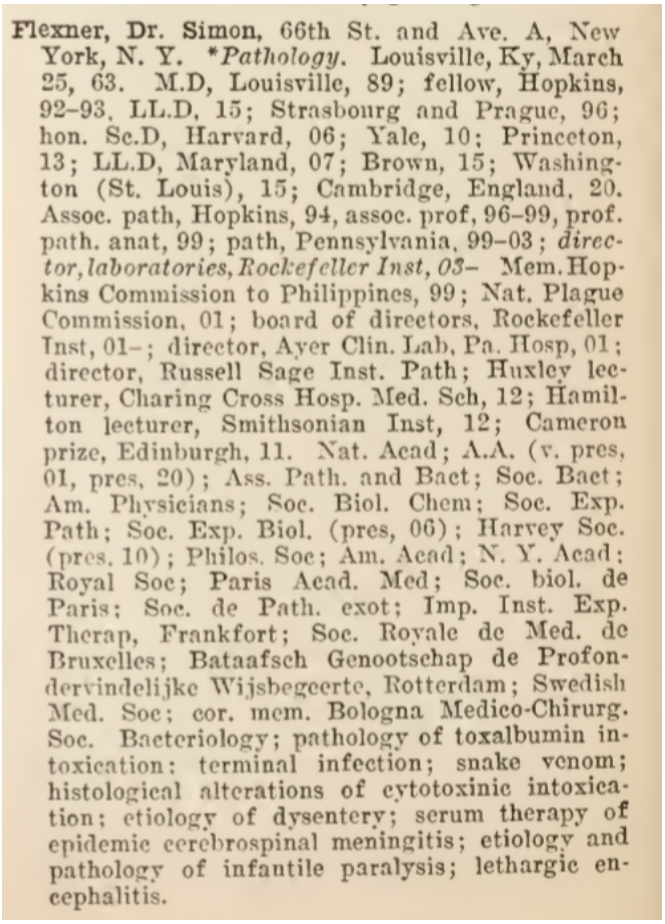
ONLINE APPENDIX
NOT FOR PUBLICATION

APPENDIX A – SUPPLEMENTARY MATERIALS AND TESTS

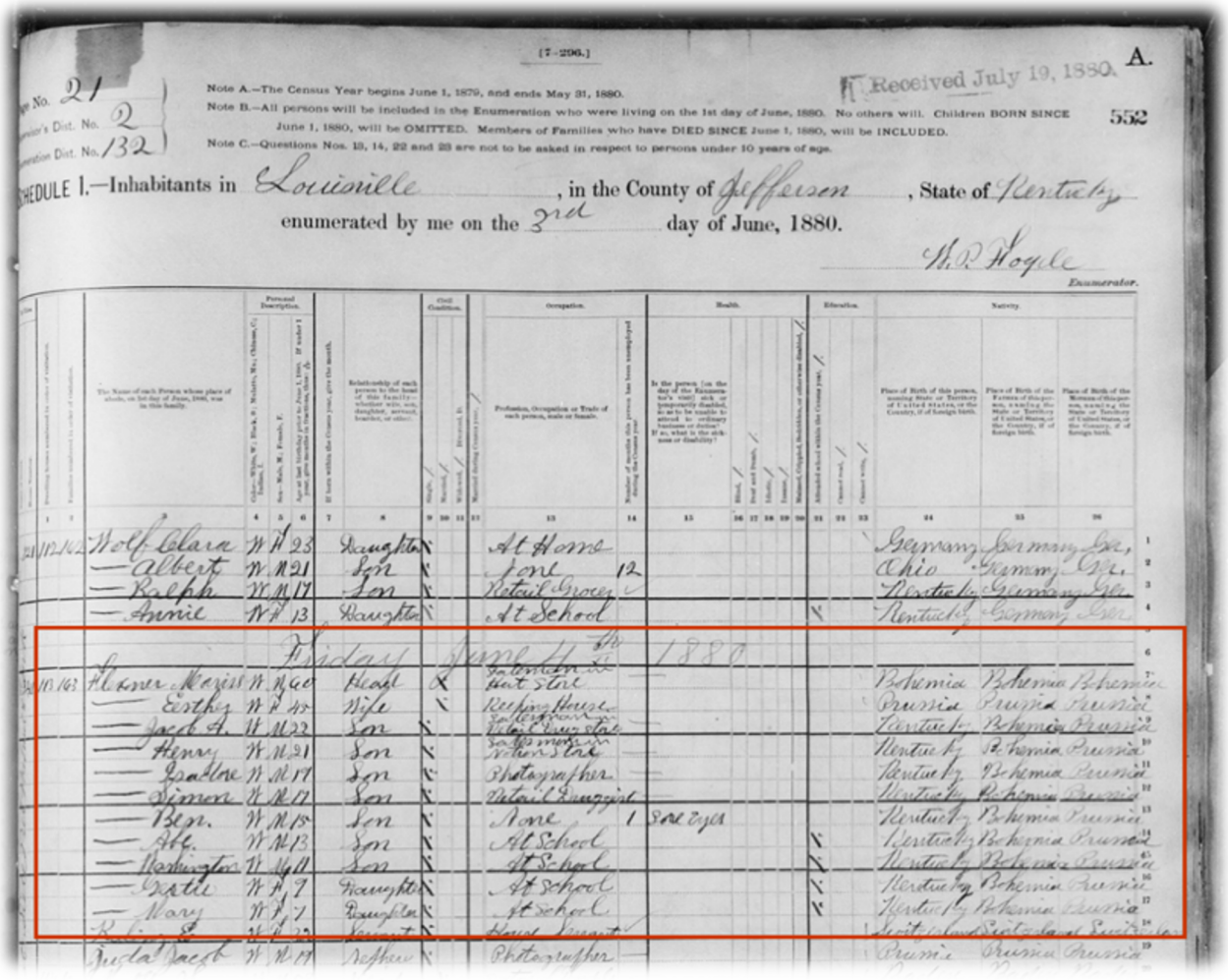
FIGURE A1 - SIMON FLEXNER IN THE MOS (1921)



Flexner, Dr. Simon, 66th St. and Ave. A, New York, N. Y. **Pathology.* Louisville, Ky, March 25, 63. M.D, Louisville, 89; fellow, Hopkins, 92–93, LL.D, 15; Strasbourg and Prague, 96; hon. Sc.D, Harvard, 06; Yale, 10; Princeton, 13; LL.D, Maryland, 07; Brown, 15; Washington (St. Louis), 15; Cambridge, England, 20. Assoc. path, Hopkins, 94, assoc. prof, 96–99, prof. path. anat, 99; path, Pennsylvania, 99–03; *director, laboratories, Rockefeller Inst, 03–* Mem. Hopkins Commission to Philippines, 99; Nat. Plague Commission, 01; board of directors, Rockefeller Inst, 01–; director, Ayer Clin. Lab, Pa. Hosp, 01; director, Russell Sage Inst. Path; Huxley lecturer, Charing Cross Hosp. Med. Sch, 12; Hamilton lecturer, Smithsonian Inst, 12; Cameron prize, Edinburgh, 11. Nat. Acad; A.A. (v. pres, 01, pres, 20); Ass. Path. and Bact; Soc. Bact; Am. Physicians; Soc. Biol. Chem; Soc. Exp. Path; Soc. Exp. Biol. (pres, 06); Harvey Soc. (pres. 10); Philos. Soc; Am. Acad; N. Y. Acad; Royal Soc; Paris Acad. Med; Soc. biol. de Paris; Soc. de Path. exot; Imp. Inst. Exp. Therap, Frankfort; Soc. Royale de Med. de Bruxelles; Bataafsch Genootschap de Profondervindelijke Wijsbegeerte, Rotterdam; Swedish Med. Soc; cor. mem. Bologna Medico-Chirurg. Soc. Bacteriology; pathology of toxalbumin intoxication; terminal infection; snake venom; histological alterations of cytotoxinic intoxication; etiology of dysentery; serum therapy of epidemic cerebrospinal meningitis; etiology and pathology of infantile paralysis; lethargic encephalitis.

