Climbing the Ivory Tower: How Socio-Economic Background Shapes Academia *

Ran Abramitzky
Stanford University

Lena Greska
University of Munich

Santiago Peréz
UC Davis

Joseph Price
BYU

Carlo Schwarz
Bocconi University

Fabian Waldinger
University of Munich

November 21, 2024

Abstract

We explore how socio-economic background shapes academia, collecting the largest dataset of US academics' backgrounds and research output ever assembled. We find that individuals from poorer backgrounds are severely underrepresented and that representation varies widely across disciplines: more math-intensive disciplines exhibit higher representation. Representation also varies across universities, with particularly low representation at elite universities. While we find no differences in the *average* number of publications, academics from poorer backgrounds are more likely to not publish at all but also to have outstanding publication records, making them riskier hires. Furthermore, academics from poorer backgrounds introduce more novel scientific concepts but are less likely to receive recognition, as measured by citations and Nobel Prize nominations and awards. We also document substantial homophily in citations by socio-economic background. Finally, the father's occupation affects discipline choice and, thus, the direction of research. Academics working in disciplines that are related to their father's occupation (e.g., children of therapists who become medicine professors) are more productive.

Keywords: Academics, socio-economic background, Science, US census

*We are grateful to Davide Cantoni, Kilian Huber and seminar audiences at Bocconi, CEU, Chicago, Cologne, NUS, LMU Munich, LSE, Mannheim, Pompeu Fabra, PSE, SOFI Stockholm, and Tilburg for insightful comments and suggestions. We are grateful to Alessandro Iaria and Sebastian Hager, who have greatly contributed to the construction of the World of Academia Database, which we use in this paper. We are also grateful to Felix Radde and Marie Spörk for outstanding research assistance.

1 Introduction

The underrepresentation of individuals from lower socio-economic backgrounds in leadership positions in government, business, and academia is becoming an increasing concern to policymakers and the general public. Two primary economic rationales underlie efforts to increase representation. First, significant disparities in the representation of societal groups raise concerns regarding fairness and equality of opportunity. Second, unequal representation can undermine efficiency, as the misallocation of talent deprives society of valuable contributions from individuals in underrepresented groups (Hsieh et al., 2019). In knowledge creation sectors, such as academia, this underrepresentation introduces an additional inefficiency: the unique lived experiences of underrepresented groups offer valuable perspectives that could diversify and enrich the scope of ideas that are explored (e.g., Thorp 2023). In essence, the absence of these individuals—*missing people*—can lead to *missing ideas*, which is particularly problematic in a world where ideas may be "getting harder to find" (Bloom et al., 2020).

Studying underrepresentation by individuals from lower socio-economic backgrounds in elite institutions has been hampered by a lack of comprehensive data on the socio-economic backgrounds of high-skilled professionals. In this paper, we overcome this challenge by compiling an extensive datasbase focused on one such high-skilled profession: U.S. academia. These novel data enable us to study the extent to which individuals from lower socioeconomic backgrounds are underrepresented in academia, whether this underrepresentation changed over time, how this underrepresentation differs by disciplines and institutions. Further, we can analyze whether the socio-economic background affects scientific output and recognition as well as the choice of research fields.

We rely on three primary data sources to assemble the most comprehensive data ever collected on the socio-economic backgrounds and research output of U.S. academics. First, we utilize complete faculty rosters from the *World of Academia Database* (Iaria et al., 2024), which provides detailed information on the name, discipline, and academic rank of nearly all academics at U.S. universities from 1900 to 1969. A key advantage of these data is that they list academics even if they do not publish or are not members of academic societies. This helps to mitigate selection biases common in studies that rely exclusively on publication or citation databases, surveys, or lists of distinguished scholars. Second, we measure the socio-economic background of academics by linking these faculty rosters to full-count U.S. census data using data via the *Census Linking Project* (CLP) (Abramitzky et al., 2021) and the *Census Tree Project* (Buckles et al. 2023). Our baseline measure of socio-economic background is the percentile rank of their father's income when they were growing up.[1] Third, we link these academics to their publication and citation records using data from the *Clarivate Web of Science*. Overall, our data enable us to

[1]The findings are similar for alternative measures of socio-economic background.

measure the socio-economic backgrounds and research output of 47,235 academics across 1,035 universities over nearly seven decades.

Using these data, we present our findings in four parts. In the first part, we explore the link between family background and the likelihood of becoming an academic. Our analysis reveals a stark underrepresentation of individuals from lower socioeconomic backgrounds in academia: those born to parents in the bottom 20% of the parental income distribution account for less than 5% of all academics. In contrast, around half of U.S. academics come from the top 20% of the income rank distribution. Children born to the highest-earning fathers are particularly overrepresented. For example, a child born to parents in the top percentile has a 56% higher chance of becoming an academic than a child from the 99th percentile. We also find that academia exhibits greater socio-economic selectivity compared to other occupations that require specialized training, such as medicine and law. Leveraging the seven decades covered by our dataset, we also investigate the long-run evolution of representation in academia. Despite the significant changes in American higher education and society over the past century, the socio-economic composition of academics has remained remarkably constant over time.

Although, on average, academics are disproportionately drawn from higher-income families, we find vast heterogeneity in the socio-economic composition of academics by discipline and university. While around 60% of academics in the humanities, architecture, archaeology, anthropology and medicine come from the top 20 per cent of the parental income distribution, only 30-40% in agriculture, pedagogy, and veterinary medicine originate from this income bracket. Additionally, we explore the relationship between skill requirements and socio-economic representation at the discipline level. We find that representation from lower socio-economic backgrounds is higher in disciplines with a stronger emphasis on quantitative relative to verbal skills.

Our rich data also enable us to explore heterogeneity in representation by university. While around two-thirds of academics in selective private universities such as Princeton, UPenn, Harvard, and Yale come from the top 20 percent of the parental income distribution, only 30-40% percent of academics in state universities such as Iowa State, University of Missouri, or the University of Virginia originate from this income bracket. These university-level differences in the socio-economic background of their academics cannot be explained by the discipline composition of the universities.

In the second part of the paper, we study how the socio-economic background correlates with academic output. We find no systematic relationship between the *average* number of publications and the parental income rank. We additionally investigate effects along different percentiles of the publication distribution. We find that individuals from lower socio-economic backgrounds are significantly more likely to never publish a paper. However, they are also more likely to have a publication count in the top 1%.

We additionally explore whether the content of publications differs by socio-economic

background. We construct a measure of novel words that a scientist introduced to the scientific community (Iaria et al., 2018). The measure proxies for the introduction of new scientific concepts that required novel scientific terms. We find that scientists whose fathers were at the 75th percentile of the income rank publish around 0.05 fewer papers (around 17 percent) with at least one novel word compared to scientists with fathers at the 25th percentile.

Overall, the results on academic output suggest that academics from lower socio-economic backgrounds are more likely to not publish at all but also to have outstanding publication records, making them somewhat riskier hires. The results on novel words suggest that they are somewhat more likely to pursue research agendas off the beaten path which may result in scientific breakthroughs but also in a higher failure rate.

In the third part, we examine the relationship between socio-economic background and recognition by other academics. First, we analyze citations to an academic's research papers, a widely-used metric for measuring recognition within the academic community. The results suggest that papers published by authors from higher socio-economic backgrounds receive somewhat more citations. The differences in citations are particularly surprising in light of the results from part two, which indicate that low-SES academics introduced more novel words. In further results, we investigate one factor that could underlie these differences: the homophily of citations based on socio-economic background. We find that authors with a 10 percentile higher average SES rank on average cite papers from authors with a 5 percentile higher average SES rank. To the best of our knowledge, we are the first to document this strong homophily in citations on the basis of socio-economic background. Given the strong overrepresentation of individuals from higher socio-economic backgrounds documented in part one of the paper, the homophily in citations provides a potential explanation for the higher number of citations that the papers of individuals from higher socio-economic backgrounds receive.

As an additional measure of recognition, we investigate Nobel Prize nominations and awards that indicate acknowledgement for exceptional scientific contributions. We find that scientists whose fathers were at the 75th percentile of the income rank are around 0.6 percentage points (or 50%) more likely to be nominated for a Nobel Prize compared to scientists with fathers at the 25th percentile. They are also 50% more likely to be awarded a Nobel Prize. These differences persist if we control for the publication and citation records of scientists.

Taken together, these results indicate that scientists from lower socio-economic backgrounds are more likely to be overlooked by the scientific community, which predominately originate from high socio-economic backgrounds.

In the fourth part of the paper, we examine the relationship between fathers' occupation and the choice of academic discipline. We develop a novel measure of overrepresentation to assess whether children of fathers in specific occupations are overrepresented in particu-

lar academic disciplines. Our findings indicate that academics tend to pursue disciplines aligned with their fathers' occupation. For example, the children of architects are more likely to become professors in architecture, children of artists are more likely to become professors of arts and design, children of bank tellers are more likely to become professors in business and management, and children of lawyers are more likely to become professors in law. Additionally, using a text embeddings model, we determine the semantic proximity of a father's occupation (e.g., "farmer") to an academic discipline (e.g., "agriculture"). This allows us to define a *patrimonial* discipline: the discipline that is closest in semantic space to the father's occupation. We then show that academics are more likely to enter their patrimonial discipline. Overall, these findings indicate that socio-economic background affects not only the probability of becoming an academic but also the specific discipline that academics pursue.

In further results, we investigate whether academics who work in their patrimonial discipline are more productive than academics who work in disciplines that are unrelated to their father's occupation. We find that academics who work in their patrimonial discipline publish 0.08 standard deviations more papers. The effect of working in the patrimonial discipline is about one-third of the large gender gap in publications that has been documented for this period (Iaria et al., 2024). These results indicate that the father's occupation not only affects discipline choice but also that academics who work in a discipline that is "close" to their father's occupation produce more output. These results are consistent with the view that children's exposure to certain topics while growing up influences their scientific output once they become academics.

While a substantial body of research has studied the underrepresentation of women and racial groups in elite institutions, the underpresentation of individuals from lower socio-economic backgrounds has been a "forgotten dimension of diversity" (Ingram, 2021).

Our paper contributes to a fast-growing literature on the backgrounds of high-skilled, "elite" professionals such as politicians (Dal Bó et al., 2017) or civil servants (Moreira and Pérez, 2022). It is particularly close to research documenting the socio-economic background of inventors (Bell et al. (2019); Aghion et al. (2018, 2023)) and academics (Morgan et al., 2022; Airoldi and Moser, 2024; Stansbury and Schultz, 2023; Stansbury and Rodriguez, 2024; Agarwal and Gaule, 2020). We contribute to this literature with the most comprehensive analysis of the socio-economic background of US academics covering the all disciplines and the near universe of universities. The unique quality of our data also allows us to shed light on the factors that underly differences in differences in hiring, productivity, recognitions, and research direction. Other related research has documented the importance of socio-economic background for the selection of students into elite universities (Chetty et al., 2020a, 2023). Further, research by Michelman et al. (2022) suggests that admission into Harvard in the early 20th century was dominated by "Old Boys' Clubs".

More recently, a newly emerging literature has begun to study the socio-economic background of academics. Our paper provides the most comprehensive analysis of the socio-economic background of US academics, disciplines, and universities. The advantage of our data is that we can study the near universe of US academics for all disciplines. Furthermore, our data enable us to trace the long-run evolution of the social backgrounds of US academics over the 1900-1969 period, a key period in the history of American higher education (from the "formative" prewar years, to the consolidation of American leadership in higher education after World War II).

Second, we contribute to the literature on how scientists' or inventors' background shapes their research and, thereby, the direction of innovation. Existing work by Koning et al. (2021); Einio et al. (2022); Kozlowski et al. (2022); Truffa and Wong (2022); Kozlowski et al. (2022) and Dossi (2024) investigates the relationship between gender, race, and the research topics of contemporary scientists in the US. Similarly, Einio et al. (2022) document that inventors from poorer backgrounds are more likely to patent "necessity" interventions. To the best of our knowledge, we provide the first systematic evidence for the impact of socio-economic background on the research of university academics. Since much innovation and basic research, as well as the training of future innovators, occurs in universities, the selection of academics likely has important knock-on effects for downstream innovation.

Third, our research is related to the literature on inefficiencies and hurdles in the knowledge-creation process. Recent studies have provided increasing evidence for the Matthew Effect (Azoulay et al., 2014; Jin et al., 2019), gatekeepers (Azoulay et al., 2019). Croix and Goñi (2024) provide evidence from pre-industrial families of scholars and document that an estimated 40% of scholars received their position through nepotism rather than merit. Studying international mathematical olympiads, Agarwal and Gaule (2020) show that participants from lower-income countries produce fewer publications and citations despite similar talents. Abstracting from the question of academic recognition of work produced outside of historical academic centers, these results already point to the crucial role of institutions in mediating the social mobility of academics. Lastly, there is a fast-growing literature on gender discrimination in academia (e.g., Card et al., 2020, 2022; Iaria et al., 2024; Koffi, 2024; Hengel, 2022). While the hurdles faced by female academics have received substantial attention, we know relatively little about the socio-economic background of academics, despite the fact that there are far fewer academics from low socio-economic backgrounds than women in academia.

2 Data

For our analysis, we construct the largest individual-level dataset of U.S. university academics ever assembled, which we combine with information on their socio-economic

background and their research output. The dataset is based on three data sources. First, we use complete faculty rosters for the near universe of US universities from the *World of Academia Database* (Iaria et al. 2024). Second, we match these data to historical U.S. Censuses based on data from the *Census Linking Project (CLP)* (Abramitzky et al. 2012, 2021) and the *Census Tree Project* (Buckles et al. 2023), which allows us to measure the socio-economic background of academics. Third, we enhance the data with publication and citation data from the *Web of Science*.

2.1 Historic Faculty Rosters from the World of Academia Database

The *World of Academia Database* contains faculty rosters for nearly all Ph.D. granting universities in the United States. We use six cross-sections covering U.S. academics in 1900, 1914, 1925, 1938, 1956, and 1969.[2] The data contains 3,449 U.S. academics who entered the database in 1900 and 65,282 U.S. academics who entered the database in 1969, reflecting the spectacular growth of the U.S. university sector during the 20th century (Table 1).

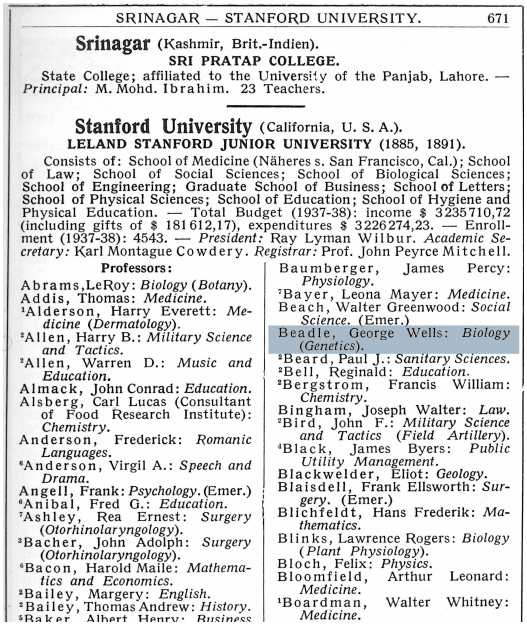
For the period of our analysis, the database provides the most comprehensive data on academics in the United States (see Iaria et al. 2024 for details and comparisons to other data sources). In addition to academics' names, universities, and academic rank (i.e., assistant, associate, or full professor), we observe their specialization, which we code into 30 disciplines. For example, the 1938 faculty roster lists George Wells Beadle as a Biology professor at Stanford University (Figure 1, panel a). He received the 1958 Nobel Prize in Physiology/Medicine for the "discovery of the role of genes in biochemical events within cells."

The *World of Academia Database* has several key features for our analysis. First, it contains *entire* faculty rosters for the vast majority of PhD granting universities in the US, which allows us to study academics even if their scientific outputs are not covered by publication databases or never became distinguished scientists. This comprehensive coverage enables us to overcome important selection biases that affect studies that rely exclusively on publication or citation databases, surveys, or lists of distinguished academics. Second, it covers all academic disciplines, including the social sciences and humanities. These features allow for a comprehensive analysis of representation in academia and enable us to study heterogeneity in representation across disciplines and universities.

[2]The data include all academics who were affiliated with a U.S. university in any of the six cross-sections. We thus also include the U.S. spells of academics who start their career abroad and move to the United States or who start their career in the United States and then move abroad. About 10 percent of the academics are only listed with initials in the faculty rosters. As the match to the census data described below uses full first names, we exclude these academics from the data. For the statistics reported in Table 1, we report their first U.S. cohort in the *World of Academia Database*.

Figure 1: Example Data Construction

(a) Sample Page: Faculty Rosters



SRINAGAR — STANFORD UNIVERSITY. 671

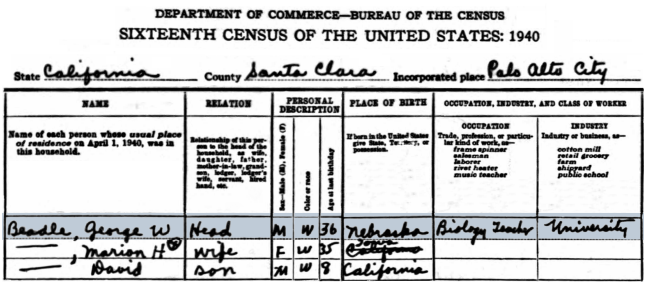
Srinagar (Kashmir, Brit.-Indien).
SRI PRATAP COLLEGE.
State College; affiliated to the University of the Panjab, Lahore. — *Principal:* M. Mohd. Ibrahim. 23 Teachers.

Stanford University (California, U. S. A.).
LELAND STANFORD JUNIOR UNIVERSITY (1885, 1891).
Consists of: School of Medicine (Näheres s. San Francisco, Cal.); School of Law; School of Social Sciences; School of Biological Sciences; School of Engineering; Graduate School of Business; School of Letters; School of Physical Sciences; School of Education; School of Hygiene and Physical Education. — Total Budget (1937-38): income $ 3235710,72 (including gifts of $ 181612,17), expenditures $ 3226274,23. — Enrollment (1937-38): 4543. — *President:* Ray Lyman Wilbur. *Academic Secretary:* Karl Montague Cowdery. *Registrar:* Prof. John Peyrce Mitchell.

Professors:

Abrams, LeRoy: *Biology (Botany).*
Addis, Thomas: *Medicine.*
[3]Alderson, Harry Everett: *Medicine (Dermatology).*
[2]Allen, Harry B.: *Military Science and Tactics.*
[2]Allen, Warren D.: *Music and Education.*
Almack, John Conrad: *Education.*
Alsberg, Carl Lucas (Consultant of Food Research Institute): *Chemistry.*
Anderson, Frederick: *Romanic Languages.*
[6]Anderson, Virgil A.: *Speech and Drama.*
Angell, Frank: *Psychology.* (Emer.)
[6]Anibal, Fred G.: *Education.*
[7]Ashley, Rea Ernest: *Surgery (Otorhinolaryngology).*
[3]Bacher, John Adolph: *Surgery (Otorhinolaryngology).*
[6]Bacon, Harold Maile: *Mathematics and Economics.*
[2]Bailey, Margery: *English.*
[5]Bailey, Thomas Andrew: *History.*
[5]Baker, Albert Henry: *Business*
Baumberger, James Percy: *Physiology.*
[7]Bayer, Leona Mayer: *Medicine.*
Beach, Walter Greenwood: *Social Science.* (Emer.)
Beadle, George Wells: *Biology (Genetics).*
[2]Beard, Paul J.: *Sanitary Sciences.*
[2]Bell, Reginald: *Education.*
[2]Bergstrom, Francis William: *Chemistry.*
Bingham, Joseph Walter: *Law.*
[2]Bird, John F.: *Military Science and Tactics (Field Artillery).*
[4]Black, James Byers: *Public Utility Management.*
Blackwelder, Eliot: *Geology.*
Blaisdell, Frank Ellsworth: *Surgery.* (Emer.)
Blichfeldt, Hans Frederik: *Mathematics.*
Blinks, Lawrence Rogers: *Biology (Plant Physiology).*
Bloch, Felix: *Physics.*
Bloomfield, Arthur Leonard: *Medicine.*
[1]Boardman, Walter Whitney: *Medicine.*

(b) Adult Census

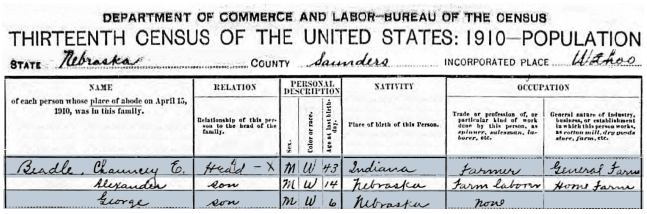


DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS
SIXTEENTH CENSUS OF THE UNITED STATES: 1940

State California County Santa Clara Incorporated place Palo Alto City

NAME	RELATION	PERSONAL DESCRIPTION			PLACE OF BIRTH	OCCUPATION, INDUSTRY, AND CLASS OF WORKER	
Name of each person whose *usual place of residence* on April 1, 1940, was in this household.	Relationship of this person to the head of the household.	Sex	Color or race	Age at last birthday	If born in the United States give State, Ter., or D. C., or possession.	OCCUPATION Trade, profession, or particular kind of work, as—	INDUSTRY Industry or business, as—
Beadle, George W	Head	M	W	36	Nebraska	Biology Teacher	University
—, Marion H	Wife	F	W	35	California		
— David	Son	M	W	8	California		

(c) Childhood Census



DEPARTMENT OF COMMERCE AND LABOR—BUREAU OF THE CENSUS
THIRTEENTH CENSUS OF THE UNITED STATES: 1910—POPULATION

STATE Nebraska COUNTY Saunders INCORPORATED PLACE Wahoo

NAME of each person whose place of abode on April 15, 1910, was in this family.	RELATION Relationship of this person to the head of the family.	PERSONAL DESCRIPTION Sex.	Color or race.	Age at last birthday.	NATIVITY Place of birth of this Person.	OCCUPATION Trade or profession of, or particular kind of work done by this person, as spinner, salesman, laborer, etc.	General nature of industry, business, or establishment in which this person works, as cotton mill, dry goods store, farm, etc.
Beadle, Chauncey E.	Head	M	W	43	Indiana	Farmer	General Farm
Alexander	son	M	W	14	Nebraska	Farm Laborer	Home Farm
George	son	M	W	6	Nebraska	none	

Notes: Panel (a) shows a sample page from the faculty roster of Stanford University from the 1938 edition of *Minerva* including the entry of the biology professor "George Wells Beadle." Panel (b) shows George W. Beadle's entry in the 1940 adult census. Panel (c) shows George Beadle's entry in his childhood census (1910) which we use to measure the occupation of his father ("farmer").

