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

Ran Abramitzky
Stanford University
NBER

Lena Greska
University of Munich

Santiago Pérez
UC Davis
NBER

Joseph Price
BYU
NBER

Carlo Schwarz
Bocconi University
CEPR

Fabian Waldinger
University of Munich
CEPR

January 29, 2026

Abstract

We explore how socio-economic background shapes academia, collecting the largest dataset of U.S. academics' backgrounds and research output. Individuals from poorer backgrounds have been severely underrepresented for seven decades (1900-1969), especially in the humanities and elite universities. Father's occupation predicts professors' discipline choice and, thus, the direction of research. While we find no differences in the average number of publications, academics from poorer backgrounds are both more likely not to publish and to have outstanding publication records. Academics from poorer backgrounds introduce more novel scientific concepts, which find application in subsequent research and patents, but are less likely to receive recognition, as measured by citations, Nobel Prize nominations, and awards.

Keywords: Academics, Socio-economic Background, Science, U.S. census

*We are grateful to Alvaro Calderón, Davide Cantoni, Gabriele Cristelli, Anna Carolina Dutra Saraiva, Kilian Huber, Santiago Paz Ramos and seminar audiences at Bocconi, CEU, Chicago, Cologne, NUS, LMU Munich, LSE, Mannheim, Pompeu Fabra, PSE, SOFI Stockholm, BI Norwegian Business School, and Tilburg as well as the NBER Summer Institute, the NBER conference on "Linking Historical Data Sources for Small Populations", and the CEPR Economic History Conference in Heidelberg 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 David Geiger, Felix Radde, Marie Spörk, and Nils Süßenbach for outstanding research assistance. Carlo Schwarz is grateful for financial support from the European Research Council (ERC) Starting Grant (Project 101164784 — CHAIN — ERC-2024-STG). Fabian Waldinger and Lena Greska are grateful for financial support from the Deutsche Forschungsgemeinschaft through CRC TRR 190 (nr. 280092119).

1 Introduction

The underrepresentation of individuals from lower socio-economic backgrounds in leadership positions in government, business, and academia has become a growing concern among policymakers and the general public. Two primary economic rationales drive efforts to increase representation. First, 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).

In this paper, we explore how socio-economic background shapes academia – from who becomes an academic through the research fields professors specialize in, to their productivity and peer recognition. For our analysis, we assemble the most comprehensive data on the socio-economic backgrounds and research output of U.S. academics ever assembled. The long-run nature and granularity of our data enable us to study how these patterns changed over time and how they differ by discipline and across universities.

We rely on various primary and secondary data sources to assemble our data. First, we utilize comprehensive 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 regardless of whether they publish or whether they are members of academic societies. This helps to mitigate selection biases common in studies that rely exclusively on publication 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. censuses. We then link academics to their family backgrounds using data from the *Census Linking Project* (CLP) (Abramitzky et al., 2021) and the *Census Tree Project* (Buckles et al., 2023). Our main measure of socio-economic background is the percentile rank of their father's predicted income when the future academics were growing up.[1] In the third step of our data collection, we link academics in six scientific disciplines – medicine, biology, biochemistry, chemistry, physics, and mathematics to their publication and citation records using data from the *Clarivate Web of Science*. In the fourth step, we link academics to data from the Nobel Nomination Archive (Nobelprize.org, 2024) and

[1]The findings are robust to alternative measures of socio-economic background, such as using the fathers' actual income available for a subsample of fathers as well as measures of socio-economic background that capture the prestige and average education level of the father's occupation.

the Nobel Foundation to measure both Nobel Prize nominations and awards. Overall, our data enable us to measure the socio-economic backgrounds of 46,139 academics (for 15,521 of whom we also have publication and citation data) across 1,026 universities over nearly seven decades.

Our paper is organized into four parts, examining key stages of academic careers and how they are shaped by socio-economic background. In the first part of the paper, we examine differential barriers to entry into academia. We find a stark underrepresentation of individuals from lower socio-economic backgrounds: those born to parents in the bottom quintile 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 quintile of the income rank distribution. Children born to the highest-earning fathers are particularly overrepresented, with those born to fathers in the 100^{th} percentile having a 56% higher chance of becoming an academic than those born to fathers in the 99^{th} percentile. The underrepresentation of low socio-economic status individuals in academia is greater than in other occupations that require specialized training, such as medicine and law.

We find that the socio-economic composition of academics has remained remarkably stable over seven decades, despite significant changes in American higher education and society – including a sharp increase in college attendance rates. This persistence stands in stark contrast with the significant increase in the representation of women in U.S. academia over the same period (e.g., Rossiter, 1982, 1998; Iaria et al., 2024).

Furthermore, linking our data to World War II enlistment records, we show that this underrepresentation cannot be fully attributed to differences in cognitive ability (proxied by Army General Classification Test scores). Even conditional on cognitive ability, individuals whose fathers were at the 75th percentile of the income distribution are about 68% more likely to become academics than individuals whose fathers were at the 25th percentile.

While academics from low socio-economic backgrounds are underrepresented in all universities, the underrepresentation of academics from low socio-economic backgrounds varies sharply by university. In selective private universities such as Princeton, Harvard, and Yale, at least 60% of academics come from families in the top quintile of the parental income distribution. In contrast "only" 30-40% of academics in state universities such as Iowa State, University of Missouri, or the University of Nebraska come from families in this quintile.

Representation also varies sharply by discipline. While around 60% of academics in the humanities come from the top quintile of the parental income distribution, around 40% of academics in mathematics and economics come from the top quintile. This heterogeneity appears to be systematically related to the types of skills required to enter a discipline. Specifically, we find that representation from lower socio-economic backgrounds is higher in disciplines with a stronger emphasis on quantitative relative to verbal skills.

In the second part of the paper, we study the extent to which the influence of parental occupation can explain differences in representation by discipline. We develop a novel measure of overrepresentation to assess whether children of fathers in specific occupations are overrepresented in particular academic disciplines. Our findings indicate that academics tend to pursue disciplines aligned with their fathers' occupations. 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 identify the discipline that is closest in semantic space to the father's occupation. We then show that academics are more likely to enter disciplines that are systematically similar to their fathers' occupations. 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. This new finding underscores the influence of parental occupation on children's career trajectories, even when children pursue occupations different from those of their parents, i.e., when children of bank tellers become *professors* in business and management. While the existing literature in economics and sociology has predominantly emphasized the tendency of children to follow directly in their parents' occupational footsteps (Rogoff, 1950; Blau and Duncan, 1967; Long and Ferrie, 2013; Xie and Killewald, 2013; Almgren et al., 2023), our result suggests that parental occupation can exert a broader impact beyond simple occupational inheritance.

In the third part of the paper, we study how socio-economic background relates to scholars' productivity. We find no systematic relationship between parental income ranks and the *average* number of publications of academics. However, individuals from lower socio-economic backgrounds are both significantly more likely to never publish and more likely to have a publication count in the top 1%.

Importantly, academics from lower socio-economic backgrounds differ in the *content* of their research. To examine potential differences in a key dimension of publication content, we develop a metric that captures the number of novel words a scientist introduced to the scientific community (Iaria et al., 2018). The measure proxies for the introduction of new scientific concepts that require novel scientific terms. We find that scientists with a low-income father (father at the 25th percentile) publish around 0.05 additional papers (or 16% more papers) with at least one novel word compared to scientists whose fathers were at the 75th percentile. We also find evidence that scientists from lower socio-economic backgrounds introduce more novel scientific concepts that are picked up in subsequent research papers and patents. These findings suggest that academics from lower socio-economic backgrounds are more likely to pursue research agendas off the beaten path, which may result in scientific breakthroughs but also in a higher failure rate, making them riskier hires.

In the fourth part of the paper, we examine the relationship between socio-economic background and recognition by other academics. We start by studying citations to academic papers, a widely used metric for measuring recognition within the academic community. We find that papers published by authors from lower socio-economic backgrounds receive fewer citations. To further explore how socio-economic background affects recognition, we investigate Nobel Prize nominations and awards — an acknowledgment of 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 than scientists with fathers at the 25th percentile. They are also 50% more likely to be awarded a Nobel Prize. These differences persist even if we control for scientists' publication and citation records.

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; Akcigit et al., 2017) and concurrent research on academics (Morgan et al., 2022; Airoldi and Moser, 2024; Stansbury and Schultz, 2023; Stansbury and Rodriguez, 2024; Novosad et al., 2024).[2]

We make two main contributions to this literature. First, we provide the most comprehensive description of the socio-economic background of U.S. academics, covering all disciplines and the near universe of universities over seven decades (1900-1969). This enables us to characterize how socio-economic backgrounds vary across disciplines and universities, how they compare to those of other elite professionals, and to trace their evolution over a key period in the history of U.S. higher education (from the "formative" prewar years, to the consolidation of American leadership in higher education after World War II). Second, we provide comprehensive evidence on how socio-economic background shapes various aspects of academics' careers, from who gets hired to discipline choice, productivity, and peers' recognition.[3]

Our paper is also related to the literature on gender discrimination in academia (e.g., Card et al., 2020, 2022; Iaria et al., 2024; Ross et al., 2022; Moser and Kim, 2022; Koffi, 2024; Babcock et al., 2017; Bagues et al., 2017). While this substantial body of research has studied the underrepresentation of women in research, the underrepresentation of

[2]Similarly, geography also shapes participation in science. Participants of the international mathematical olympiads from lower-income countries are less likely to enroll in PhD programs and produce fewer publications and citations despite similar talents (Agarwal and Gaule, 2020).

[3]Other related research has documented the importance of socio-economic background for the selection of *students* into elite universities (Chetty et al., 2020; Michelman et al., 2022; Chetty et al., 2023; Abramitzky et al., 2024).

individuals from lower socio-economic backgrounds has been a "forgotten dimension of diversity" (Ingram, 2021), which we examine in this paper.

Finally, we contribute to the literature on how scientists' or inventors' background shapes their research focus 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); Dossi (2024); Croix and Goñi (2024) investigates how gender and race impact the research focus of scientists. One of the few papers that studies how socio-economic background affects the direction of research is a recent contribution by Einio et al. (2022). They 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 of how the socio-economic background shapes the research of university academics. Since most 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.

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, their research output, and measures of recognition by the scientific community. The dataset is based on four data sources. First, we use complete faculty rosters for the near universe of U.S. universities from the *World of Academia Database* (Iaria et al. 2024). Second, we match these data to historical U.S. censuses (Ruggles et al., 2024). Using links from the *Census Linking Project (CLP)* (Abramitzky et al. 2012, 2021), the *Census Tree Project* (Buckles et al. 2023) and the *IPUMS Multigenerational Longitudinal Panel (MLP)* (Ruggles et al., 2019) we are able to trace academics to their childhood homes, which enables us to measure the socio-economic background of academics. Third, we enhance the data with publication and citation data from the *Web of Science* to observe the academics' research output and its content. Fourth, we link academics to information on Nobel nominations and awards from the Nobel Foundation.

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.[4] For example, the data contain 3,441 U.S. academics

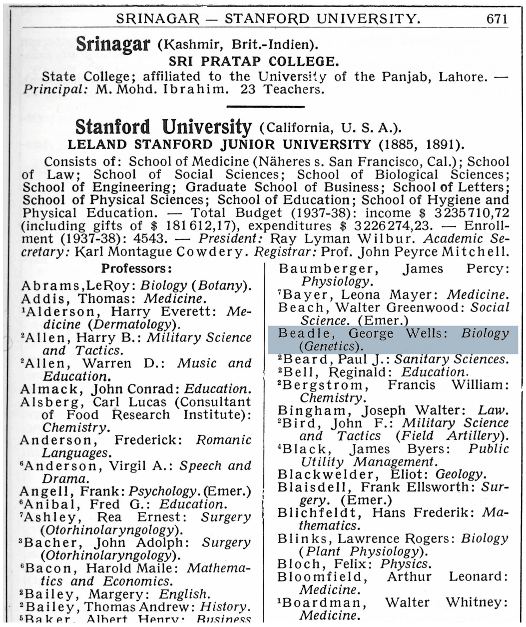
[4]The data include all academics who were affiliated with a U.S. university in at least one of the six cross-sections. We thus also include the U.S. spells of academics who start their career abroad and move to the United States or who start their career in the United States and then move abroad. About 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*.

who entered the database in 1900 and 65,340 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 36 disciplines.[5] 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."

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.).

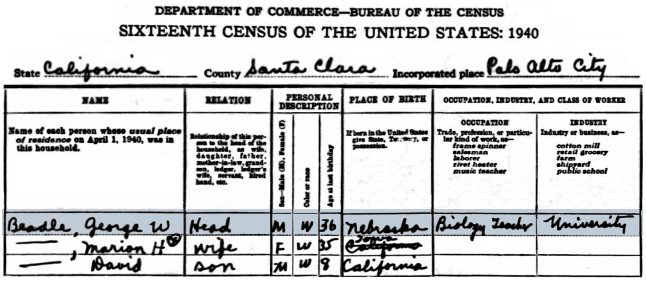
LELAND STANFORD JUNIOR UNIVERSITY (1885, 1891).

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.*
[1]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.*
[2]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

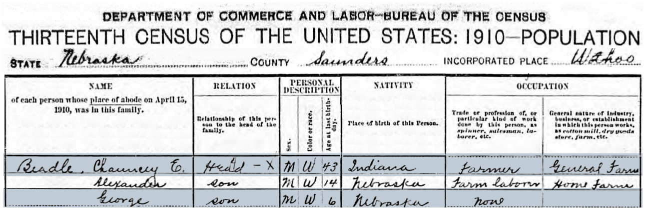

DEPARTMENT OF COMMERCE—BUREAU OF THE CENSUS

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, as wife, daughter, father, mother-in-law, grandson, lodger, lodger's wife, servant, hired hand, etc.	Sex—Male (M), Female (F)	Color or race	Age at last birthday	If born in the United States give State, Territory, or possession.	OCCUPATION Trade, profession, or particular kind of work, as— *frame spinner salesman laborer rivet heater music teacher*	INDUSTRY Industry or business, as— *cotton mill retail grocery farm shipyard public school*
Beadle, George W	Head	M	W	36	Nebraska	Biology Teacher	University
—, Marion H	wife	F	W	35	~~California~~ Iowa		
— 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	RELATION	PERSONAL DESCRIPTION			NATIVITY	OCCUPATION	
of each person whose place of abode on April 15, 1910, was in this family.	Relationship of this person to the head of the family.	Sex.	Color or race.	Age at last birthday.	Place of birth of this Person.	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 – X	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 Handbuch der Gelehrten Welt* which underlies the *World of Academia Database.* The page includes the entry of 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 race, age, state of residence and occupation of his father ("farmer").

The *World of Academia Database* offers several key features that are integral to our analysis. First, it contains *entire* faculty rosters for the vast majority of PhD granting universities in the United States, which allows us to study academics even if they never published 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. For instance,

[5]For the vast majority of universities, the data report all academics who are assistant professors and above. Lecturers and similar academic staff are usually not reported.

lists of distinguished academics might underestimate the number of academics from lower SES-backgrounds if such academics are less likely to be recognized by their peers (as we document below). Second, our dataset encompasses all academic disciplines, including the social sciences and humanities. This broad scope enables us to conduct a comprehensive analysis of representation in academia, examining variations across universities and disciplines.

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, the Census Tree Project, and IPUMS Multigenerational Longitudinal Panel (MLP) 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 have been released to the public.

We link academics in the faculty rosters to their contemporaneous censuses based on the full name of the academic, their census occupation, and their geographic location.[6] We define a 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")[7]

[6]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 of using the "professor" occupational category from the census records.

[7]Here, we both use the IPUMS occupation coding (occ1950, see IPUMS (2024a)) as well as the original string responses recorded by the census (occstr). This enables us to also match individuals whose occupation or industry was coded as "not yet classified". Typically, occupations are unclassified due to transcription or spelling errors.

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 1940, lived in Palo Alto City, 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[8]

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 applying criteria 1–6, approximately 70% of potential matches indicate an industry in the census that aligns with their academic position. For instance, individuals may be listed in industry 888 - Educational Services. Similarly, medical professors are often listed in industry 869 - Hospitals. In contrast, the remaining 30% are listed in industries that do not closely correspond to their academic roles (e.g. 246 - Construction) or fall into an unclassified category. To enhance the reliability of these matches, we introduce a seventh criterion that leverages the specific industry and occupation strings reported in the census:

7. the potential matches must report industry and occupation strings in the census that are consistent with becoming a professor

For the seventh criterion, all potential matches with a misaligned industry are independently reviewed by two research assistants, who classify each link as either correct or incorrect. For instance, the Stanford physics professor Frederick John Rogers was linked to a census record listing the industry as 0 - none reported. The research assistants examined the associated occupation ("Assoct Projessor [sic]") and industry ("physico at Stanford [sic]")

[8]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	2,485	72.2	1,726	69.5	50.2
1914	5,899	4,487	76.1	3,073	68.5	52.1
1925	6,401	4,731	73.9	3,188	67.4	49.8
1938	23,458	17,792	75.8	12,338	69.3	52.6
1956	53,243	28,814	54.1	17,052	59.2	32.0
Total	92,442	58,309	63.1	37,377	64.1	40.4
Extended sample: 1900-1969 cohorts						
			⋮			
1969	65,340	17,306	26.5	8,762	50.6	13.4
Total	157,782	75,615	47.9	46,139	61.0	29.2

Notes: The table reports the number of academics and linking rates to their "adult" and "childhood" censuses by cohorts of academics. Academics are reported in the first cohort in which they appear in the faculty rosters of the *World of Academia Database*.

strings from the record and determined the match to be correct.[9] In contrast, Vanderbilt University biology professor George W. Martin was linked to a census record listing the industry as 636 - Food stores, except dairy products. The research assistants examined the associated occupation ("druggist") and industry ("own store") strings and classified the link as incorrect. For the analysis, we only retain matches that both research assistants classified as correct.[10]

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 academics who entered the *World of Academia* database in 1969 to an adult census. With this in mind, the main sample in our

[9]The misspellings in the occupation and industry fields result from the transcription of handwritten census records.

[10]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.

analysis is restricted to academics who we first observe in 1956 or earlier facutly roster cohorts. However, we also consider an extended sample in which we attempt to match all academics in our data (including those who enter the faculty rosters in 1969).

Of the 92,442 academics in the main sample, we link 58,309 (63%) to a contemporaneous census (Table 1).[11] Manual inspections suggest that transcription mistakes of the historical handwritten census records account for many 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 75,615 (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.

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).

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 the corresponding childhood censuses. First, we use the links available from the Census Linking Project (CLP, Abramitzky et al. 2012, 2021).[12] We then combine these links with links from the Census Tree Project (CT, Buckles et al. 2023) for the 1900-1940 adult censuses and IPUMS Multigenerational Longitudinal Project (MLP) (Ruggles et al., 2019) for the 1950 adult census.[13] In addition to enabling us to increase the sample size, the additional links allow us to link to the childhood records of some female academics, which are less frequently captured by traditional linking methods.[14]

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

[12]Specifically, we use the "ABE-exact" links. As of November 2024, the Census Linking Project has not released links between the 1950 census and earlier censuses. Therefore, we create our own crosswalks for the 1950 census using the ABE algorithm in its "exact standard" version.

[13]In the rare cases in which these links point to different individuals, we privilege links made by the ABE exact algorithm. There are few such cases because there is a very high rate of conditional agreement between ABE links and those made by machine learning algorithms, i.e. when both methods identify a link, the links are identical in close to 100% of cases (Abramitzky et al., 2021).

[14]The share of female academics in the faculty rosters is only 13% in the main sample and 14% in the extended sample (see also Iaria et al. 2024). Overall, linking rates for female academics are 28% for the

To maximize the likelihood of capturing an academic's parental background, we link adult census records to all potential childhood censuses. Childhood censuses are defined as those in which future academics are observed as children under the age of 22 and residing with their parents. In cases where an academic is linked to multiple childhood censuses, we prioritize the census in which the academic is youngest.[15]

Our exemplary academic, George Wells Beadle, can be linked to his childhood census of 1910. At the time, he was six years old and listed in the census as the son of Chauncey E. Beadle, who was 43 years old and worked as a farmer (Figure 1, panel c). The information on the father's occupation will be the key information to reconstruct George Wells Beadle's socio-economic background.

For the main sample, we are able to link 37,377 (or 64% of the adult census) records to a childhood census (Table 1). For the extended sample, we can link 46,139 (or 61%) of the adult census records to a childhood census.[16] These linking rates are high compared to linking rates in existing research because we rely on a combination of linking algorithms and since we link to multiple potential childhood censuses.