Notes: The entry for Dr. Simon Flexner in the *American Men of Science* (MoS, Cattell 1921). Flexner was an experimental pathologist at Penn, whose achievements include early analyses of polio and the development of a serum treatment for meningitis.

FIGURE A2 - SIMON FLEXNER AS A 17-YEAR-OLD IN THE CENSUS OF 1880



Notes: This image (from FamilySearch.org) shows the entry for Simon Flexner in the US Census of 1880. In addition to Simon Flexner, it includes the entry for Simon's younger brother Abraham ("Abe.") Flexner, who would become the author of the 1910 Flexner Report that ushered in a reform of US medical training.

TABLE A1 – COMPARING TRAITS OF CENSUS-MATCHED AND OTHER SCIENTISTS

	1870 Census			1880 Census			1900 Census		
	Match	Other	p-value	Match	Other	p-value	Match	Other	p-value
Panel A: All scientists									
Age	8.14	9.29	0.000	11.51	12.59	0.000	25.04	25.64	0.000
	(0.19)	(0.23)		(0.16)	(0.22)		(0.17)	(0.20)	
Elite undergrad	0.28	0.23	0.016	0.27	0.24	0.148	0.26	0.24	0.764
	(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)	
Elite graduate	0.36	0.34	0.252	0.39	0.39	0.522	0.41	0.42	0.885
	(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)	
1+ pubs.	0.71	0.68	0.112	0.73	0.69	0.000	0.71	0.69	0.012
	(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)	
Pre-1921 pubs.	12.98	11.69	0.584	12.90	9.90	0.003	9.73	9.76	0.002
	(0.54)	(0.62)		(0.44)	(0.39)		(0.25)	(0.37)	
Citations	38.67	37.59	0.725	48.12	35.26	0.049	41.61	38.77	0.458
	(3.27)	(4.12)		(2.84)	(2.84)		(1.93)	(2.44)	
Panel B: Scientists whom we observe as children in the census									
Age	6.36	6.57	0.315	8.04	8.00	0.812	13.03	13.10	0.647
	(0.14)	(0.16)		(0.10)	(0.15)		(0.08)	(0.12)	
Elite undergrad	0.29	0.25	0.060	0.28	0.27	0.303	0.24	0.25	0.942
	(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)	
Elite graduate	0.37	0.35	0.418	0.42	0.41	0.762	0.46	0.43	0.101
	(0.01)	(0.02)		(0.01)	(0.01)		(0.01)	(0.02)	
1+ pubs	0.76	0.76	0.735	0.80	0.74	0.000	0.80	0.79	0.581
	(0.01)	(0.01)		(0.01)	(0.01)		(0.01)	(0.01)	
Pre-1921 pubs	13.15	12.79	0.698	12.77	10.94	0.017	6.02	6.60	0.213
	(0.58)	(0.72)		(0.49)	(0.50)		(0.25)	(0.42)	
Citations	40.86	44.24	0.584	51.77	41.59	0.064	36.15	36.78	0.890
	(3.60)	(5.26)		(3.44)	(3.89)		(2.41)	(4.28)	

Notes: Mean and standard deviation of the characteristics of matched and unmatched scientists, by census wave. *Pre-1921 pubs* counts a scientist's publications before 1921, when the MoS (1921) was published; *citations* are the count of citations of these publications. *Elite undergrad* and *elite graduate* are indicators for scientists who hold a degree from an *Ivy-plus* university (Brown, Columbia, Cornell, Chicago, Dartmouth, Duke, Harvard, MIT, Pennsylvania, Princeton, Stanford, and Yale, as in Chetty et al. 2019. We perform a two-tailed t-test of the difference in means for matched and unmatched scientists and report p-values for that test.