2.2 Measuring Parental Socio-Economic Background

To measure academics' parental socio-economic background, we link the faculty rosters to historical full-count U.S. censuses (Ruggles et al., 2024) using a two-step procedure. In the first step, we link the cross-sections of academics to a contemporaneous U.S. census ("adult census"). In the second step, we use census crosswalks from the Census Linking Project and the Census Tree Project to construct back-links to each academic's childhood census records to measure parental background.

Linking Faculty Rosters to Contemporaneous U.S. Censuses "Adult Census"

In the first step, we link all academics who appear in the faculty rosters to the two closest contemporaneous censuses. For example, we link the 1925 faculty roster to both the 1920 and 1930 censuses. The only exceptions are the 1956 and 1969 faculty rosters, which can be linked to only one census (the 1950 census) since neither the 1960 nor the 1970 full count censuses are currently publicly available.

We link academics in the faculty rosters to the contemporaneous censuses based on the full name of the academic, the census occupation, and the location.[3] We define a

[3]It is important to note, that a relatively small share of professors are listed under the occupation "professor" in the census. Biology professors, for example, are listed as "professor", "biologist", or "biology teacher." This highlights the importance of using faculty rosters to capture university professors instead

potential match as someone:

1. who has the exact same first and last name in the census and in the faculty rosters

2. whose implied age is between 20 and 100 (based on their age in the census) at the time we observe them in the corresponding faculty rosters

3. who indicates an occupation in the census that aligns with a professorship in a specific discipline (e.g., biology professors may be listed with the occupations "professor", "biologist", or "biology teacher")[4]

We consider all matches that satisfy criteria 1-3 above. If criteria 1-3 only return one potential match between the census and the faculty rosters, we consider the observation pair as matched, and the procedure continues with step 7 (described below). For example, we can link the faculty roster entry of George Wells Beadle to the 1940 census. The unique match in the census reports that he was 36 years old in 1935, lived in Pasadena, California, and worked as a "Biology Teacher" at a "University" (Figure 1, panel b).

If there are multiple potential matches, we disambiguate them using the following additional criteria:

4. the potential match in the census lives in a county within 150 kilometers of the university reported in the faculty rosters[5]

5. the potential match has the same middle name initial(s) in the census and the faculty rosters

6. the potential match reports an occupation in the census which aligns more closely with their discipline (i.e., if there are two potential matches for a biology professor, one listed in the census as "professor" and the other one as "biology professor," we select the latter observation)

We then keep all matches that are unique after disambiguating them using at least one of the criteria 4-6. After enforcing criteria 1-6, about 70% of the potential matches report an industry in the census that aligns with an academic position. For example, they report "education" as their industry in the census (or medical professors report "hospital"). For the remaining 30% of potential matches for which the industry does not as closely align (e.g., they report "construction"), We further improve the quality of the link by enforcing a seventh criterion, which is based on the industry reported in the census:

of using the "professor" occupational category from the census records.

[4]In this step, we augment IPUMS coding of occupations using the original occupation. Doing so enables us to classify individuals whose occupation or industry is listed as "not yet classified". This typically happens in cases in which the occupations or industries are misspelled in the original records.

[5]For academics that are affiliated with multiple universities, we calculate the distance between each of their universities and the county and use the minimum distance for disambiguation.

Table 1: Linking Rates

Cohort	Academics entering faculty rosters	Matched to Adult Census: Total	Matched to Adult Census: % Faculty roster	Matched to Childhood Census: Total	Matched to Childhood Census: % Adult census	Matched to Childhood Census: % Faculty roster
Main sample: 1900-1956 cohorts						
1900	3,441	2,529	73.5	1,768	69.9	51.4
1914	5,899	4,525	76.7	3,110	68.7	52.7
1925	6,401	4,775	74.6	3,229	67.6	50.4
1938	23,458	17,911	76.4	12,446	69.5	53.1
1956	53,243	29,048	54.6	17,285	59.5	32.5
Total	92,442	58,788	63.6	37,838	64.4	40.9
Extended sample: 1900-1969 cohorts						
		⋮				
1969	65,340	17,944	27.5	9,397	52.4	14.4
Total	157,782	76,732	48.6	47,235	61.6	29.9

7. the potential matches must report an industry in the census that is consistent with becoming a professor

For this seventh criterion, all potential matches with a misaligned industry are independently hand-checked by two research assistants who classify them as correct or incorrect. We only retain the matches that both research assistants classify as correct.[6]

For each academic that we successfully link to a contemporaneous census, we extract the birth year and the birth state from the adult census. These variables are crucial to link academics to their childhood censuses (see below for more details). For example, we extract George Beadle's birthyear (1903 or 1904, based on the 36 years of age that he reports) and his birth state ("Nebraska") from his 1940 census record (Figure 1, panel b).

Throughout the paper, we show results for two different samples:

1. *Main Sample:* 1900-1956 faculty rosters
2. *Extended Sample:* 1900-1969 faculty rosters

We use two different samples because the full count censuses for 1960 and 1970 are not yet available. It is, therefore, challenging to link individuals who entered the *World of*

[6]In cases where we match an academic to multiple census years, we additionally check whether these matches are internally consistent, i.e. that the main demographic information used for backlinking is the same across all matches. For example, an academic matched to a person aged 45 in the 1910 census should match to a person aged 55 in the 1920 census. Our research assistants hand-check all observations for which this is not the case and remove incorrect matches.

Academia database in 1969 to an adult census. With this in mind, the main sample in our analysis is restricted to academics who we first observe in or prior to 1956. However, we also consider an extended sample in which we attempt to match all academics in our data (including those who enter the data in 1969).

Of the 92,500 academics in the main sample, we link 58,788 (64%) to a contemporaneous census (Table 1).[7] Manual inspections suggest that transcription mistakes of the historical handwritten census records account for the majority of missed links. Furthermore, as we require unique matches based on our linking criteria, we also miss links if matches between the faculty rosters and the census record are not unique. In the extended sample we link 76,732 (48%) to a contemporaneous census (Table 1). Linking rates are lower for the 1956 and 1969 cohorts for two main reasons. First, these cohorts can only be matched to the 1950 census. Linking to just one adult census lowers the linking rate, as linking to two censuses enables us to deal with idiosyncratic transcription errors occurring in one census but not the other. Second, these cohorts likely include individuals who were not yet academics in 1950 and, hence, cannot be matched on the basis of their census occupation to an adult census.

Linking to the Childhood Census to Measure Socio-Economic Background

To construct measures of the socio-economic background of academics, we use census-to-census crosswalks to link the adult census record to childhood censuses. Specifically, we use all ABE-exact census links established by the Census Linking Project (Abramitzky et al. 2012, 2021).[8] To increase match rates between the adult census and the childhood census, we complement the ABE-exact links with crosswalks established through machine-learning linking algorithms. To establish backlinks for the 1900-1940 adult censuses, we use the census links from the Census Tree Project database (Buckles et al. 2023). Because Census Tree Project backlinks are not yet available for the 1950 census, we use the IPUMS Multigenerational Longitudinal Project (MLP) (Ruggles et al., 2019) to establish backlinks for the 1950 adult census. The machine-learning links also enable us to link to childhood censuses of female academics.[9]

To maximize the probability of observing an academic's parental background, we link adult censuses to all possible childhood censuses, which we define as censuses in which we observe future academics as children below the age of 22 and living with their parents.

[7] Below, we provide evidence that linked academics are similar to academics who we are unable to link, thereby alleviating selection concerns.

[8] Since the Census Linking Project has not released (as of October 2024) links between the 1950 census and earlier censuses, we create our own crosswalks for the 1950 census using the ABE algorithm in its exact standard version.

[9] The share of female academics in the faculty rosters is only 13% in 1900 to 1956 and 14% in 1900 and 1969 (see also Iaria et al. 2024). Overall, linking rates for female academics are 28% for the main sample and 21% for the extended sample, compared to 43% and 31% for male academics. All results remain unchanged in a sample of male academics.

Whenever we link an academic to multiple childhood censuses, we use the childhood census in which the academic is the youngest.[10]

For example, we link George Wells Beadle to his childhood census of 1910. At the time, he was six years old and listed in the census as the son of Chauncey Beadle, who was 43 years old and worked as a farmer (Figure 1, panel c).

For the main sample, we are able to link 37,838 (or 64% of the adult census) records to a childhood census (Table 1). For the extended sample, we can link 47,235 (or 62%) of the adult census records to a childhood census.[11] These linking rates are high compared to linking rates in existing research, because we rely on various linking algorithms and because we link to multiple potential childhood censuses.

Overall, we link 37,838 (or 40.9%) of the main sample to their childhood census. These linked academics form the basis for our analysis. We investigate selection in the links between academics and their census records. For this analysis, we correlate the department rank (measured as the average number of citations of all academics in a department, see Hager et al. 2024) with the linking rate at the department level. As shown in Figure A.1, Panel (a,b), we find no systematic relationship between department quality and the linking rates (p-value=0.71). As a further check, we investigate the correlation between the linking rates and the average income of a last name. We again find no systematic association between these variables (Figure A.1, Panel b). Together, these results give us confidence that there does not appear to be any strong and systematic selection into our analysis sample.

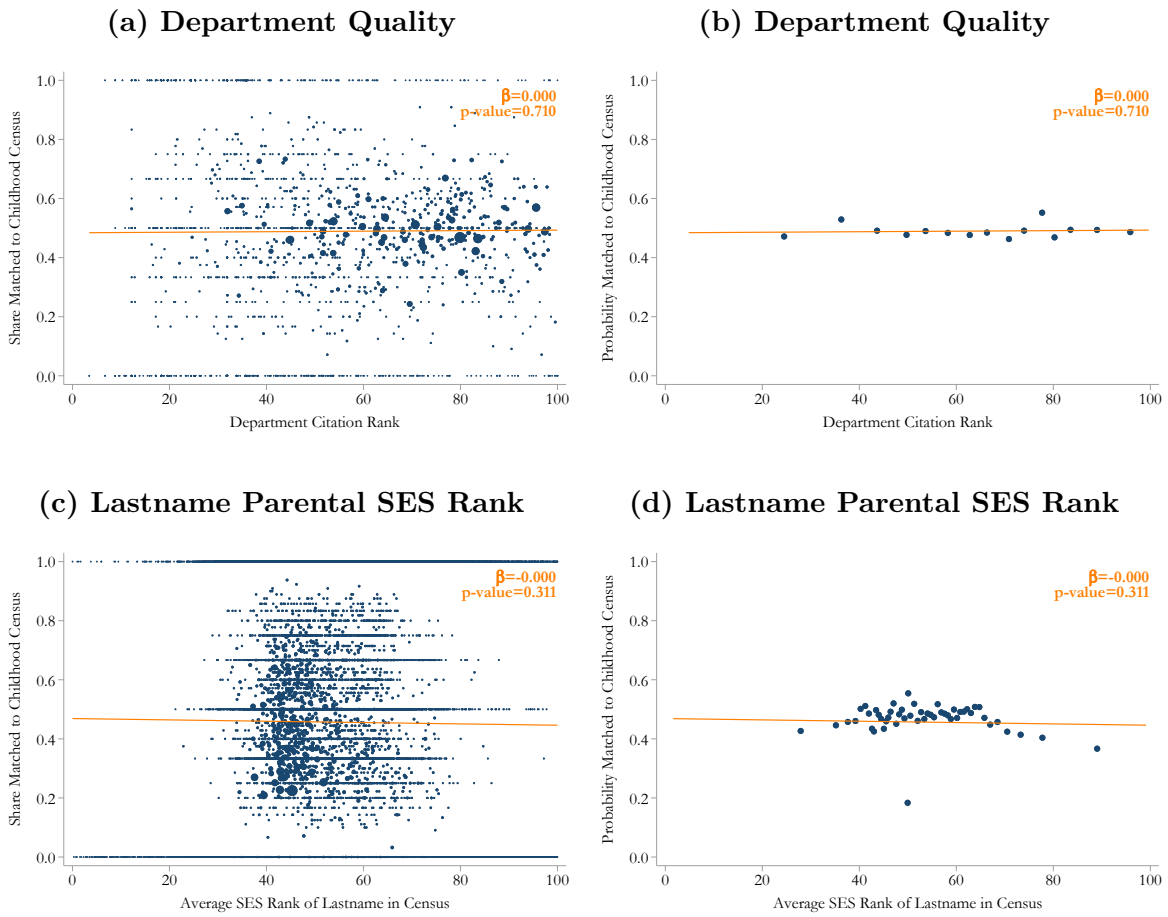
Constructing Parental SES ranks

For our baseline results, we rely on father's occupational income scores to proxy for socio-economic background. We do so because other measures of parental socio-economic status (such as parental income or parental education) are not available in pre-1940 US censuses. Specifically, we create a summary measure of the socio-economic background by constructing parental "income scores" for each academic (Abramitzky et al. 2021). We use data on wage income from the 1940 census (the first year for which individual-level income is observed in the U.S. census) to map the father's occupation into an "income score". We estimate the following regression for all working-age (20-70 years old) men in

[10] As we link some academics to multiple adult censuses, a small fraction of them have backlinks to different individuals in a childhood census. For instance, someone who is present in the 1914 faculty roster could, in principle, be matched to both the 1910 and the 1920 adult censuses and it may be the case that the 1920-1880 backlinks might point to a different individual than the 1910-1880 backlinks. Whenever this is the case, we keep the backlink corresponding to the adult census closest in time to the childhood census (in the previous example, we would keep the link based on the 1910-1880 crosswalk).

[11] For academics who moved to the United States to study or when they were already academics, we cannot link them to a childhood census by construction. Of the 76,732 academics who we link to an adult census, 6,769 or 7.9% are foreign-born.

Figure 2: Correlation of Linking Rates With Department Quality and Lastname Parental SES Rank



Notes: Panel (a) shows the correlation between a department's citation rank and the probability of linking a scientist to a childhood census. Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's SES Rank based on the entire U.S. census and the probability of linking an academic to a childhood census. Panel (d) shows a binned scatter plot of the same relationship. Bins are chosen according to Cattaneo et al. (2024).

the 1940 census:

$$\ln(\text{Income}_j) = \beta_0 + \beta_1 \text{Occupation}_j \times \text{State FE} + \beta_2 \text{Age}_j + \beta_3 \text{Age}_j^2 + \beta_4 \text{Race}_j + \epsilon_j \quad (1)$$

where $ln(Income)_j$ measures the income of individual j in 1940. Occupation × State FE is a separate fixed effect for each three-digit occupation interacted with the state of residence of individual j. In addition, we also included a second-order polynomial in age as well as race fixed effects. Because the 1940 census includes information on income from wages but not on other sources of income, we adjust the income of self-employed individuals (including farmers) using the method developed by Collins and Wanamaker (2022).[12]

[12]For a few occupations and states, the number of individuals in certain occupation by state cells is low. Furthermore, some census occupation codes change across census years (see IPUMS (2024a)). In these cases, we use coarser fixed effects to predict income ranks. See Appendix A.1. for details.

We then use the estimated coefficients from equation (1) to predict income for all fathers with at least one child below the age of 21 in each childhood census. We then calculate the income ranks of academics' fathers within the distribution of income ranks among *all* children (for both academics and non-academics) with the same birth year. In robustness tests reported below, we construct alternative SES ranks based on income predictions that do not differ by state.

2.3 Linking Scientists with Publications and Citations

To study how the socio-economic background affects scientific output and the direction of science, we link academics in six scientific disciplines – medicine, biology, biochemistry, chemistry, physics, and mathematics with publication and citation data from the *Clarivate Web of Science*. We focus on these disciplines for two main reasons. First, they have particularly good coverage in the Web of Science. Second, by the start of the 20th century, they had already established a culture of publishing in scientific journals, and the publishing process was similar to today's.

The linking algorithm is based on the procedure developed by Iaria et al. (2024), which links publications and citations based on the academic's surname, first name, or initials (depending on whether first names are available), country, city, and discipline.[13] To improve match quality, we harmonize affiliations across the faculty rosters and the *Web of Science* with the *Google Maps API*.

2.4 Linking Scientists with Nobel Prize Data

To measure recognition by the scientific community, we also collect data on nominations for the physics, chemistry, and physiology or medicine Nobel Prizes from Nobelprize.org (2024), which contains all nominations for the Nobel Prize in physics and chemistry from 1901 to 1970, and all nominations for the Nobel Prize in physiology or medicine from 1901 to 1953. We hand-link the academics in the faculty rosters to the entries in the Nobel Nomination archive to identify individuals who were nominated for the Nobel Prize. We also hand-link all Nobel Prize winners to our faculty rosters. Table 2 provides summary statistics for the most important variables in our data.

[13]To reduce false positives, matches are always based on the primary discipline (e.g., physics) of each academic).

Table 2: Summary Statistics

Panel A: 1900 – 1956			
Variable	Mean	SD	Observations
SES Rank	72.81	24.85	37,838
Age at Entry into Faculty Rosters	45.31	10.13	37,838
Female	0.09		37,838
Publications	4.66	9.51	12,765
Papers with Novel Words	0.30	1.12	11,962
Nominated for Nobel Prize	0.01		12,765
Awarded Nobel Prize	0.00		12,765
Panel B: 1900 – 1969			
SES Rank	72.16	25.06	47,235
Age at Entry into Faculty Rosters	47.34	10.95	47,235
Female	0.10		47,235
Publications	4.91	10.63	15,519
Papers with Novel Words	0.29	1.12	14,716
Nominated for Nobel Prize	0.01		15,519
Awarded Nobel Prize	0.00		15,519

Notes: The table reports summary statistics. Panel A reports information for the main sample, which includes academics who enter the faculty rosters by the 1956 cohort. Panel B reports information for the extended sample, which includes academics who enter the faculty rosters by the 1969 cohort. Data on acacemics come from the *World of Academia Database*. SES ranks are constructed based on U.S. census micro data. Data on publications and citations come from the *Web of Science*. Publications are measured in a ± 5-year window around the year of entering the faculty roster data. Papers with novel words measures the number of publications published in a ± 5-year window around the year of entering the faculty roster data that introduce a novel word (see section 2 for details). Nominated for Nobel Prize is an indicator whether a scientist was ever nominated for a Nobel Prize, and Awarded Nobel Prize is an indicator for winning the Nobel Prize.

3 Socio-Economic Background and the Probability of Becoming an Academic

In the first part of the paper, we investigate the relationship between socio-economic background and the probability of becoming an academic.

Many anecdotes suggest that even exceptionally talented individuals from lower socio-economic backgrounds often face challenges in pursuing academic careers. For example, in his *Recollections*, Nobel Prize winner George Beadle stated that: "It was tacitly assumed I would eventually take over the family farm. [...] Father was not keen on the college idea, being convinced that a farmer did not need all that education. But determination won, and I enrolled at the University of Nebraska College of Agriculture, fully intending to return to the farm." (Beadle, 1974).

In the following, we investigate whether individuals like Goerge Beadle were the exception or to what extent talented individuals were to pursue an academic career independently of their socio-economic background.

3.1 Representation of Academics by Socio-Economic Background

We calculate the share of U.S. academics that come from each percentile of the parental SES rank distribution. It is important to note that the parental SES rank should be interpreted as an omnibus measure of socio-economic background, which captures a combination of different factors. We do not argue that, for example, a lack of parental income is the sole or even predominant driver behind our findings.

An equal distribution based on SES rank would imply that 1% of academics stem from each percentile. We highlight this benchmark with a horizontal line in Figure 3. In stark contrast to this equal representation benchmark, we show that people from higher socio-economic backgrounds are markedly overrepresented in academia with the degree of overrepresentation strongly increasing with higher parental SES ranks (Figure 3, panel a). Overall, around half of all academics come from the top 20% of the SES rank distribution. The overrepresentation is particularly large for very high percentiles of the SES rank distribution. For example, individuals born to parents in the 95th percentile are more than three times as likely to become academics than one would expect under the equal representation benchmark. We also observe a strong discontinuity at the 100th percentile. People from the highest percentile of the socio-economic background distribution are more than five times as likely to become academics than one would expect under the equal representation benchmark. Strikingly, even when compared to individuals from the 99th percentile, they have a 1.6 times higher chance of becoming an academic.[14]

The results are similar if we predict parental SES ranks only based on the father's individual characteristics and his occupation, without using state of residence fixed effects in the income prediction, Figure 3, panel b). In additional robustness checks, we report the share of academics by other measures of socio-economic background (hisclass and Duncan Socioeconomic Index (SEI)). See Figures B.3 and B.4.

