(1.011)	(1.011)	(1.011)
early_teamsize_top		1.206***	0.941***	0.941***
		(1.011)	(1.012)	(1.012)
Genderfemale:support				1.092***
				(1.015)
Observations	137,578	137,578	137,578	137,578
Log Likelihood	-1,076,422.000	-1,074,905.000	-1,070,069.000	-1,070,052.000
LR Test	1,201.837*** (df = 62)	4,236.148*** (df = 65)	13,908.570*** (df = 66)	13,942.450*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 7. Weibull regressions from Social Sciences (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.941***	0.954***	0.971***	0.990
	(1.007)	(1.007)	(1.007)	(1.008)
support			0.418***	0.436***
			(1.008)	(1.012)
early_product_top		1.913***	1.671***	1.673***
		(1.014)	(1.014)	(1.014)
early_c5_top		1.289***	1.310***	1.310***
		(1.012)	(1.012)	(1.012)
early_teamsize_top		0.816***	1.061***	1.061***
		(1.012)	(1.012)	(1.012)
Genderfemale:support				0.927***
				(1.015)
Constant	13.777***	12.878***	12.965***	12.926***
	(1.099)	(1.098)	(1.095)	(1.095)
Observations	137,578	137,578	137,578	137,578
Log Likelihood	-348,934.300	-347,293.700	-342,057.100	-342,044.600
chi^2	4,598.411*** (df = 62)	7,879.714*** (df = 65)	18,352.820*** (df = 66)	18,377.800*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 8. Cox proportional hazard regressions from Social Sciences (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.066***	1.053***	1.036***	1.012
	(1.011)	(1.011)	(1.011)	(1.013)
support			2.361***	2.254***
			(1.011)	(1.019)
early_product_top		0.436***	0.511***	0.511***
		(1.024)	(1.025)	(1.025)
early_c5_top		0.931***	0.891***	0.891***
		(1.018)	(1.018)	(1.018)
early_teamsize_top		1.317***	0.967*	0.968*
		(1.017)	(1.018)	(1.018)
Genderfemale:support				1.074***
				(1.023)
Observations	65,755	65,755	65,755	65,755
Log Likelihood	-404,963.400	-404,080.400	-401,511.500	-401,506.600
LR Test	727.797*** (df = 13)	2,493.900*** (df = 16)	7,631.626*** (df = 17)	7,641.554*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 9. Weibull regressions from Social Sciences (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.946***	0.956***	0.970***	0.988
	(1.009)	(1.009)	(1.008)	(1.010)
support			0.492***	0.510***
			(1.009)	(1.015)
early_product_top		2.008***	1.719***	1.720***
		(1.020)	(1.020)	(1.020)
early_c5_top		1.063***	1.100***	1.100***
		(1.015)	(1.014)	(1.014)
early_teamsize_top		0.789***	1.026*	1.026*
		(1.014)	(1.014)	(1.014)
Genderfemale:support				0.944***
				(1.018)
Constant	13.937***	13.412***	14.722***	14.554***
	(1.019)	(1.019)	(1.018)	(1.018)
Observations	65,755	65,755	65,755	65,755
Log Likelihood	-126,382.000	-125,458.900	-122,712.600	-122,707.500
chi^2	2,428.691*** (df = 13)	4,274.890*** (df = 16)	9,767.566*** (df = 17)	9,777.672*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