TABLE A2 - LOGIT OF THE ODDS OF STARDOM AS A FUNCTION OF CHILDHOOD SES (SERVANTS), PUBLICATIONS, AND OTHER TRAITS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Servants	0.564***	0.643***	0.511***	0.458***	0.510***	0.530***	0.511***	0.507***	0.636***
	(0.146)	(0.149)	(0.161)	(0.162)	(0.162)	(0.163)	(0.161)	(0.161)	(0.150)
Industry job		-1.040***	-1.047***	-1.032***	-0.949***	-1.043***	-1.047***	-1.051***	-1.056***
		(0.165)	(0.184)	(0.185)	(0.184)	(0.186)	(0.184)	(0.184)	(0.167)
Publications			0.130**	0.135**	0.141**	0.127**	0.130**	0.130**	
			(0.062)	(0.062)	(0.062)	(0.062)	(0.062)	(0.062)	
Citations			0.413***	0.406***	0.401***	0.416***	0.413***	0.414***	
			(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	(0.045)	
Elite undergrad				0.334***					
				(0.114)					
Elite grad					0.446***				
					(0.110)				
N siblings						-0.053			
						(0.036)			
First-born						-0.149			
						(0.128)			
Foreign-born parents							-0.011		
							(0.182)		
Illiterate mother								-0.595	
								(0.671)	
Patent									0.071
									(0.122)
Connect									0.580
									(0.470)
% Δ in odds	75.70%	90.30%	66.70%	58.12%	66.55%	69.86%	66.72%	65.97%	88.92%

Notes: The variable *industry job* is an indicator for scientists who report no academic positions in the MOS (1921). The variable *patent* indicates scientists with at least 1 patent by 1921; *connect* indicates scientists who are connected with an existing star by a joint patent before 1921. *Elite undergrad* and *elite grad* distinguish scientists who have completed a degree at an Ivy-Plus university (Chetty et al. 2019). *First-born* scientists is an indicator for scientists who are the oldest child in their household. *Foreign-born parents* indicates scientists with at least one foreign-born parent. An *illiterate mother* can neither read nor write. All estimates include age, discipline, and census year FE; robust standard errors in parentheses.

TABLE A3 – DIFFERENCES IN THE ODDS OF STARDOM FOR ACADEMIC VS. INDUSTRY SCIENTISTS

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
High SES	0.394***	0.377***	0.467	0.546*				
	(0.101)	(0.112)	(0.308)	(0.317)				
Servants					0.619***	0.428**	0.878**	0.993**
					(0.160)	(0.175)	(0.399)	(0.412)
Publications		0.149**		0.017		0.178***		-0.204
		(0.064)		(0.181)		(0.069)		(0.218)
Citations		0.442***		0.279**		0.417***		0.447***
		(0.046)		(0.137)		(0.049)		(0.168)
% Δ in odds	48.35%	45.86%	59.54%	72.68%	85.62%	53.45%	140.51%	169.96%
Sample	Academic	Academic	Industry	Industry	Academic	Academic	Industry	Industry
N	3,323	3,323	698	698	3,372	3,372	718	718