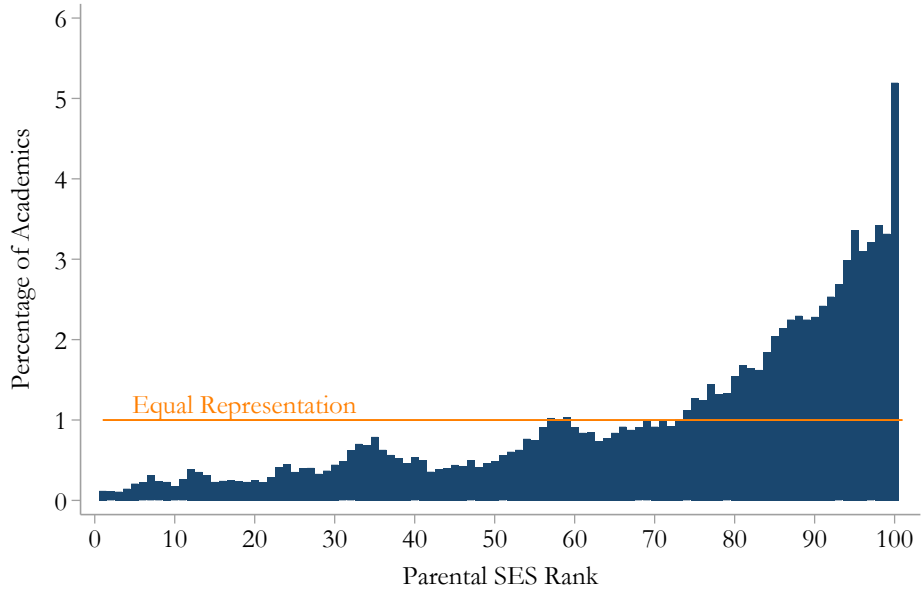
3.2 Representation Over Time

The large differences in the probability of becoming academic translate into a highly skewed composition of academia. As a next step, we analyze whether these representation patterns change over time (Figure 4). The share of academics from the top 20% of the SES rank distribution for the birth cohorts born after 1920 is 52.6%, almost identical to the share of 52.3% in the pre-1870 birth cohorts. Similarly, the share of academics from the bottom 20% of the SES rank distribution is around 4-5% and hardly changes over time. This stability is surprising in light of the sharp rise in educational attainment that took place in the US over our period of study, which could have widened the pool

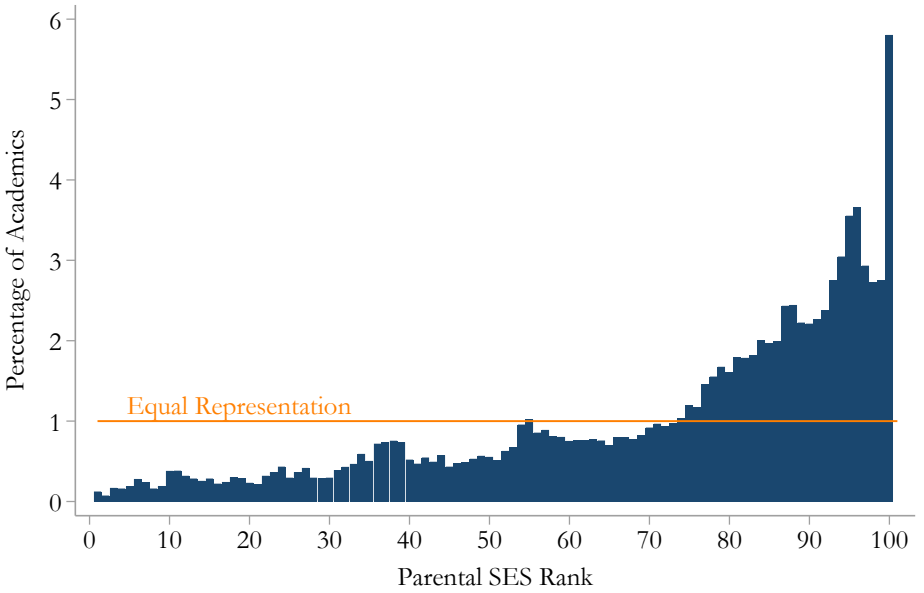
[14]The particularly high probability of becoming academic for individuals born to parents in the 100th percentile could partly be driven by the fact that for some census years and states, professors are ranked in the 100th percentile of the parental income distribution. However, if we remove individuals with fathers who list "professor" as their census occupation, we find a similar pattern (Appendix Figure B.1).

Figure 3: Representation by Socio-Economic Background

(a) Baseline Parental Income Prediction



(b) Parental Income Prediction Without Regional Variation

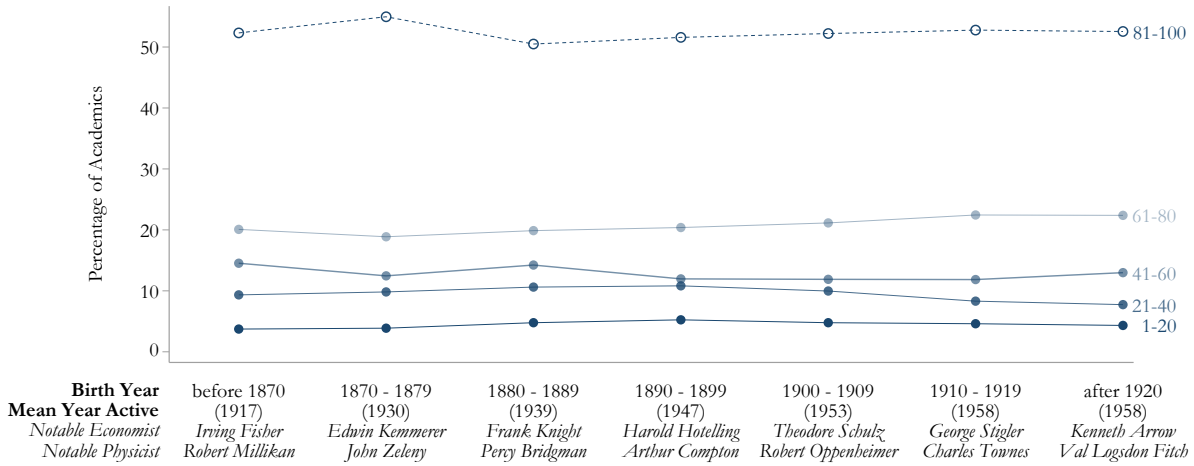


Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each bar represents the percentage of all academics whose fathers are from specific income percentile rank. For example, the right-most bar shows that around 5 percent of academics have fathers who were in the top 1 percent of the predicted income distribution. The horizontal line represents a hypothetical equal representation benchmark.

of potential academics: On average, Americans born in 1920 completed three additional years of schooling than those born in 1870 (Goldin and Katz, 2009).

Together, these results suggest that there appear to be significant barriers that

Figure 4: Representation by Socio-Economic Background Over Time



Notes: The figure shows the representation of academics based on their socio-economic background over time. Each line represents the percentage of all academics whose fathers are from specific income percentile ranks. For example, the top line indicates the percentage of academics whose fathers were in the top 20 percent of the predicted income distribution.

prevent individuals from low socio-economic backgrounds from participating in academia. Such barriers can take many different forms. Examples could include, among others, differences in ability, education, income, network ties, or institutional knowledge. In the following, we shed some light on some parental and institutional differences that appear to contribute to the selection of academics by socio-economic background.

3.3 Representation in Academia versus Other High-Skilled Professions

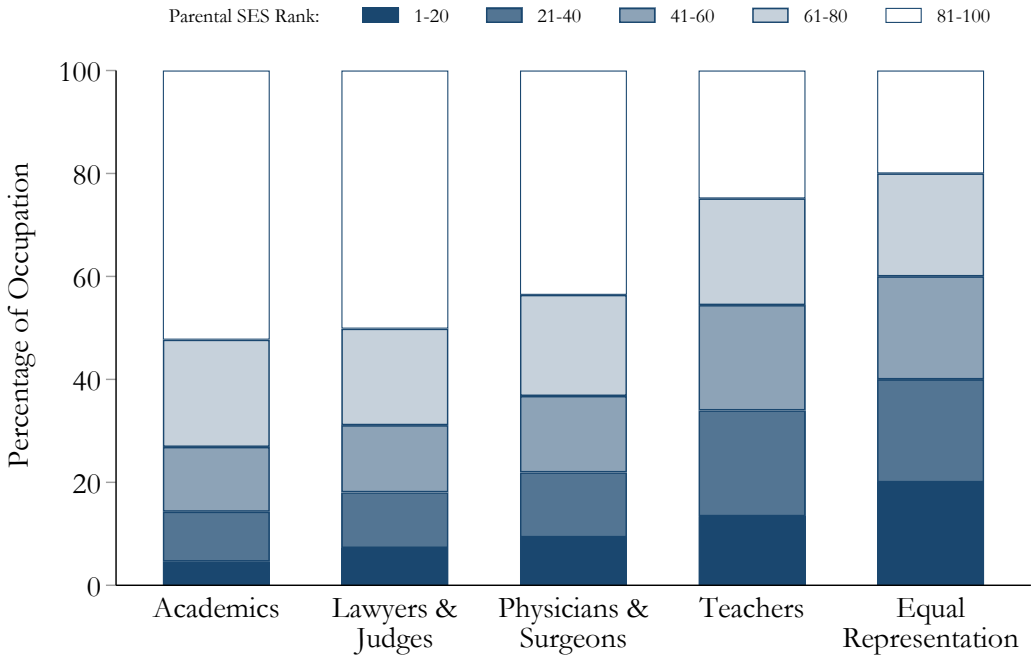
The previous findings raise the question of whether academia is an outlier compared to other high-skilled professions. After all, the low representation of lower-SES individuals in academia might simply reflect the fact that entering such occupation requires credentials (such as a college degree), which might be expensive to obtain. We thus compare the socio-economic background of academics to the background of individuals in other professions.

More specifically, we study the parental socio-economic background of lawyers, doctors, and teachers using comparable data from the census micro-data (see Appendix A.2. for details).

While lawyers and doctors are also recruited from higher socio-economic backgrounds, academics are more selected (Figure 5). For example, 52% of academics come from the top 20% of the SES rank distribution, while only 50% of laywers and judges, and 44% of medical doctors come from the top 20% of the SES rank distribution. In academia, the representation of individuals from the bottom 20% is particularly low: only 5% of academics come from the bottom 20% percent while 7% of laywers, and 9% of doctors come

from the bottom 20%. Teachers, in contrast, exhibit a much weaker degree of selection based on socio-economic background.

Figure 5: Comparison to other High-Skilled Professions



Notes: The figure compares the representation of academics based on their socio-economic background to representation in other professions. The representation in other professions is based on U.S. census samples of lawyers & judges, physicians & surgeons, and teachers that match the sample of academics (see Appendix A.2. for details).

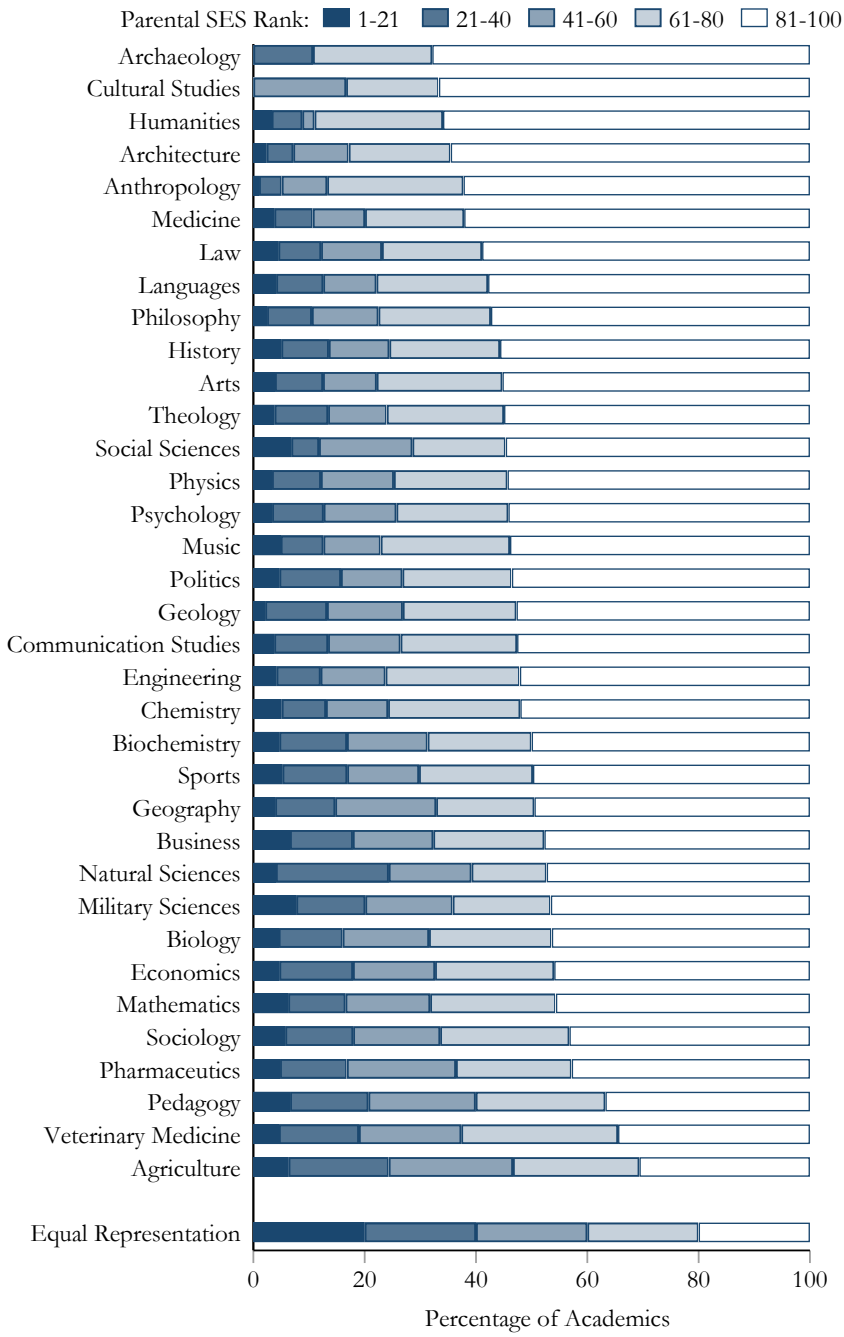
3.4 Representation in Academic Disciplines and Universities

In the next set of results, we investigate whether individuals from lower socio-economic backgrounds face similar hurdles in all disciplines and universities or if some disciplines or universities exhibit a higher representation of individuals from such backgrounds.

Disciplines

While individuals from higher socio-economic backgrounds are overrepresented in all disciplines, there are large differences across disciplines (Figure 6). Agriculture, veterinary medicine, pedagogy, pharmaceutics and sociology are the disciplines with the highest representation of individuals from lower socio-economic backgrounds. In contrast, archaeology, cultural studies, the humanities, architecture, anthropology, medicine, and law have the lowest representation. Contrary to the common perception of economists, economics is more representative than the median discipline.

Figure 6: Representation by Discipline



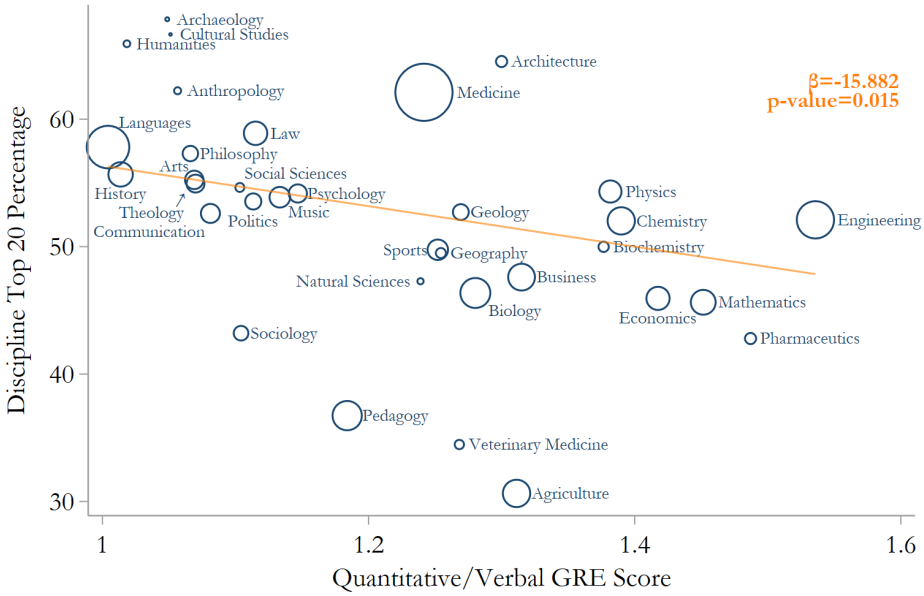
Notes: The figure shows the representation of academics based on their socio-economic background by academic discipline. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

Particularly striking are the differences in the socio-economic composition of medicine relative to veterinary medicine and sociology relative to law. Even though these disciplines arguably require similar training and skills, the institutional setting in each of these disciplines appears to lead to large differences in the participation of people from different

socio-economic backgrounds.

A broad pattern that emerges from Figure 6 is that disciplines that require more sophisticated language skills exhibit lower representation from individuals from poorer backgrounds. In comparison, disciplines that require more mathematics skills, on average, exhibit higher representation. To investigate this hypothesis, we correlate discipline-level representation with the language versus mathematics skills requirement in each discipline. We proxy the language versus mathematics requirement with the ratio of quantitive over verbal GRE scores for students intending to pursue graduate studies in the respective discipline.[15] The findings suggest that representation from lower socio-economic backgrounds is indeed higher in disciplines that require more quantitative relative to verbal skills (Figure 7).

Figure 7: Discipline Mathematics vs. Language Requirements and Representation



Notes: The figure shows the share of academics from the top quintile of the distribution of socio-economic background by academic discipline in relation to the importance of quantitative relative to verbal skills in the discipline. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. We proxy the importance of mathematics relative to language skills with the ratio of the average GRE quantitative score to the average verbal reasoning GRE of test takers intending to pursue a graduate degree in the respective discipline. GRE score data come from (ETS) (2009), Extended Table 4. The size of the circles indicates the number of academics in the respective discipline in our data.

[15]The Educational Testing Service, which administers GRE tests, publishes three-year average test scores of seniors and nonenrolled college graduates in 3 categories (verbal reasoning, quantitative reasoning and analytical writing) for 290 intended graduate majors in their *GRE Guide to the Use of Scores*. We aggregate these to reflect our coding of academic specializations into disciplines. We take the data for the 2005-2008 cohorts of test-takers ((ETS) (2009)), the oldest available via the Internet Archive Wayback Machine.

Universities

Next, we study heterogeneity by university. As the faculty rosters contain more than 1,000 U.S. universities, we can only plot a small subset of the universities. To choose this subset, we show all universities for which we measure the socio-economic background of at least 10 academics per cohort and of, on average, 100 academics for each of the five cohorts in the main sample. We also show all universities in the Ivy Plus group, as defined by Chetty et al. (2020a).[16]

There are striking differences in representation by university (Figure 8). The most "socioeconomically selective" universities are elite private universities such as those in the Ivy League – Harvard, Princeton, UPenn, and Yale. At the lower end of selectivity (among this subset of universities) are public, often land-grant, universities like the University of Nebraska, the University of Missouri, and Iowa State University.
We investigate which university characteristics are correlated with socioeconomic selectivity, by estimating the following regression on the full sample of universities:

$$\begin{aligned} Faculty\ Top\ SES\ Share_u = \beta_0 + \beta_1 \text{Ivy Plus}_u + \beta_2 \text{Elite Private}_u + \beta_3 \text{Elite Public}_u \\ + \beta_4 \text{Discipline Shares} + \text{State FE} + \epsilon_i \end{aligned} \quad (2)$$

The dependent variable $Faculty\ Top\ SES\ Share_u$ measures the share of academics of university u who come from the top 20, top 10, top 5, or top 1 percent of the parental SES rank. Ivy Plus_u is an indicator that equals one if university u is an Ivy Plus university as defined by Chetty et al. (2020a). Elite Private_u is an indicator that equals one if university u is an elite private institution which is not in the Ivy Plus category (e.g., New York University) and Elite Public_u is an indicator that equals one if the university is an elite public institution (e.g., Berkeley).[17]

The regression results show that Ivy Plus universities recruit their faculty from higher socio-economic backgrounds than other elite private institutions. These findings hold for the share of faculty from the top 20, top 10, top 5, and even top 1 percent. While the average university in our sample recruits 3 percent of their academics from the top 1 percent, the share is about 3.1 percentage points higher in Ivy Plus universities (Table 3, column 12). Public elite institutions recruit their faculty from similar socio-economic backgrounds as all other universities (Table 3).

Universities' selectivity might partly reflect the discipline composition of a university. For example, Harvard does not have an agriculture department. To account for differences in the discipline compositions of universities, we add controls for the share of academics

[16]The Ivy Plus group contains the following universities: Brown, Columbia, Cornell, Dartmouth, Harvard, U Penn, Princeton, Yale, Stanford, MIT, Chicago, and Duke.

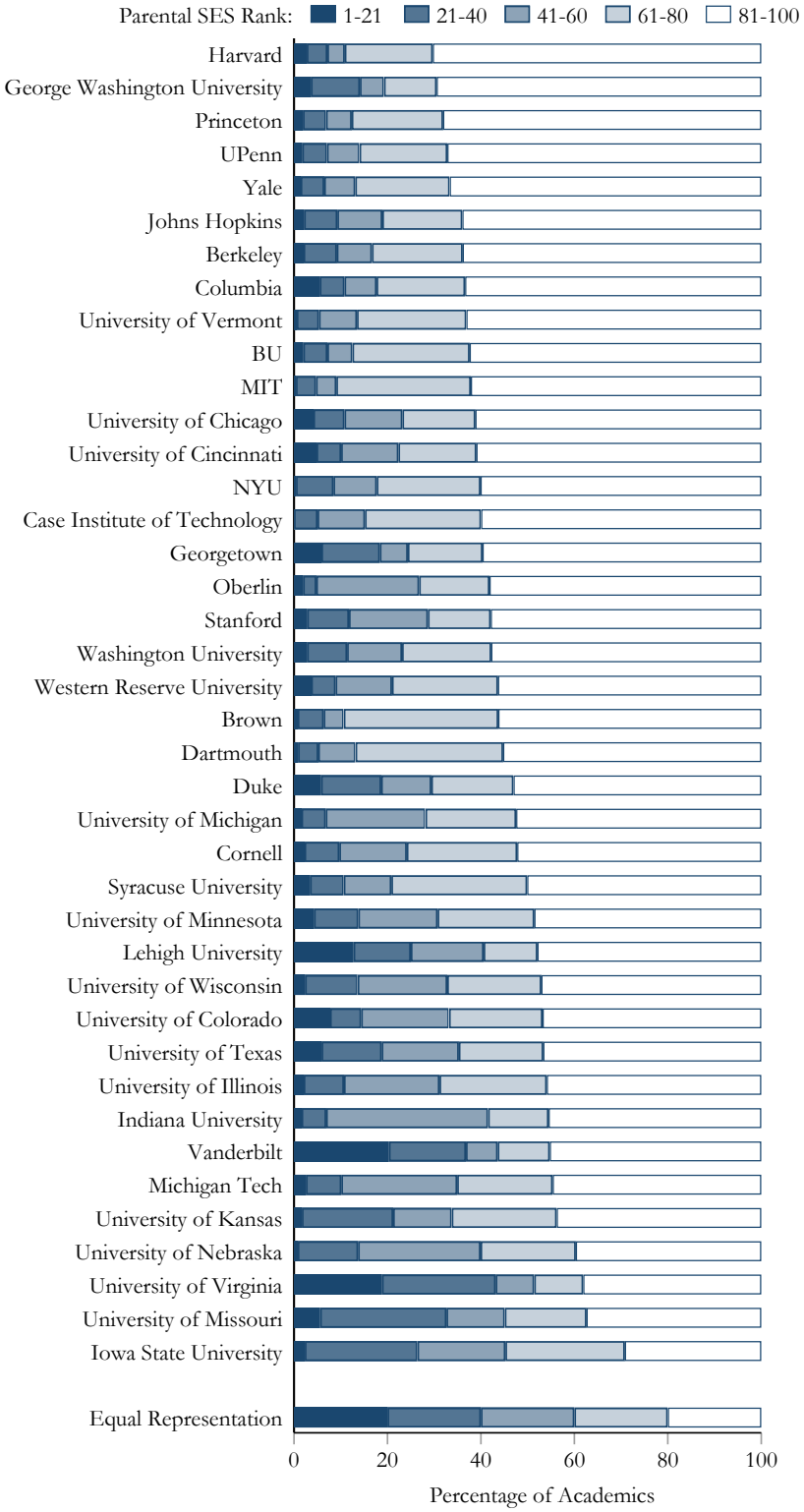
[17]Elite Private includes all private universities in Chetty et al. (2020a)'s "elite universities". Elite Public includes all public universities in Chetty et al. (2020a)'s "elite universities" as well as all universities in their "Highly-Selective Public" category.