Overall, we successfully link 37,377 (or 40%) individuals from the main sample to their childhood census. These linked academics form the basis for our analysis. To assess potential selection introduced by our linking procedure, 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. We find no systematic relationship between department quality and the linking rate (Figure 2, Panels (a) and (b), p-value=0.69).[17] As a further check, we investigate the correlation between the linking rates and the average income associated with a last name.[18] We find no systematic association between these variables (Figure 2, Panels (c) and (d), p-value=0.36). Together, these results indicate that our linking procedure does not introduce systematic selection.

main sample and 21% for the extended sample, compared to 42% and 31% for male academics. All results remain unchanged in a sample of male academics.

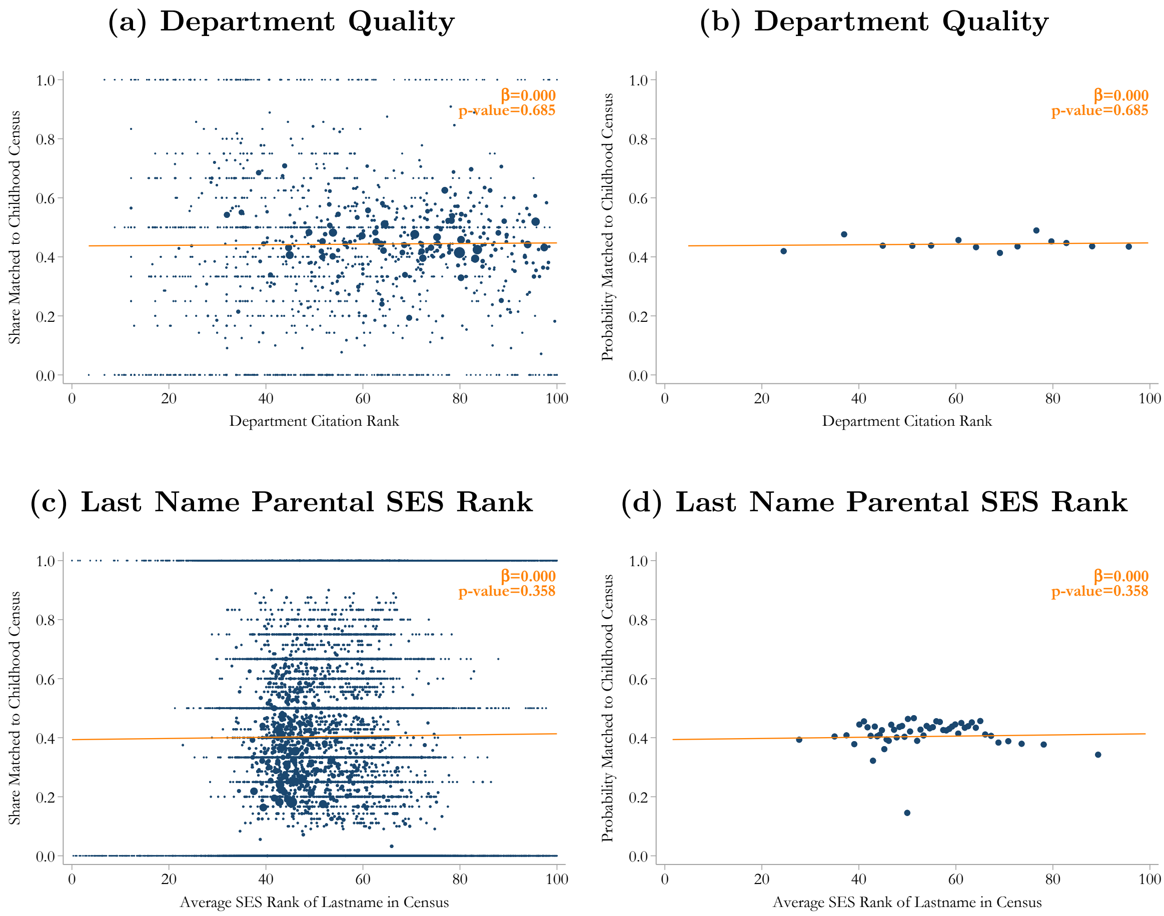
[15]As we link some academics to multiple adult censuses that can be linked to different childhood censuses, a small fraction of them have backlinks to different individuals in a childhood census. For example, an individual listed in the 1914 faculty roster could theoretically be matched to both the 1910 and 1920 adult censuses, and the 1920–1880 backlinks might identify a different individual than the 1910–1880 backlinks. In such cases, we retain the backlink associated with the adult census that is closest to the childhood census. In the given example, we would prioritize the link based on the 1910–1880 crosswalk.

[16]For academics who moved to the United States to study or when they were already academics, we cannot link them to a childhood census in the United States. Of the 75,615 academics who we link to an adult census, 6,769 or 7.9% are foreign-born. Foreign-born academics are part of the dataset if they migrated as children and can be observed in at least one childhood census after moving to the United States.

[17]We report equivalent figures for the extended sample in Figure A.1, Panels (a) and (b). There is a small and marginally significant positive correlation between department quality and matching rates in the extended sample.

[18]We measure the average income of a last name in the census using an analogous procedure to the one described in the next subsection.

Figure 2: Correlation of Linking Rates With Department Quality and Last Name 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 main sample. Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's parental 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). Appendix Figure A.1 shows the equivalent figures for the extended sample.

Constructing Parental SES ranks

For our baseline results, we rely on father's occupational income scores as a proxy for socio-economic background, because other measures of parental socio-economic status such as parental income or parental education are not available in pre-1940 U.S. censuses. We construct parental "income scores" for each academic, following Abramitzky et al. (2021). Specifically, we use data on wage income from the 1940 census (the first U.S. census to include individual-level income) and estimate the following regression for all working-age (20-70 years old) men in 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 $\ln(\text{Income}_j)$ measures the income of individual j in 1940. $\text{Occupation}_j \times$ State FE is a separate fixed effect for each census occupation code interacted with the state of residence of individual j. In addition, we also include 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 farmers using the approach developed by Collins and Wanamaker (2022).[19]

We then use the estimated coefficients from equation (1) to predict income for fathers in all census years. We use these predicted incomes to rank fathers relative to *all* fathers, including the fathers of non-academics, with children born in the same year.

Alternative Measures of Parental Socio-Economic Status

In robustness tests, we demonstrate that our results are robust to using various alternative measures of socio-economic status (SES). First, we use the *actual* income of academics' fathers as reported in the 1940 census to calculate parental income ranks for the subsample of academics whose fathers can be observed in the 1940 census. Second, we construct alternative ranks based on income predictions that do not differ across states. Third, following the approach in Abramitzky et al. (2012), we use information on mean wages by occupation from the 1901 Cost of Living Survey. Finally, we use Hisclass (van Leeuwen and Maas, 2011) and Duncan's Socioeconomic Index (SEI) as alternative measures of socio-economic status that also capture the prestige and education requirements of the father's occupation.

2.3 Linking Scientists with Publications and Citations

To investigate how socio-economic background influences scientific output and the direction of research, we link academics from 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 early 20th century, these disciplines had already established a culture of publishing in scientific journals, with publishing processes resembling contemporary practices. In contrast, disciplines such as the humanities and social sciences predominantly relied on book publishing during this period.

We use the procedure developed by Iaria et al. (2024) to link publications and citations to the faculty rosters. The procedure uses the academic's last name, first name, or initials (depending on whether first names are available), country, city, and discipline.[20] To improve match quality, we harmonize affiliations across the faculty rosters and the *Web of Science* with the *Google Maps API*.

[19] In cases where the number of individuals within certain occupation-by-state cells is low, or where census occupation codes change across years (see IPUMS 2024a), we apply coarser fixed effects to predict income ranks. See Appendix A.1. for details.

[20] To reduce false positives, matches are restricted to the academic's primary discipline (e.g., physics).

2.4 Linking Scientists with Nobel Prize Data

To measure recognition by the scientific community, we hand-link data on nominations for the physics, chemistry, and physiology or medicine Nobel Prizes from the Nobel Nomination Archive (Nobelprize.org, 2024). This database 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 also hand-link all Nobel Prize winners to our faculty rosters. Table 2 provides summary statistics for the most important variables in our data.

Table 2: Summary Statistics

Panel A: 1900 – 1956			
Variable	Mean	SD	Observations
Parental SES Rank	72.83	24.84	37,377
Age at Entry into Faculty Rosters	45.34	10.11	37,377
Female	0.09		37,377
Publications	4.67	9.51	12,767
Papers with Novel Words	0.32	1.18	11,972
Nominated for Nobel Prize	0.01		12,767
Awarded Nobel Prize	0.00		12,767
Panel B: 1900 – 1969			
Parental SES Rank	72.18	25.06	46,139
Age at Entry into Faculty Rosters	47.28	10.92	46,139
Female	0.10		46,139
Publications	4.91	10.63	15,521
Papers with Novel Words	0.32	1.22	14,726
Nominated for Nobel Prize	0.01		15,521
Awarded Nobel Prize	0.00		15,521

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 academics come from the *World of Academia Database*. Parental SES ranks are constructed based on U.S. Census microdata. Data on publications come from the *Web of Science*. Publications are measured in a ± 5-year window around the year of entering the faculty rosters. Papers with novel words measures the number of papers published in a ± 5-year window around the year of entering the faculty rosters that introduce at least one novel word. 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 explore whether individuals like George Beadle represent rare exceptions or if talented individuals were able to pursue academic careers regardless of their socio-economic background.

3.1 Representation of Academics by Socio-Economic Background

We visualize the share of U.S. academics that come from each percentile of the parental SES rank distribution (Figure 3). It is important to note that the parental SES rank should be interpreted as an omnibus measure of socio-economic background capturing a combination of different factors such as parental income but also education and other traits of the socio-economic background that are correlated with income. We do not argue that any single factor, such as a lack of parental income, is the sole or even dominant driver of our findings.

An equal distribution based on parental SES ranks would imply that 1% of academics stem from each percentile. We illustrate 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 increasing particularly sharply for higher parental SES ranks (Figure 3, panel a). Overall, approximately half of all academics come from the top 20% of the parental SES rank distribution. The degree of overrepresentation is particularly large for very high percentiles of the parental SES rank distribution. For example, individuals born to parents in the 95th percentile are more than three times as likely to become academics as one would expect under the equal representation benchmark.

The disparity is even more striking at the highest percentile. Individuals from the 100th percentile of the socio-economic background distribution are more than five times as likely to become academics as one would expect under the equal representation benchmark. Strikingly, even when compared to individuals from the 99th percentile, those from the 100th percentile have a 1.6 times higher chance of becoming an academic.[21]

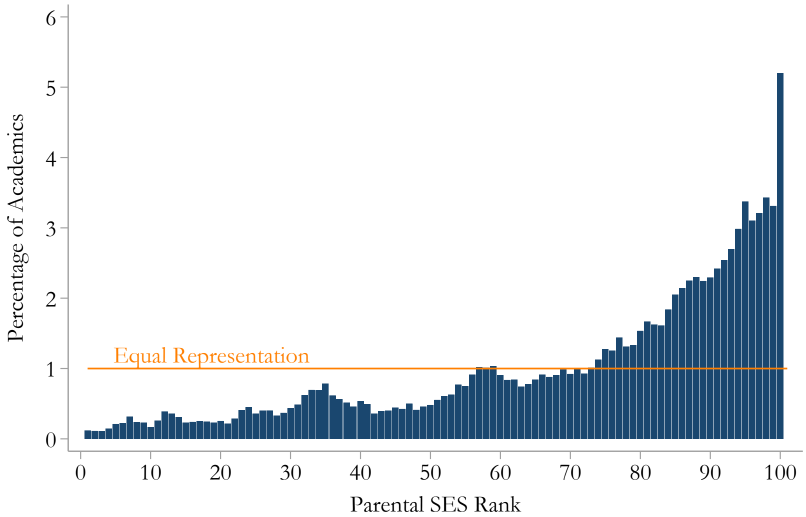
Robustness to Alternative Measures of Parental Socio-Economic Status

These findings are robust to using various alternative measures of parental socio-economic status. First, the results are similar if we predict parental SES ranks solely based on the father’s individual characteristics and his occupation, excluding state of residence fixed

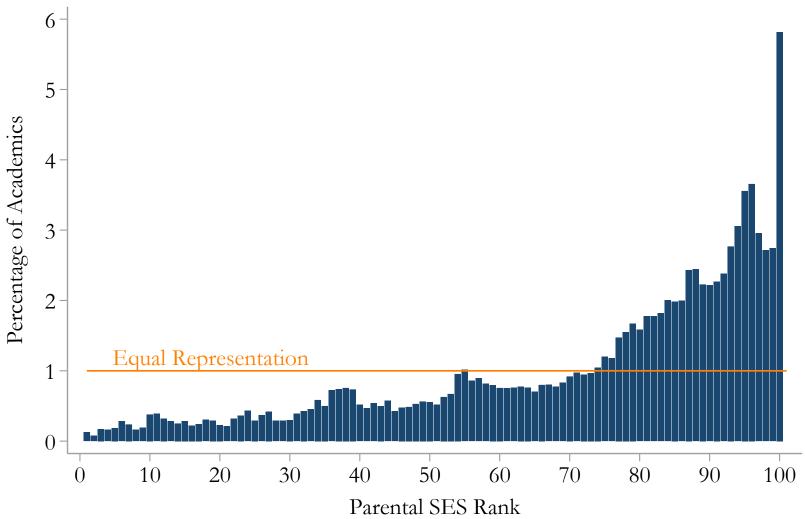
[21]This extreme overrepresentation at the 100th percentile may partially reflect that, in certain census years and states, professors themselves are classified in the highest parental income percentile. However, even after excluding individuals whose fathers report “professor” as their occupation in the census, the overall pattern remains similar (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 for the main sample. 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 a specific income percentile rank. For example, the right-most bar shows that around 5 percent of academics have fathers who were in the 100th percentile of the predicted income distribution. The horizontal line represents a hypothetical equal representation benchmark. Appendix Figure B.2 shows equivalent figures for the extended sample.

effects in the income prediction (Figure 3, panel b). This adjustment rules out that our findings are driven by state-level differences in incomes within the same occupation.

Second, the results are similar if we use the *actual* income of academics' fathers as reported in the 1940 census to measure socio-economic status for the subsample of academics whose fathers can be observed in the 1940 census (Appendix Figure B.3).[22]

[22]This analysis is limited to academics who entered the faculty rosters in 1956 and 1969 and who were young enough to still live with their parents in 1940 so that we can observe their father's income. Less than ten percent of the extended sample fall into this category. Income in the 1940 census is top-coded at

This exercise helps address the concern that parental ranks based on predicted income may not capture the actual income ranks of fathers of academics who worked in certain occupations.

Third, the results are similar if we use an alternative measure of socio-economic status based on mean income by occupation from the 1901 Cost of Living Survey (Abramitzky et al., 2012). The 1901 Cost of Living survey was conducted earlier and allows us to address concerns that the ordering of occupational income may have changed over time. However, the 1901 Cost of Living survey was only conducted in 99 cities across the US and only captures mean income by occupation. Hence, it is coarser and likely subject to more measurement error than our baseline measure, which uses income for the entire U.S. population and includes adjustments based on respondents' race and state of residence. While the measure of socio-economic status based on the 1901 Cost of Living survey is coarser, we obtain similar findings (Appendix Figure B.6).

In additional robustness checks, we report the share of academics by other measures of socio-economic background, which also capture occupational prestige and average education levels of the father's occupation (Hisclass and Duncan Socioeconomic Index (SEI), Appendix Figures B.4 and B.5). These findings confirm that academics are disproportionately drawn from high socio-economic backgrounds.

3.2 Representation Over Time

The large differences in the probability of becoming an academic translate into a highly skewed socio-economic composition of academia. As a next step, we analyze whether these representation patterns changed over time (Figure 4). The share of academics from the top quintile of the parental 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 quintile of the parental SES rank distribution is around 4-5% and hardly changes over time. This persistence is striking, given the substantial expansion in educational attainment in the United States during this period.[23]

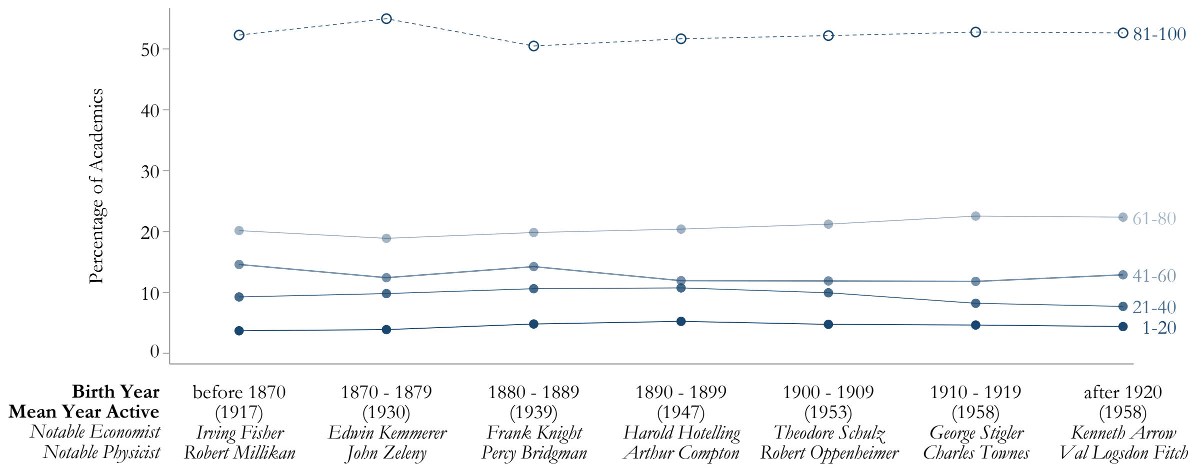
3.3 The Role of Cognitive Ability

The previous results show that individuals from lower socio-economic backgrounds remain underrepresented among academics over the seven decades covered by our data. This suggests that there are significant and persistent factors that prevent individuals from low socio-economic backgrounds from pursuing careers in academia.

$5,000. Around 2% of fathers in the 1940 census have an income of $5,000 and above. The highest reported bar in Appendix Figure B.3 therefore reports the percentage of academics from the top 2 percentiles of actual income.

[23]For example, U.S. Americans born in 1920, on average, completed three additional years of schooling compared to those born in 1870 (Goldin and Katz, 2009).

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 for the main sample. Each line represents the percentage of all academics whose fathers are from a specific income quintile. For example, the top line indicates the percentage of academics whose fathers were in the top quintile of the predicted income distribution. Appendix Figure B.7 shows the equivalent figure for the extended sample.

One important potential confounder is that individuals from lower socioeconomic backgrounds may be underrepresented in academia because entry into academia requires high cognitive ability. If parental income and cognitive ability are positively correlated, the underrepresentation may simply reflect lower cognitive ability among individuals from lower socioeconomic backgrounds. To assess this possibility, we use a sample of individuals for whom it is possible to observe a close proxy of cognitive ability, namely Army General Classification Test (AGCT) scores. While not formally an IQ test, there is a strong correlation (0.83) between AGCT and IQ scores (Ferrie et al., 2012).

The Army General Classification Test (AGCT) was administered to all individuals who enlisted in the U.S. military during World War II. However, enlistment records report AGCT scores only for those who enlisted in spring 1943, because during that period clerks entered AGCT scores in the field otherwise used to record enlistees' weight (Ferrie et al., 2012). Accordingly, we restrict the analysis to individuals who enlisted between March and June 1943.

Since the World War II enlistment records have been digitized and include identifying information (name, year of birth, state of birth, and race), we can match them to the 1940 U.S. population census and our data on academics. Specifically, we use the crosswalk between World War II enlistment records and the 1940 Census constructed by Breen et al. (2023).

As above, we restrict the sample to individuals for whom we observe parental SES rank. We further restrict the sample to enlistees with available AGCT scores. We can thus observe future academics, as well as all other individuals, along with their AGCT score and parental SES rank. This sample includes 80,027 individuals, of whom about 0.01

percent later become academics. We use these data to estimate the following regression:

$$\mathbb{1}(Becomes\ Academic_i) = \theta \cdot Parental\ SES\ Rank_i + \beta AGCT_i + \epsilon_i \quad (2)$$

where $\mathbb{1}(Becomes\ Academic_i)$ is an indicator that is one if person i becomes an academic. To improve readability, we multiply the outcome variable by 100; thus, the dependent variable can be interpreted as a probability expressed as a percentage. $Parental\ SES\ Rank_i$ ranges from 0 to 100, capturing the percentile of the income rank of individual i's father. $AGCT_i$ are six indicators that control nonparametrically for a percentile range of individual i's *AGCT* score (e.g., above 99th percentile, 95-99th percentile).