B5.3 Biology

Table 10. Cox proportional hazard regressions from Biology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.162***	1.151***	1.108***	1.132***
	(1.004)	(1.004)	(1.004)	(1.005)
support			1.920***	1.976***
			(1.005)	(1.006)
early_product_top		0.652***	0.722***	0.723***
		(1.008)	(1.008)	(1.008)
early_c5_top		0.858***	0.863***	0.863***
		(1.007)	(1.007)	(1.007)
early_teamsize_top		0.983**	0.842***	0.842***
		(1.007)	(1.007)	(1.007)
Genderfemale:support				0.940***
				(1.009)
Observations	334,111	334,111	334,111	334,111
Log Likelihood	-2,864,308.000	-2,862,405.000	-2,853,009.000	-2,852,985.000
LR Test	5,916.985*** (df = 62)	9,722.522*** (df = 65)	28,513.970*** (df = 66)	28,562.950*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 11. Weibull regressions from Biology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.850***	0.859***	0.897***	0.873***
	(1.004)	(1.004)	(1.004)	(1.005)
support			0.504***	0.486***
			(1.005)	(1.006)
early_product_top		1.576***	1.401***	1.398***
		(1.008)	(1.008)	(1.008)
early_c5_top		1.181***	1.170***	1.169***
		(1.007)	(1.007)	(1.007)
early_teamsize_top		1.017**	1.192***	1.192***
		(1.007)	(1.007)	(1.007)
Genderfemale:support				1.082***
				(1.009)
Constant	18.778***	17.839***	19.450***	19.605***
	(1.061)	(1.061)	(1.059)	(1.059)
Observations	334,111	334,111	334,111	334,111
Log Likelihood	-841,897.500	-839,833.300	-829,580.200	-829,540.900
chi^2	18,879.150*** (df = 62)	23,007.480*** (df = 65)	43,513.780*** (df = 66)	43,592.280*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 12. Cox proportional hazard regressions from Biology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.194***	1.184***	1.159***	1.209***
	(1.006)	(1.006)	(1.006)	(1.008)
support			2.005***	2.117***
			(1.007)	(1.009)
early_product_top		0.594***	0.674***	0.676***
		(1.014)	(1.014)	(1.014)
early_c5_top		0.879***	0.870***	0.871***
		(1.011)	(1.011)	(1.011)
early_teamsize_top		1.014	0.835***	0.835***
		(1.011)	(1.011)	(1.011)
Genderfemale:support				0.899***
				(1.013)
Observations	165,518	165,518	165,518	165,518
Log Likelihood	-1,193,563.000	-1,192,618.000	-1,187,215.000	-1,187,180.000
LR Test	1,618.452*** (df = 13)	3,508.962*** (df = 16)	14,315.300*** (df = 17)	14,385.960*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 13. Weibull regressions from Biology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.860***	0.867***	0.885***	0.854***
	(1.005)	(1.005)	(1.005)	(1.006)
support			0.568***	0.543***
			(1.005)	(1.007)
early_product_top		1.535***	1.371***	1.368***
		(1.011)	(1.011)	(1.011)
early_c5_top		1.114***	1.119***	1.119***
		(1.009)	(1.009)	(1.009)
early_teamsize_top		0.987	1.155***	1.155***
		(1.009)	(1.009)	(1.009)
Genderfemale:support				1.092***
				(1.010)
Constant	11.886***	11.297***	12.925***	13.139***
	(1.010)	(1.010)	(1.010)	(1.010)
Observations	165,518	165,518	165,518	165,518
Log Likelihood	-334,791.400	-333,810.100	-328,097.500	-328,058.900
chi^2	5,041.518*** (df = 13)	7,004.077*** (df = 16)	18,429.220*** (df = 17)	18,506.540*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