Table 3: Correlates of University SES-Selectivity

Dependent Variable:	*Faculty Top SES Share*											
	20%			*10%*			*5%*			*1%*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: 1900 – 1956												
Ivy Plus	9.650**	9.373***	3.407	12.932***	11.059***	6.564***	11.326***	10.361***	5.303***	5.636***	5.716**	5.062**
	(4.250)	(2.294)	(2.536)	(2.818)	(1.969)	(1.889)	(2.139)	(2.064)	(1.777)	(1.028)	(2.239)	(2.131)
Private Elite	9.967***	5.925*	3.328	9.249***	5.473**	3.270	7.839***	4.358**	2.939	2.496***	1.205	1.036
	(3.051)	(3.091)	(3.430)	(2.340)	(2.289)	(2.584)	(1.634)	(1.652)	(1.768)	(0.739)	(0.921)	(0.962)
Public Elite	8.126***	9.005**	7.382*	2.876	2.774	3.199	1.802	1.278	1.658	1.169*	0.805	0.783
	(2.886)	(3.676)	(4.128)	(1.937)	(2.367)	(3.136)	(1.440)	(1.781)	(2.143)	(0.601)	(0.635)	(1.010)
R^2	0.02	0.10	0.19	0.02	0.11	0.19	0.03	0.13	0.22	0.03	0.10	0.16
Observations	756	756	756	756	756	756	756	756	756	756	756	756
Dependent Variable Mean	48.113	48.113	48.113	27.040	27.040	27.040	13.846	13.846	13.846	3.377	3.377	3.377
Panel B: 1900 – 1969												
Ivy Plus	13.950***	10.161***	5.309*	15.135***	11.551***	7.022***	11.165***	9.562***	5.257***	4.639***	4.356***	3.093**
	(3.681)	(2.948)	(2.699)	(2.802)	(2.530)	(2.032)	(2.700)	(1.702)	(1.278)	(0.515)	(1.489)	(1.491)
Private Elite	11.462***	5.590*	3.836	10.203***	5.439**	3.573	7.449***	4.400**	3.394*	2.043***	0.883	0.913
	(3.095)	(3.111)	(3.663)	(2.496)	(2.415)	(2.789)	(1.449)	(1.672)	(1.738)	(0.632)	(0.981)	(0.971)
Public Elite	5.477	0.203	1.457	4.560	0.395	2.422	3.376	0.590	1.639	0.317	-0.516	-1.254
	(4.456)	(3.266)	(2.967)	(3.489)	(3.160)	(2.347)	(2.561)	(2.275)	(1.764)	(0.705)	(0.632)	(0.804)
R^2	0.01	0.09	0.16	0.02	0.11	0.19	0.02	0.10	0.19	0.01	0.05	0.10
Observations	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035	1,035
Dependent Variable Mean	44.992	44.992	44.992	24.649	24.649	24.649	12.857	12.857	12.857	3.439	3.439	3.439
Discipline Controls		Yes	Yes		Yes	Yes		Yes	Yes		Yes	Yes
State FEs			Yes			Yes			Yes			Yes

The table reports the estimates of Equation (2). The dependent variable in columns 1-3, 4-6, 7-9, and 10-12 measures the share of faculty in university u who come from the top 20, top 10, top 5, top 1 percent of the parental SES rank distribution, respectively. Ivy Plus_u is an indicator that equals one if university u is an Ivy Plus university as defined by Chetty et al. (2020b). Elite Privateu is an indicator that equals one if university u is an elite private institution which is not in the Ivy Plus category (e.g., New York University) and Elite Publicu is an indicator that equals one if the university is an elite public institution (e.g., Berkeley). Standard errors are clustered at the state-level. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Figure 8: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

in each discipline. The results remain very similar (columns 2 and 5). The results also remain comparable, if we control for state fixed effects (columns 3 and 6), suggesting that the pattern is not solely driven by geographical factors.

4 Socio-Economic Background, Scientific Publications, and Novel Scientific Concepts

In the second part of the analysis, we investigate whether and how socio-economic background influences productivity after entering academia. In particular, we aim to understand if scientific productivity and novelty differ by socio-economic background.

4.1 Scientific Publications

We first explore differences in the number of publications by socio-economic background. As described above, this analysis focuses on the six scientific disciplines: medicine, biology, biochemistry, chemistry, physics, and mathematics, with good coverage of publications and citations in publication and citation databases. We estimate the following regression:

$$Publications_i = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \tag{3}$$

where $Publications_i$ captures different measures of scientific publications that scientist i published in a $\pm$ 5-year interval around the first cohort that the scientist entered the faculty rosters. For scientists entering the faculty rosters in 1956, we measured publications from 1951 to 1961. We estimate results for publication counts and standardized publications. We standardize publication counts to have a mean of zero and a standard deviation of one by discipline and cohort. The standardized publications ease interpretation and account for differences in publications across disciplines and over time. $Parental\ SES\ Rank_i$ ranges from 0 to 100 and measures the percentile of the income rank of scientist i's father. The coefficient θ captures if there are any differences in scientific output by socio-economic background. $\mathbf{X_i}$ are controls that account for differences in scientific output by age (observed in the census), gender, cohort, discipline, or state. As the parental SES rank is based on father's occupation, childhood state, and birth year of the academic, we cluster standard errors at the level of father's occupation, childhood state, and birth year to account for potential correlations of regression residuals.

Number of Publications

The results from the regression estimates of eq. (3) are presented in Table 4. We find no systematic relationship between the background and the average number of publications, independently of the set of fixed effects that we include as regression controls (see columns

1-3). This result holds for the baseline sample (Panel A) and the extended sample (Panel B). As described before, to account for differences in publication cultures across disciplines and over time, we also estimate models that use publications that we standardize at the cohort and discipline level. These results again confirm that there is no systematic relationship between the socio-economic background of academics and the average number of publications (Table 4, columns 4-6).

Table 4: Socio-Economic Background and Publications

Dependent Variable:	*Publications*			*Standardized Publications*			*No Publications*		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A: 1900 – 1956									
Parental SES Rank	0.00782*	0.00440	-0.00299	0.00040	0.00012	0.00006	-0.00112***	-0.00094***	-0.00051***
	(0.00425)	(0.00424)	(0.00423)	(0.00041)	(0.00041)	(0.00042)	(0.00019)	(0.00019)	(0.00018)
R^2	0.04	0.06	0.09	0.04	0.06	0.06	0.05	0.07	0.12
Observations	12,765	12,765	12,765	12,765	12,765	12,765	12,765	12,765	12,765
Dependent Variable Mean	4.664	4.664	4.664	0.011	0.011	0.011	0.418	0.418	0.418
Panel B: 1900 – 1969									
Parental SES Rank	0.00418	0.00158	-0.00628	0.00016	-0.00005	-0.00014	-0.00102***	-0.00085***	-0.00043**
	(0.00408)	(0.00407)	(0.00407)	(0.00036)	(0.00036)	(0.00036)	(0.00017)	(0.00017)	(0.00017)
R^2	0.03	0.05	0.08	0.03	0.06	0.06	0.04	0.06	0.12
Observations	15,519	15,519	15,519	15,519	15,519	15,519	15,519	15,519	15,519
Dependent Variable Mean	4.911	4.911	4.911	-0.013	-0.013	-0.013	0.421	0.421	0.421
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes			Yes

Notes: The table reports the estimates of equation (3) The dependent variable measures publications in a ± 5-year window around the cohort when academic i enters the faculty rosters. We standardize publications to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explantory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Demographic controls include age, age squared and an indicator for whether the academic is female. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the academic. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

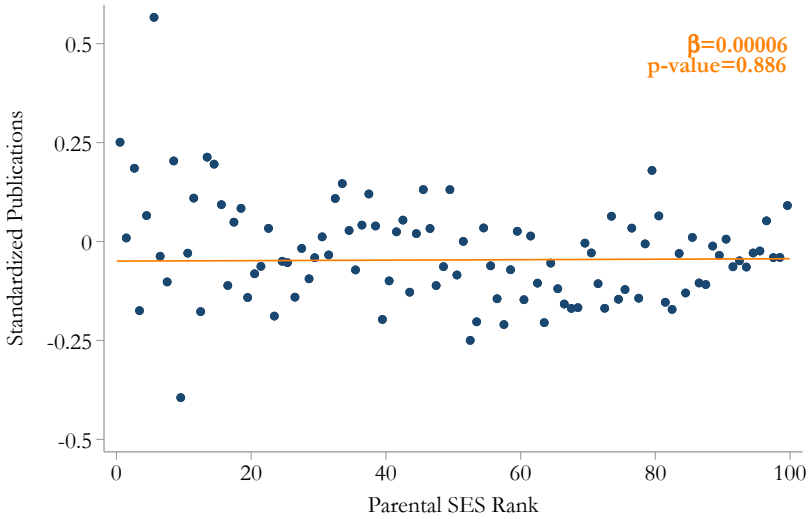
We also visualize the the relationship between parental income ranks (x-axis) and standardized publications (y-axis) in a binned scatterplot (see Figure 9). The figure provides additional evidence that there is no systematic relationship between the *average* number of publications and the parental income rank.

Probability of Zero Publications

As a substantial number of academics never publish in a journal covered by the Web of Science, we investigate the relationship between socio-economic background and the probability of never publishing in a journal covered by the Web of Science.

Therefore, we re-estimate equation (3) with an alternative dependent variable that equals one if academic i does not publish any papers in the ±5 year window surrounding their entry into the faculty rosters, and zero otherwise. We find that individuals from lower socio-economic backgrounds are significantly more likely to never publish a paper. Specifically, the probability of not publishing at all is approximately 4 percentage points (or around 10 percent) higher for scientists whose fathers were at the 75th percentile of the income rank, compared to scientists with fathers at the 25th percentile (Table 4, column

Figure 9: Socio-Economic Background and Average Number of Publications



Notes: The figure shows a binned scatterplot of the relationship between academics' socio-economic background and publications. Publications are standardized within cohort and discipline. We show 100 quantiles and use the covariate adjustment (equivalent to column (4) in Table 4) as proposed in Cattaneo et al. (2024).

7, significant at the 1 percent level). The result is around half as large if we include the complete set of fixed effects but remains highly significant. It is also robust in the extended sample which includes academics who enter the faculty rosters in 1969 (Table 4, columns 8-9 and Panel B).

The Distribution of Publications

The previous results suggest that although academics from lower socio-economic backgrounds, on average, produce a comparable total number of publications, they exhibit a higher likelihood of having no publications at all. This suggests that academics from lower socio-economic backgrounds must publish relatively more in higher percentiles of the publication distribution.

To test this hypothesis, we estimate equation (3) with alternative dependent variables:

$$\mathbb{1}(Publication\ Percentile\ Range_i = q) = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \quad (4)$$

where the dependent variable $\mathbb{1}(Publication\ Percentile\ Range_i = q)$ is indicator that equals one if scientist i's publication record is in a certain percentile range. Since scientific productivity is well-known to be highly skewed (see e.g., Lotka 1926), we use the following percentile ranges q of the publication distribution: bottom 50% (which coincides with not publishing at all for most disciplines and cohorts), $> 50-70$th, $> 70-90$th, $> 90-95$th, $> 95-97$th, $97-99$th, and > 99th percentile of the publication distribution. As publication patterns differ greatly across disciplines (chemists and medical researchers publish a lot

more papers than mathematicians) and across cohorts (academics in later cohorts publish more papers), we calculate these percentiles at the discipline cohort level.

The regression results are reported in Appendix Table C.1 and we plot the estimated coefficients in Figure 10. The first coefficient from the left indicates that academics from higher parental income ranks are less likely to have a publication count in the bottom 50% of the publication distribution. The second coefficient ($> 50 - 70$) indicates that academics from higher parental income ranks are as likely to have a publication count between the 50th and the 70th percentile of the publication distribution. The third coefficient ($> 70 - 90$) indicates that academics from higher parental income ranks are more likely to have a publication count between the 70th and the 90th percentile of the publication distribution than academics from lower parental income ranks. For the next percentile ranges, the coefficients are not significantly different from zero. In contrast, the last coefficient ($> 99 - 100$), indicates that academics from higher parental income ranks are less likely to have a publication count in the top 1% of the publication distribution. In other words, individuals from lower socio-economic backgrounds are more likely to have a publication count in the top 1%. Specifically, the probability of having a publication record in the top 1% is approximately 0.35 percentage points (or around 44 percent) lower for scientists whose fathers were at the 75th percentile of the income rank, compared to scientists with fathers at the 25th percentile. This large effect, in percentage terms, is highly relevant as a long-standing literature in the sociology of science has highlighted that the most productive scientists have a disproportionate impact on the advancement of science (e.g., Lotka 1926, Merton 1957).

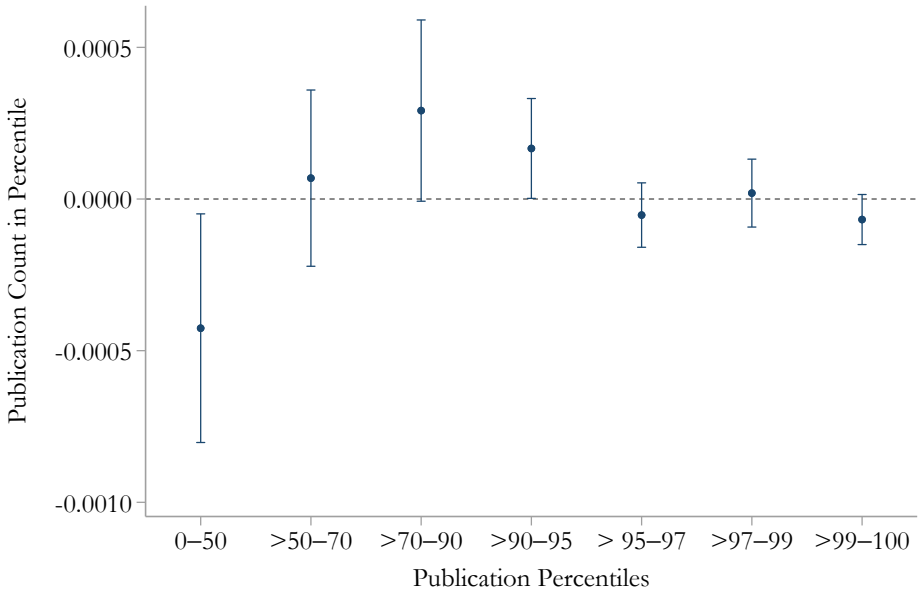
Overall, the results on the distribution of publications suggest that academics from lower socio-economic backgrounds are somewhat "riskier" hires as they are more likely to have zero publications but also more likely to be in the top 1% of the publication distribution.

4.2 Novel Scientific Concepts

In the next sub-section, we explore whether and how the content of publications differs by socio-economic background and whether we find additional evidence that suggests that academics from lower socio-economic backgrounds may work on riskier research agendas.

To explore these hypotheses, we follow the methodology developed by Iaria et al. (2018) and construct a measure of novel words that a scientist introduced to the scientific community. The measure proxies for the introduction of new scientific concepts that required novel scientific terms. We define novel words as words that were first used in the title of a paper and that had not been used in any prior paper title covered by the entire Web of Science.

Figure 10: Socio-Economic Background and the Distribution of Publications



Notes: The figure plots the estimated coefficients for θ for seven regressions of Equation 3. In each of the seven regressions, the dependent variable is an indicator of whether a scientist's number of publications lies within the relevant percentiles of the publication distribution, measured at the cohort and discipline-level. We report coefficients from regressions using the covariate and fixed effects equivalent to column (3) in Table 4. Complete regression results are reported in Appendix Table C.1.

As the coverage of the Web of Science begins in 1900, we compute the novel words measure for paper titles published from 1910 onwards. This allows for a 10-year period to measure words appearing in academic papers, before designating a new word as novel. Thus, we cannot measure the introduction of novel words for scientists who enter the faculty rosters in 1900. To ensure that we do not consider words that were already in use in other domains, we exclude frequently used words, as well as numbers, from the data.[18]

One example of a novel scientific term is "Sulfolobus acidocaldarius," a bacterium in hot springs in Yellowstone National Park. It was discovered by Thomas D. Brock. Brock grew up on a farm, and his father had never received a formal education. Later in life Brook became a professor of microbiology at Western Reserve University, Indiana University, and the University of Wisconsin. His research on thermophilic microorganisms led to the discovery of an organism thriving at 70 °C (160 °F). The ability of an enzyme

[18]We exclude the 20,000 most frequently used words in English-language books contained in the Project Gutenberg database as of April, 16 2006 (available at `https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists#English`). Project Gutenberg currently contains the full text of over 70,000 books. Because the database contains books whose copyright has expired, the typical book in the database was published before 1923. The most frequently used words, therefore, reflect historical language use that is relevant to the period of our analysis. The results are robust to excluding only 10,000 or all 36,663 frequently used words that are reported in Project Gutenberg. For the main results, we do not remove all frequently used words because words such as quantum (on position 17,132) may have existed before but might have taken on a new meaning with the publication of a scientific paper. For more details on the novel scientific words measure, see Iaria et al. (2018).

to tolerate high temperatures would, 20 years later, lead to the invention of a procedure called polymerase chain reaction (PCR), which led to the award of the 1993 Nobel Prize in Chemistry to Kary Mullis and Michael Smith.

To study how the parental SES rank is related to introducing novel scientific terms, we estimate the following regression:

$$Novel\ Words_i = \omega \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \tag{5}$$

where $Novel\ Words_i$ measures the number of papers with at least one novel word that scientist i published in the $\pm$ 5-year interval around entering the faculty rosters. I.e., for scientists entering the faculty rosters in 1956, we measure the number of papers published between 1951 and 1961 that introduced at least one novel word. To ease interpretation, and to account for differences in the number of novel words introduced in different disciplines and over time, we standardize the novel word counts to have a mean of zero and a standard deviation of one by discipline and cohort. As before, $Parental\ SES\ Rank_i$ ranges from 0 to 100 and measures the percentile of the income rank of scientist i's father. $\mathbf{X_i}$ are controls that account for differences in scientific output by age, cohort, and discipline.

Table 5: Socio-Economic Background and Novelty

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00089*	-0.00101**	-0.00100**	-0.00073*	-0.00090**	-0.00090**
	(0.00048)	(0.00047)	(0.00048)	(0.00043)	(0.00044)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,970	11,970	11,970	11,970	11,970	11,970
Dependent Variable Mean	0.301	0.301	0.301	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00076*	-0.00084**	-0.00085**	-0.00074**	-0.00085**	-0.00087**
	(0.00042)	(0.00041)	(0.00042)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,724	14,724	14,724	14,724	14,724	14,724
Dependent Variable Mean	0.291	0.291	0.291	-0.011	-0.011	-0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (5). The dependent variable measures the number of publications which introduce at least one novel word and were published in a $\pm$ 5-year window around the cohort when academic i enters the faculty rosters. We exclude the 20211 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

The baseline specification controls for age, gender, childhood state fixed effects and

cohort fixed effects. We find that scientists from higher socio-economic backgrounds introduce fewer novel words (Table 5, column 1, significant a the 10% level). The result is similar if we control for university state and discipline fixed effects (Table 5, columns 2-3). Specifically, scientists whose fathers were at the 75th percentile of the income rank publish around 0.05 fewer papers (around 17 percent) with at least one novel word compared to scientists with fathers at the 25th percentile.

The result is also robust to standardizing the novel words measure at the level of disciplines and cohorts (Table 5, columns 4-6) and in the extended sample (Table 5, panel B).

5 Socio-Economic Background and Recognition

In the following section, we examine the relationship between socio-economic background and recognition by other academics. First, we analyze citations to an academic's research papers, a widely-used metric for measuring recognition within the academic community. Next, we investigate Nobel Prize nominations and awards as indicators of recognition for exceptional scientific contributions.

5.1 Citations

To estimate the relationship between socio-economic background and citations, we switch to an analysis at the paper level, which allows us to abstract from potential differences in the number of publications. The data includes all papers linked to at least one author for whom we can measure the parental SES rank. We estimate the following regression:

$$Citations_p = \gamma \cdot Avg.\ Parental\ SES\ Rank_p + \mathbf{X'_p}\beta + \epsilon_p \tag{6}$$

where $Citations_p$ measures the number of citations that paper p received until 2010. To account for differences in citations across disciplines and over time, we standardize citations at the level of disciplines and the year of publication.[19] $Avg\ Parental\ SES\ Rank_p$ measures the average SES rank of the fathers (ranging from 0 to 100) of all authors of paper p that we can link to a childhood census. $\mathbf{X_i}$ are controls for the characteristics of the paper. Whenever we measure characteristics at the author level, we average them for all authors of paper p that we can link to a childhood census. As the parental SES rank is based on the father's occupation, childhood state, and birth year, we cluster standard errors at the level of father's occupation, childhood state, and birth year to account for potential correlations of regression residuals.

[19]To capture the whole distribution of citations for the standardization, we use citations to all papers linked to U.S. academics in the faculty rosters and not only the citations to papers of U.S. academics, which we can link to a childhood census.