Table 3: Socio-Economic Background and Probability of Becoming an Academic, Controlling for AGCT Scores

Dependent Variable:	*Becomes Academic*		
	Full Sample	AGCT Sample	
	(1)	(2)	(3)
Parental SES Rank	0.0024***	0.0029***	0.0013***
	(0.0001)	(0.0005)	(0.0004)
AGCT >25-50 percentile			-0.0126
			(0.0120)
AGCT >50-75 percentile			-0.0160
			(0.0146)
AGCT >75-90 percentile			0.0783**
			(0.0338)
AGCT >90-95 percentile			0.2025**
			(0.0802)
AGCT >95-99 percentile			0.8809***
			(0.1720)
AGCT >99 percentile			1.8469***
			(0.5024)
R^2	0.0002	0.0007	0.0067
Observations	6,838,403	80,027	80,027
Dependent Variable Mean	0.095	0.095	0.095

Notes: The table reports estimates of equation (2). The dependent variable is an indicator that is one if the individual i becomes an academic, multiplied by 100. The main explanatory variable is the father's SES rank, measured as the percentile in the predicted income distribution of scientist i's father. Column 1 shows the relationship between these two variables in the full sample of individuals who can be matched to the 1940 census. The remaining columns are limited to individuals with available *AGCT* (Army General Classification Test) scores. Significance levels: $***$ p<0.01, $**$ p<0.05, and $*$ p<0.1.

In column (1), we report results for the full 1940 census sample of men with observed parental SES, regardless of whether we can link them to an AGCT score. Consistent with our earlier results, we find a strong positive relationship between parental SES rank and the probability of becoming an academic. We then restrict the sample to men for whom we observe both parental SES and an AGCT score. The estimated association is very similar in this subsample (column (2)), which suggests that selection into the AGCT-linked sample is limited.

In column 3, we control nonparametrically for an individual's AGCT score. As expected, the probability of becoming an academic increases sharply with cognitive ability. This is particularly pronounced for higher AGCT percentiles, and especially for individuals above the 99th percentile. Including this control reduces the estimated coefficient on parental SES rank by 56.31%, consistent with a positive correlation between parental SES rank and AGCT performance.

However, the association between parental income rank and the probability of becoming an academic remains statistically significant and economically large. In particular, men whose fathers are at the 75th percentile of the income-rank distribution are 0.065 percentage points (68%) more likely to become academics than otherwise similar men whose fathers are at the 25th percentile.

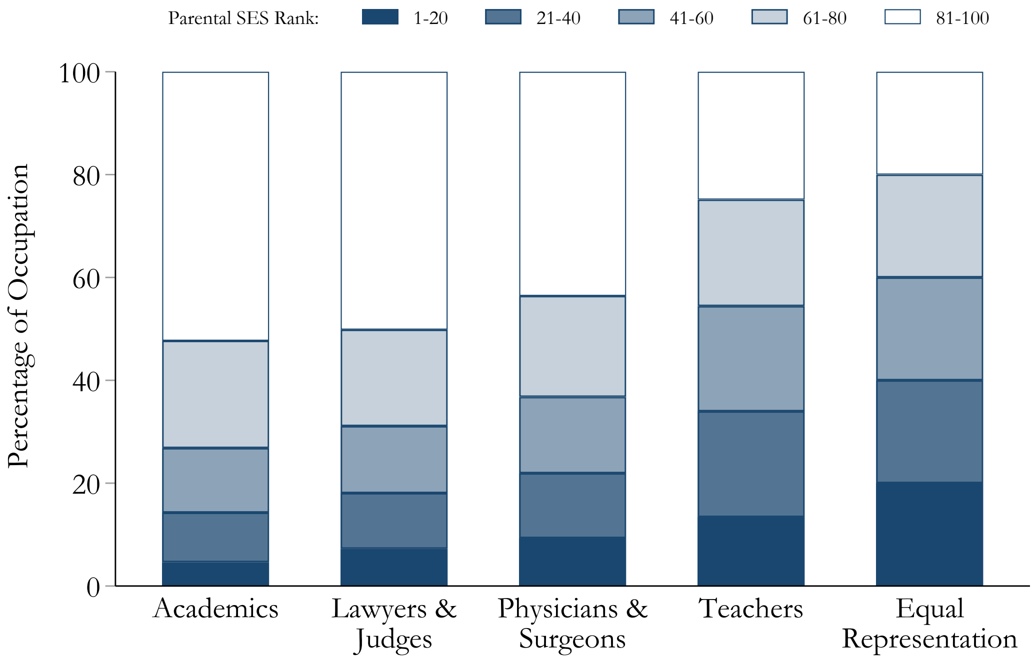
3.4 Representation in Academia versus Other Professions

A question arising from the previous findings is whether academia is an outlier compared to other professions. The small share of individuals from low socio-economic backgrounds in academia might simply reflect the fact that entering a profession requires credentials (e.g., a college degree), which might be expensive to obtain. To explore this, we compare the socio-economic backgrounds of academics to those of other professionals – lawyers and judges, physicians and surgeons, and teachers – using comparable data from the census (see Appendix A.2. for details). While lawyers and doctors also disproportionately come from high socio-economic backgrounds, the degree of selection in academia is even more pronounced (Figure 5). For example, 52% of academics come from the top quintile of the parental SES rank distribution, while "only" 50% of laywers and judges, and 44% of medical doctors come from the top quintile of the parental SES rank distribution. At the other end of the spectrum, representation from the bottom quintile of the parental SES rank distribution is especially low in academia: only 5% of academics come from the bottom quintile, while 7% of laywers, and 9% of doctors come from the bottom quintile. Teachers, in contrast, exhibit a much weaker degree of selection based on socio-economic background.

3.5 Representation by University

In the next set of results, we investigate whether individuals from lower socio-economic backgrounds are similarly underrepresented in all universities or if certain universities exhibit a higher degree of representation of individuals from these backgrounds. As the

Figure 5: Comparison to other Professions



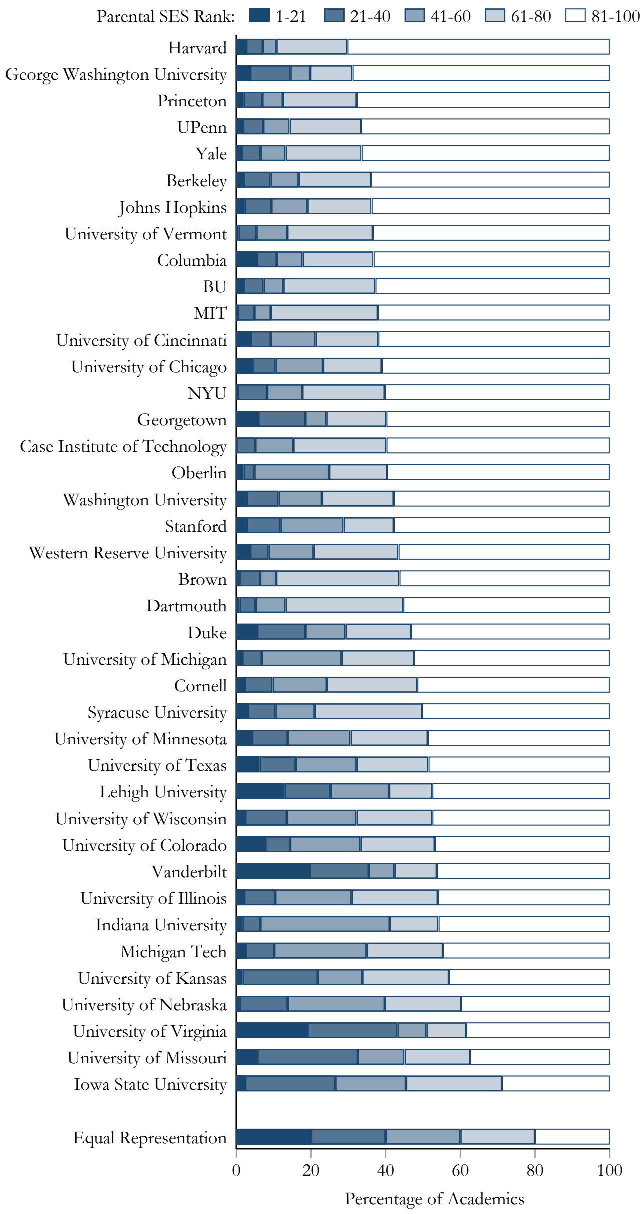
Notes: The figure compares the representation of academics based on their socio-economic background to the representation in other professions for the main sample. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Each color shows the percentage of individuals in an occupation whose fathers were in a specific quintile of the predicted income distribution. E.g., the white bar shows the percentage of individuals whose father was in the top quintile of the predicted income distribution. 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). Appendix Figure B.8 shows the equivalent figure for the extended sample.

faculty rosters contain more than 1,000 U.S. universities, we show examples for a small subset of these universities. We choose examples of universities for which we measure the socio-economic background of academics in each of the five cohorts plus all universities in the Ivy Plus group, as defined by Chetty et al. (2020).[24]

We find striking differences in representation by university (Figure 6, which is sorted in descending order based on the proportion of faculty with fathers from the top 20%). The most "socio-economically selective" universities are elite private universities such as those in the Ivy League – Harvard, Princeton, UPenn, and Yale. In contrast, universities with lower levels of "social selectivity" within this subset are predominantly public institutions, such as the University of Nebraska, the University of Missouri, and Iowa State University. These differences highlight significant variation in socio-economic representation across universities.

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

Figure 6: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university for the main sample. 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 quintile of the predicted income distribution. Appendix Figure B.9 shows the equivalent figure for the extended sample.

To more systematically investigate which university characteristics are correlated with socio-economic selectivity, we estimate the following regression for all universities:

$$\begin{aligned} \textit{Faculty Top SES Share}_u = \beta_0 + \beta_1 \text{Ivy Plus}_u + \beta_2 \text{Elite Private}_u + \beta_3 \text{Elite Public}_u \\ + \beta_4 \text{Discipline Shares} + \text{State FE} + \epsilon_i \end{aligned} \quad (3)$$

The dependent variable $\textit{Faculty Top SES Share}_u$ measures the share of academics of university u who come from the top 20, top 10, top 5, or top 1 % of the parental SES rank distribution. Ivy Plus_u is an indicator that equals one if university u is an Ivy Plus university as defined by Chetty et al. (2020). 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).[25]

The results indicate that Ivy Plus universities recruit faculty from significantly higher socio-economic backgrounds compared to other elite private institutions. These findings hold for the share of faculty from the top 20, top 10, top 5, and even top 1 %. While the average university in our sample recruits 3.4 % of their academics from the top 1 %, the share is about 5.2 percentage points higher in Ivy Plus universities (Table 4, column 12). In contrast, public elite institutions recruit their faculty from lower socio-economic backgrounds than Ivy Plus universities (Table 4).

The selectivity of universities may, in part, reflect their discipline composition. For example, Harvard does not have an agriculture department, which could influence the selectivity of its faculty. As demonstrated in the next section, representation varies substantially across disciplines. To address these differences, we add controls for the share of academics in each discipline. The results remain very similar (columns 2, 5, 8, and 11). The differences across university types are similar even though somewhat smaller if we control for state fixed effects (columns 3, 6, 9, and 12). This suggests that the observed patterns are not solely driven by geographical factors.

3.6 Representation by Discipline

While individuals from higher socio-economic backgrounds are overrepresented in all disciplines, there are large differences across disciplines (Figure 7). Agriculture, veterinary medicine, pedagogy, sociology, and pharmaceutics are the disciplines with the highest representation of individuals from lower socio-economic backgrounds. In contrast, the humanities, archaeology, architecture, cultural studies, medicine, anthropology, and law have the lowest representation.[26] Contrary to the common perception of economists, economics is more representative than the median discipline.

Figure 7 suggests that disciplines that require more sophisticated language skills have less representation from individuals of lower socio-economic backgrounds. In comparison, disciplines that require more mathematics skills 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 to verbal Graduate Record Exami-

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

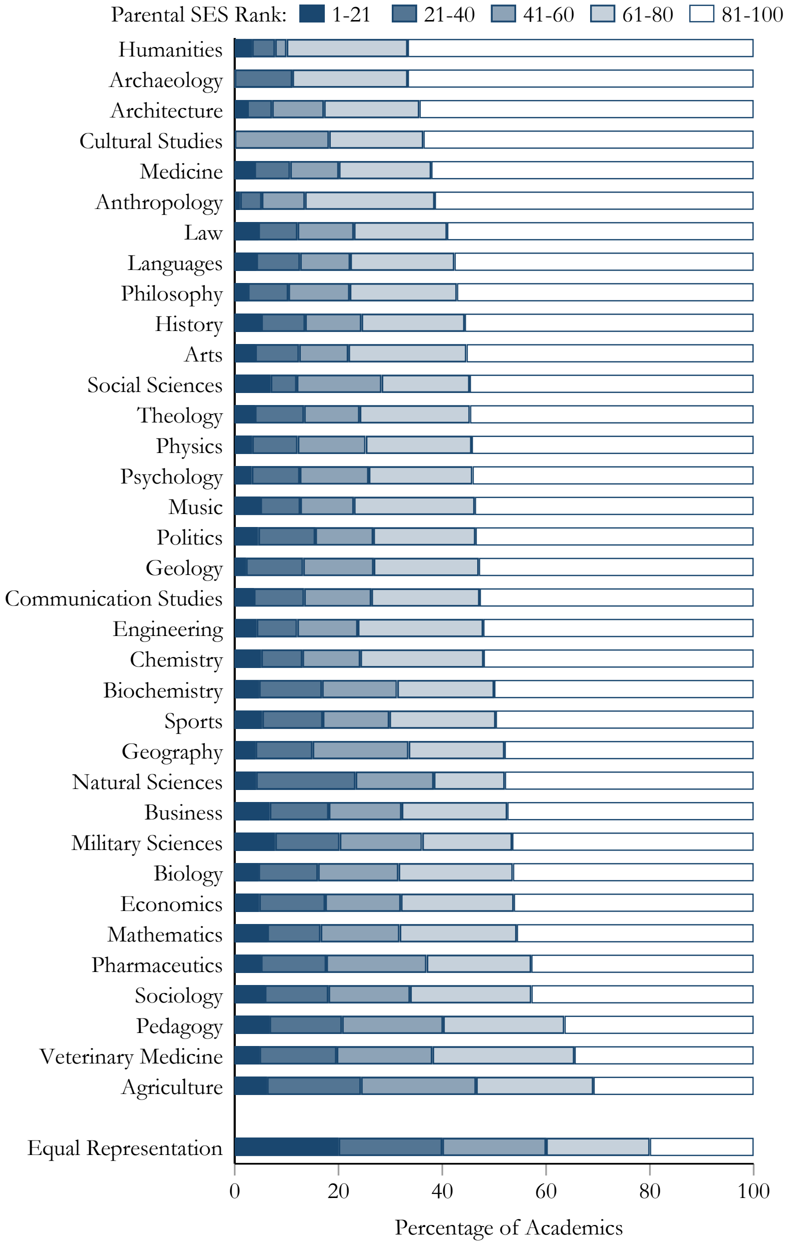
[26]Academics who list humanities, social sciences, and natural science as their discipline in the faculty rosters are less specialized and often teach at liberal arts colleges.

Table 4: 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	0.103**	0.096***	0.037	0.135***	0.115***	0.068***	0.115***	0.106***	0.054***	0.058***	0.059***	0.052***
	(0.045)	(0.025)	(0.028)	(0.031)	(0.020)	(0.020)	(0.023)	(0.019)	(0.017)	(0.008)	(0.020)	(0.019)
Private Elite	0.100***	0.059*	0.032	0.093***	0.054**	0.032	0.078***	0.044**	0.030	0.025***	0.012	0.010
	(0.031)	(0.031)	(0.035)	(0.023)	(0.023)	(0.026)	(0.016)	(0.016)	(0.018)	(0.007)	(0.009)	(0.009)
Public Elite	0.083***	0.092**	0.076*	0.028	0.027	0.030	0.019	0.015	0.018	0.011*	0.007	0.007
	(0.029)	(0.037)	(0.041)	(0.019)	(0.023)	(0.031)	(0.015)	(0.018)	(0.022)	(0.006)	(0.007)	(0.011)
R^2	0.02	0.10	0.19	0.02	0.10	0.18	0.03	0.13	0.22	0.03	0.10	0.16
Observations	755	755	755	755	755	755	755	755	755	755	755	755
Dependent Variable Mean	0.481	0.481	0.481	0.270	0.270	0.270	0.138	0.138	0.138	0.034	0.034	0.034
Panel B: 1900 – 1969												
Ivy Plus	0.146***	0.111***	0.059**	0.153***	0.124***	0.077***	0.112***	0.098***	0.054***	0.048***	0.044***	0.031**
	(0.040)	(0.031)	(0.030)	(0.031)	(0.026)	(0.023)	(0.029)	(0.018)	(0.014)	(0.005)	(0.014)	(0.014)
Private Elite	0.114***	0.056*	0.039	0.097***	0.049*	0.030	0.071***	0.039**	0.029	0.020***	0.008	0.009
	(0.031)	(0.031)	(0.037)	(0.025)	(0.025)	(0.029)	(0.014)	(0.017)	(0.018)	(0.006)	(0.010)	(0.010)
Public Elite	0.057	0.002	0.014	0.043	0.000	0.024	0.033	0.003	0.017	0.003	-0.006	-0.012
	(0.045)	(0.033)	(0.031)	(0.035)	(0.032)	(0.023)	(0.026)	(0.023)	(0.017)	(0.007)	(0.007)	(0.008)
R^2	0.01	0.08	0.16	0.02	0.10	0.18	0.02	0.08	0.17	0.01	0.05	0.10
Observations	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026	1,026
Dependent Variable Mean	0.449	0.449	0.449	0.249	0.249	0.249	0.130	0.130	0.130	0.035	0.035	0.035
Discipline Controls		Yes	Yes		Yes	Yes		Yes	Yes		Yes	Yes
State FEs			Yes			Yes			Yes			Yes

Notes: The table reports the estimates of Equation (3). The dependent variable measures the share of faculty in university u who come from the top 20 (columns 1-3), top 10 (columns 4-6), top 5 (columns 7-9), or top 1 percent (columns 10-12) 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. (2020). 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 university u 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 7: Representation by Discipline



Notes: The figure shows the representation of academics based on their socio-economic background by academic discipline for the main sample. 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 quintile of predicted income. Appendix Figure B.10 shows the equivalent figure for the extended sample.

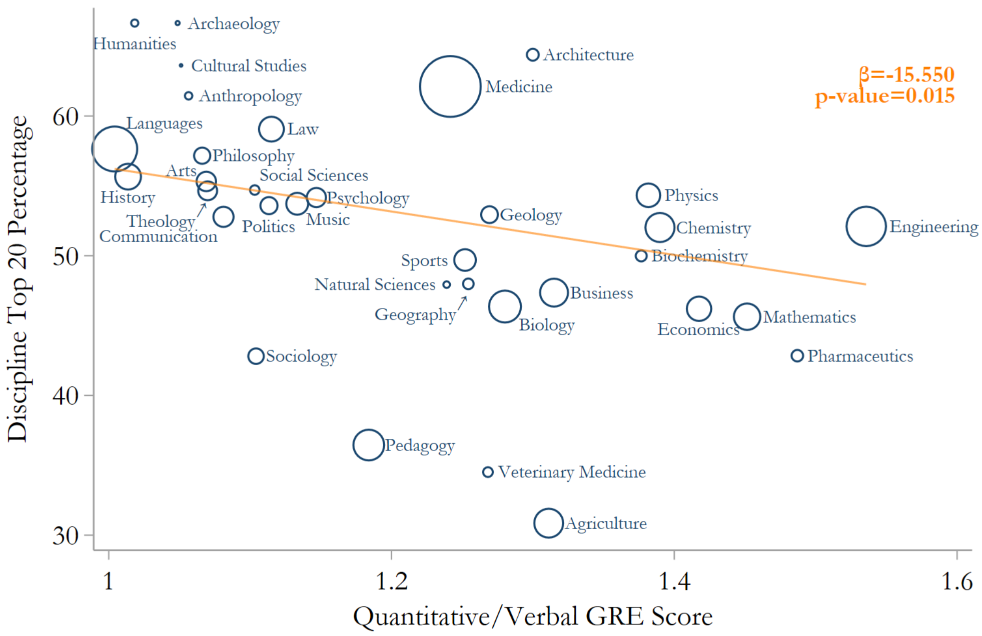
nation (GRE) scores for students intending to pursue graduate studies in each discipline.[27]

[27]The Educational Testing Service (ETS), which administers the GRE, publishes three-year average test scores of seniors and nonenrolled college graduates in three 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 majors into the corresponding disciplines from the faculty rosters and calculate the average quantitative versus verbal GRE score in each discipline. The data used in this analysis is based on the 2005–2008 cohorts of test-takers, obtained from the oldest available edition of the guide available via the Internet Archive Wayback Machine (ETS, 2009).