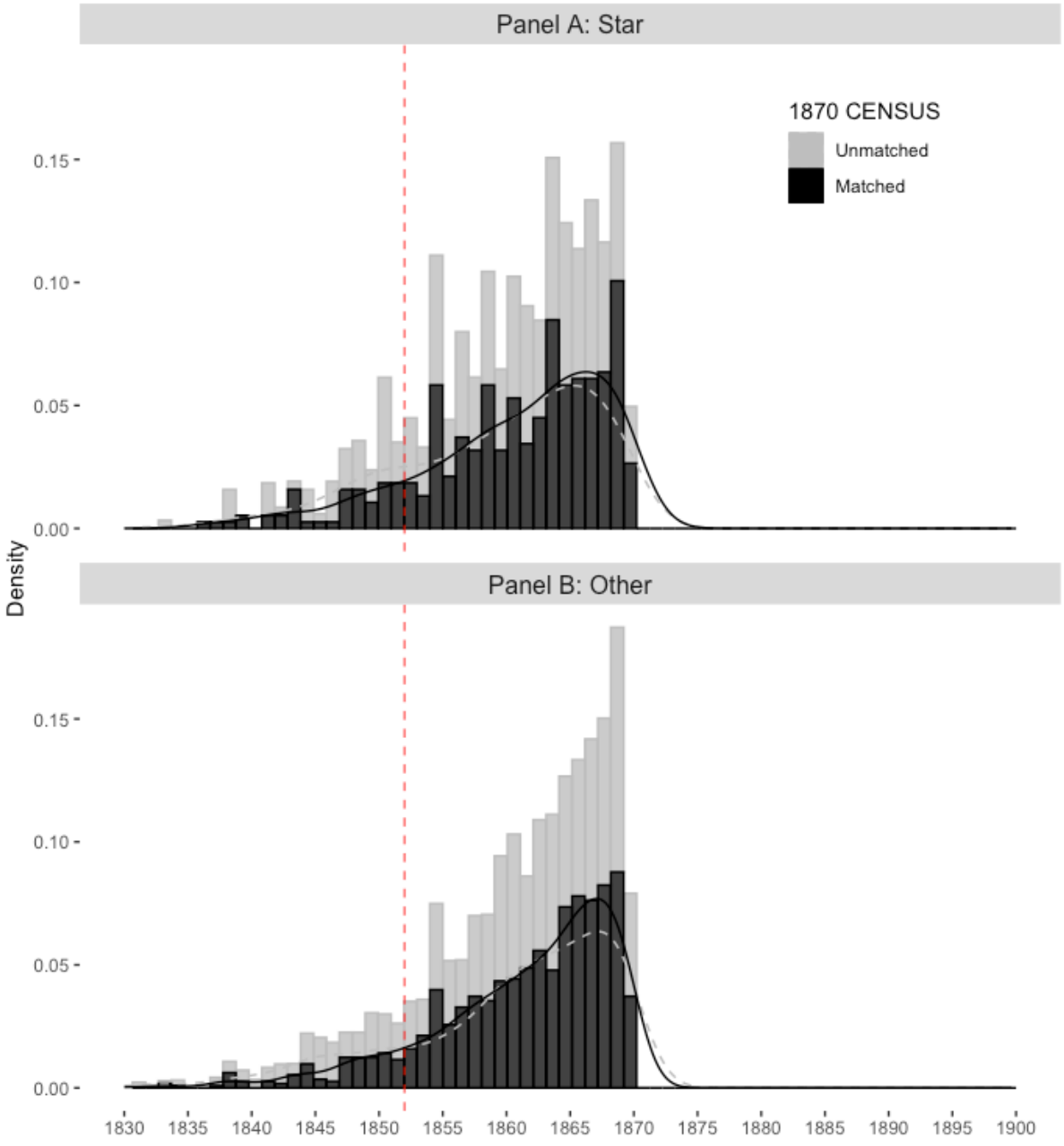
Notes: Differences in the odds of stardom for academic and industry scientists from a logistic regression of equation 1 (without controls for publications, odd columns) and equation 3 (controlling for publications, even columns). Academic scientists are those who held an academic position at least once, industry scientists are those who never worked in academia. *Publications* and *citations* are transformed using the inverse hyperbolic sine. All estimates include age, discipline, and census year FE; robust standard errors in parentheses.