B5.4 Chemistry

Table 14 Cox proportional hazard regressions from Chemistry (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.221***	1.214***	1.174***	1.204***
	(1.006)	(1.006)	(1.006)	(1.007)
support			1.656***	1.694***
			(1.006)	(1.007)
early_product_top		0.661***	0.718***	0.719***
		(1.010)	(1.010)	(1.010)
early_c5_top		0.944***	0.962***	0.963***
		(1.008)	(1.008)	(1.008)
early_teamsize_top		0.868***	0.755***	0.756***
		(1.009)	(1.009)	(1.009)
Genderfemale:support				0.934***
				(1.012)
Observations	186,360	186,360	186,360	186,360
Log Likelihood	-1,735,619.000	-1,734,411.000	-1,730,756.000	-1,730,740.000
LR Test	5,061.049*** (df = 62)	7,476.185*** (df = 65)	14,786.600*** (df = 66)	14,819.780*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 15. Weibull regressions from Chemistry (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.799***	0.805***	0.838***	0.811***
	(1.006)	(1.006)	(1.006)	(1.007)
support			0.580***	0.564***
			(1.006)	(1.007)
early_product_top		1.568***	1.424***	1.422***
		(1.010)	(1.010)	(1.010)
early_c5_top		1.066***	1.043***	1.043***
		(1.009)	(1.009)	(1.009)
early_teamsize_top		1.168***	1.351***	1.349***
		(1.009)	(1.009)	(1.009)
Genderfemale:support				1.092***
				(1.012)
Constant	19.338***	18.254***	19.666***	19.771***
	(1.057)	(1.057)	(1.056)	(1.056)
Observations	186,360	186,360	186,360	186,360
Log Likelihood	-505,801.500	-504,451.500	-500,346.400	-500,319.900
chi^2	11,903.790*** (df = 62)	14,603.790*** (df = 65)	22,813.950*** (df = 66)	22,867.090*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 16. Cox proportional hazard regressions from Chemistry (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.197***	1.188***	1.178***	1.214***
	(1.009)	(1.009)	(1.009)	(1.012)
support			1.790***	1.842***
			(1.009)	(1.012)
early_product_top		0.608***	0.682***	0.683***
		(1.018)	(1.018)	(1.018)
early_c5_top		0.905***	0.937***	0.937***
		(1.014)	(1.014)	(1.014)
early_teamsize_top		0.874***	0.708***	0.709***
		(1.014)	(1.015)	(1.015)
Genderfemale:support				0.927***
				(1.018)
Observations	75,134	75,134	75,134	75,134
Log Likelihood	-569,965.100	-569,447.400	-567,542.800	-567,534.100
LR Test	448.287*** (df = 13)	1,483.720*** (df = 16)	5,292.778*** (df = 17)	5,310.206*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 17. Weibull regressions from Chemistry (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.853***	0.859***	0.868***	0.846***
	(1.007)	(1.007)	(1.007)	(1.009)
support			0.614***	0.599***
			(1.007)	(1.009)
early_product_top		1.525***	1.373***	1.371***
		(1.015)	(1.015)	(1.015)
early_c5_top		1.090***	1.056***	1.056***
		(1.012)	(1.012)	(1.012)
early_teamsize_top		1.123***	1.334***	1.332***
		(1.012)	(1.012)	(1.012)
Genderfemale:support				1.069***
				(1.015)
Constant	8.748***	8.208***	9.459***	9.550***
	(1.013)	(1.013)	(1.013)	(1.014)
Observations	75,134	75,134	75,134	75,134
Log Likelihood	-159,194.000	-158,641.900	-156,586.100	-156,575.900
chi^2	1,353.206*** (df = 13)	2,457.499*** (df = 16)	6,569.052*** (df = 17)	6,589.445*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

B5.5 Psychology

Table 18. Cox proportional hazard regressions from Psychology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.090***	1.074***	1.058***	1.039***
	(1.007)	(1.008)	(1.007)	(1.009)
support			2.435***	2.353***
			(1.008)	(1.012)
early_product_top		0.538***	0.615***	0.614***
		(1.015)	(1.015)	(1.015)
early_c5_top		0.787***	0.771***	0.771***
		(1.013)	(1.013)	(1.013)
early_teamsize_top		1.234***	0.942***	0.942***
		(1.012)	(1.013)	(1.013)
Genderfemale:support				1.063***
				(1.016)
Observations	116,346	116,346	116,346	116,346
Log Likelihood	-886,690.500	-885,322.500	-880,343.000	-880,335.600
LR Test	935.024*** (df = 62)	3,671.098*** (df = 65)	13,630.030*** (df = 66)	13,644.910*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 19. Weibull regressions from Psychology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.908***	0.924***	0.941***	0.954***
	(1.008)	(1.008)	(1.008)	(1.009)
support			0.390***	0.400***
			(1.009)	(1.013)
early_product_top		1.954***	1.661***	1.662***
		(1.016)	(1.015)	(1.015)
early_c5_top		1.309***	1.322***	1.323***
		(1.013)	(1.013)	(1.013)
early_teamsize_top		0.793***	1.059***	1.059***
		(1.013)	(1.013)	(1.013)
Genderfemale:support				0.954***
				(1.016)
Constant	13.414***	12.640***	12.780***	12.751***
	(1.116)	(1.115)	(1.110)	(1.110)
Observations	116,346	116,346	116,346	116,346
Log Likelihood	-294,723.400	-293,243.900	-287,852.100	-287,847.800
chi^2	3,599.727*** (df = 62)	6,558.876*** (df = 65)	17,342.400*** (df = 66)	17,350.960*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 20. Cox proportional hazard regressions from Psychology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.121***	1.107***	1.093***	1.092***
	(1.012)	(1.012)	(1.012)	(1.015)
support			2.562***	2.560***
			(1.012)	(1.021)
early_product_top		0.420***	0.512***	0.512***
		(1.027)	(1.028)	(1.028)
early_c5_top		0.936***	0.898***	0.898***
		(1.020)	(1.020)	(1.020)
early_teamsize_top		1.357***	0.962**	0.962**
		(1.019)	(1.019)	(1.019)
Genderfemale:support				1.001
				(1.025)
Observations	54,238	54,238	54,238	54,238
Log Likelihood	-321,076.900	-320,296.400	-317,624.100	-317,624.100
LR Test	602.618*** (df = 13)	2,163.607*** (df = 16)	7,508.358*** (df = 17)	7,508.360*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 21. Weibull regressions from Psychology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.906***	0.916***	0.929***	0.929***
	(1.010)	(1.010)	(1.010)	(1.012)
support			0.462***	0.463***
			(1.010)	(1.016)
early_product_top		2.073***	1.709***	1.709***
		(1.023)	(1.022)	(1.022)
early_c5_top		1.059***	1.092***	1.092***
		(1.016)	(1.016)	(1.016)
early_teamsize_top		0.769***	1.029*	1.029*
		(1.016)	(1.015)	(1.015)
Genderfemale:support				1.000
				(1.020)
Constant	15.005***	14.424***	16.283***	16.282***
	(1.021)	(1.021)	(1.020)	(1.021)
Observations	54,238	54,238	54,238	54,238
Log Likelihood	-103,058.500	-102,242.700	-99,390.320	-99,390.320
chi^2	1,984.912*** (df = 13)	3,616.555*** (df = 16)	9,321.375*** (df = 17)	9,321.375*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