The findings suggest that representation from lower socio-economic backgrounds is indeed higher in disciplines that require more quantitative relative to verbal skills (Figure 8). The estimates imply that an increase in relative quantitative versus verbal skills by 0.5 (approximately the difference between history and mathematics) is associated with a 7.8 percentage point decrease in the share of academics from the top quintile of the parental SES rank distribution.

However, Figure 7 also highlights striking differences in representation even when comparing disciplines that arguably require similar skills. For instance, there are large differences in the socio-economic composition of medicine relative to veterinary medicine and of sociology relative to law. This suggests that factors beyond skill requirements also impact representation across disciplines.

Figure 8: Discipline Mathematics vs. Language Requirements and Representation



Notes: The figure shows the percentage 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 main sample. 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 GRE verbal reasoning score 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. We also report the coefficient and p-value from a discipline-size weighted regression of this relationship. Appendix Figure B.11 shows the equivalent figure for the extended sample.

4 Socio-Economic Background and Discipline Choice

In the second part of the analysis, we examine whether fathers' occupation affects academics' choice of discipline. This enables us to study a different facet of socio-economic background that goes beyond fathers' positions in the SES rank distribution.

4.1 Measuring Discipline-Level Overrepresentation by Father's Occupation

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

$$\text{Overrepresentation}_{do} = \frac{P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o)}{P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)}, \quad (4)$$

where $P(\text{Discipline}_i = d)$ is the probability of academic i working in discipline d, $P(\text{Father's Occupation}_i = o)$ 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.[28]

The measure isolates the relationship between a father's occupation and an academic discipline by accounting for baseline differences in the probabilities of choosing specific disciplines and having fathers in certain occupations. If there was no systematic relationship between father's occupation and the choice of discipline (i.e., the probabilities are independent), $\text{Overrepresentation}_{od} = 1$, since $P(\text{Discipline}_i = d, \text{Father's Occupation}_i = o) = P(\text{Discipline}_i = d) \cdot P(\text{Father's Occupation}_i = o)$. If a certain father's occupation is overrepresented in a specific discipline, the measure is greater than one. Inversely, in case of underrepresentation, the measure is smaller than one.

For example, we can calculate the overrepresentation of farmers' children among professors of agricultural science. The numerator measures the probability that an academic whose father was a farmer works as a professor of agricultural science (in our data this probability is 0.024). The denominator is the product of two probabilities: the probability of being a professor of agriculture among all academics (in our data: 0.043), and the probability that any academic's father was a farmer (in our data: 0.232). Thus the overrepresentation index for professors of agriculture who are farmer's children is $0.024/(0.043 \cdot 0.232) = 2.4$. 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 2.4 times more likely to be the child of a farmer, compared to academics overall. Thus, the measure quantifies the extent to which children of farmers are disproportionately represented in agricultural sciences.

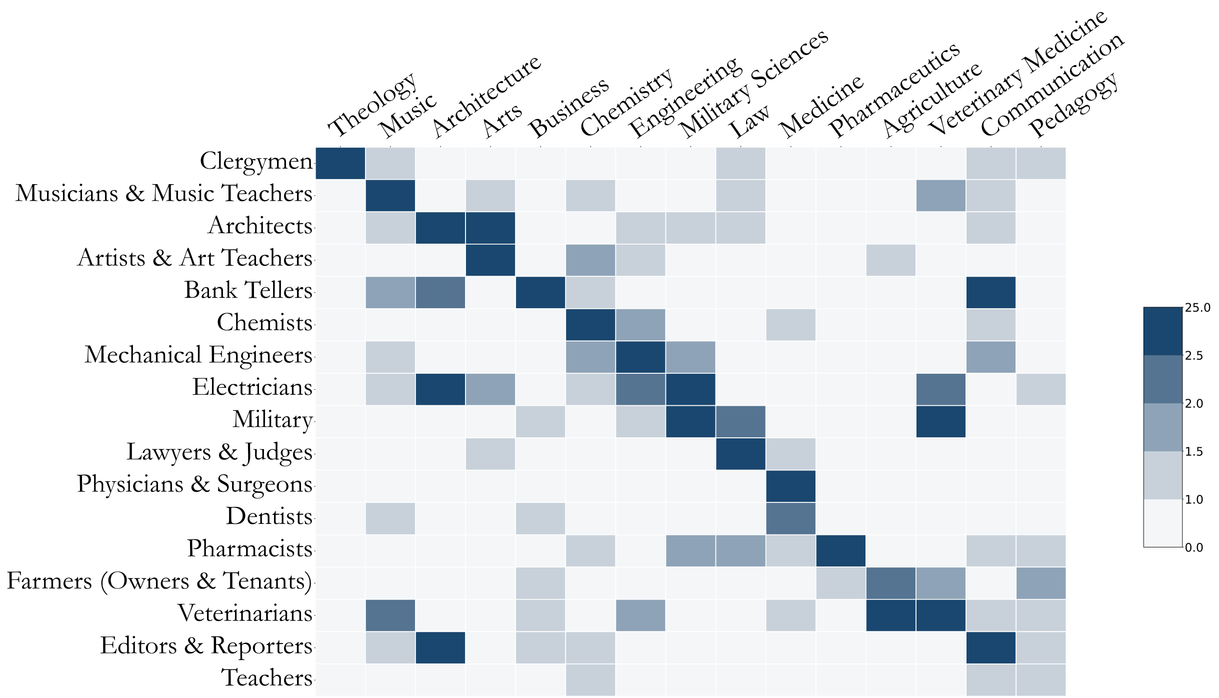
We calculate this measure for all pairs of father's occupations (130 different occupations in the data) and academic disciplines (34 disciplines), i.e., we calculate $130 \times 34 = 4,420$ overrepresentation indices.[29] We visualize examples of such pairs in Figure 9. The figure plots the father's occupation on the vertical axis and the academic discipline on the

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

[29]As the overrepresentation index is sensitive to outliers in small disciplines and occupations, we restrict the sample to disciplines for which we can observe the occupation of the father for at least 15 academics and to fathers' occupations in which at least 15 children become academics in any discipline.

horizontal axis. The blue shading indicates quartiles of the overrepresentation index, with darker blues indicating stronger overrepresentation.

Figure 9: Father's Occupation and Discipline Choice



Notes: The figure shows the relationship between father's occupation (rows) and their children's choice of academic discipline when the children become professors (columns) for selected father's occupation - discipline pairs for the main sample. Darker shades indicate higher levels of overrepresentation, as measured by equation (4). Appendix Figure C.1 shows the equivalent figure for the extended sample.

The figure suggests a strong connection between the father's occupation and their children's choice of academic discipline. For example, children of architects disproportionately become professors of architecture and arts, while children of artists and art teachers become professors in arts-related disciplines. Children of lawyers, medical doctors, or pharmacists predominantly become professors of law, medicine, and pharmaceutics, respectively. Children of editors and reporters become professors in communication studies, which encompasses journalism as a sub-discipline. Interestingly, this pattern extends to children of fathers in non-professional occupations. For example, children of bank tellers are overrepresented among professors in business and management. Meanwhile, children of teachers, who often teach various school subjects, exhibit a more evenly distributed representation across academic disciplines.

4.2 Predicting Semantically Close Academic Disciplines

Figure 9 presents a selected subset of father's occupation-discipline pairs, that we handpicked from the data chosen to illustrate notable patterns in the data. To systematically investigate the relationship between a father's occupation and their child's academic discipline choice, we construct an external measure of similarity between each father's occupation and each academic discipline. Crucially, this measure does not depend on observed transitions between specific paternal occupations and their children's academic disciplines. Instead, it is based solely on textual similarities, capturing how "close" a father's occupation is to a particular academic discipline. This method provides a systematic way to explore the relationship between the father's occupation and the discipline for all father's occupation-discipline pairs.

Specifically, we use text embeddings to measure the *semantic* similarity between the text string of all fathers' occupations and the text string of all academic disciplines. Embeddings convert text into fixed-length vector representations that encode both syntactic and semantic relationships learned from 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 intricate relationships between words. We use the "all-MiniLM-L6-v2" model, which has been trained on data from scientific papers, Wikipedia, Reddit, and many other sources.[30] The model represents each father's occupation string and each discipline string as a vector of length $n = 384$.

Following standard practice in natural language processing, we measure the similarity between each father's occupation and all academic disciplines using the cosine similarity of their vector representations:

$$Cosine\ Similarity(x, y) = \frac{\sum_{i=1}^{n} x_i \cdot y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}}, \tag{5}$$

where x represents the vector of a father's occupation and y represents the vector of a discipline, both 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. As outlined above, this measure is based solely on the textual representations of occupation and discipline strings without incorporating any information on professors' actual academic discipline choices. For example, as expected, the closest discipline in semantic space for the occupation "architect" is "architecture" (cosine similarity 0.77). Similarly, the closest discipline in semantic space for the occupation "buyers/shippers of farm products" is "agriculture" (cosine similarity 0.53).[31]

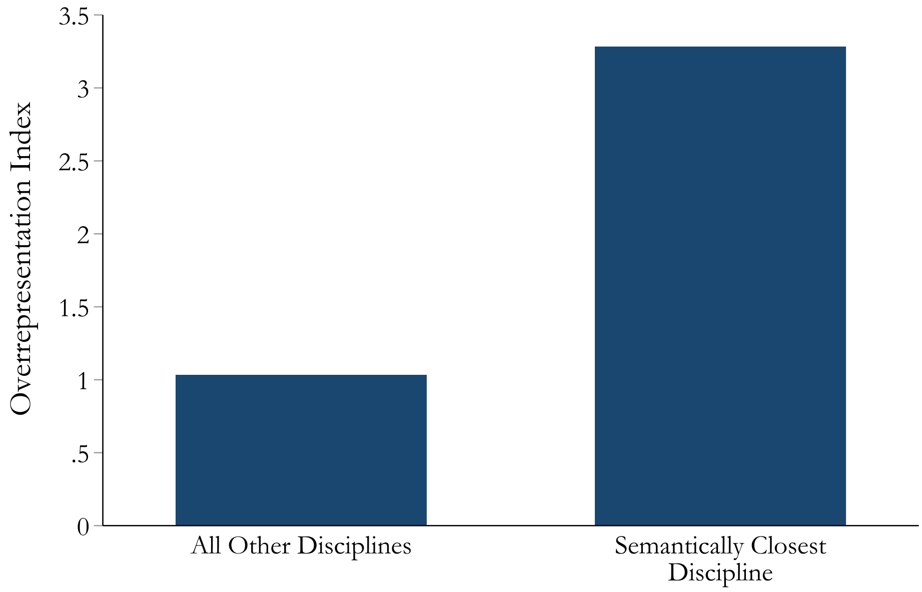
[30] The "all-MiniLM-L6-v2" model is one of the most commonly used sentence embedding models. For example, it was the third most downloaded model on huggingface.com as of July 2024. The findings do not depend on the choice of a specific model.

[31] To ensure that we predict close disciplines that are genuinely close in semantic space, we 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. For instance, while the discipline most similar to the occupation "private household worker" is law, the similarity falls below the mean cosine similarity threshold. As a

4.3 Overrepresentation in Semantically Close Disciplines

After identifying the semantically closest academic discipline for each occupation, we compute the average overrepresentation index (equation 4) for the discipline-occupation pair that is closest in semantic space. Additionally, we calculate the corresponding average for all other discipline-occupation pairs. This enables us to measure whether academics are *systematically* overrepresented in disciplines that are "close" to their father's occupation.

Figure 10: Overrepresentation in Semantically Closest Discipline



Notes: The figure shows overrepresentation as measured by equation (4) 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 the main sample. For more details, see section 4.2 and section 4.3. Appendix Figure C.2 shows the equivalent figure for the extended sample.

The average overrepresentation index is 3.28 in the semantically closest discipline (Figure 10). In contrast, the overrepresentation index is 1.03 for all other disciplines, indicating that for disciplines that are not semantically close to fathers' occupations, academics are represented as good as random.

Overall, these results provide further evidence that socio-economic background not only affects the likelihood of entering academia but also the choice of discipline. This new finding shows the influence of parental occupation on children's career trajectories, even when children pursue occupations different from those of their parents. Potential explanations for this phenomenon include increased interest stemming from the transmission of family values, early exposure to a particular field, or differential access to resources and opportunities, such as privileged information on how to succeed in a given discipline.

result, children of "private household workers" have no semantically closest discipline. Importantly, our findings remain robust when we redefine semantically close disciplines using a threshold of one standard deviation above the mean or when we eliminate the minimum cosine similarity requirement altogether (see Appendix Figure C.3).

Combined with the findings in the previous part of the paper, these results suggest that the unequal selection of academics based on socio-economic background could have repercussions for the composition of academic disciplines. Overrepresentation of individuals from certain parental occupations in academia could skew the composition of academic disciplines, leading to imbalances in the supply of talent. This misalignment may advantage some disciplines over others, not due to societal demand for knowledge, but rather due to the unequal distribution of opportunities.

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

In the third part of the analysis, we investigate whether and how socio-economic background influences productivity after entering academia. In particular, we study whether scientific productivity and novelty differ by socio-economic background. It is important to note that any potential differences by socio-economic background reflect a combination of selection effects and hurdles to publish.

5.1 Scientific Publications

We first explore differences in the number of publications by socio-economic background. As described in the data section, this analysis focuses on six scientific disciplines: medicine, biology, biochemistry, chemistry, physics, and mathematics, which are well-represented in academic publication databases. We estimate the following regression:

$$Publications_i = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \tag{6}$$

where $Publications_i$ captures different measures of scientific publications that scientist i published in a ± 5-year interval centered on the year that the scientist entered the faculty rosters, i.e., for scientists entering the faculty rosters in 1956, we measure publications from 1951 to 1961.[32] 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.[33] Standardized publications ease interpretation and account for differences in publications across disciplines and over time. $Parental\ SES\ Rank_i$ ranges from 0 to 100, capturing the percentile of the income rank of scientist i's father. The coefficient θ captures the relationship between socio-economic background and scientific output. We also include a set of controls, $\mathbf{X_i}$, to account for differences in scientific output

[32]We use a ± five-year window because many scientists do not publish every year. Results are very similar if we measure publications using alternative time-windows (e.g., a ± three-year window). Results are also similar, if we measure publications for all cohorts in which an academic was active.

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

by age, gender, cohort, discipline, or state. Since the parental SES rank is based on father's occupation, childhood state, and birth year of the scientist, 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

We find no systematic relationship between the socio-economic background and the *average* number of publications, regardless of the set of fixed effects that we include as regression controls (Table 5, columns 1-3). This result holds in the main sample (Panel A) and in the extended sample (Panel B). As described before, to account for differences in publication practices across disciplines and over time, we also estimate models using standardized publications. These results further confirm that there is no systematic relationship between the socio-economic background of scientists and the average number of publications (Table 5, columns 4-6).

Table 5: 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	0.00783*	0.00441	-0.00299	0.00040	0.00012	0.00006	-0.00113***	-0.00094***	-0.00052***
	(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	12,767	12,767	12,767	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	4.666	4.666	4.666	0.011	0.011	0.011	0.418	0.418	0.418
Panel B: 1900 – 1969									
Parental SES Rank	0.00419	0.00158	-0.00628	0.00016	-0.00005	-0.00014	-0.00102***	-0.00085***	-0.00043**
	(0.00408)	(0.00408)	(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	15,521	15,521	15,521	15,521	15,521	15,521	15,521	15,521	15,521
Dependent Variable Mean	4.912	4.912	4.912	-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

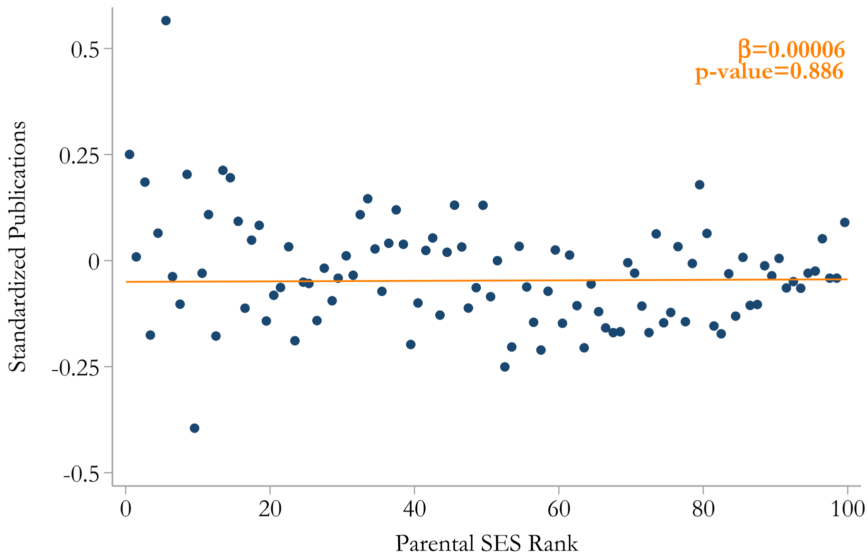
Notes: The table reports estimates of equation (6). In columns (1)-(3), the dependent variable measures publications in a ± 5-year window around the cohort when scientist i enters the faculty rosters. In columns (4)-(6), the dependent variable measures publications, standarized at the level of disciplines and cohorts. In columns (7)-(9), the dependent variable is an indicator that equals one if the scientist does not publish at all in the ± 5-year window around entering 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 scientist i's father. Demographic controls include age, age squared and an indicator for whether the scientist is female. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the scientist. Significance levels: $***$ p<0.01, $**$ p<0.05, and $*$ p<0.1.

We also visualize the relationship between parental income ranks (horizontal axis) and standardized publications (vertical axis) in a binned scatterplot (Figure 11). 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

A considerable share of scientists never publish in journals indexed by the Web of Science, which predominantly includes high-quality journals (Hager et al., 2024). To examine the likelihood of never publishing in a Web of Science-indexed journal, we estimate variants of

Figure 11: Socio-Economic Background and Average Number of Publications



Notes: The figure shows a binned scatterplot of the relationship between scientists' socio-economic background and publications. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. Publications are standardized within cohort and discipline. We show 100 quantiles and use the covariate adjustment (equivalent to column (4) in Table 5) as proposed in Cattaneo et al. (2024).

equation (6) with an alternative dependent variable that equals one if scientist 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 higher socio-economic backgrounds are significantly less likely to never publish (Table 5, column 7). For example, the probability of not publishing at all is approximately 4 percentage points (or around 10 percent) lower for scientists whose fathers were at the 75th percentile of the income distribution, compared to those with fathers at the 25th percentile. While the magnitude of this effect is halved when including the full set of fixed effects, it remains highly statistically significant. The result is also robust in the extended sample (Table 5, columns 7-9 and Panel B).