TABLE A4 –LOGIT ESTIMATES OF THE ODDS OF BEING A STAR AS A FUNCTION OF FAMILY WEALTH

	Above median	95th percentile	Above 20k (97th percentile)
Personal estate	0.166	0.450	0.473*
	(0.132)	(0.300)	(0.262)
N	1,305	1,305	1,305
% change in odd	18.07%	56.85%	60.48%
Age and disciplines FE	Y	Y	Y

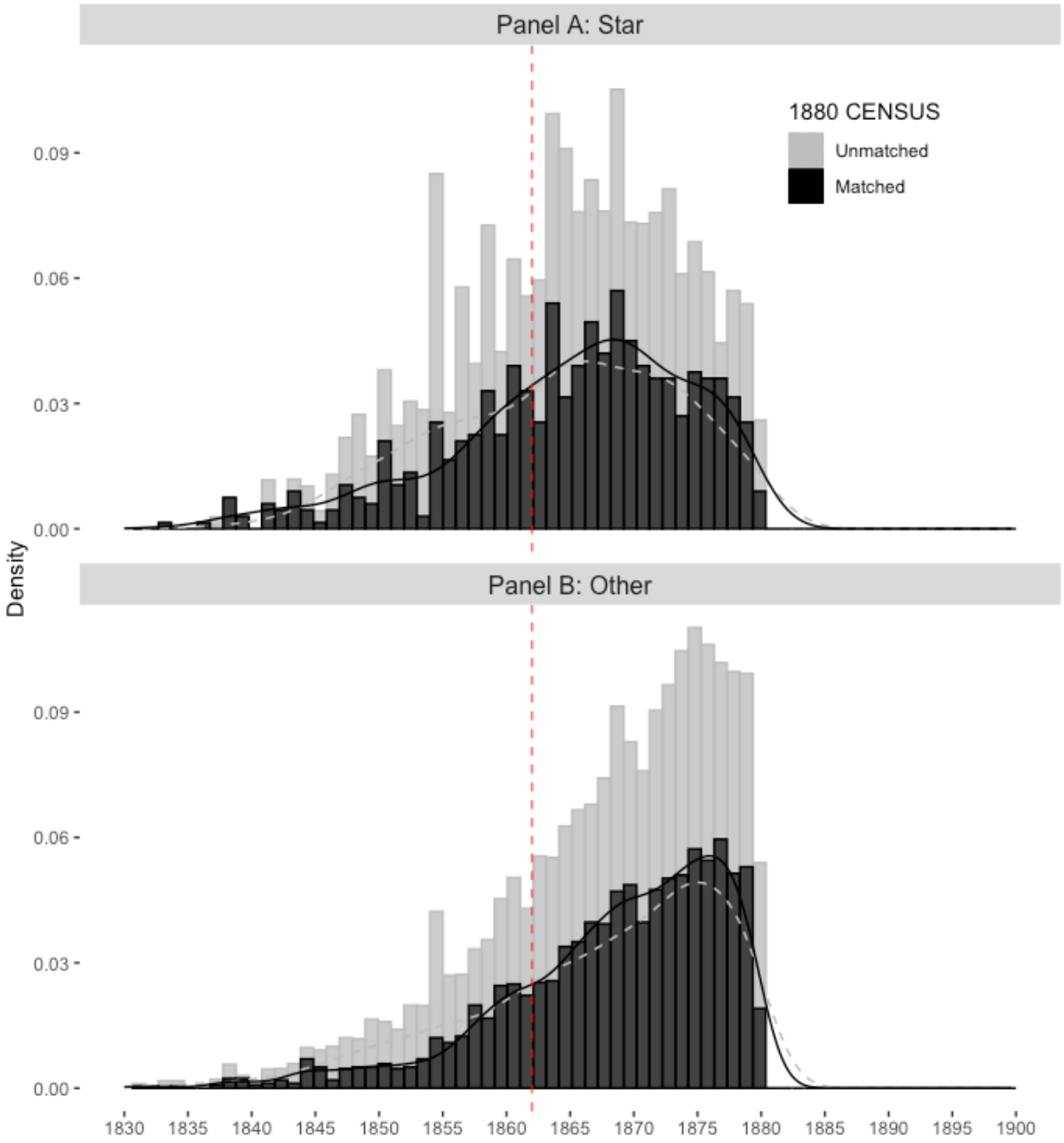
Notes: We re-estimate the odds of being star as a function of their father's personal estate for 1,305 scientists whom we observe as children in the 1870 census and whose fathers are farmers. Column 1 shows estimates for fathers whose personal estate was above the median ($1000); column 2 reports the 95th percentile, and column 3 shows the same for personal estates above $20,000 (roughly the 97th percentile).