The results indicate that papers published by authors from higher socio-economic backgrounds receive somewhat more citations (Table 6, panel A, column 1). While the coefficient is positive in all specification it is only significantly different from zero in some of them.

Table 6: Parental SES Rank and Paper-Level Citations

Dependent Variable:	*Standardized Citations*				
	(1)	(2)	(3)	(4)	(5)
Panel A: 1900 – 1956					
Average Parental SES Rank	0.00057***	0.00041**	0.00040**	0.00041**	0.00041**
	(0.00020)	(0.00020)	(0.00020)	(0.00018)	(0.00018)
R^2	0.04	0.05	0.05	0.15	0.15
Observations	58,549	58,549	58,549	58,549	58,549
Dependent Variable Mean	-0.024	-0.024	-0.024	-0.024	-0.024
Panel B: 1900 – 1969					
Average Parental SES Rank	0.00053***	0.00045**	0.00044**	0.00042***	0.00041***
	(0.00018)	(0.00018)	(0.00018)	(0.00016)	(0.00015)
R^2	0.03	0.04	0.04	0.15	0.16
Observations	76,014	76,014	76,014	76,014	76,014
Dependent Variable Mean	-0.024	-0.024	-0.024	-0.024	-0.024
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes
Publication Year FEs	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes	Yes	Yes
Discipline FEs			Yes	Yes	Yes
Journal FEs			Yes	Yes	Yes
Author Count FEs					Yes

Notes: The table reports the estimates of Equation (6). The dependent variable measures the number of citations received by paper p until 2010. We standardize citations at the level of displines and years, to account for differences in citations patterns. The main explanatory variable is the average SES rank of the fathers of all authors of paper p that can be linked to a childhood cenus. We measure the SES rank of fathers with the percentile in the predicted income distribution the father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Homophily of Citations by Income Rank

The differences in citations are surprising in light of the previous findings, which indicated that low-SES academics introduced more novel words. We shed more light on one of the factors that could underlie these differences: the homophily of citations based on socio-economic-background.

For this analysis, we use the fact that the Web of Science allows us to reconstruct the entire citation network. This allows us to analyze the socio-economic background of the authors of articles cited by a paper p. For each paper p, we reconstruct the average parental SES rank of all cited authors. Naturally, we can only measure the socio-economic background of cited authors who can be linked to their childhood census. Additionally, we cannot measure the socio-economic background of cited authors who are not in our

database either because they are Ph.D. students or because they are from other countries. Put differently, the sample consists of a subsample of papers from the previous analysis for which we can measure the socio-economic background of at least one cited author. We then estimate the following regression:

$$Avg.\ SES\ Rank\ Cited\ Authors_p = \nu \cdot Avg.\ SES\ Rank\ Citing\ Authors_p + \mathbf{X'_p}\beta + \epsilon_p \quad (7)$$

where $Avg.\ SES\ Rank\ Cited\ Authors_p$ is the average SES rank of the authors of paper cited by paper p. $Avg.\ SES\ Rank\ Citing\ Authors_p$ is the average SES rank of the authors of paper p. The vector $\mathbf{X'_p}\beta$ contains controls for the average age, age^2, the female share, fixed effects for the childhood state, and fixed effects of the discipline of the citing paper p. Standard errors are again clustered at the level of father's occupation, childhood state, and birth year for the citing paper.

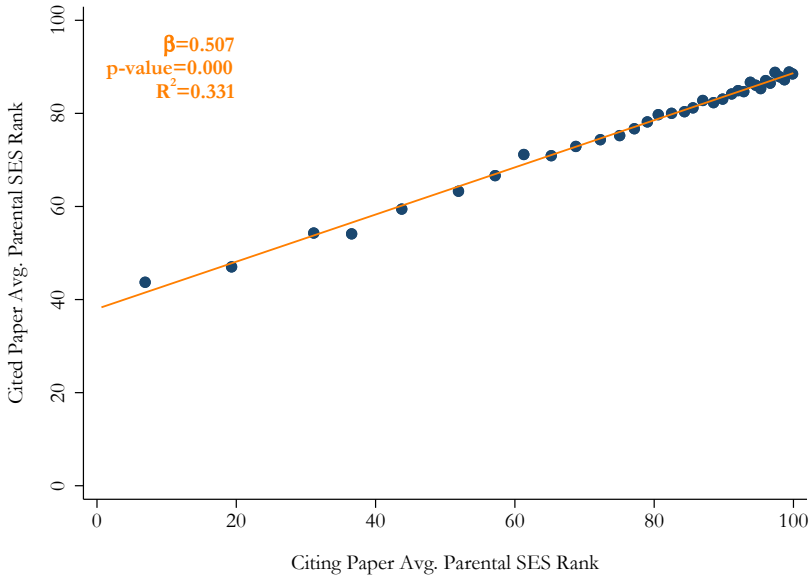
We show the regression estimates from eq. (7) in Table 7. We find a strong positive association between the average parental SES rank citing and cited paper. Authors with a 10 percentile higher average SES rank on average cite papers from authors with a 5 percentile higher average SES rank. This pattern is highly robust even when we control for the university state, discipline, journal, and author counts. It is particularly surprising that even within the same discipline and journal, authors from higher SES ranks appear to cite papers from other high-SES authors.

Table 7: Homophily in Citations

Dependent Variable:	*Average Parental SES Rank Cited Authors*				
	(1)	(2)	(3)	(4)	(5)
Panel A: 1900 – 1956					
Citing Paper Avg. Parental SES Rank	0.503***	0.501***	0.499***	0.493***	0.493***
	(0.011)	(0.010)	(0.010)	(0.010)	(0.010)
R^2	0.35	0.36	0.36	0.38	0.38
Observations	28,459	28,459	28,459	28,459	28,459
Dependent Variable Mean	76.58	76.58	76.58	76.58	76.58
Panel B: 1900 – 1969					
Citing Paper Avg. Parental SES Rank	0.528***	0.526***	0.524***	0.512***	0.512***
	(0.010)	(0.009)	(0.009)	(0.009)	(0.009)
R^2	0.37	0.38	0.38	0.41	0.41
Observations	43,652	43,652	43,652	43,652	43,652
Dependent Variable Mean	75.02	75.02	75.02	75.02	75.02
Demographic Controls	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes
Publication Year FEs	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes	Yes	Yes
Discipline FEs			Yes	Yes	Yes
Journal FEs			Yes	Yes	Yes
Author Count FEs					Yes

We also visualize the underlying variation of this regression in a binscatter plot (see Figure 11). The figure suggests that the relationship between the socio-economic background of the authors of the cited paper holds along the entire SES background distribution. Further, the parental SES background of citing authors explains a third of all variation in the variation in SES background of the cited papers ($R^2 = 0.331$).
To the best of our knowledge, we are the first to document this strong homophily in citations on the basis of socio-economic background. The homophily could stem from differences in topics that individuals of different socio-economic backgrounds research or from interactions at certain departments or conferences that are predominately attended by individuals from certain socio-economic backgrounds. Given the strong overrepresentation of individuals from higher socio-economic backgrounds in academia, which we have documented, the homophily in citations provides a potential explanation for the higher number of citations that the papers of individuals from higher socio-economic backgrounds receive.

Figure 11: Homophily in Citations



Notes: The figure shows

5.2 Nobel Prize: Nominations and Awards

Nobel Prize Nominations

Next, we study an alternative measure of recognition that captures whether elite scientists consider another academic's body of research worthy of a nomination for a Nobel Prize (Iaria et al. (2018)). In this period, nominations for the Nobel Prize were made by a small group of elite scientists and therefore might be subject to biases. We study this question

using the following regression:

$$\mathbb{1}\{Nobel\ Nomination/Award_i\} = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \tag{8}$$

where $\mathbb{1}\{Nobel\ Nomination/Award_i\}$ is an indicator for whether scientist i was ever nominated for a Nobel Prize (*Nobel Nomination*) or an indicator for whether the scientist has ever been awarded a Nobel Prize (*Nobel Award*). As before, $Parental\ SES\ Rank_i$ ranges from 0 to 100 and measures the percentile of the income rank of scientist i's father. $\mathbf{X_i}$ are controls as defined above.

We find that individuals from higher parental SES ranks are more likely to be nominated for a Nobel Prize. A scientist from the 75th percentile of the income distribution, on average, has a 0.06 percentage point higher probability of being nominated, an increase of about 50 percent (Table 8, column 1). The results are robust to controlling for the state of the academic's university and the discipline (Table 8, columns 2-3). The results are also robust to controlling for both publications and citations (Table 8, columns 4-6), indicating that academics from poorer backgrounds are less likely to be nominated for a Nobel Prize even conditional on publications and citations. Overall, these results suggest that academics from lower socio-economic backgrounds receive less recognition by their peers when it comes to being nominated for the Nobel Prize.

Table 8: Socio-Economic Background and Nobel Prize Nominations

Dependent Variable:	*Nobel Nomination*					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1900 – 1956						
Parental SES Rank	0.00011***	0.00010**	0.00012***	0.00010**	0.00009**	0.00012***
	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)	(0.00004)
R^2	0.01	0.02	0.03	0.08	0.08	0.10
Observations	12,765	12,765	12,765	12,765	12,765	12,765
Dependent Variable Mean	0.012	0.012	0.012	0.012	0.012	0.012
Panel B: 1900 – 1969						
Parental SES Rank	0.00010***	0.00009**	0.00010***	0.00009***	0.00009**	0.00010***
	(0.00003)	(0.00003)	(0.00004)	(0.00003)	(0.00003)	(0.00003)
R^2	0.01	0.02	0.03	0.07	0.07	0.08
Observations	15,519	15,519	15,519	15,519	15,519	15,519
Dependent Variable Mean	0.011	0.011	0.011	0.011	0.011	0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Publication & Citation Controls				Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of equation (3). The dependent variable is an indicator whether a scientist was ever nominated for a nobel prize. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Demographic controls include age, age squared and an indicator for whether the academic is female. Publication and citation controls are a scientist's standardized total publication and citation counts. We standardize publication and citation counts to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the academic. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Nobel Prize Awards

In Table 9, we also investigate the probability of winning a Nobel Prize. We again find that scientists from higher parental SES ranks are more likely to be nominated for a Nobel Prize. A scientist from the 75th percentile of the income distribution has a 0.015 percentage point higher probability of winning a noble prize, an increase of about 50 percent relative to the mean. This finding persists even conditional on a scientist's publication and citation record.

Table 9: Socio-Economic Background and Nobel Prize Awards

Dependent Variable:	*Nobel Award*					
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1900 – 1956						
Parental SES Rank	0.00003	0.00002	0.00003*	0.00002	0.00002	0.00003*
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
R^2	0.01	0.01	0.02	0.03	0.03	0.04
Observations	12,765	12,765	12,765	12,765	12,765	12,765
Dependent Variable Mean	0.003	0.003	0.003	0.003	0.003	0.003
Panel B: 1900 – 1969						
Parental SES Rank	0.00003*	0.00003*	0.00004**	0.00003*	0.00003*	0.00003**
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
R^2	0.01	0.01	0.02	0.02	0.02	0.03
Observations	15,519	15,519	15,519	15,519	15,519	15,519
Dependent Variable Mean	0.002	0.002	0.002	0.002	0.002	0.002
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Publication & Citation Controls				Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of equation (3). The dependent variable is an indicator whether a scientist was ever nominated for a nobel prize. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Demographic controls include age, age squared and an indicator for whether the academic is female. Publication and citation controls are a scientist's standardized total publication and citation counts. We standardize publication and citation counts to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the academic. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

6 Socio-Economic Background and Discipline Choice

In the third part of the paper, we investigate the importance of socio-economic background for the choice of academic discipline beyond its impact on the probability of becoming an academic. In particular, we study whether individuals from different socio-economic backgrounds, measured by father's occupation, specialize in different disciplines.

6.1 Overrepresentation of Individuals with Certain Father's Occupations in Specific Disciplines

For this analysis, we construct an overrepresentation index that measures whether individuals with fathers in certain occupations are overrepresented in specific academic disciplines:

$$\text{Overrepresentation}_{do} = \frac{P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o)}{P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)}, \tag{9}$$

where $P(\text{Discipline}_i = d)$ is the probability of academic i working in discipline d, similarly $P(\text{Father's Occupation}_i = o)$ is the probability of academic i having a father with occupation o and $P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o)$ is the joint probability.[20]

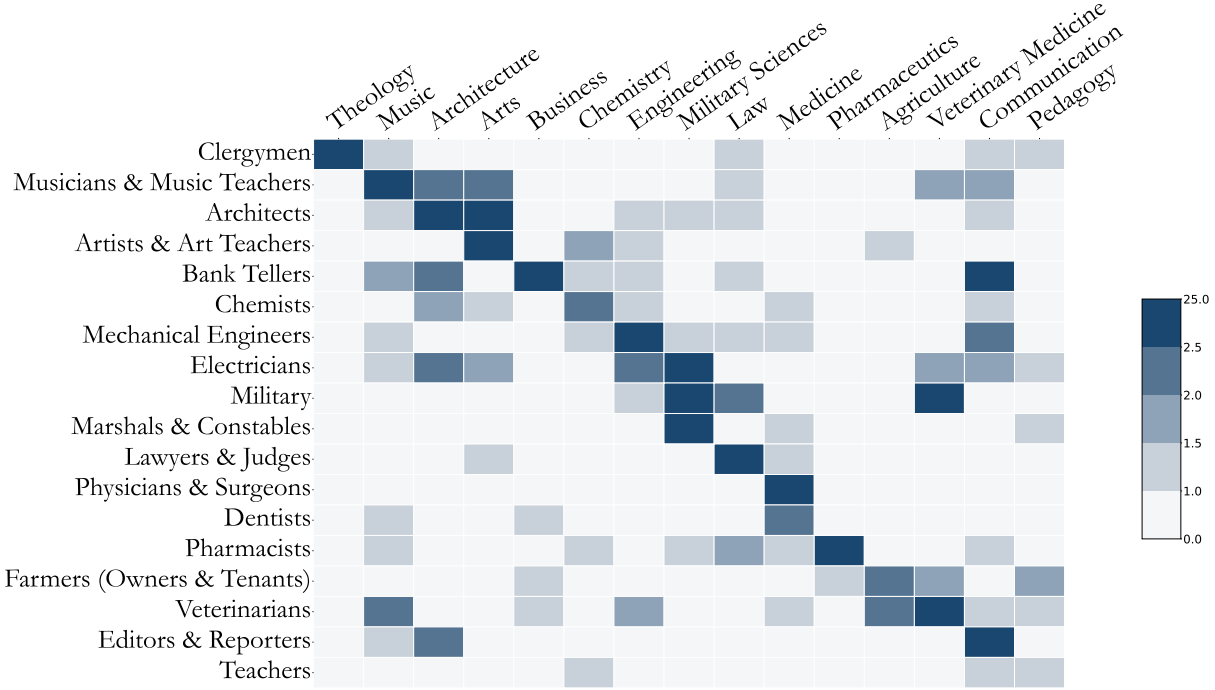
The measure abstracts from baseline differences in the probability of choosing a specific academic discipline and from baseline differences in the probability of having a father in a certain occupation. If there was no systematic relationship between father's occupation and the discipline (i.e., the probabilities are independent), $P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o) = P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)$ and $\text{Overrepresentation}_{od} = 1$. For the father's occupation-discipline pairs that are overrepresented in the data, the measure is greater than one. Inversely, for the father's occupation-discipline pairs that are underrepresented in the data, the measure is smaller than one.

For example, we can calculate the overrepresentation of farmers' children in the academic discipline of agriculture. The numerator measures the probability that an academic whose father was a farmer works as a professor of agriculture (in our data this probability is 0.024). The denominator is the product of the probability of being a professor of agriculture, among all disciplines (in our data: 0.043), and the probability of having a father who was a farmer, among all father's occupations (in our data: 0.232). Thus the overrepresentation index for professors of agriculture who are farmer's children is 2.3. In other words, 56% ($0.024/0.043 \times 100$) of all agricultural scientists are the children of farmers, while only 23% of all academics are children of farmers, making agricultural scientists 2.3 times more likely to be the child of a farmer than all academics.

We calculate this measure for all pairs of father's occupations (130) and academic disciplines (34), i.e., we calculate $130 \times 34 = 4,420$ overrepresentation indices. We visualize examples of such pairs in Figure 12. The figure plots the father's occupation on the vertical axis and the academic discipline on the horizontal axis. The blue shading indicates the quartiles of overrepresentation distribution, with darker blues indicating stronger overrepresentation.

[20]The measure is related to pointwise mutual information, a common measure in information theory.

Figure 12: Father's Occupation and Discipline Choice



Notes: The figure shows the relationship between father's occupation (rows) and their children's academic discipline choice (columns) for selected father's occupation - discipline pairs. Darker shades indicate higher levels of overrepresentation, as measured by equation (9).

The figure suggests a strong connection between the father's occupation and their children's choice of discipline. For example, children of architects are overrepresented in the academic disciplines of architecture and arts. Children of artists and art teachers are overrepresented in the academic discipline of arts. Children of lawyers, medical doctors, or pharmacists choose law, medicine, and pharmaceutics as their disciplines, respectively. Children of editors and reports choose communication studies, which also includes journalism as a sub-discipline. Interestingly, the connection exists even among children with fathers in non-professional occupations. For instance, children of bank tellers are overrepresented in the academic discipline business. Children of teachers who teach various school subjects spread more equally across disciplines.

These findings highlight that academics are more likely to work in disciplines that are close to their father's occupation. This suggests a strong intergenerational transmission of the affinity to disciplines.

6.2 Predicting Semantically Close Academic Disciplines

To reduce clutter, Figure 12 only presents a selected subset of father's occupation – discipline pairs. To more systematically explore whether a father's occupation affects discipline choice, we use natural language processing techniques to construct an external measure of

semantic similarity between each father's occupation and each academic discipline. We then investigate whether children of fathers with a certain occupation are more likely to enter disciplines where the text string of the father's occupation (e.g., "farmer") is close in semantic space to the text string of the discipline (e.g., "agriculture").

Specifically, we measure semantic similarity using embeddings. Embeddings transform a text into a fixed-length vector representation that encodes the syntactic and semantic relationships in the training data. The resulting vectors can then be used for text similarity calculations, as similar sentences are located close to each other in the vector space. Intuitively, if the word "farmer" is used in similar contexts to the word "agriculture", the model will identify these words as being semantically similar. Embedding models are trained by applying advanced machine learning techniques, such as deep learning transformer models, to vast corpora of text such that the model learns the relationship between words.

We use the "all-MiniLM-L6-v2" model, which has been trained on data from Wikipedia, scientific papers, Reddit, and many other sources.[21] The model represents each father's occupation string as well as each discipline string as a vector of length $n = 384$.

As is standard in natural language processing, we then measure the similarity of the text string of the father's occupation and the text string of the discipline using the cosine similarity of the two vector representations:

$$Cosine\ Similarity(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{10}$$

where x represents the vector of father's occupation and y represents the vector of the discipline, derived from the sentence embedding model.

Using this measure of semantic similarity, we predict the closest discipline in semantic space for each father's occupation. Similarly, we predict the second closest discipline in semantic space. Note, this measure of semantic similarity is entirely based on the occupation and discipline strings and does not use any information on the actual discipline choice of professors. E.g., not surprisingly the closest discipline in semantic space for the occupation "architect" is "architecture" (cosine similarity 0.77). The second closest discipline in semantic space is "engineering" (cosine similarity 0.53). Similarly, the closest discipline in semantic space for the occupation "buyers/shipers of farm products" is "agriculture" (cosine similarity 0.53). The second closest discipline in semantic space is "business" (cosine similarity 0.41).[22]

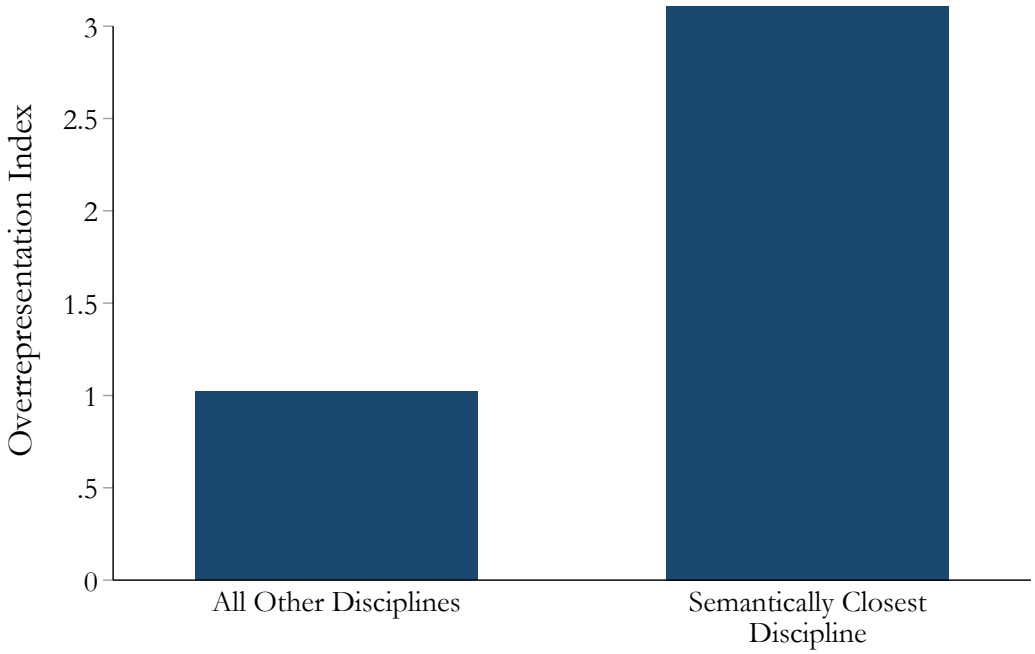
[21]The "all-MiniLM-L6-v" model is one of the most commonly used sentence embedding models. For example, it was the third most downloaded model on huggingface.com as of July 2024. The findings do not depend on the choice of a specific model.

[22]To ensure that we predict close disciplines that are actually close in semantic space, we only classify an occupation-discipline pair as semantically close if their cosine similarity is at least two standard deviations above the mean of all cosine similarities. Without this condition, the most similar discipline to the occupation "private household worker", for example, would be "law" as there is no other discipline that is

6.3 Overrepresentation in Semantically Close Disciplines

Having identified the semantically closest academic discipline for each occupation, we then calculate the average overrepresentation index (equation 9) across all such discipline-occupation pairs, as well as across all other discipline-occupation pairs in the data (Figure 13).