The Distribution of Publications

The preceding results suggest that while scientists from lower socio-economic backgrounds, on average, produce a comparable total number of publications, they are significantly more likely to have no publications at all. This suggests that scientists from lower socio-economic backgrounds must publish relatively more in higher percentiles of the publication distribution. To test this hypothesis, we estimate equation (6) with alternative dependent variables:

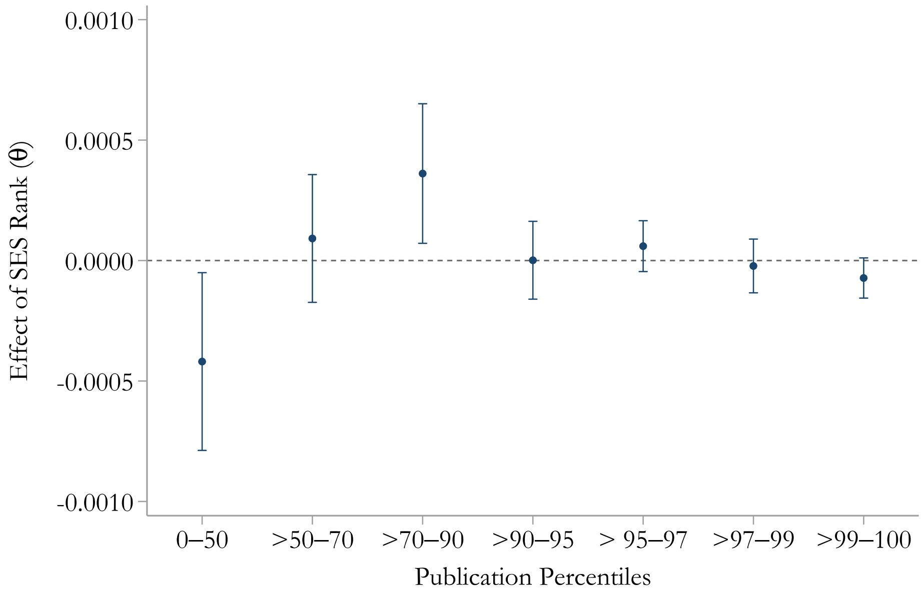
$$\mathbb{1}(Publication\ Percentile\ Range_i = q) = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \quad (7)$$

where the dependent variable $\mathbb{1}(Publication\ Percentile\ Range_i = q)$ is an indicator that equals one if scientist i's publication record falls within a specified percentile range q.

Since scientific productivity is well-known to be highly skewed (see e.g., Lotka 1926), we define the following percentile ranges of the publication distribution: $0 - 50$th (which coincides with not publishing at all for many 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. To account for variations in publication patterns across disciplines (e.g., chemists and medical researchers publish more than mathematicians) and cohorts (e.g., later cohorts tend to publish more), these percentiles are calculated at the discipline-cohort level. Appendix Table D.1 reports the number of publications required to achieve each percentile across disciplines and cohorts.

We estimate seven different regressions using an indicator for publishing in a certain publication percentile (Appendix Table D.2). We visualize the estimated coefficients in Figure 12.

Figure 12: Socio-Economic Background and the Distribution of Publications



Notes: The figure plots the estimated coefficients for θ for seven regressions of Equation 7. In each of the seven regressions, the dependent variable is an indicator of whether a scientist's number of publications falls within the relevant percentiles of the publication distribution, measured at the cohort and discipline-level. We report coefficients from regressions using the covariates and fixed effects equivalent to column (3) in Table 5. The corresponding regression results are reported in Appendix Table D.2.

The first coefficient from the left indicates that scientists from higher parental income ranks are less likely to have a publication count in the bottom 50% of the publication distribution. In contrast, the second coefficient ($> 50 - 70$) indicates that scientists from higher parental income ranks are as likely as scientists from lower parental income ranks to have a publication count between the 50th and the 70th percentile of the publication distribution. The third coefficient ($> 70 - 90$) indicates that scientists 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 scientists 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 scientists from higher parental income ranks are less likely to have a publication count in the top 1% of the publication distribution (p-value=0.089). This suggests that individuals from lower socio-economic backgrounds are disproportionately more likely to have publication records 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 distribution compared to scientists with fathers at the 25th percentile. This large effect, in percentage terms, is particularly relevant as a long-standing literature has highlighted that the most productive scientists have a disproportionate impact on the advancement of science (e.g., Lotka 1926, Merton 1957).

Overall, the findings on the distribution of publications suggest that scientists from lower socio-economic backgrounds may represent relatively "riskier" hires. They are more likely to have no publications at all but are also disproportionately represented in the top 1% of the publication distribution.

5.2 Novel Scientific Concepts

In the next subsection, we study whether and how the content of publications differs by socio-economic background and explore additional evidence of whether scientists from lower socio-economic backgrounds may pursue riskier research agendas.

To explore these hypotheses, we adopt the methodology developed by Iaria et al. (2018) and measure the number of novel words introduced by a scientist to the scientific community. The measure proxies for the introduction of new scientific concepts that require novel scientific terms. Specifically, we define novel words as words that were used for the first time in the title of a paper and had not been used in any prior paper title in the entire Web of Science database (not just the papers published by the scientists in our sample).

As the coverage of the Web of Science begins in 1900, we compute novel words for paper titles published from 1910 onwards. This approach allows for a 10-year window to identify words appearing in scientific papers before designating a word as novel. Consequently, 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 number words (e.g., thirty-five), from the data.[34]

[34] We exclude 211 number words plus the 36,663 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. This ensures that we exclude frequently used words that reflect historical language use relevant to the period of analysis. The results are robust to excluding only 10,000 or 20,000 frequently used words reported in Project Gutenberg (Appendix Table D.3 and D.4).

One example of a novel scientific term is "microbeam," which was used and developed by Raymond E. Zirkle to study the effects of ionizing radiation on living cells.[35] Zirkle, who is widely regarded as the pioneer in the field of radiation biology, grew up on a farm in northern Oklahoma. "As a young boy, his only source of education were one-room country schoolhouses in Oklahoma and southern Missouri. He gained exposure to the outside world and science through reading books." (Atomic Heritage Foundation, 2022). During WW2, Zirkle became a principal investigator in the Manhattan Project biological research program, where he worked on assessing the risk of radiation. From 1944 onwards, he worked at the University of Chicago, where he served as director of the Institute of Radiobiology and Biophysics. In 1952, he also became the first president of the Radiation Research Society. Beginning in the 1990s, Zirkle's pioneering work contributed to a significant improvement of radiation therapy for cancer treatment (Wu and Hei, 2018).

To examine how socio-economic background is associated with the introduction of novel scientific terms, we estimate the following regression:

$$Novel\ Words_i = \omega \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \quad (8)$$

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. For example, 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 facilitate interpretation and to account for differences in the number of novel words introduced in different disciplines and over time, we also estimate regressions with an alternative dependent variable that measures novel words that are standardized at the level of disciplines and cohorts. 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 introducing novel words by age, cohort, and discipline.
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 6, column 1, significant a the 10% level). The result is similar and becomes significant at the 5% level if we control for university state and discipline fixed effects (Table 6, columns 2-3). Specifically, scientists whose fathers were at the 75th percentile of the income rank publish around 0.05 fewer papers (around 16% less) with at least one novel word compared to those whose fathers were at the 25th percentile. The results are robust to standardizing the novel words measure at the level of disciplines and cohorts (Table 6, columns 4-6) and in the extended sample (Table 6, panel B).

We also exclude words as novel, if they only appear in a single paper. In many cases, such "words" are misspelled in the Web of Science and thus do not reflect a novel scientific concept.

[35]The first noted use of micro-irradiation for the analysis of cells stems from work by Tschachotin (1912) who used ultraviolet light.

Table 6: 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	-0.00092*	-0.00106**	-0.00105**	-0.00074*	-0.00092**	-0.00092**
	(0.00050)	(0.00049)	(0.00050)	(0.00043)	(0.00044)	(0.00044)
R^2	0.01	0.02	0.05	0.02	0.02	0.03
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.320	0.320	0.320	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00077*	-0.00086*	-0.00087*	-0.00070*	-0.00083**	-0.00085**
	(0.00044)	(0.00044)	(0.00044)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.317	0.317	0.317	-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 (8). In columns (1)-(3), the dependent variable measures the number of publications that 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. For the results reported in columns (4)-(6), 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.

Novel Words Used in Subsequent Research

Individuals from lower socio-economic backgrounds introduce more novel scientific concepts, but it may be the case that these concepts are rarely used in subsequent research. To test this hypothesis, we construct alternative dependent variables that indicate whether a scientist introduces a novel word that appears at least three times, five times, or ten times in subsequent papers (published by any scientist in the world).

We find that scientists from lower socio-economic backgrounds are not only more likely to introduce novel words that appear once but also more likely to publish papers that introduce novel words that subsequently appear in the titles of other academic papers (Table 7, column 1). Compared to scientists with fathers at the 25th percentile of the income rank distribution, scientists with fathers at the 75th percentile publish around 4% of a standard deviation fewer papers with at least one novel word that appeared in at least three subsequent papers. The magnitudes are very similar if we impose higher thresholds for subsequent usage (Table 7, columns 2-3). This suggests that scientists from low socio-economic backgrounds are indeed more likely to introduce novel ideas that are adopted by the broader scientific community.

Table 7: Socio-Economic Background and Novelty (Mentions in Subsequent Papers and Patents)

Dependent Variable:	Std. Papers with Novel Word Appearing in:					
	Subequent Papers			*Subequent Patents*		
	3×	5×	10×	3×	5×	10×
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: 1914 – 1956						
Parental SES Rank	-0.00089**	-0.00084**	-0.00084**	-0.00076*	-0.00075*	-0.00064*
	(0.00042)	(0.00041)	(0.00041)	(0.00039)	(0.00039)	(0.00039)
R^2	0.02	0.02	0.02	0.02	0.02	0.02
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	-0.003	-0.001	-0.003	-0.007	-0.009	-0.009
Panel B: 1914 – 1969						
Parental SES Rank	-0.00072**	-0.00070*	-0.00078**	-0.00074**	-0.00073**	-0.00063*
	(0.00036)	(0.00036)	(0.00036)	(0.00034)	(0.00034)	(0.00034)
R^2	0.02	0.02	0.02	0.02	0.02	0.01
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	-0.009	-0.007	-0.009	-0.014	-0.014	-0.015
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	Yes	Yes
Discipline FEs	Yes	Yes	Yes	Yes	Yes	Yes

Notes: The table reports estimates of variants of Equation (8). In columns (1)-(3), the dependent variable measures the number of publications that introduce at least one novel word (published in a ± 5-year window around the cohort when scientist i enters the faculty rosters) that was subsequently mentioned in the titles of at least 3, 5, or 10 papers. In columns (4)-(6), the dependent variable measures the number of publications that introduce at least one novel word (published in a ± 5-year window around the cohort when scientist i enters the faculty rosters) that was subsequently mentioned in the full text of at least 3, 5, or 10 U.S. patents. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist 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.

Novel Words Used in Subsequent Patents

To evaluate whether the new basic science also influenced technological progress, we examine the application of novel scientific concepts in the development of new technologies, as reflected in patent texts. Specifically, we follow Iaria et al. (2018) and measure how many subsequent patents mentioned the novel words that were introduced to the scientific community.

To construct this measure, we analyze the full text of over 2.5 million patents granted by the U.S. Patent Office between 1920 and 1979.[36] We identify occurrences of the novel scientific terms in the full text of all U.S. patents. Specifically, for a paper published in year t, the measure counts the number of times the novel scientific word appears in subsequent patents that were granted between year t and 1979. For example, the novel scientific word "microbeam," which was introduced in 1954, appeared in 15 subsequent

[36] The Patent Grant OCR Text files stem from a cooperation between Google and the USPTO are available here.

patents, including US3406044, assigned to Monsanto, and US3488494, assigned to the US Atomic Energy Commission. The patents broadly cover applications of microbeams for radiation microscopy.

We re-estimate equation 8 with an alternative dependent variable that measures the number of novel words that scientist i introduced in a paper that were subsequently used in at least three (or alternatively five, or ten) patents. Scientists from lower socio-economic backgrounds are significantly more likely to introduce novel words that appear in subsequent patents (Table 7, columns 4-6). For example, compared to scientists with fathers at the 25th percentile of the income rank distribution, scientists whose fathers were at the 75th percentile publish around 4% of a standard deviation fewer papers with at least one novel word that appeared in least three subsequent patents (column 4). We find relatively similar effects if we impose higher thresholds for subsequent appearances in patents (columns 5-6).

Overall, these findings suggest that scientists from lower socio-economic backgrounds contribute more novel scientific concepts that are relevant for follow-on academic research. Moreover, they are also more likely to generate concepts with technological relevance, as evidenced by their subsequent use in patents.

6 Socio-Economic Background and Recognition

In the last part of the paper, we examine the relationship between socio-economic background and recognition by other academics. First, we analyze citations to a scientist's research papers, a widely-used metric for measuring recognition within the scientific community. Next, we investigate Nobel Prize nominations and awards as indicators of recognition for exceptional scientific contributions.[37]

6.1 Citations

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

$$Citations_p = \gamma \cdot Avg.\ Parental\ SES\ Rank_p + \mathbf{X_p'}\beta + \epsilon_p \tag{9}$$

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

[37] Citations and Nobel nominations and awards reflect both recognition and inherent differences in quality (Hager et al., 2024).

at the level of disciplines and the year of publication.[38] Since the distribution of citations is highly skewed and contains outliers,[39] we also estimate results where we winsorize citation counts at the 99th percentile of the discipline-and-publication year-specific distribution of citations. $Avg\ Parental\ SES\ Rank_p$ 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 observe characteristics at the author level, we aggregate them for all authors of paper p that we can link to a childhood census.[40] 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 the author team's fathers' occupations, childhood states, and birth years to account for potential correlations of regression residuals.

Table 8: Parental SES Rank and Paper-Level Citations

Dependent Variable:	*Standardized Citations*					*Winsorized Std. Citations*				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Panel A: 1900 – 1956										
Average Parental SES Rank	0.00080**	0.00060*	0.00058*	0.00061*	0.00061*	0.00085***	0.00068***	0.00067***	0.00067***	0.00067***
	(0.00033)	(0.00033)	(0.00033)	(0.00032)	(0.00032)	(0.00026)	(0.00026)	(0.00026)	(0.00024)	(0.00023)
R^2	0.03	0.04	0.04	0.10	0.10	0.03	0.04	0.04	0.13	0.14
Observations	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549	58,549
Dependent Variable Mean	0.012	0.012	0.012	0.012	0.012	-0.021	-0.021	-0.021	-0.021	-0.021
Panel B: 1900 – 1969										
Average Parental SES Rank	0.00081***	0.00068**	0.00067**	0.00068**	0.00067**	0.00076***	0.00066***	0.00065***	0.00063***	0.00063***
	(0.00029)	(0.00029)	(0.00029)	(0.00027)	(0.00027)	(0.00022)	(0.00022)	(0.00022)	(0.00020)	(0.00020)
R^2	0.02	0.03	0.03	0.10	0.10	0.02	0.04	0.04	0.14	0.14
Observations	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014	76,014
Dependent Variable Mean	0.011	0.011	0.011	0.011	0.011	-0.021	-0.021	-0.021	-0.021	-0.021
Demographic Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Childhood State FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Publication Year FEs	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Uni State FEs		Yes	Yes	Yes	Yes		Yes	Yes	Yes	Yes
Discipline FEs			Yes	Yes	Yes			Yes	Yes	Yes
Journal FEs				Yes	Yes				Yes	Yes
Author Count FEs					Yes					Yes

Notes: The table reports the estimates of Equation (9). 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 (columns 1-5), and winsorize standardized citations at the 99th percentile to account for extreme outliers (columns 6-10). 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 proxy the SES rank of fathers with the percentile in the predicted income distribution the father. Demographic controls include age, age squared and the share of female authors. Standard errors are clustered at the level of the author teams' fathers' occupation, childhood states, and birth years. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

We find that papers authored by teams from higher socio-economic backgrounds receive more citations (Table 8, panel A, column 1, significant at the 5% level). Specifically, papers authored by individuals whose fathers, on average, are ranked at the 25th percentile of the income rank distribution, receive approximately 0.04 standard deviations fewer citations compared to papers authored by individuals with fathers ranked at the 75th percentile. For example, in medicine, this translates to a paper receiving 2 to 3.5 (13% of the mean) more citations per paper.

[38] For the purpose of standardization, we use the full distribution of citations to *all* papers linked to U.S. scientists, rather than the distribution of citations for papers of those scientists, whose parental background can be observed.

[39] For example, a 1955 medical paper received 61 standard deviations more citations than the average medical paper in that year.

[40] Specifically, we average continuous variables, i.e., we control for the mean age and the share female authors of paper p, and create a separate fixed effect for each combination of childhood states as well as university states of the author team.

The results remain robust, albeit slightly smaller in magnitude, when we include fixed effects for the author team's university state and discipline combination, as well as journal fixed effects. In columns 5 and 10, we add fixed effects for both the total number of authors and the number of authors for which we observe a parental SES rank. When we account for extreme outliers in citations by windsorizing at the 99th percentile (columns 6-10), we estimate coefficients of similar magnitude which are highly significant.

6.2 Nobel Prize: Nominations and Awards

Nobel Prize Nominations

Next, we study an alternative measure of scientific recognition that captures whether fellow scientists regard a scientist's body of research deserving for a Nobel Prize nomination (Iaria et al., 2018). During this period, Nobel Prize nominations were made by a select group of elite scientists, making the nominations a marker of peer recognition by the scientific elite. We study this question using the following regression:

$$\mathbb{1}\{Nobel\ Nomination_i\} = \theta \cdot Parental\ SES\ Rank_i + \mathbf{X_i'}\beta + \epsilon_i \tag{10}$$

where $\mathbb{1}\{Nobel\ Nomination_i\}$ is an indicator for whether scientist i was ever nominated for a Nobel Prize. 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.

Individuals from higher parental SES ranks are more likely to be nominated for a Nobel Prize. Specifically, scientists with fathers at the 75th percentile of the income rank distribution have a 0.06 percentage point (or 50%) higher probability of being nominated compared to scientists with fathers at the 25th percentile (Table 9, column 1).

The results are robust to controlling for the state of the scientist's university and the discipline (Table 9, columns 2-3). The results are also robust to controlling for both publications and citations (Table 9, columns 4-6), indicating that scientists from poorer backgrounds are less likely to be nominated for a Nobel Prize, even if they have the same number of publications and citations.

Nobel Prize Awards

We also investigate the relationship between the parental income rank and the probability of *winning* a Nobel Prize. We estimate a variant of Equation (10) with an indicator for winning the Nobel Prize as the dependent variable. We find that scientists from higher parental SES ranks are more likely to win a Nobel Prize (Table 10). Although some coefficients are not statistically significant, the point estimates remain largely consistent across specifications, regardless of the fixed effects and controls included in the regression. Specifically, scientists with fathers at the 75th percentile of the income rank distribution have a 0.015 percentage point (or 50%) higher probability of winning a Nobel Prize compared to scientists with fathers at the 25th percentile (Table 10).

Table 9: 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	12,767	12,767	12,767	12,767	12,767	12,767
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	15,521	15,521	15,521	15,521	15,521	15,521
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 (10). 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 scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist 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 scientist. Significance levels: *** p<0.01, ** p<0.05, and * p<0.1.

This finding is robust to controlling for the scientist's publication and citation record. The results reported in columns 4–6 of Table 9 and Table 10 include linear controls for publications and citations. However, it is plausible that nominations for the Nobel Prize, as well as the likelihood of winning, are particularly influenced by having exceptional publication and citation records. As shown above, socio-economic background is correlated with both publications and citations, and this correlation may differ across various points in the publication and citation distributions. To assess whether this correlation drives our Nobel Prize results, we test the robustness of our findings by flexibly controlling for publication and citation records. Specifically, we include indicator variables for each percentile of the publication and citation distributions. The point estimates remain largely unchanged. However, the results for winning the Nobel Prize become somewhat less statistically significant (Appendix Table D.5).

Overall, these results suggest that socio-economic background plays a significant role in shaping peer recognition, as measured by Nobel Prize nominations and awards, with scientists from less privileged backgrounds receiving disproportionately less recognition from the scientific elite.

Table 10: 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	12,767	12,767	12,767	12,767	12,767	12,767
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	15,521	15,521	15,521	15,521	15,521	15,521
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 estimates of a variant of equation (10). The dependent variable is an indicator whether a scientist was awarded the Nobel Prize. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist 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 scientist. Significance levels: $***$ p<0.01, $**$ p<0.05, and $*$ p<0.1.

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 universities and disciplines. Further, we find that father's occupation is systematically related to the choice of discipline. Once in academia, socio-economic background is not related to the number of publications, on average, but scientists 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. We also find that scientists from lower socio-economic backgrounds 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 scientists from lower socio-economic backgrounds receive less recognition by the scientific community, as proxied by citations and Nobel Prize nominations and awards. 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.

Abramitzky, R., J. K. Kowalski, S. Pérez, and J. Price (2024). The GI Bill, Standardized Testing, and Socioeconomic Origins of the US Educational Elite Over a Century. *National Bureau of Economic Research Working Paper*.

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. *National Bureau of Economic Research Working Paper*.

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

Akcigit, U., J. Grigsby, and T. Nicholas (2017). The rise of american ingenuity: Innovation and inventors of the golden age. *National Bureau of Economic Research Working Paper*.

Almgren, M., J. Kramer, and J. Sigurdsson (2023). It runs in the family: Occupational choice and the allocation of talent. Technical report, Mimeo.