B5.6 Sociology

Table 22. Cox proportional hazard regressions from Sociology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	0.928***	0.930***	0.927***	0.911***
	(1.017)	(1.017)	(1.017)	(1.018)
support			1.907***	1.718***
			(1.030)	(1.044)
early_product_top		0.624***	0.649***	0.648***
		(1.037)	(1.037)	(1.037)
early_c5_top		0.843***	0.812***	0.813***
		(1.028)	(1.028)	(1.029)
early_teamsize_top		1.064**	0.894***	0.892***
		(1.029)	(1.030)	(1.030)
Genderfemale:support				1.216***
				(1.057)
Observations	21,232	21,232	21,232	21,232
Log Likelihood	-145,864.500	-145,739.200	-145,531.000	-145,524.800
LR Test	446.501*** (df = 62)	697.216*** (df = 65)	1,113.506*** (df = 66)	1,125.994*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 23. Weibull regressions from Sociology (Entire Periods)

	(1)	(2)	(3)	(4)
Genderfemale	1.076***	1.074***	1.075***	1.095***
	(1.016)	(1.016)	(1.016)	(1.017)
support			0.531***	0.587***
			(1.029)	(1.041)
early_product_top		1.589***	1.525***	1.525***
		(1.035)	(1.035)	(1.035)
early_c5_top		1.185***	1.227***	1.226***
		(1.027)	(1.027)	(1.027)
early_teamsize_top		0.941**	1.117***	1.119***
		(1.027)	(1.029)	(1.029)
Genderfemale:support				0.829***
				(1.054)
Constant	15.106***	14.077***	13.996***	13.974***
	(1.196)	(1.195)	(1.193)	(1.193)
Observations	21,232	21,232	21,232	21,232
Log Likelihood	-53,758.460	-53,625.490	-53,405.120	-53,398.770
chi^2	1,055.973*** (df = 62)	1,321.904*** (df = 65)	1,762.654*** (df = 66)	1,775.351*** (df = 67)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 24. Cox proportional hazard regressions from Sociology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	0.908***	0.901***	0.894***	0.879***
	(1.024)	(1.024)	(1.024)	(1.025)
support			1.958***	1.794***
			(1.038)	(1.060)
early_product_top		0.516***	0.542***	0.542***
		(1.055)	(1.055)	(1.055)
early_c5_top		0.923**	0.872***	0.871***
		(1.042)	(1.042)	(1.042)
early_teamsize_top		1.171***	0.933*	0.931*
		(1.040)	(1.043)	(1.043)
Genderfemale:support				1.154**
				(1.074)
Observations	11,517	11,517	11,517	11,517
Log Likelihood	-64,838.370	-64,732.020	-64,591.780	-64,589.750
LR Test	213.755*** (df = 13)	426.453*** (df = 16)	706.942*** (df = 17)	710.993*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