Figure 13: Overrepresentation in Semantically Closest Discipline



Notes: The figure shows overrepresentation as measured by equation (9) in the father's occupation-discipline pair that is semantically closest, e.g., "farmer" and "agriculture" and all other father's occupation - discipline pairs. For more details, see section 6.2 and section 6.3.

The average overrepresentation index is 3.29 in the semantically closest discipline. In contrast, the overrepresentation index is 1.03 for all other disciplines, indicating that for disciplines that are not related to fathers' occupations, academics are indeed represented as good as random.

Overall, these results provide evidence that parental occupation affects the discipline choice of academics, indicating that socio-economic background not only affects the probability of becoming an academic but also the choice of discipline. Possible explanations could include heightened interest due to exposure to a particular field of study, family values, or access to resources and opportunities, including better information on how to "make it" in a given discipline. Combined with the results in the previous part of the paper, these results suggest that the unequal selection of academics based on socio-

closer in semantic space. The results are robust to defining semantically close disciplines if their cosine similarity is only one standard deviation above the mean, or without enforcing a minimum cosine similarity (see Appendix Figure D.3).

economic background could have repercussions on the composition of academic disciplines. If professors from certain parental occupations are overrepresented in academia, certain disciplines face higher supply of talent which will benefit some disciplines over others independent of the actual societal need for knowledge from certain domains.

6.4 Scientific Output and Working in Patrimonial Disciplines

Next, we explore differences in output for scientists who work in the discipline that is closest to their father's occupation (*patrimonial* discipline). As above, we use the semantic similarity of fathers' occupation and academics' disciplines to define patrimonial discipline choices (see section 6.2). We augment equation 3 with an indicator for whether the academic works in a scientific discipline that is "close" in semantic space to their fathers' occupation:

$$Output_i = \delta \cdot Patrimonial\ Discipline_i + \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \qquad (11)$$

The parameter δ captures the effect of working in a patrimonial discipline over and above the effect of parental socio-economic status, i.e., the children of "physicians and surgeons" or "therapists and healers" who become medical researchers might publish because they work in a discipline that is close to their father's occupation even conditional on their father's socio-economic background.

The results from this analysis are presented in Table 10. We find that scientists who work in their patrimonial discipline, i.e., the discipline that is semantically closest to their father's occupation, publish on average 0.07 standard deviations more papers (equivalent to around 1.5 additional papers in a ±5-year period around the cohort of first observing the scientist) than other academics (Table 10, Panel A, column 2). This effect remains similar if we control for additional fixed effects and also holds in the extended sample (Table 10, Panel B).

The results in the previous section suggest that the patrimonial socio-economic background affects scientific productivity in ways that go beyond the father's income.

7 Conclusion

This paper examines the role of socio-economic background in shaping the careers of academics and their research output. We show that people from higher socio-economic backgrounds are more likely to become academics and that there is large heterogeneity in representation at the level of disciplines and universities. Once in academia, the socio-economic background is not related to the number of publications, on average, but individuals from lower socio-economic backgrounds are more likely to not publish at all as well as more likely to have outstanding publication records, making them somewhat

Table 10: Patrimonial Disciplines and Publications

Dependent Variable:	*Publications*			*Standardized Publications*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1900 – 1956						
Patrimonial Discipline	1.62926***	1.53895***	0.61450	0.07296**	0.07074**	0.06127*
	(0.36900)	(0.36485)	(0.37376)	(0.03658)	(0.03592)	(0.03669)
SES Rank	0.00346	0.00035	-0.00444	0.00020	-0.00007	-0.00008
	(0.00452)	(0.00451)	(0.00446)	(0.00043)	(0.00043)	(0.00043)
R^2	0.04	0.06	0.09	0.04	0.06	0.06
Observations	12,765	12,765	12,765	12,765	12,765	12,765
Dependent Variable Mean	4.664	4.664	4.664	0.011	0.011	0.011
Panel B: 1900 – 1969						
Patrimonial Discipline	1.66260***	1.59392***	0.49175	0.06638**	0.06577**	0.04859
	(0.37239)	(0.37007)	(0.37760)	(0.03310)	(0.03263)	(0.03328)
SES Rank	-0.00011	-0.00247	-0.00741*	-0.00001	-0.00021	-0.00025
	(0.00427)	(0.00426)	(0.00421)	(0.00038)	(0.00038)	(0.00038)
R^2	0.03	0.05	0.08	0.03	0.06	0.06
Observations	15,519	15,519	15,519	15,519	15,519	15,519
Dependent Variable Mean	4.911	4.911	4.911	-0.013	-0.013	-0.013
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of equation (11). The dependent variable in columns 1-3 measures publications in a ± 5-year window around the cohort when academic i enters the faculty rosters. The dependent variable in columns 4-6 measures standardized publications. We standardize publicationsto have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The first main explanatory variable measures the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. The second main explanatory variable is an indicator that equals 1 if the child is an academic in the patrimonial discipline that is predicted by the father's occupation (based on the semantic similarity of father's occupation and the discipline). Demographic controls include age, age squared and an indicator for whether the academic is female. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the academic. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1. Patrimonial Disciplines are: Dentists, Physicians and Surgeons, Therapists and Healers – Medicine; Biological Scientists – Biology; Chemical Engineers, Chemistry-Professors, Chemists – Chemistry.

riskier hires. The results on novel words suggest that they are somewhat more likely to pursue research agendas off the beaten path which may result in scientific breakthroughs but also in a higher failure rate. We also find evidence that academics from lower socio-economic background receive less recognition by the scientific community, as measured by citations and Nobel Prize nominations and awards. Lastly, we find that father's occupation is systematically related to the choice of discipline and that academics who work in a discipline that is close to their father's occupation are more productive.

Overall, the paper highlights the importance of understanding the role of socio-economic background in shaping the academic workforce and the creation of new knowledge.

References

Abramitzky, R., L. Boustan, K. Eriksson, J. Feigenbaum, and S. Pérez (2021). Automated Linking of Historical Data. *Journal of Economic Literature 59*(3), 865–918.

Abramitzky, R., L. Boustan, E. Jacome, and S. Perez (2021). Intergenerational Mobility of Immigrants in the United States over Two Centuries. *American Economic Review 111*(2), 580–608.

Abramitzky, R., L. P. Boustan, and K. Eriksson (2012). Europe's Tired, Poor, Huddled Masses: Self-selection and Economic Outcomes in the Age of Mass Migration. *American Economic Review 102*(5), 1832–1856.

Agarwal, R. and P. Gaule (2020). Invisible Geniuses: Could the Knowledge Frontier Advance Faster? *American Economic Review: Insights 2*(4), 409–24.

Aghion, P., U. Akcigit, A. Hyytinen, and O. Toivanen (2018). On the Returns to Invention within Firms: Evidence from Finland. *AEA Papers and Proceedings 108*, 208–212.

Aghion, P., U. Akcigit, A. Hyytinen, and O. Toivanen (2023). Parental Education and Invention: the Finnish Enigma. *mimeo National Bureau of Economic Research*.

Airoldi, A. and P. Moser (2024). Inequality in Science: Who Becomes a Star? *mimeo NYU Stern*.

Azoulay, P., C. Fons-Rosen, and J. S. G. Zivin (2019). Does Science Advance One Funeral at a Time? *American Economic Review 109*(8), 2889–2920.

Azoulay, P., T. Stuart, and Y. Wang (2014). Matthew: Effect or Fable? *Management Science 60*(1), 92–109.

Beadle, G. W. (1974). Recollections. *Annual Review of Biochemistry 43*(1), 1–14.

Bell, A., R. Chetty, X. Jaravel, N. Petkova, and J. Van Reenen (2019). Who becomes an Inventor in America? The Importance of Exposure to Innovation. *The Quarterly Journal of Economics 134*(2), 647–713.

Bloom, N., C. I. Jones, J. Van Reenen, and M. Webb (2020). Are ideas getting harder to find? *American Economic Review 110*(4), 1104–1144.

Buckles, K., A. Haws, J. Price, and H. E. Wilbert (2023). Breakthroughs in Historical Record Linking Using Genealogy Data: The Census Tree Project. *mimeo National Bureau of Economic Research*.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2020). Are Referees and Editors in Economics Gender Neutral? *The Quarterly Journal of Economics 135*(1), 269–327.

Card, D., S. DellaVigna, P. Funk, and N. Iriberri (2022). Gender Differences in Peer Recognition by Economists. *Econometrica 90*(5), 1937–1971.

Cattaneo, M. D., R. K. Crump, M. H. Farrell, and Y. Feng (2024). On Binscatter. *American Economic Review 114*(5), 1488–1514.

Chetty, R., D. J. Deming, and J. N. Friedman (2023). Diversifying Society's Leaders? The Causal Effects of Admission to Highly Selective Private Colleges. *mimeo National Bureau of Economic Research*.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2020a). Income Segregation and Intergenerational Mobility Across Colleges in the United States. *The Quarterly Journal of Economics 135*(3), 1567–1633.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2020b). Income segregation and intergenerational mobility across colleges in the united states. *The Quarterly Journal of Economics 135*(3), 1567–1633.

Collins, W. J. and M. H. Wanamaker (2022). African American Intergenerational Economic Mobility since 1880. *American Economic Journal: Applied Economics 14*(3), 84–117.

Croix, D. d. l. and M. Goñi (2024). Nepotism vs. Intergenerational Transmission of Human Capital in Academia (1088-1800). *Journal of Economic Growth*, 1–46.

Dal Bó, E., F. Finan, O. Folke, T. Persson, and J. Rickne (2017). Who Becomes a Politician? *The Quarterly Journal of Economics 132*(4), 1877–1914.

Dossi, G. (2024). Race and Science. *mimeo UCL*.

Einio, E., J. Feng, and X. Jaravel (2022). Social Push and the Direction of Innovation. *mimeo CEP LSE*.

(ETS), E. T. S. (2009). Gre guide to the use of scores 2009-2010. `https://web.archive.org/web/20121222214014/http://ets.org/Media/Tests/GRE/pdf/gre_0910_guide_extended_table4.pdf`. Extended Table 4, archived on December 22, 2012.

Goldin, C. and L. F. Katz (2009). *The race between education and technology*. harvard university press.

Hager, S., C. Schwarz, and F. Waldinger (2024). Measuring Science: Performance Metrics and the Allocation of Talent. *American Economic Review forthcoming*.

Hengel, E. (2022). Publishing while Female: Are Women Held to Higher Standards? Evidence from Peer Review. *The Economic Journal 132*(648), 2951–2991.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The Allocation of Talent and US Economic Growth. *Econometrica 87*(5), 1439–1474.

Iaria, A., C. Schwarz, and F. Waldinger (2018). Frontier Knowledge and Scientific Production: Evidence from the Collapse of International Science. *The Quarterly Journal of Economics 133*(2), 927–991.

Iaria, A., C. Schwarz, and F. Waldinger (2024). Gender Gaps in Academia: Global Evidence Over the Twentieth Century. *mimeo LMU Munich*.

Ingram, P. (2021). The forgotten dimension of diversity. *Harvard Business Review 99*(1), 58–67.

IPUMS (2024a). Census Occupation Codes, 1950 Basis. `https://usa.ipums.org/usa-action/variables/OCC1950#codes_section`, Last accessed on 2024-09-07.

IPUMS (2024b). Integrated Occupation and Industry Codes and Occupational Standing Variables in the IPUMS. `https://usa.ipums.org/usa/chapter4/chapter4.shtml`, Last accessed on 2024-09-07.

Jin, G. Z., B. Jones, S. F. Lu, and B. Uzzi (2019). The Reverse Matthew Effect: Consequences of Retraction in Scientific Teams. *The Review of Economics and Statistics 101*(3), 492–506.

Koffi, M. (2024). Innovative Ideas and Gender Inequality. *mimeo University of Toronto*.

Koning, R., S. Samila, and J.-P. Ferguson (2021). Who do We Invent for? Patents by Women Focus More on Women's Health, but Few Women Get to Invent. *Science 372*(6548), 1345–1348.

Kozlowski, D., V. Larivière, C. R. Sugimoto, and T. Monroe-White (2022). Intersectional Inequalities in Science. *Proceedings of the National Academy of Sciences 119*(2).

Lotka, A. J. (1926). The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences 16*(12), 317–323.

Merton, R. K. (1957). Priorities in Scientific Discovery: a Chapter in the Sociology of Science. *American Sociological Review 22*(6), 635–659.

Michelman, V., J. Price, and S. D. Zimmerman (2022). Old Boys' Clubs and Upward Mobility among the Educational Elite. *The Quarterly Journal of Economics 137*(2), 845–909.

Moreira, D. and S. Pérez (2022). Who Benefits from Meritocracy? *mimeo National Bureau of Economic Research*.

Morgan, A. C., N. LaBerge, D. B. Larremore, M. Galesic, J. E. Brand, and A. Clauset (2022). Socioeconomic Roots of Academic Faculty. *Nature Human Behaviour 6*(12), 1625–1633.

Nobelprize.org (2024). Nomination Archive. `http://www.nobelprize.org/nomination/archive/`.

Ruggles, S., C. Fitch, R. Goeken, J. D. Hacker, J. Helgertz, E. Roberts, M. Sobek, K. Thompson, J. R. Warren, and J. Wellington (2019). IPUMS Multigenerational Longitudinal Panel.

Ruggles, S., C. A. Fitch, R. Goeken, J. D. Hacker, M. A. Nelson, E. Roberts, M. Schouwiler, and M. Sobek (2024). IPUMS Ancestry Full Count Dataset: Version 4.0.

Stansbury, A. and K. Rodriguez (2024). The Class Gap in Career Progression: Evidence from US Academia. *mimeo MIT*.

Stansbury, A. and R. Schultz (2023). The Economics Profession's Socioeconomic Diversity Problem. *Journal of Economic Perspectives 37*(4), 207–230.

Thorp, H. H. (2023). It Matters Who Does Science. *Science 380*(6648), 873–873.

Truffa, F. and A. Wong (2022). Undergraduate Gender Diversity and Direction of Scientific Research. *mimeo Northwestern University*.

van Leeuwen, M. and I. Maas (2011). *Hisclass: A Historical International Social Class Scheme*. G - Reference,Information and Interdisciplinary Subjects Series. Leuven University Press.

Appendix

The Appendix presents details on data collection and additional results:

- Appendix A provides further details on the construction of the data.
- Appendix B reports robustness checks and additional findings related to section section 3.
- Appendix C reports robustness checks and additional findings related to section section 4.
- Appendix D reports robustness checks and additional findings related to section section 6.

A Appendix: Additional Details on Data

A.1. Constructing SES Ranks

As described in the main paper, we use the 1940 census to predict income. In section 2.2 we describe how we use interactions of fathers' occupation and home state to predict their income (see Equation 1). In practise, this approach faces two issues:

1. Rare Occupations
2. Changing Occupation Coding

We address these issues by adjusting the income prediction for fathers in affected occupations.

Rare Occupations For a few occupations and states, the number of individuals in certain occupation by state cells in 1940 is low, potentially leading to inaccurate predictions for affected Occupation × State FEs. For example, only four working age male actors reported their income in the 1940s census in Montana, and only one in Wyoming. For occupation by state cells with less than 10 observations, we use the following regression to predict income:

$$\begin{aligned}\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation}_i \times \text{Region FE} + \beta_2 \text{State FE} \\ + \beta_3 \text{Age}_i + \beta_4 \text{Age}_i^2 + \beta_5 \text{Race}_i + \epsilon_i \end{aligned} \tag{A.1}$$

Instead of interacting occupations with states, we interact them with census regions, and estimate a separate state fixed effect for all occupations. Predictions for fathers in the same occupation by state cells in earlier census years are based on this regression. That

is, actors in Montana in 1880 are assigned a prediction based on all 1940 actors in the Mountain Division, plus a Montana fixed effect.
For even rarer occupations, those that have less than 10 observations in a given occupation by census *region* cell, we adjust our prediction further:

$$\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation}_i + \beta_2 \text{State FE} + \beta_3 \text{Age}_i + \beta_4 \text{Age}_i^2 + \beta_5 \text{Race}_i + \epsilon_i \quad \text{(A.2)}$$

Rather than estimating region-specific occupational wage profiles, we now base our prediction on national averages. Only two occupation by region cells are subject to this adjustment: Milliners and Loom Fixers, both in the Mountain Division.

Changing Occupation Coding Historically, the Census Bureau has sometimes changed the codes corresponding to specific occupations. For example, the code for actors (and actresses) was 13 from 1850 to 1900, 828 in 1910 and 1920, 192 in 1930, 020 in 1940 and 001 in 1950. To ease comparability across census years, all earlier census occupation codings were also coded into the 1950s classification scheme by IPUMS.[23] We exclusively use the integrated 1950 occupation classification in this paper. The harmonization process resulted in some 1950 occupation codes being present in earlier samples, but not in 1940.[24] For example, the 1950 occupation classification includes one code for mining engineers and one code for metallurgical engineers, whereas the 1940 occupation classification pools the two engineering fields. 1930 and 1920 contain mining engineers again. For 1950 occupation codes that were not present in 1940, we predict fathers' income via this regression:

$$\ln(\text{Income}_i) = \beta_0 + \beta_1 \text{Occupation Group}_i \times \text{State FE} + \beta_2 \text{Age}_i + \beta_3 \text{Age}_i^2 + \beta_4 \text{Race}_i + \epsilon_i, \quad \text{(A.3)}$$

where an Occupation Group is the broad one-digit occupational category of an occupation.[25]

A.2. Constructing Comparison Group Samples for Other Professions

Since we do not have access to a panel similar to the World of Academia Database for lawyers and judges, physicians and surgeons, and teachers we need to construct these samples from the U.S. Census. This is done in four steps:

[23] See Integrated Occupation and Industry Codes and Occupational Standing Variables in the IPUMS.
[24] At the writing of this paper, accurate income data from the 1950s census was not available yet.
[25] Professional, Technical; Farmers; Managers, Officials, and Proprietors; Clerical and Kindred workers; Sales workers; Craftsmen; Operatives; Service workers (private household); Service workers (not household); Farm Laborers; Laborers (non-farm). See 1950 Occupation Codes and Groups.

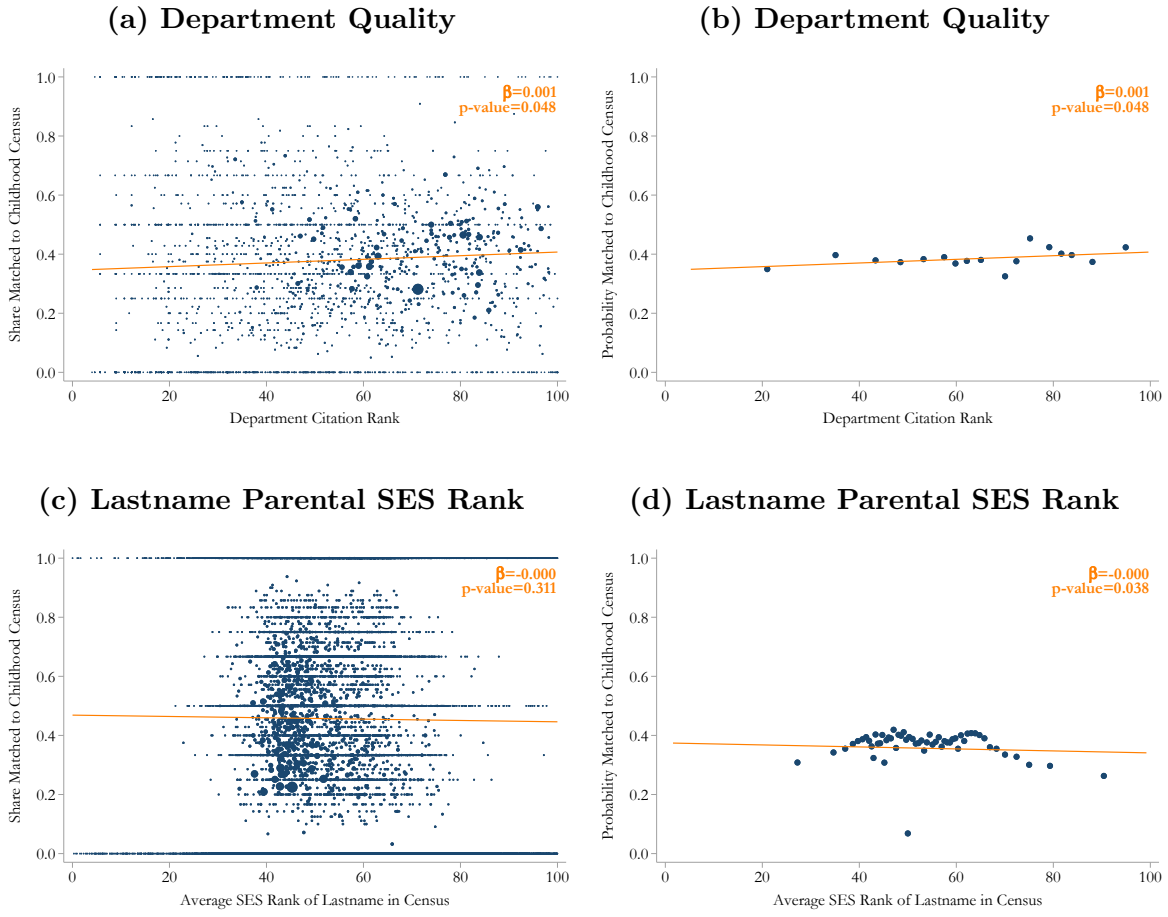
1. From each available full count census in the coverage period of the World of Academia Database (1900-1970), extract all observations with 1950 occupation code 55 (Lawyers & Judges), 75 (Physicians & Surgeons) and 93 (Teachers).

2. Remove all matched academics from this sample.

3. Use Census Linking Project crosswalks to create a panel from the repeated cross-sections of lawyers & judges, physicians & surgeons, and teachers.

4. Match childhood socio-economic backgrounds as described in Sections 2.2 and 2.2.

Figure A.1: Extended Sample 1900-1969: Correlation of Linking Rates With Department Quality and Lastname Parental SES Rank



Notes: Panel (a) shows the correlation between a department's citation rank and the probability of linking a scientist to a childhood census for the extended sample (1900-1969). Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's SES Rank based on the entire U.S. census and the probability of linking an academic to a childhood census for the extended sample (1900-1969). Panel (d) shows a binned scatter plot of the same relationship. Bins are chosen according to Cattaneo et al. (2024).