Atomic Heritage Foundation (2022). Raymond E. Zirkle. `https://ahf.nuclearmuseum.org/ahf/profile/raymond-e-zirkle/`. accessed December 9, 2024.

Babcock, L., M. P. Recalde, L. Vesterlund, and L. Weingart (2017). Gender differences in accepting and receiving requests for tasks with low promotability. *American Economic Review 107*(3), 714–47.

Bagues, M., M. Sylos-Labini, and N. Zinovyeva (2017). Does the gender composition of scientific committees matter? *American Economic Review 107*(4), 1207–38.

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.

Blau, P. M. and O. D. Duncan (1967). *The American Occupational Structure.* John Wiley and Sons.

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.

Breen, C. F., M. Osborne, and J. R. Goldstein (2023). Censoc: Public linked administrative mortality records for individual-level research. *Scientific Data 10*(1), 802.

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

Economic Research Working Paper.

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. *National Bureau of Economic Research Working Paper*.

Chetty, R., J. N. Friedman, E. Saez, N. Turner, and D. Yagan (2020). 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, Extended Table 4. available at `https://web.archive.org/web/20121222214014/http://ets.org/Media/Tests/GRE/pdf/gre_0910_guide_extended_table4.pdf`, archived on December 22, 2012.

Ferrie, J. P., K. Rolf, and W. Troesken (2012). Cognitive disparities, lead plumbing, and water chemistry: Prior exposure to water-borne lead and intelligence test scores among world war two us army enlistees. *Economics & Human Biology 10*(1), 98–111.

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 114*(12), 4052–90.

Hsieh, C.-T., E. Hurst, C. I. Jones, and P. J. Klenow (2019). The Allocation of Talent and U.S. 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. available at 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. available at https://usa.ipums.org/usa/chapter4/chapter4.shtml, Last accessed on 2024-09-07.

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).

Long, J. and J. Ferrie (2013). Intergenerational occupational mobility in great britain and the united states since 1850. *American Economic Review 103*(4), 1109–1137.

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? *National Bureau of Economic Research Working Paper*.

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.

Moser, P. and S. Kim (2022). Women in Science. Lessons from the Baby Boom. *mimeo NYU and University of Pennsylvania*.

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

Novosad, P., S. Asher, C. Farquharson, and E. Iljazi (2024). Access to opportunity in the sciences: Evidence from the nobel laureates. *Centre for Economic Policy Research Discussion Paper*.

Rogoff, N. (1950). *Recent trends in occupational mobility.* Ph. D. thesis, The University of Chicago.

Ross, M., B. Glennon, R. Murciano-Goroff, E. Berkes, B. Weinberg, and J. Lane (2022, August). Women are credited less in science than men. *Nature 608*(7921), 135–145.

Rossiter, M. W. (1982). *Women Scientists in America: Struggles and Strategies to 1940*, Volume 1. JHU Press.

Rossiter, M. W. (1998). *Women scientists in America: Before affirmative action, 1940-1972*, Volume 2. JHU Press.

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*.

Tschachotin, S. (1912). Die mikroskopische strahlen-stichmethode eine zelloperationsmethode. *Biol. Zentralblatt 32*, 623–630.

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

Wu, J. and T. K. Hei (2018). Focus small to find big–the microbeam story. *International journal of radiation biology 94*(8), 782–788.

Xie, Y. and A. Killewald (2013). Intergenerational occupational mobility in great britain and the united states since 1850: Comment. *American Economic Review 103*(5), 2003–2020.

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 3.
- Appendix C reports robustness checks and additional findings related to section 4.
- Appendix D reports robustness checks and additional findings related to section 5.

A Appendix: Additional Details on Data

A.1. Constructing Parental SES Ranks – Details

As described in the main paper, we use the 1940 census to predict income. We use interactions of fathers' occupation and home state to predict father's income for all childhood censuses (see Equation 1). For some census years and occupations, this approach faces two issues:

1. Rare ocupations
2. Changing occupation coding

To overcome these issues, we adjust 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. We thus adjust the income prediction for occupation × state with less than 10 observations, by estimating the following regression to predict income:

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

I.e., instead of interacting occupations with states, we interact them with census regions, and estimate a separate state fixed effect.

For even rarer occupations, i.e., those with less than 10 observations in a certain occupation by census *region* cell, we adjust our prediction further:

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

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 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 (IPUMS, 2024b). We exclusively use the integrated 1950 occupation classification in this paper.

The Census Bureau harmonization process implies that some 1950 occupation codes are present in earlier census years, but not in 1940. For example, the 1950 occupation classification includes codes for "mining engineers" and for "metallurgical engineers", whereas the 1940 occupation classification pools the two engineering fields. In contrast the 1930 and 1920 censuses contain a separate occupation code for "mining engineers."

To address the issue of occupation codes that are aggregated for the 1940 census but disaggregated for earlier censuses, we predict fathers' income via the following regression:

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

where an Occupation Group is the broad one-digit occupational category of an occupation.[41]

Note, that this issue only affects 1.9 % of academics in our data.

A.2. Constructing Comparison Group Samples for Other Professions

To compare representation among academics to other professions, we construct samples for lawyers and judges, physicians and surgeons, and teachers, from U.S. Censuses. We proceed in three steps:

1. From each available full-count census corresponding to the coverage period of the World of Academia Database (1900-1950), we extract all observations with occupation code 55 (Lawyers & Judges), 75 (Physicians & Surgeons) and 93 (Teachers).[42]

2. We use Census Linking Project to de-duplicate individuals who appear in multiple censuses and keep only one observation per individual.

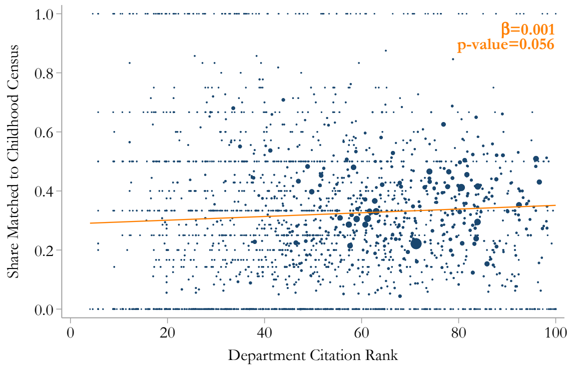
[41] Professional, Technical; Farmers; Managers, Officials, and Proprietors; Clerical and Kindred workers; Sales workers; Craftsmen; Operatives; Service workers (private household); Service workers (not household); Farm Laborers; Laborers (non-farm). See IPUMS (2024a).

[42] As discussed in the main text, some academics are not listed as professors but, e.g., as lawyers or surgeons in the U.S. census, we therefore remove all matched academics from this sample.

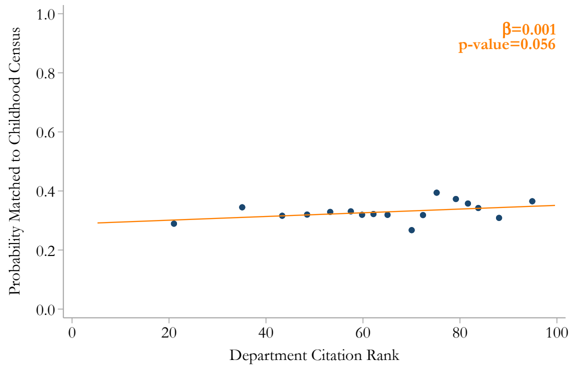
3. We then link these observations to their childhood census and construct parental SES ranks as described in Section 2.2.

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

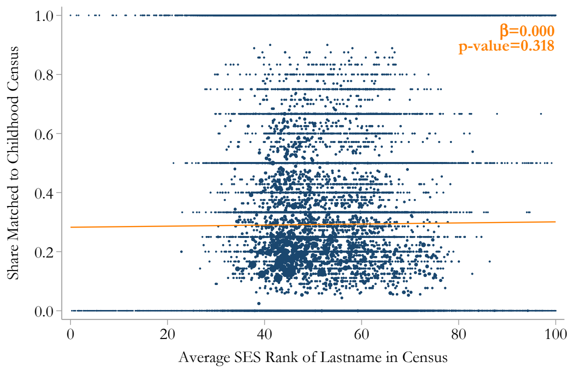
(a) Department Quality



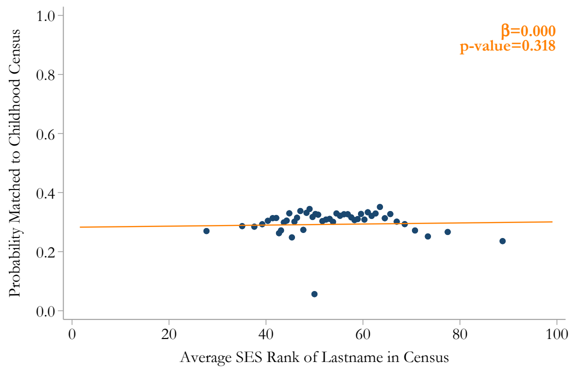
(b) Department Quality



(c) Last Name Parental SES Rank



(d) Last Name Parental SES Rank



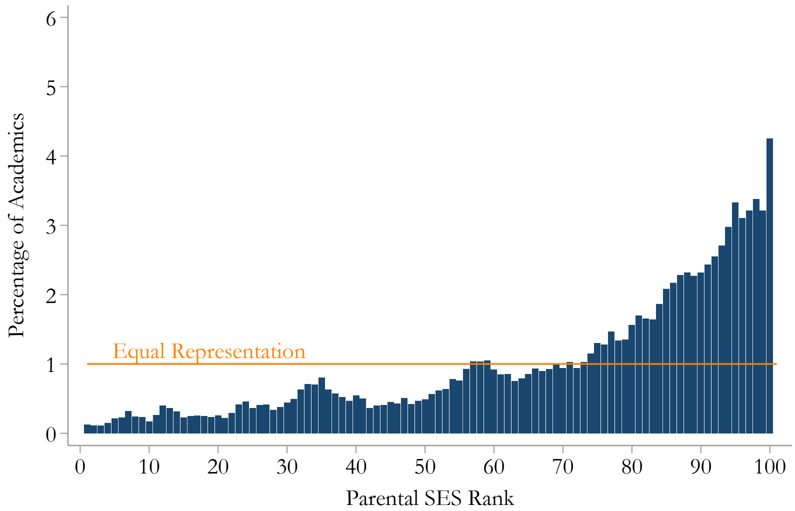
Notes: Panel (a) shows the correlation between a department's citation rank and the probability of linking a scientist to a childhood census for the extended sample (1900-1969). Panel (b) shows a binned scatter plot of the same relationship. Panel (c) shows the correlation between a last name's SES Rank 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 Socio-Economic Background and the Probability of Becoming an Academic: Additional Results

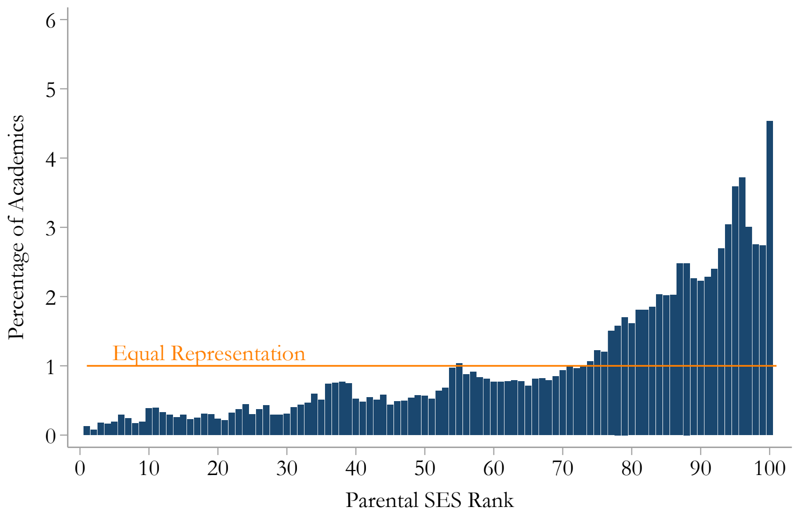
Representation of Academics by Socio-Economic Background

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

(a) With Regional Variation



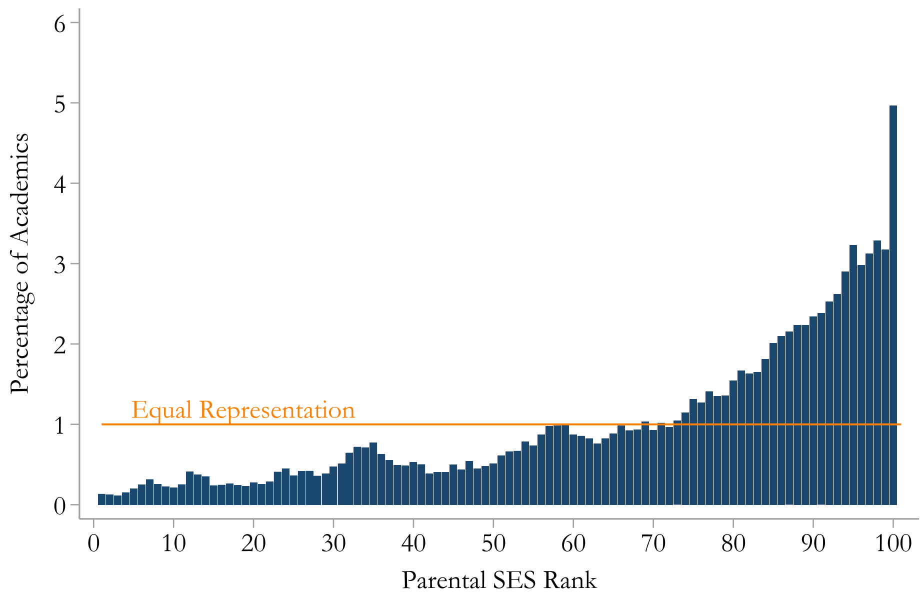
(b) Without Regional Variation



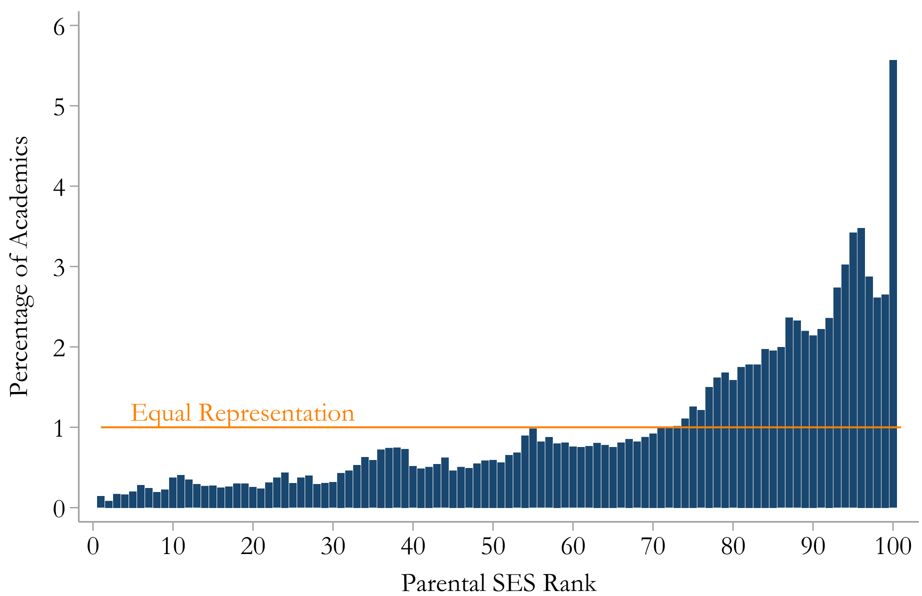
Notes: The figure shows the representation of academics based on their socio-economic background, excluding academics who are children of professors. We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. The horizontal line represents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

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

(a) With Regional Variation

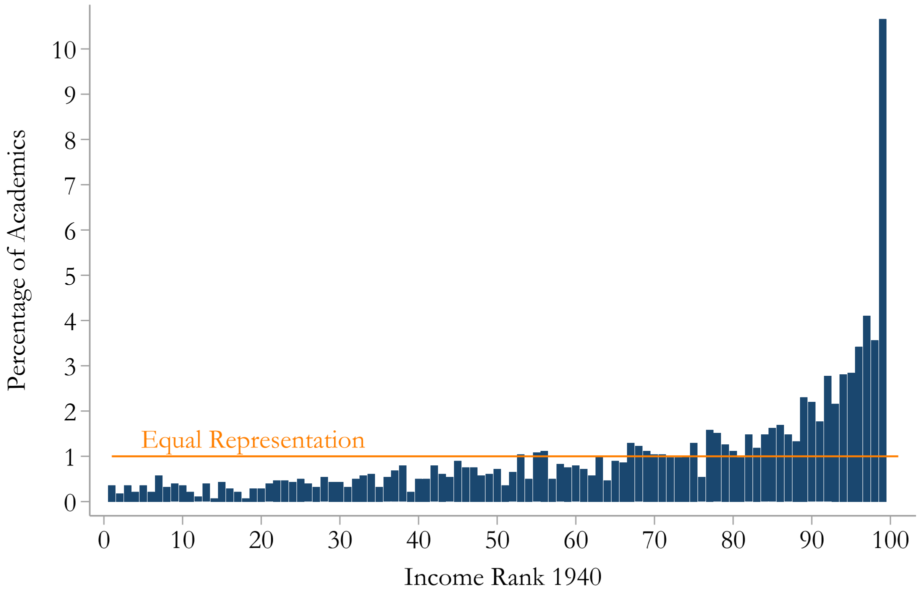


(b) Without Regional Variation



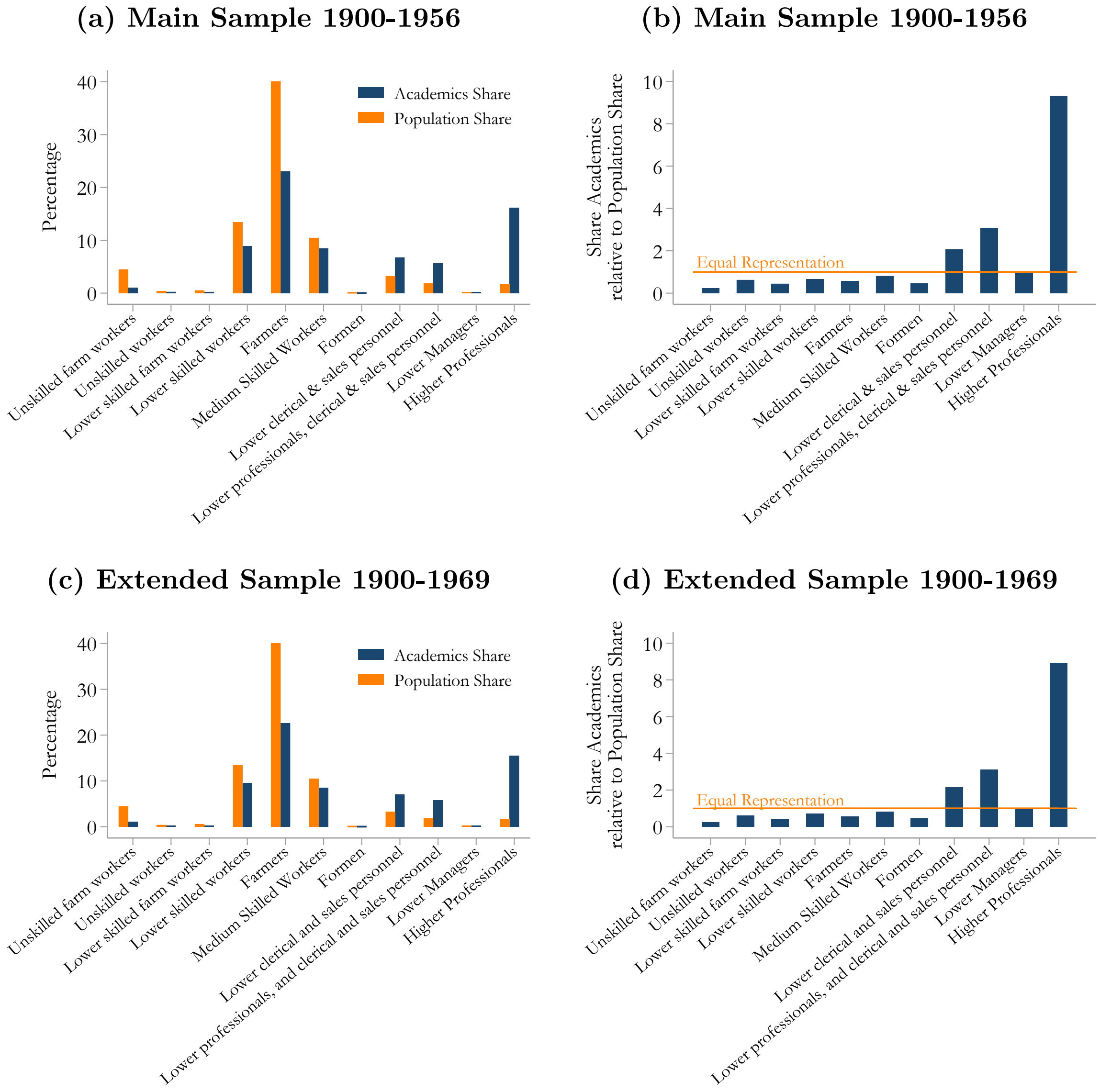
Notes: The figure shows the representation of academics based on their socio-economic background for the extended sample (1900-1969). We proxy socio-economic background with the father's income rank based on predicted income as described in section 2.2. The horizontal line represents a hypothetical equal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

Figure B.3: Extended Sample 1900-1969: Representation by Socio-Economic Background Measured by Actual Income of the Father



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 the reported income in the 1940 census. The horizontal line represents a hypothetical equal representation benchmark.