Table 25. Weibull regressions from Sociology (2000-2012)

	(1)	(2)	(3)	(4)
Genderfemale	1.083***	1.090***	1.095***	1.111***
	(1.019)	(1.019)	(1.019)	(1.020)
support			0.572***	0.615***
			(1.031)	(1.048)
early_product_top		1.729***	1.650***	1.651***
		(1.045)	(1.044)	(1.044)
early_c5_top		1.065*	1.115***	1.115***
		(1.034)	(1.034)	(1.034)
early_teamsize_top		0.876***	1.058*	1.060*
		(1.032)	(1.034)	(1.034)
Genderfemale:support				0.889**
				(1.059)
Constant	10.415***	10.082***	10.187***	10.126***
	(1.041)	(1.041)	(1.041)	(1.041)
Observations	11,517	11,517	11,517	11,517
Log Likelihood	-23,159.420	-23,048.710	-22,899.690	-22,897.580
chi^2	525.179*** (df = 13)	746.608*** (df = 16)	1,044.655*** (df = 17)	1,048.860*** (df = 18)

Exponentiated coefficients; Standard errors in parentheses. p<0.1; **p<0.05; ***p<0.01

REFERENCES

1. H. Etzkowitz, C. Kemelgor, B. Uzzi, *Athena unbound: The advancement of women in science and technology*. (Cambridge University Press, 2000).
2. J. S. Long, M. F. Fox, Scientific careers: Universalism and particularism. *Annual review of sociology*, 45-71 (1995).
3. A. E. Preston, *Leaving science*. (Russell Sage Foundation, 2004).
4. H. Zuckerman, J. R. Cole, Women in American science. *Minerva*, 82-102 (1975).
5. C. R. Sugimoto, V. Larivière, *Equity for Women in Science: Dismantling Systemic Barriers to Advancement*. (Harvard University Press, 2023).
6. J. R. Cimpian, T. H. Kim, Z. T. McDermott, Understanding persistent gender gaps in STEM. *Science* **368**, 1317-1319 (2020).
7. S.-J. Leslie, A. Cimpian, M. Meyer, E. Freeland, Expectations of brilliance underlie gender distributions across academic disciplines. *Science* **347**, 262-265 (2015).
8. E. Reuben, P. Sapienza, L. Zingales, How stereotypes impair women's careers in science. *Proceedings of the National Academy of Sciences* **111**, 4403-4408 (2014).
9. J. Huang, A. J. Gates, R. Sinatra, A.-L. Barabási, Historical comparison of gender inequality in scientific careers across countries and disciplines. *Proceedings of the National Academy of Sciences*, (2020).
10. J. D. Adams, G. C. Black, J. R. Clemmons, P. E. Stephan, Scientific teams and institutional collaborations: Evidence from US universities, 1981–1999. *Research policy* **34**, 259-285 (2005).
11. S. Wuchty, B. F. Jones, B. Uzzi, The increasing dominance of teams in production of knowledge. *Science* **316**, 1036-1039 (2007).
12. M. Weber. (CA: University of California Press, 1978).
13. D. S. Pugh, D. J. Hickson, C. R. Hinings, C. Turner, Dimensions of organization structure. *Administrative science quarterly*, 65-105 (1968).
14. J. P. Walsh, Y.-N. Lee, The bureaucratization of science. *Research Policy* **44**, 1584-1600 (2015).
15. W. O. Hagstrom, Traditional and modern forms of scientific teamwork. *Administrative Science Quarterly*, 241-263 (1964).
16. E. J. Hackett, Science as a vocation in the 1990s: The changing organizational culture of academic science. *The journal of higher education* **61**, 241-279 (1990).
17. F. Xu, L. Wu, J. Evans, Flat teams drive scientific innovation. *Proceedings of the National Academy of Sciences* **119**, e2200927119 (2022).
18. Y.-N. Lee, J. P. Walsh, Rethinking Science as a Vocation: One Hundred Years of Bureaucratization of Academic Science. *Science, Technology, & Human Values* **0**, 01622439211026020 (2021).
19. S. Milojević, F. Radicchi, J. P. Walsh, Changing demographics of scientific careers: The rise of the temporary workforce. *Proceedings of the National Academy of Sciences* **115**, 12616-12623 (2018).
20. B. Maddox, *Rosalind Franklin: The dark lady of DNA*. (HarperCollins New York, 2002).
21. Z. Lin, Y. Yin, L. Liu, D. Wang, SciSciNet: A large-scale open data lake for the science of science research. *Scientific Data* **10**, 315 (2023).
22. L. Roberts, S. Rockman, A. Hui, Historiographies of science and labor: From past perspectives to future possibilities. *History of Science* **61**, 448-474 (2023).
23. J. S. Long, M. F. Fox, Scientific careers: Universalism and particularism. *Annual review of sociology* **21**, 45-71 (1995).
24. M. F. Fox, Women and scientific careers. *Handbook of science and technology studies*, 205-223 (1995).
25. K. Imai, L. Keele, D. Tingley, A general approach to causal mediation analysis. *Psychological Methods* **15**, 309-334 (2010).
26. L. L. Hargens, Patterns of scientific research. *Washington, DC: American Sociological Association*, (1975).