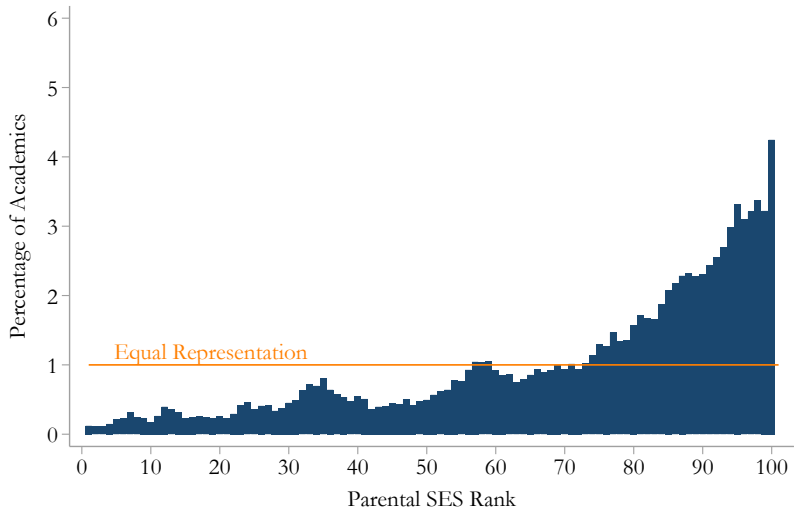
B Additional Results

B.1. Socio-Economic Background and the Probability of Becoming an Academic: Additional Results

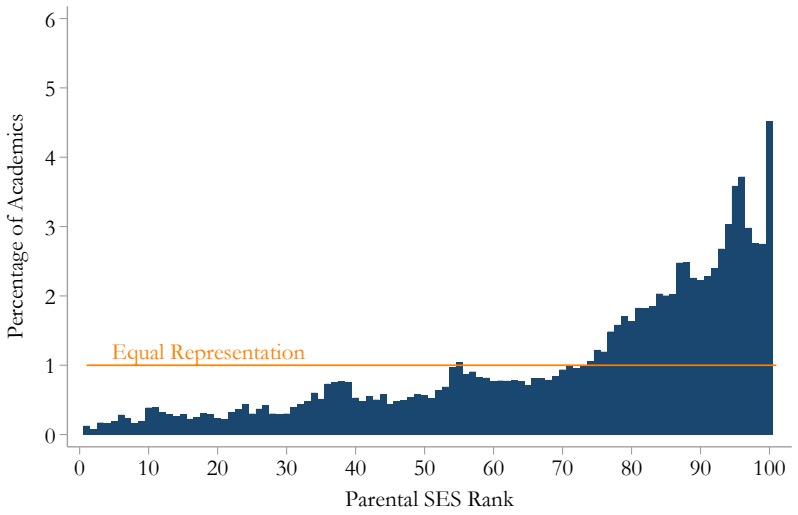
Representation of Academics by Socio-Economic Background

Figure B.1: Representation by Socio-Economic Background, Excluding Children of Professors

(a) With Regional Variation



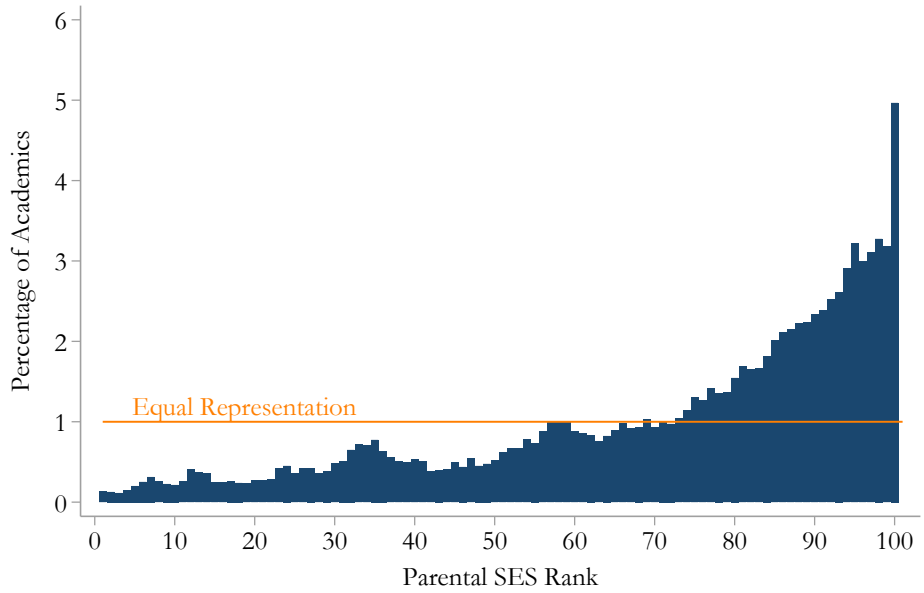
(b) Without Regional Variation



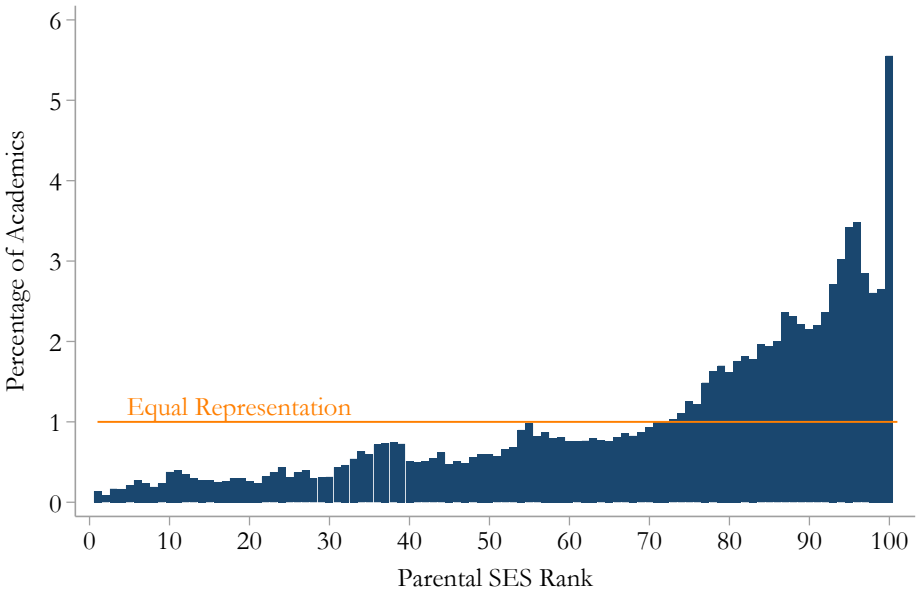
Notes: The figure shows the representation of academics based on their socio-economic background, excluding academics who are children of professors. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. The horizontal line represents a hypothetical equal representation from all income ranks.

Figure B.2: Extended Sample 1900-1969: Representation by Socio-Economic Background

(a) With Regional Variation



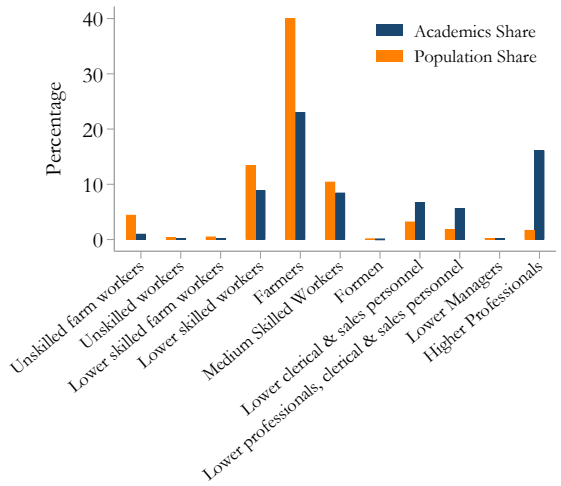
(b) Without Regional Variation



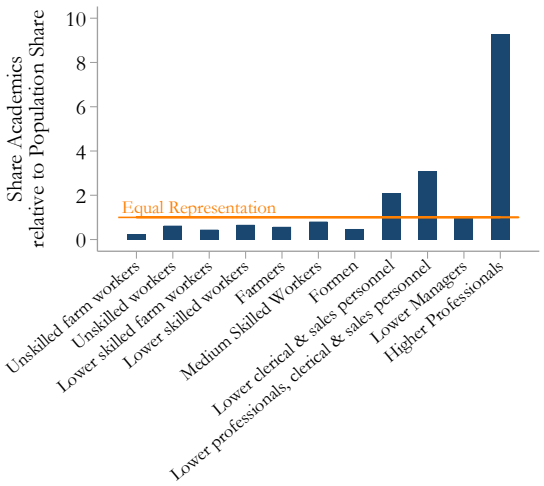
Notes: The figure shows the representation of academics based on their socio-economic background for the extended sample (1900-1969). We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. The horizontal line represents a hypothetical equal representation from all income ranks.

Figure B.3: Representation by Socio-Economic Background, Alternative Measures of SES: HISCLASS

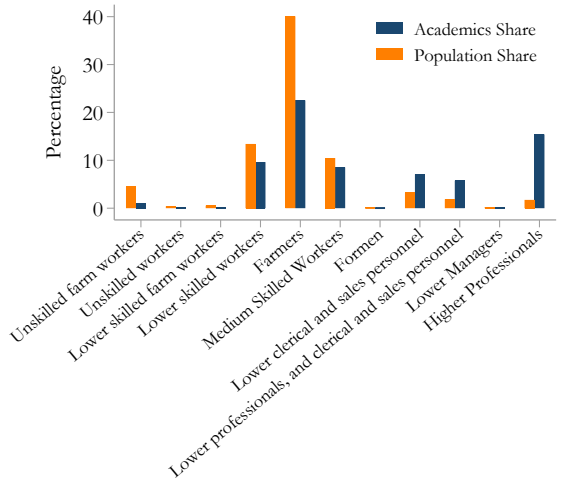
(a) Main Sample 1900-1956



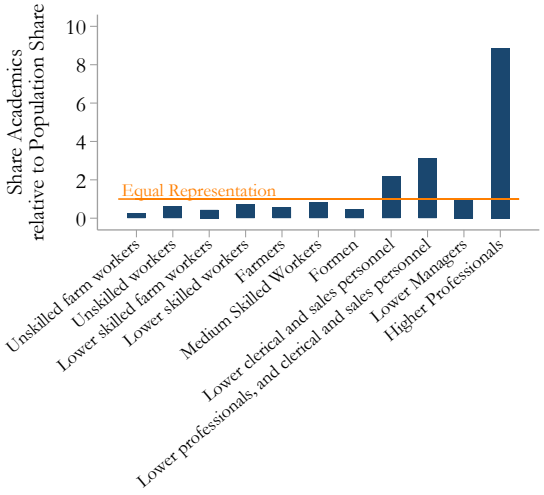
(b) Main Sample 1900-1956



(c) Extended Sample 1900-1969



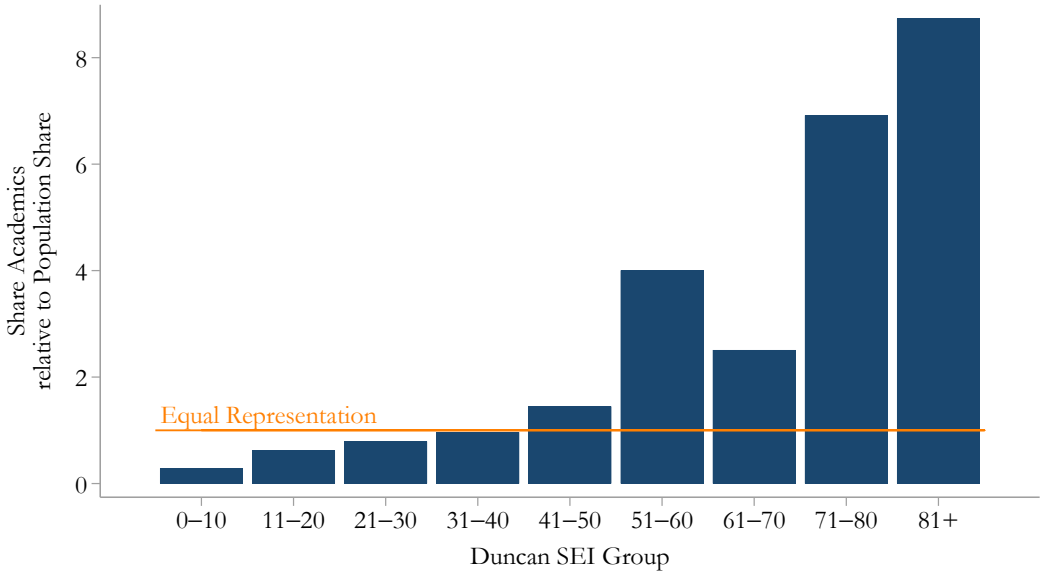
(d) Extended Sample 1900-1969



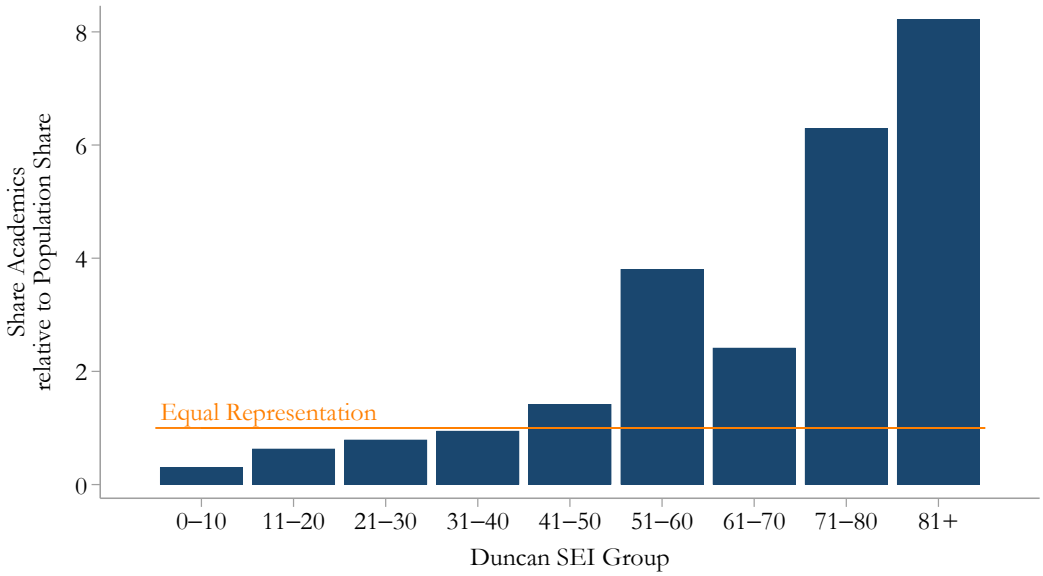
Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with HISCLASS, a measure of the social standing of a father's occupation (van Leeuwen and Maas, 2011). In panels a) and c), the orange bars indicate the share of individuals from a particular HISCLASS in the census. Compared to the census, academics are disproportionately children of fathers in higher status occupations (higher professionals). Panels b) and d) show the share of academics from a HISCLASS relative to the share of the population from the same HISCLASS. The horizontal line represents a hypothetical equal representation of these HISCLASS' in the population of academics.

Figure B.4: Representation by Socio-Economic Background, Alternative Measures of SES: Duncan Socioeconomic Index

(a) Main Sample 1900-1956



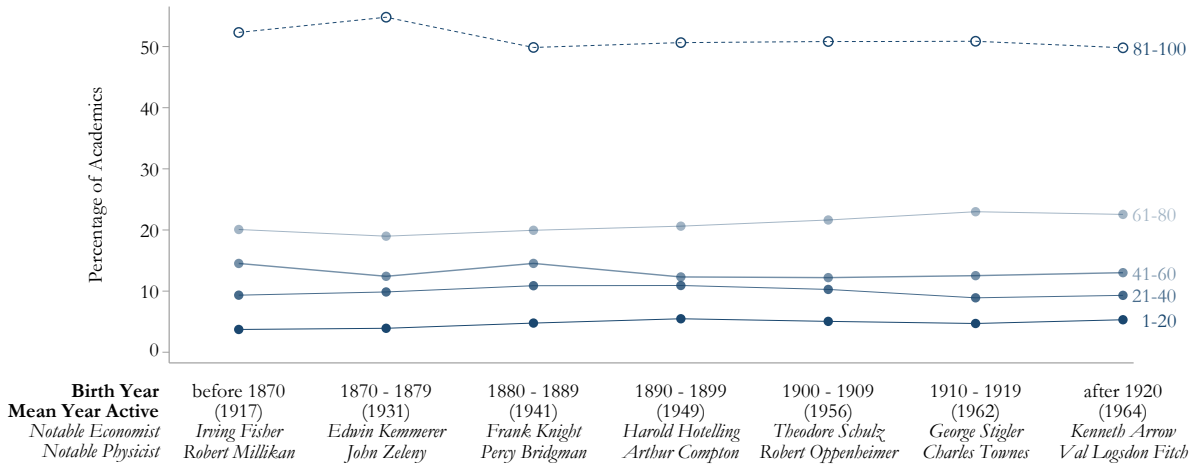
(b) Extended Sample 1900-1969



Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with the Duncan Socioeconomic Index (SEI), a measure of the social standing of a father's occupation. SEI reflects the income level and educational attainment of an occupation in 1950. For details, see IPUMS (2024b). SEI is an ordinal measure of occupational social status with gaps, which we group into 9 categories. For example, the top category, 81+, contains SEI 81-87 (no gaps), 90, 92, 93 and 96. SEI 89 does not exist in the census data of the relevant period. The horizontal line represents a hypothetical equal representation of these SEI categories in the population of academics.

Representation Over Time

Figure B.5: Extended Sample 1900-1969: Representation by Socio-Economic Background Over Time

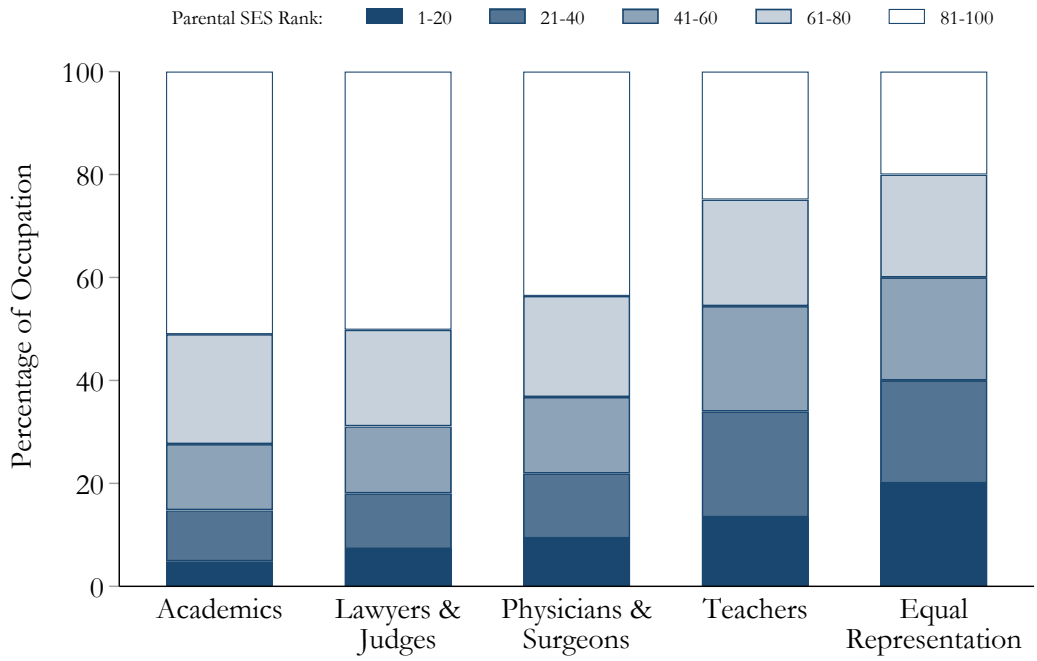


Notes: The figure shows the representation of academics based on their socio-economic background over time. Each line represents the percentage of all academics whose fathers are from specific income percentile ranks. For example, the top line indicates the percentage of academics whose fathers are in the top 20 income percentile ranks.