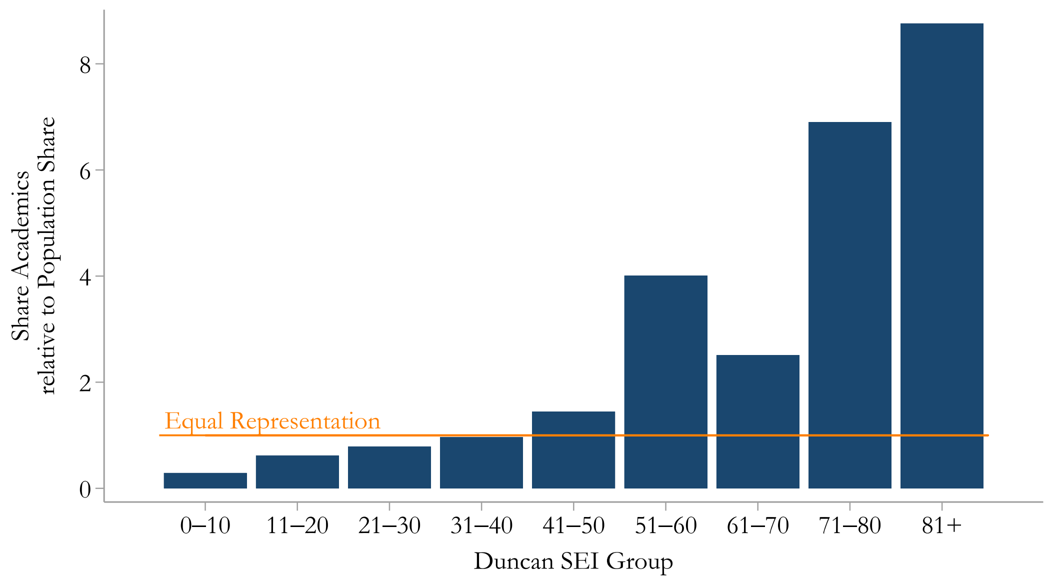
Figure B.4: Representation by Socio-Economic Background, Alternative Measures of SES: HISCLASS



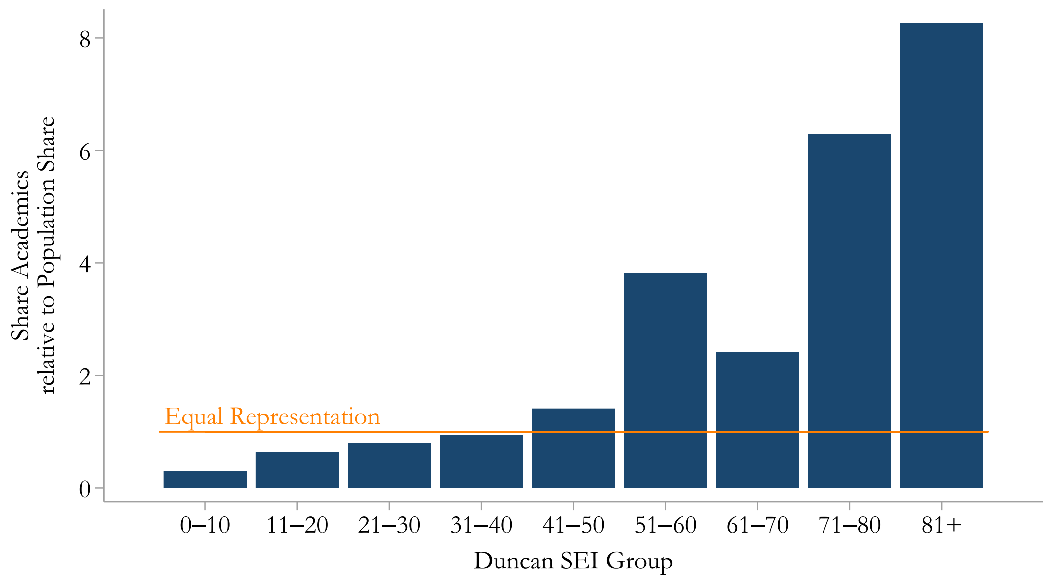
Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with HISCLASS, a measure of the social standing of a father's occupation (van Leeuwen and Maas, 2011). In panels a) and c), the orange bars indicate the share of individuals 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 benchmark.

Figure B.5: Representation by Socio-Economic Background, Alternative Measures of SES: Duncan Socioeconomic Index

(a) Main Sample 1900-1956



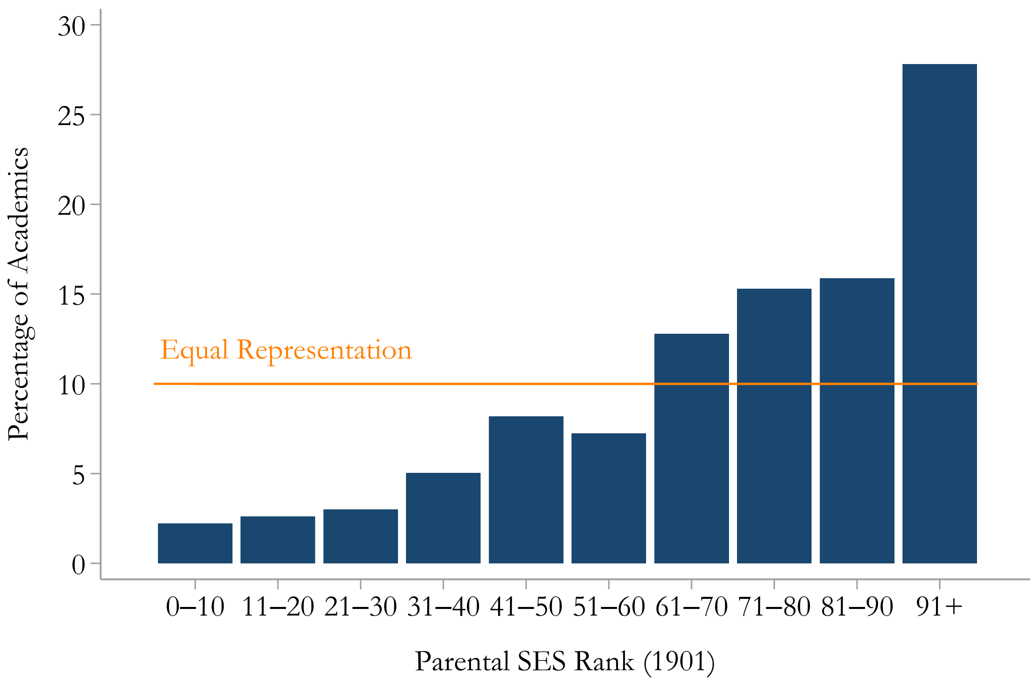
(b) Extended Sample 1900-1969



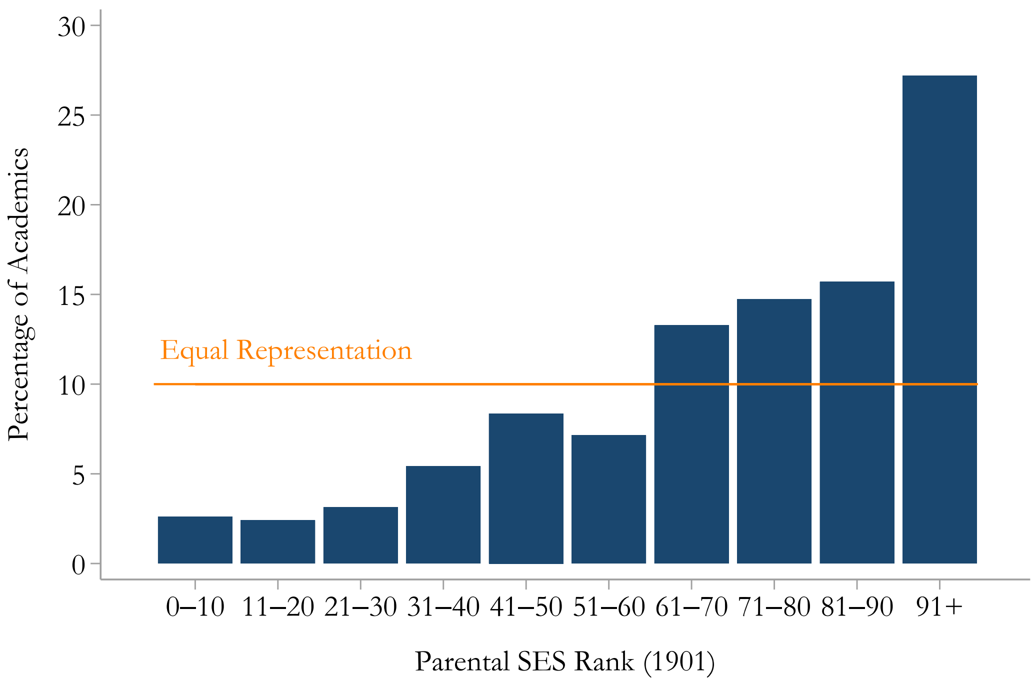
Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with the Duncan Socioeconomic Index (SEI), a measure of the social standing of a father's occupation. SEI reflects the income level and educational attainment of an occupation in 1950. For details, see IPUMS (2024b). SEI is an ordinal measure of occupational social status. The measure ranges from 0 to 96, but some values are missing. For example, SEI 89 does not exist in the census data of the relevant period. We group the measure into 9 categories. For example, the top category, 81+, contains SEI 81-87, 90, 92, 93 and 96. The horizontal line represents a hypothetical equal representation benchmark.

Figure B.6: Representation by Socio-Economic Background, Alternative Measures of SES: 1901 Cost of Living Survey

(a) Main Sample 1900-1956



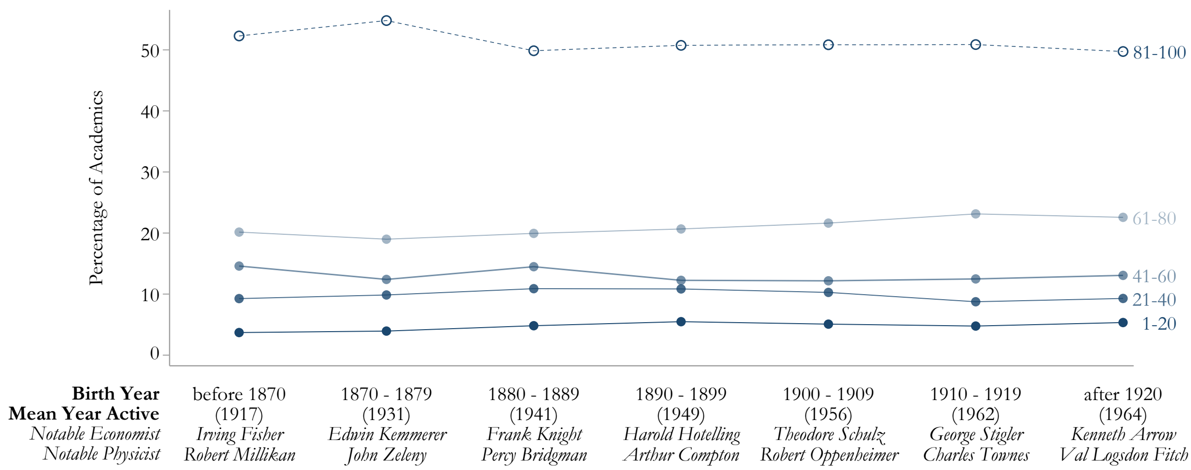
(b) Extended Sample 1900-1969



Notes: The figure shows the representation of academics based on their socio-economic background. We proxy socio-economic background with income measures from the 1901 cost of living survey. The horizontal line represents a hypothetical equal representation benchmark.

Representation Over Time

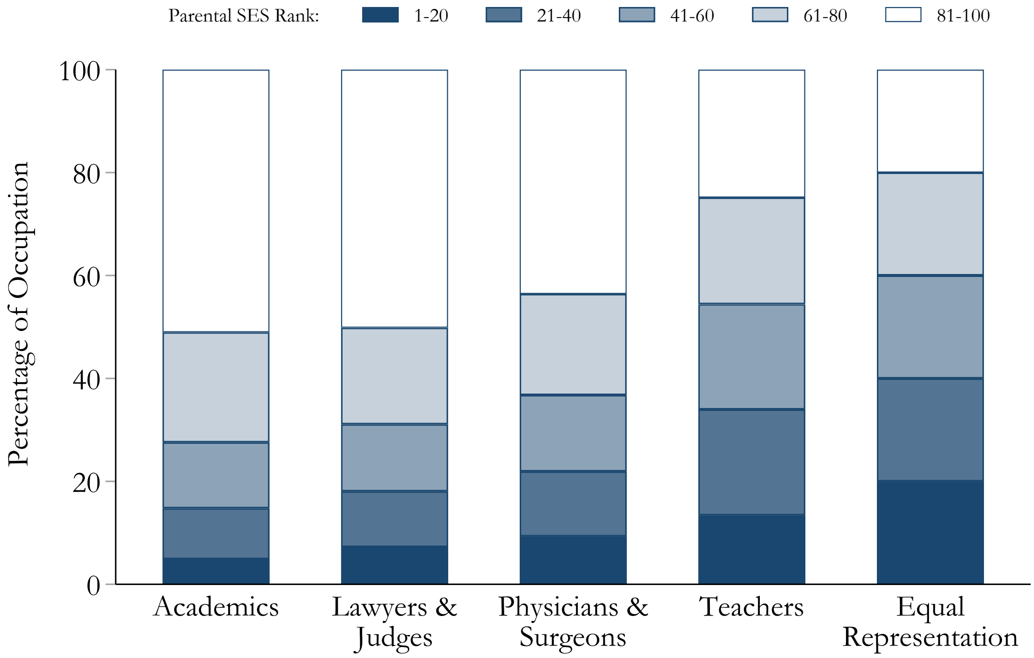
Figure B.7: Extended Sample 1900-1969: Representation by Socio-Economic Background Over Time



Notes: The figure shows the representation of academics based on their socio-economic background over time for the extended sample. 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 quintile of the predicted income rank distribution.

Representation in Academia versus Other Professions

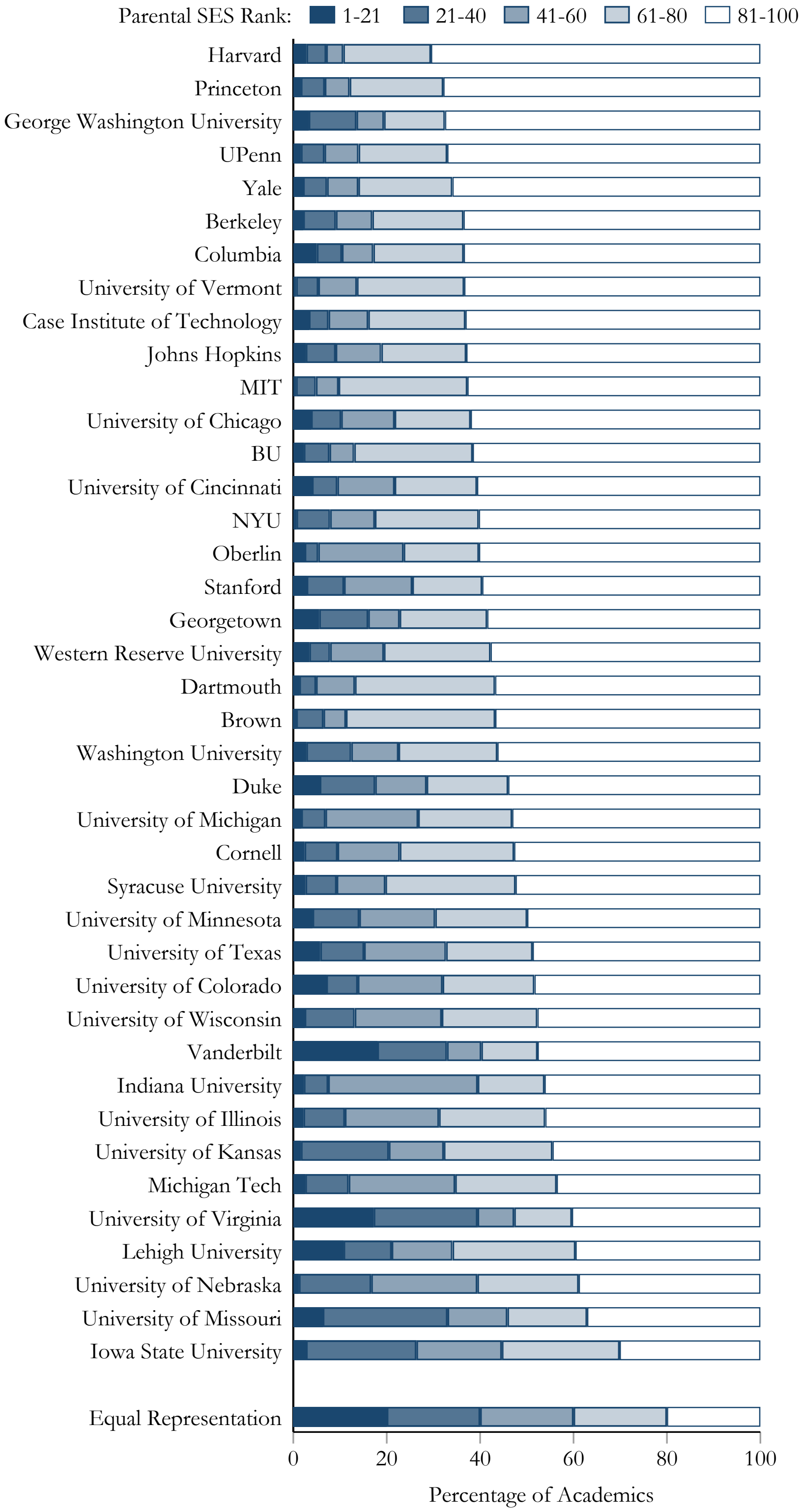
Figure B.8: Extended Sample 1900-1969: Comparison to other Professions



Notes: The figure compares the representation of academics based on their socio-economic background to representation in other professions for the extended sample. 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).