27. S. Fuchs, *The professional quest for truth: A social theory of science and knowledge*. (Suny Press, 1992).
28. B. Macaluso, V. Larivière, T. Sugimoto, C. R. Sugimoto, Is science built on the shoulders of women? A study of gender differences in contributorship. *Academic Medicine* **91**, 1136-1142 (2016).
29. J. P. Walsh, Y.-N. Lee, in *ResearchProfessional News*. (2022).
30. M. Fox, in *Higher Education: Handbook of Theory and Research,* J. Smart, Ed. (2008), pp. 73-103.
31. D. Peterson, All that is solid: Bench-building at the frontiers of two experimental sciences. *American Sociological Review* **80**, 1201-1225 (2015).
32. S. Jabbehdari, J. P. Walsh, Authorship norms and project structures in science. *Science, Technology, & Human Values* **42**, 872-900 (2017).
33. I. Van Buskirk, A. Clauset, D. B. Larremore, An Open-Source Cultural Consensus Approach to Name-Based Gender Classification. *Proceedings of the International AAAI Conference on Web and Social Media* **17**, 866-877 (2023).
34. NSF, "Doctorate Recipients from U.S. Universities: 2013. Special Report NSF 15-304.," (National Science Foundation, 2013).
35. S. Milojević, F. Radicchi, J. P. Walsh, Reply to Hanlon: Transitions in science careers. *Proceedings of the National Academy of Sciences* **116**, 17625 (2019).
36. V. Larivière, D. Pontille, C. R. Sugimoto, Investigating the division of scientific labor using the Contributor Roles Taxonomy (CRediT). *Quantitative Science Studies* **2**, 111-128 (2021).
37. V. Larivière *et al.*, Contributorship and division of labor in knowledge production. *Social Studies of Science* **46**, 417-435 (2016).
38. Y. Lin, C. B. Frey, L. Wu, Remote collaboration fuses fewer breakthrough ideas. *Nature* **623**, 987-991 (2023).
39. T. K. Kelly, W. P. Butz, S. Carroll, D. M. Adamson, G. Bloom, "The US scientific and technical workforce: improving data for decisionmaking," (RAND CORP SANTA MONICA CA, 2004).
40. S. Milojević, F. Radicchi, J. P. Walsh, Reply to Hanlon: Transitions in science careers. *Proceedings of the National Academy of Sciences* **116**, 17625-17626 (2019).
41. S. M. Hanlon, Scientists who leave research to pursue other careers in science are still scientists. *Proceedings of the National Academy of Sciences* **116**, 17624-17624 (2019).
42. R. K. Merton, *The sociology of science: Theoretical and empirical investigations*. (University of Chicago press, 1973).
43. D. R. Cox, Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)* **34**, 187-202 (1972).
44. P. D. Allison, *Event history and survival analysis: regression for longitudinal event data*. (SAGE publications, 2014), vol. 46.
45. D. F. Moore, *Applied survival analysis using R*. (Springer, 2016).
46. T. Therneau, A package for survival analysis in S. *R package version* **2**, (2015).
47. K. Imai, L. Keele, D. Tingley, T. Yamamoto, Unpacking the black box of causality: Learning about causal mechanisms from experimental and observational studies. *American Political Science Review* **105**, 765-789 (2011).
48. D. Tingley, T. Yamamoto, K. Hirose, L. Keele, K. Imai, Mediation: R package for causal mediation analysis. (2014).