FIGURE A3 – DISTRIBUTION OF STARS AND OTHER SCIENTISTS ACROSS BIRTH YEARS FOR SCIENTISTS MATCHED WITH THE 1870 CENSUS AND UNMATCHED SCIENTISTS



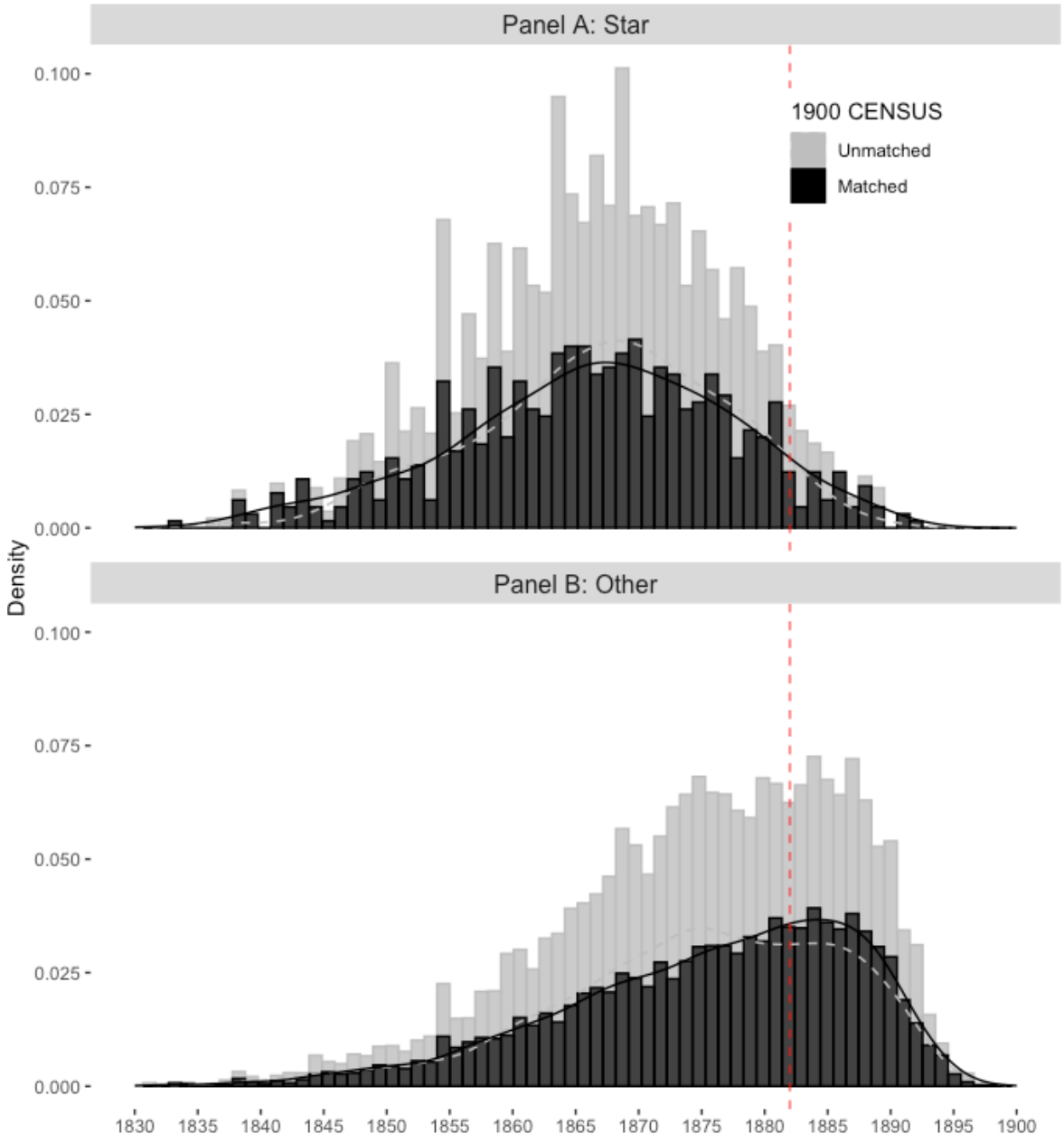
Notes: To investigate the quality of the census matching, we compare the distribution of stars (Panel A) and other scientists (Panel B) for scientists whom we have matched with the 1870 census to the same distributions for other, unmatched scientists. Data include 1,484 scientists whom we match with the census of 1870 and 1,174 unmatched US-born male scientists born before 1870. Scientists born after 1852 (to the right of the red dashed line) were minors in the 1870 census. 25% of the matched and 25% of the unmatched scientists are stars.