Representation in Academia versus Other High-Skilled Professions

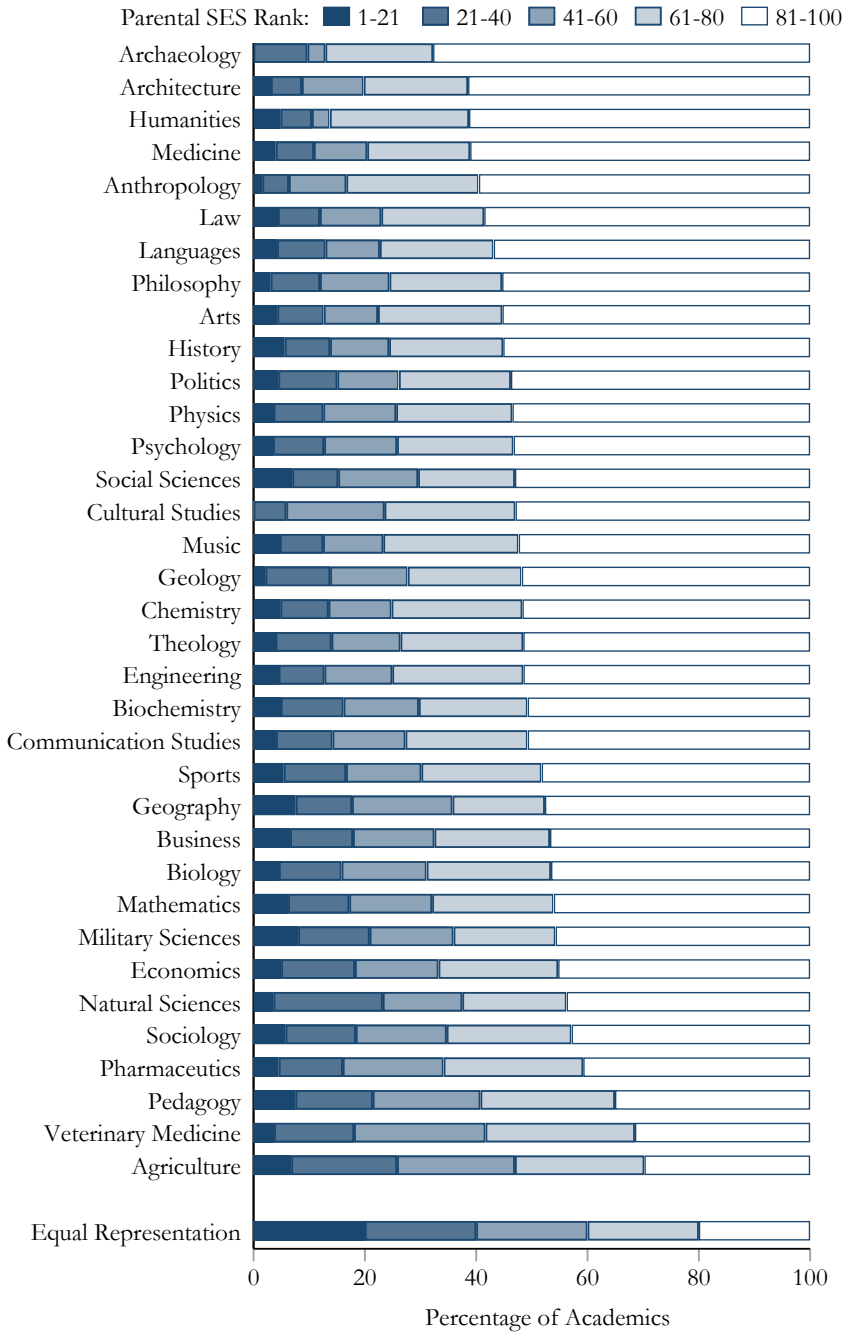
Representation in Academic Disciplines and Universities: Additional Results

Figure B.6: Extended Sample 1900-1969: Comparison to other High-Skilled Professions



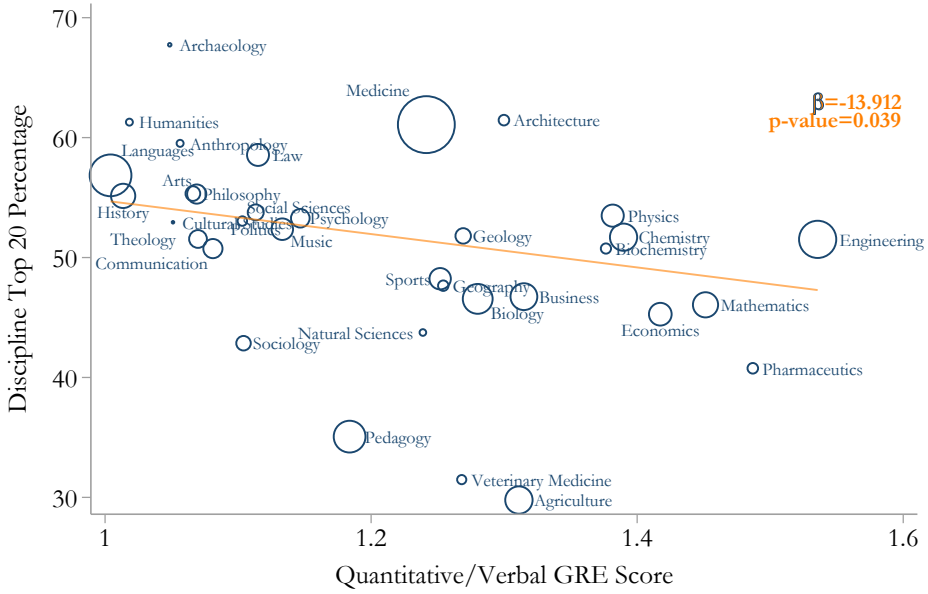
Notes: The figure compares the representation of academics based on their socio-economic background to representation in other professions. The representation in other professions is based on U.S. census samples of lawyers & judges, physicians & surgeons, and teachers that match the sample of academics (see Appendix A.2. for details).

Figure B.7: Extended Sample 1900 - 1969: Representation by Discipline



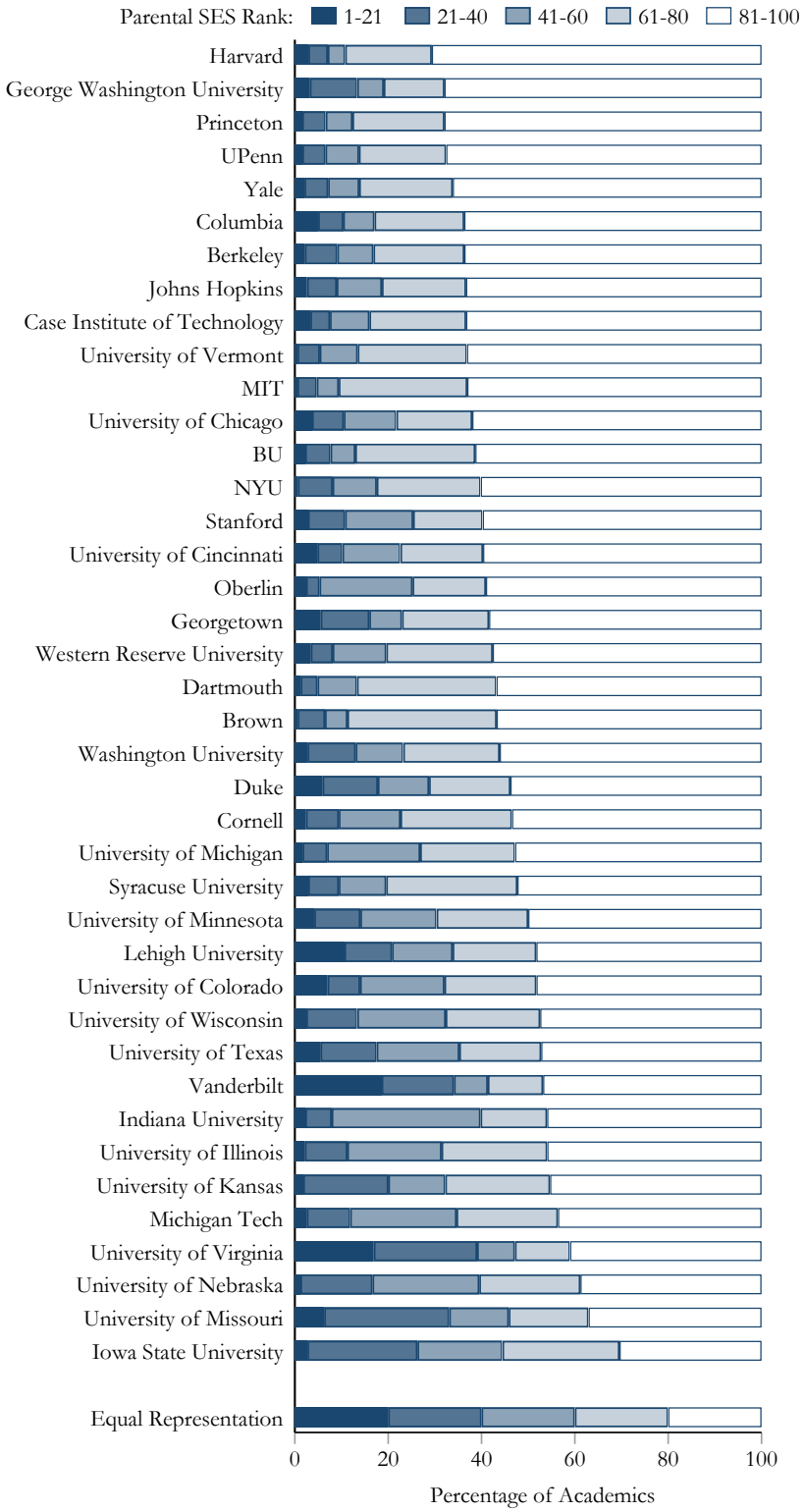
Notes: The figure shows the representation of academics based on their socio-economic background by academic discipline. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

Figure B.8: Extended Sample 1900 - 1969: Discipline Mathematics vs. Language Requirements and Representation



Notes: The figure shows the share of academics from the top quintile of the distribution of socio-economic background by academic discipline in relation to the importance of quantitative relative to verbal skills in the discipline for the extended sample (1900-1969). We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. We proxy the importance of mathematics relative to language skills with the ratio of the average GRE quantitative score to the average verbal reasoning GRE of test takers intending to pursue a graduate degree in the respective discipline. GRE score data come from (ETS) (2009), Extended Table 4. The size of the circles indicates the number of academics in the respective discipline in our data.

Figure B.9: Extended Sample 1900 - 1969: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each color shows the percentage of academics whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of academics whose father was in the top 20 percentiles of predicted income.

C Socio-Economic Background, Scientific Publications, and Novel Scientific Concepts: Additional Results

Table C.1: Socio-Economic Background and the Distribution of Publications

Dependent Variable:	*Publication Count in Percentile*						
	0 − 50	*> 50 − 70*	*> 70 − 90*	*> 90 − 95*	*> 95 − 97*	*> 97 − 99*	*> 99 − 100*
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A: 1914 – 1956							
Parental SES Rank	-0.00043**	0.00007	0.00029*	0.00017**	-0.00005	0.00002	-0.00007
	(0.00019)	(0.00015)	(0.00015)	(0.00008)	(0.00005)	(0.00006)	(0.00004)
R^2	0.09	0.04	0.03	0.02	0.01	0.01	0.01
Observations	12,765	12,765	12,765	12,765	12,765	12,765	12,765
Dependent Variable Mean	0.541	0.171	0.194	0.047	0.018	0.021	0.008
Panel B: 1914 – 1969							
Parental SES Rank	-0.00029*	0.00006	0.00028**	0.00010	-0.00007	-0.00002	-0.00006
	(0.00017)	(0.00013)	(0.00014)	(0.00007)	(0.00005)	(0.00005)	(0.00004)
R^2	0.08	0.03	0.03	0.02	0.01	0.01	0.02
Observations	15,519	15,519	15,519	15,519	15,519	15,519	15,519
Dependent Variable Mean	0.557	0.168	0.185	0.045	0.017	0.019	0.008
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Discipline FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports the estimates of Equation (3).The dependent variable is an indicator whether an academics publication count falls into a certain range of publication percentiles. Publication counts are an academic's total number of publications that were published in a ± 5-year window around the cohort when academic i enters the faculty rosters. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

Table C.2: Socio-Economic Background and Novelty: Excluding the 10,000 Most Common Words

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00085*	-0.00097**	-0.00096**	-0.00068	-0.00085*	-0.00084*
	(0.00048)	(0.00048)	(0.00048)	(0.00043)	(0.00043)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,970	11,970	11,970	11,970	11,970	11,970
Dependent Variable Mean	0.305	0.305	0.305	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00073*	-0.00082**	-0.00082**	-0.00070*	-0.00081**	-0.00082**
	(0.00042)	(0.00042)	(0.00042)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,724	14,724	14,724	14,724	14,724	14,724
Dependent Variable Mean	0.295	0.295	0.295	-0.012	-0.012	-0.012
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (5). The dependent variable measures the number of publications which introduce at least one novel word and were published in a ± 5-year window around the cohort when academic i enters the faculty rosters. We exclude the 10,000 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: $***$ p<0.01, $**$ p<0.05, and $*$ p<0.1.

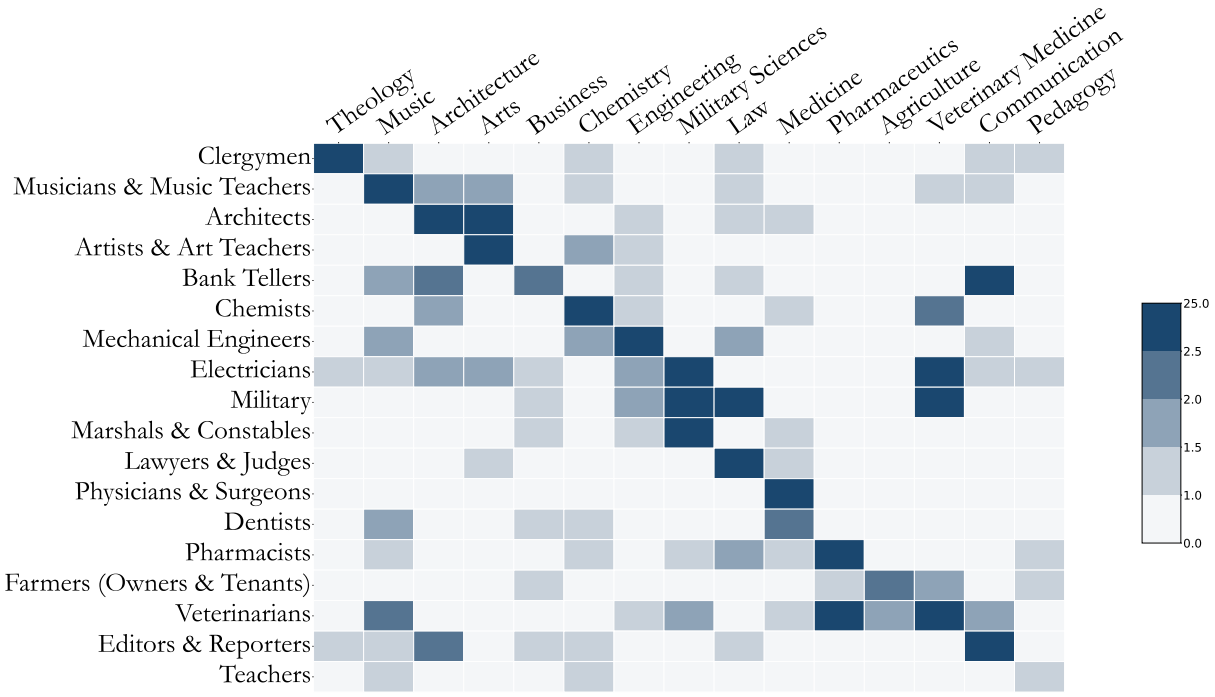
Table C.3: Socio-Economic Background and Novelty: Excluding the 36,872 Most Common Words

Dependent Variable:	*Papers with Novel Words*			*Std. Papers with Novel Words*		
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00095**	-0.00107**	-0.00106**	-0.00081*	-0.00098**	-0.00099**
	(0.00048)	(0.00047)	(0.00047)	(0.00043)	(0.00044)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,970	11,970	11,970	11,970	11,970	11,970
Dependent Variable Mean	0.296	0.296	0.296	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00080*	-0.00088**	-0.00088**	-0.00079**	-0.00090**	-0.00092**
	(0.00042)	(0.00041)	(0.00041)	(0.00038)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.04	0.01	0.02	0.02
Observations	14,724	14,724	14,724	14,724	14,724	14,724
Dependent Variable Mean	0.287	0.287	0.287	-0.011	-0.011	-0.011
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes		Yes	Yes
Discipline FEs			Yes			Yes

Notes: The table reports the estimates of Equation (5). The dependent variable measures the number of publications which introduce at least one novel word and were published in a ± 5-year window around the cohort when academic i enters the faculty rosters. We exclude the 36,872 most common words. We standardize the novel word measure to have a mean of 0 and a standard deviation of 1 within disciplines and cohorts. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of academic i's father. Standard errors are clustered at the level of father's occupation, childhood state, and birth year. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

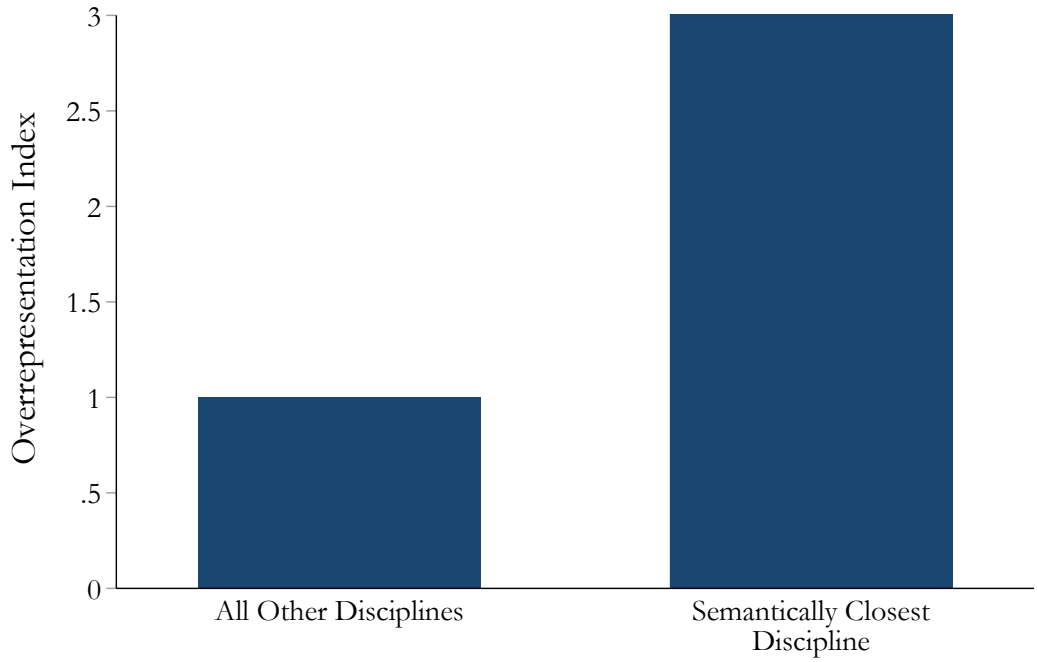
D Socio-Economic Background and Discipline Choice

Figure D.1: Extended Sample 1900-1969: Father's Occupation and Discipline Choice



Notes: The figure shows the relationship between father's occupation (rows) and the children's academic discipline choice (columns) for selected father's occupation - discipline pairs. Darker shades indicate more extreme levels of overrepresentation as measured by eq. (9).

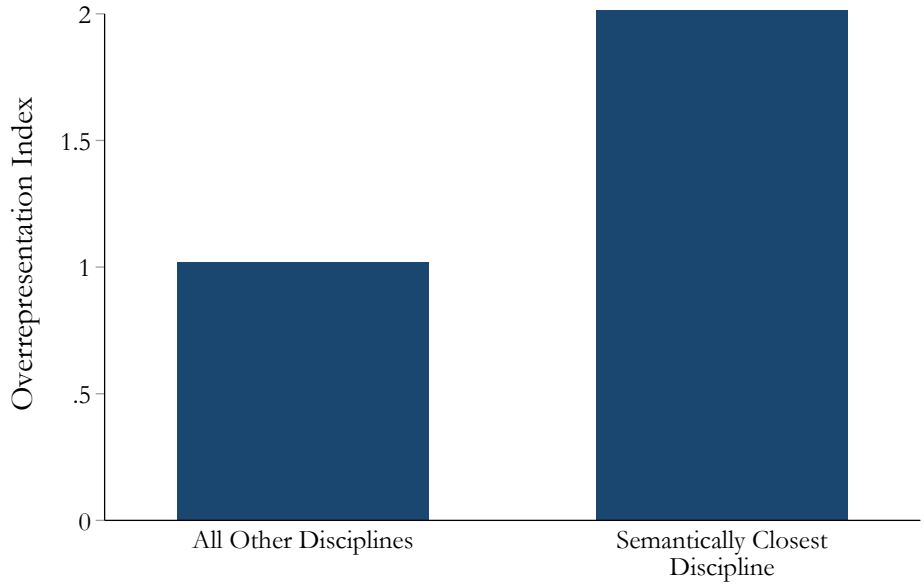
Figure D.2: Extended Sample 1900-1969: Overrepresentation in Semantically Closest Discipline



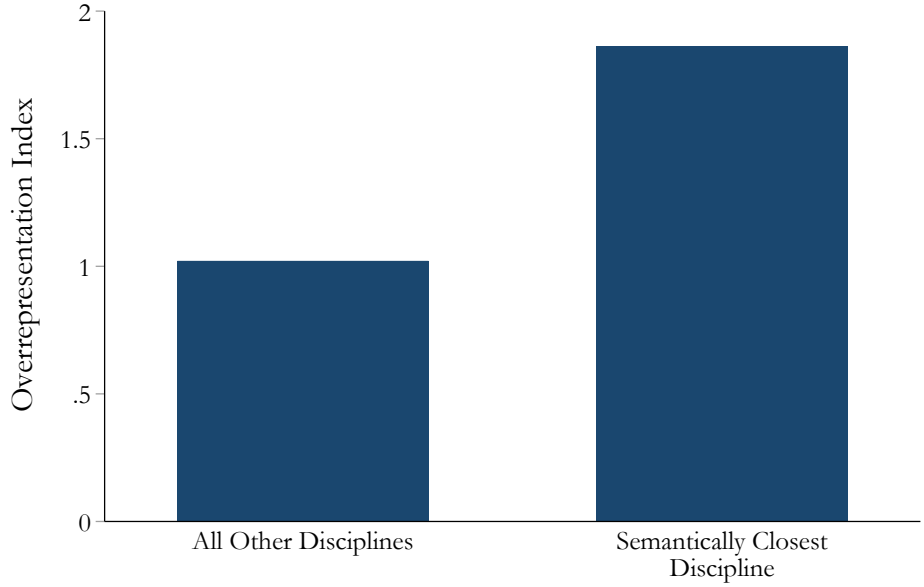
Notes: The figure shows overrepresenation as measured by eq. (9) in the father's occupation-discipline pair whose name (e.g., "agriculture") is semantically closest to the text string of the father's occupation (e.g., "farmer") as well as all other the relationship between father's occupation (rows) and the children's academic discipline choice (columns) for selected father's occupation - discipline pairs. Darker shades indicate more extreme levels of overrepresentation as measured by .

Figure D.3: Robustness –Overrepresentation in Semantically Closest Discipline

(a) One SD Cosine Similarity Cutoff



(b) No Cosine Similarity Cutoff



Notes: The figure shows overrepresentation as measured by equation 9 in the father's occupation-discipline pair whose name (e.g., "agriculture") is semantically closest to the text string of the father's occupation (e.g., "farmer") as well as all other father's occupation-discipline pairs. Panel a defines the closest discipline as the discipline that is semantically closest, and the cosine similarity is at least one standard deviation above the mean of all cosine similarities of all father's occupation-discipline pairs. Panel b defines the closest discipline as the discipline that is semantically closest without enforcing a further cutoff on the cosine similarity.