Representation by University: Additional Results

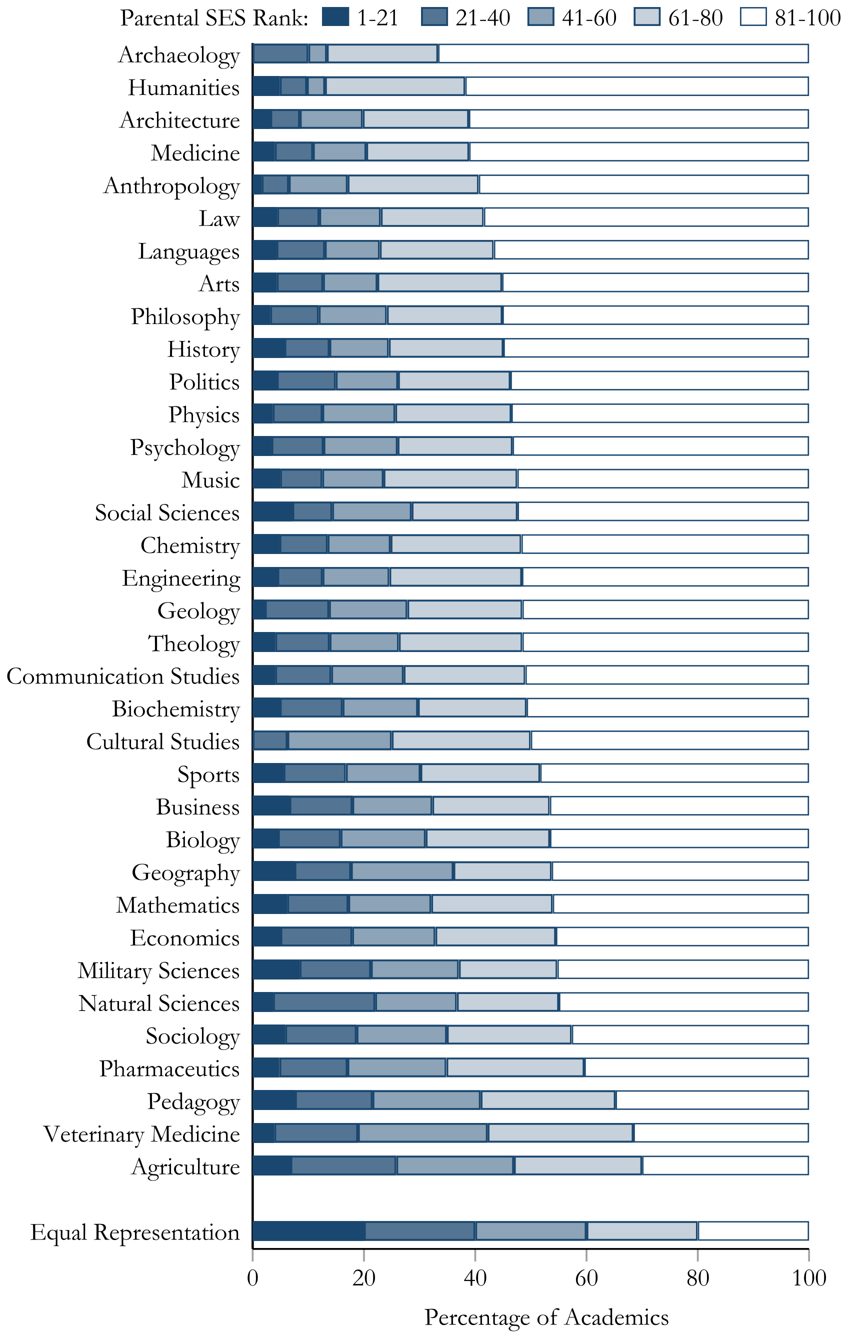
Figure B.9: Extended Sample 1900 - 1969: Selection by University



Notes: The figure shows the representation of academics based on their socio-economic background by university for the extended sample. 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.

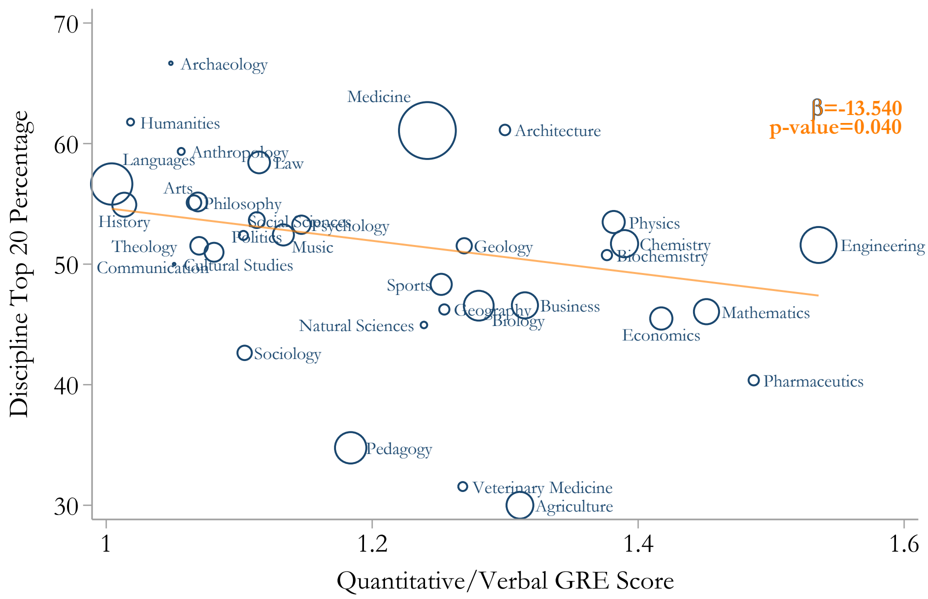
Representation by Discipline: Additional Results

Figure B.10: Extended Sample 1900 - 1969: Representation by Discipline



Notes: The figure shows the representation of academics based on their socio-economic background by academic discipline for the extended sample. 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.

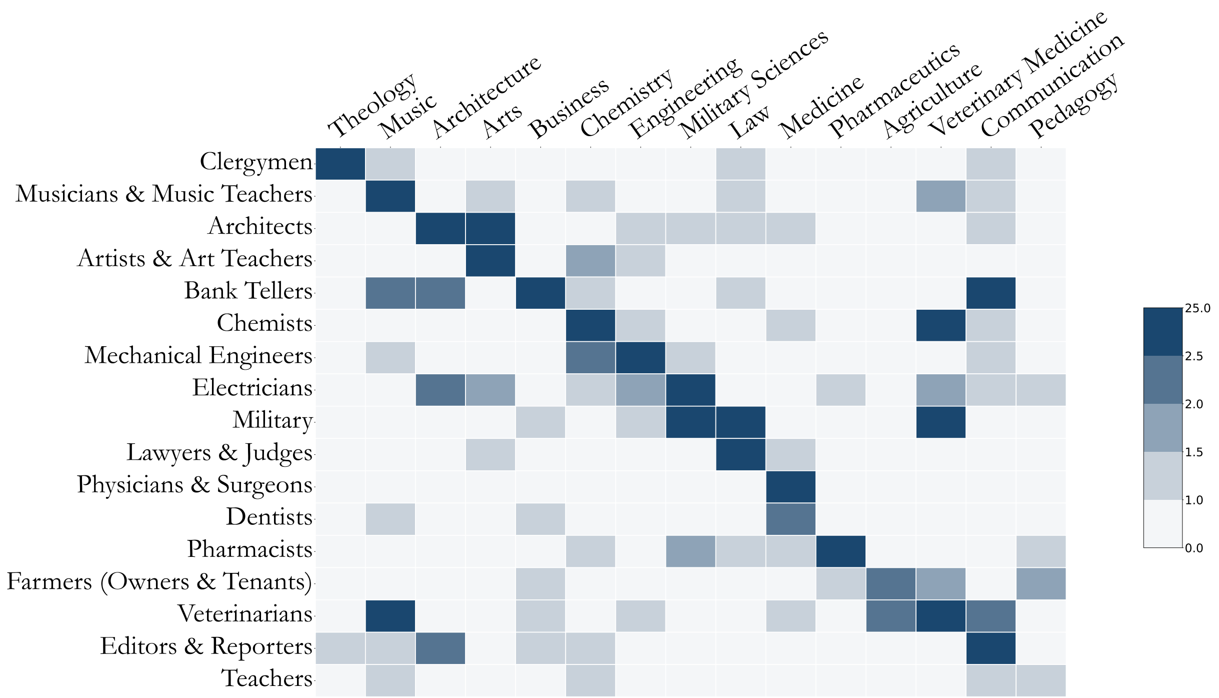
Figure B.11: Extended Sample 1900 - 1969: Discipline Mathematics vs. Language Requirements and Representation



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.

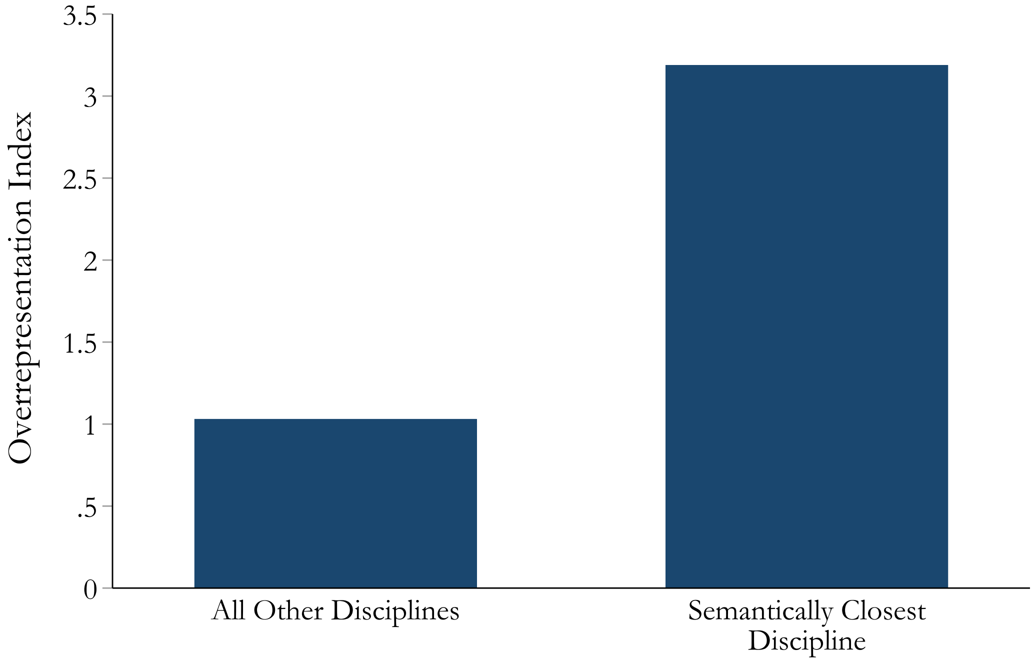
C Socio-Economic Background and Discipline Choice: Additional Results

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



Notes: The figure shows the relationship between father's occupation (rows) and their children's choice of academic discipline when the children become professors (columns) for selected father's occupation - discipline pairs for the extended sample. Darker shades indicate higher levels of overrepresentation, as measured by equation (4).

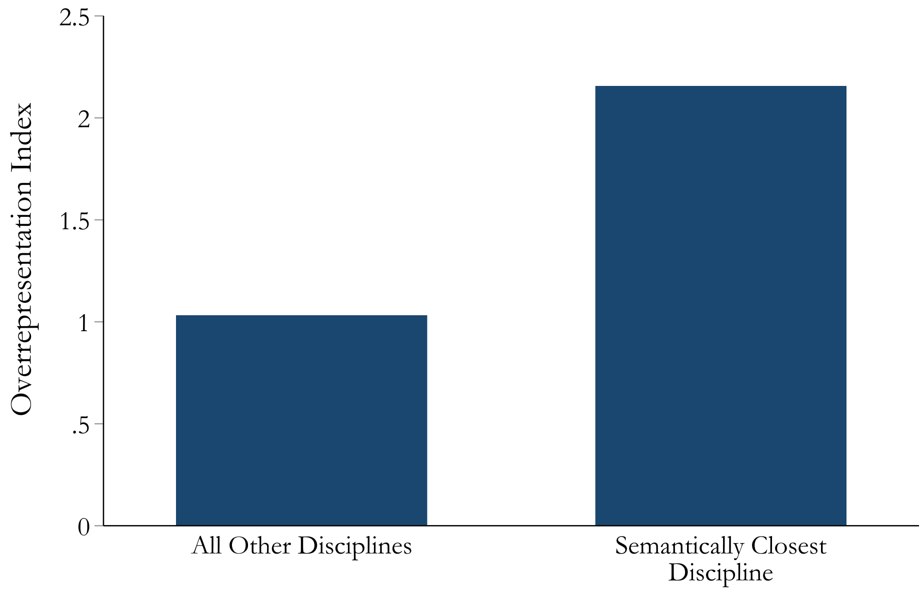
Figure C.2: Extended Sample 1900-1969: Overrepresentation in Semantically Closest Discipline



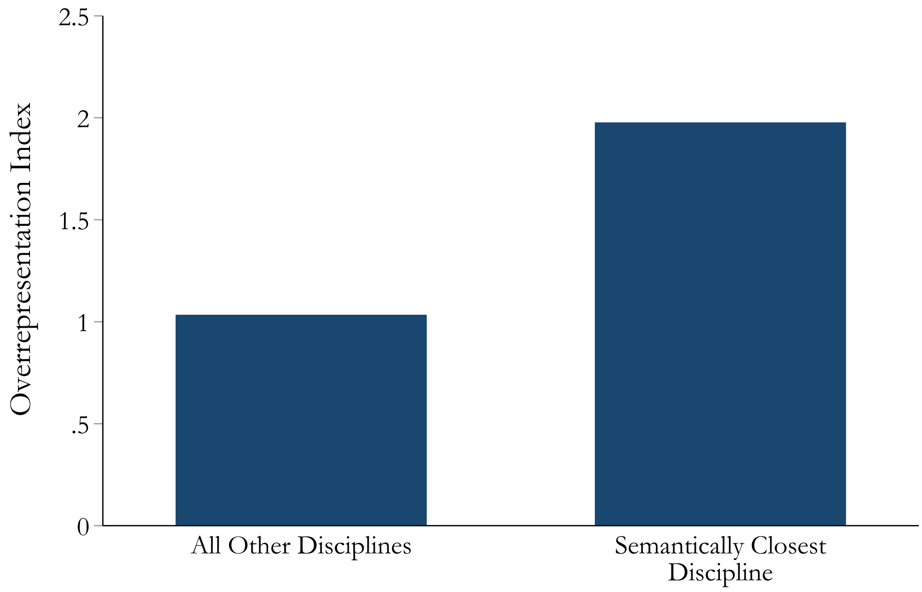
Notes: The figure shows overrepresentation as measured by equation (4) in the father's occupation-discipline pair that is semantically closest, e.g., "farmer" and "agriculture" and all other father's occupation - discipline pairs for the extended sample. For more details, see section 4.2 and section 4.3.

Figure C.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 4 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 the 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.

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

Table D.1: The Distribution of Publications Across Cohorts and Disciplines

		Publication Percentiles					
Discipline	Cohort	*50th*	*70th*	*90th*	*95th*	*97th*	*99th*
Biochemistry							
	1900	1	6	6	6	6	6
	1914	8	13	44	54	58	58
	1925	3.8	9	18	28	30	40
	1938	3	5.5	15.5	22.5	25	57
	1956	4	10	21.5	30	36	50
	1969	5	11	30	41	52	70
Biology							
	1900	0	1	4	7	10	26
	1914	1	2.5	9	13	18	25
	1925	0	2	7	11	13	19
	1938	0	2	6	10	13	20
	1956	1	2	8	11	15	21
	1969	1	4	12	18	22.5	33
Chemistry							
	1900	0	1	6	11	15	58
	1914	1	3	13	19.3	24.5	50.5
	1925	1	3	13	23	27	54
	1938	1	4	16	24	31	63
	1956	1.5	6	21	33	42	64
	1969	2	6	24	39	51	76
Mathematics							
	1900	0	0	3	6	6	13
	1914	0	0	5	8	11	17
	1925	0	0	2	6.5	9	19
	1938	0	0	4	8	12	18.5
	1956	0	0	5	8	11	17
	1969	0	2	9	13	16	24
Medicine							
	1900	0	1	6	9	12	18
	1914	1	3	11	16	21	32
	1925	1	4	13	21	25	42.5
	1938	1	5	15	22	28	44
	1956	2.5	7	20	31	40	59
	1969	3	9	26	40	52	86.9
Physics							
	1900	0	1	7	15	19	37
	1914	1	3	10	12	19	32
	1925	0	2	8	17	24	40
	1938	1	3	10	16	19	30
	1956	1	5	14	21	26	38
	1969	3	9	21	31	39	58

Notes: The table shows the number of publications that place a scientist in a certain percentile by cohort and discipline. Publications do not need to be integers because we assign publications proportionally to scientists who have the same last name, initial(s), discipline, and geographic identifier (e.g., city). E.g., there may be multiple biologists with the name J. Smith in Cambridge.

Table D.2: 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	-0.00042**	0.00009	0.00036**	0.00000	0.00006	-0.00002	-0.00007*
	(0.00019)	(0.00014)	(0.00015)	(0.00008)	(0.00005)	(0.00006)	(0.00004)
R^2	0.09	0.03	0.03	0.02	0.02	0.02	0.01
Observations	12,767	12,767	12,767	12,767	12,767	12,767	12,767
Dependent Variable Mean	0.586	0.141	0.180	0.044	0.020	0.019	0.010
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	15,521	15,521	15,521	15,521	15,521	15,521	15,521
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 eq. (7). The dependent variable is an indicator whether an academic's 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 D.3: 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	-0.00083*	-0.00098*	-0.00096*	-0.00063	-0.00081*	-0.00080*
	(0.00050)	(0.00050)	(0.00050)	(0.00043)	(0.00043)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.330	0.330	0.330	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00071	-0.00081*	-0.00082*	-0.00062*	-0.00075**	-0.00076**
	(0.00045)	(0.00045)	(0.00045)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.326	0.326	0.326	-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 (8). In columns (1)-(3), the dependent variable measures the number of publications that 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. For the results reported in columns (4)-(6), 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 D.4: Socio-Economic Background and Novelty: Excluding the 20,000 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	-0.00086*	-0.00101**	-0.00099**	-0.00068	-0.00085**	-0.00085*
	(0.00050)	(0.00050)	(0.00050)	(0.00043)	(0.00043)	(0.00044)
R^2	0.01	0.02	0.05	0.01	0.02	0.03
Observations	11,972	11,972	11,972	11,972	11,972	11,972
Dependent Variable Mean	0.325	0.325	0.325	-0.003	-0.003	-0.003
Panel B: 1914 – 1969						
Parental SES Rank	-0.00074*	-0.00083*	-0.00084*	-0.00066*	-0.00079**	-0.00081**
	(0.00045)	(0.00044)	(0.00045)	(0.00037)	(0.00038)	(0.00038)
R^2	0.01	0.02	0.05	0.01	0.02	0.02
Observations	14,726	14,726	14,726	14,726	14,726	14,726
Dependent Variable Mean	0.322	0.322	0.322	-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 (8). In columns (1)-(3), the dependent variable measures the number of publications that introduce at least one novel word and were published in a ± 5-year window around the cohort when scientist i enters the faculty rosters. For the results presented in columns (4)-(6), we standardize the novel words measure at the level of disciplines and cohorts. We exclude the 20,000 most common words. The main explanatory variable is the SES rank of the father, as measured by the percentile in the predicted income distribution of scientist 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 D.5: Socio-Economic Background and Nobel Prize Nominations/Awards (Non-Parametric Control for Publications and Citations)

Dependent Variable:	*Nobel Nomination*		*Nobel Wins*	
	(1)	(2)	(3)	(4)
Panel A: 1900 – 1956				
Parental SES Rank	0.00012***	0.00011***	0.00003*	0.00003
	(0.00004)	(0.00004)	(0.00002)	(0.00002)
R^2	0.10	0.11	0.04	0.05
Observations	12,767	12,767	12,767	12,767
Dependent Variable Mean	0.012	0.012	0.003	0.003
Panel B: 1900 – 1969				
Parental SES Rank	0.00010***	0.00010***	0.00003**	0.00003**
	(0.00003)	(0.00003)	(0.00002)	(0.00002)
R^2	0.08	0.10	0.03	0.04
Observations	15,521	15,521	15,521	15,521
Dependent Variable Mean	0.011	0.011	0.002	0.002
Demographic Controls	Yes	Yes	Yes	Yes
Publication & Citation Controls	Yes		Yes	
Publication & Citation Controls (Non-parametric)		Yes		Yes
Childhood State FEs	Yes	Yes	Yes	Yes
Cohort FEs	Yes	Yes	Yes	Yes
Uni State FEs	Yes	Yes	Yes	Yes
Discipline FEs	Yes	Yes	Yes	Yes

Notes: The table reports the estimates of equation (10). In columns (1) and (2), the dependent variable is an indicator whether scientist i was ever nominated for a Nobel Prize. In columns (3) and (4), the dependent variable is an indicator whether scientist i was ever awarded 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 scientist i's father. Demographic controls include age, age squared, and an indicator for whether the scientist is female. In columns (1) and (3), publication and citation controls are scientist i's standardized 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. In columns (2) and (4), we control non-parametrically for publications by including a set of indicator variables that equal one if a scientist i's publication record falls into a certain percentile of the publication distribution. Similarly, we add indicators for the percentile of the citation record. Standard errors are clustered at the level of father's occupation, childhood state, and birth year of the scientist. Significance levels: $***$ p<0.01, $**$ p<0.05, and $*$ p<0.1.