FIGURE A4 – DISTRIBUTION OF STARS AND OTHER SCIENTISTS ACROSS BIRTH YEARS FOR SCIENTISTS MATCHED WITH THE 1880 CENSUS AND UNMATCHED SCIENTISTS



Notes: To investigate the quality of the census matching, we compare the distribution of stars (Panel A) and other scientists (Panel B) for scientists whom we have matched with the 1880 census to the same distributions for other, unmatched scientists. Data include 3,189 scientists whom we match with the census of 1880 and 1,820 unmatched US-born male scientists born before 1880. Scientists born after 1862 (to the right of the red dashed line) were minors in 1880. 19% of the matched and 21% of the unmatched scientists are stars.

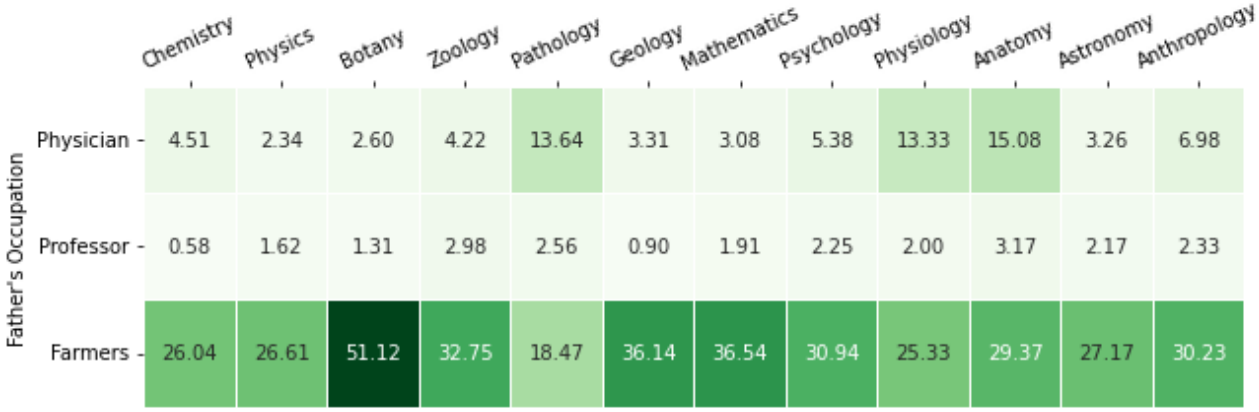
FIGURE A5 – DISTRIBUTION OF STARS AND OTHER SCIENTISTS ACROSS BIRTH YEARS FOR SCIENTISTS MATCHED WITH THE 1900 CENSUS AND UNMATCHED SCIENTISTS



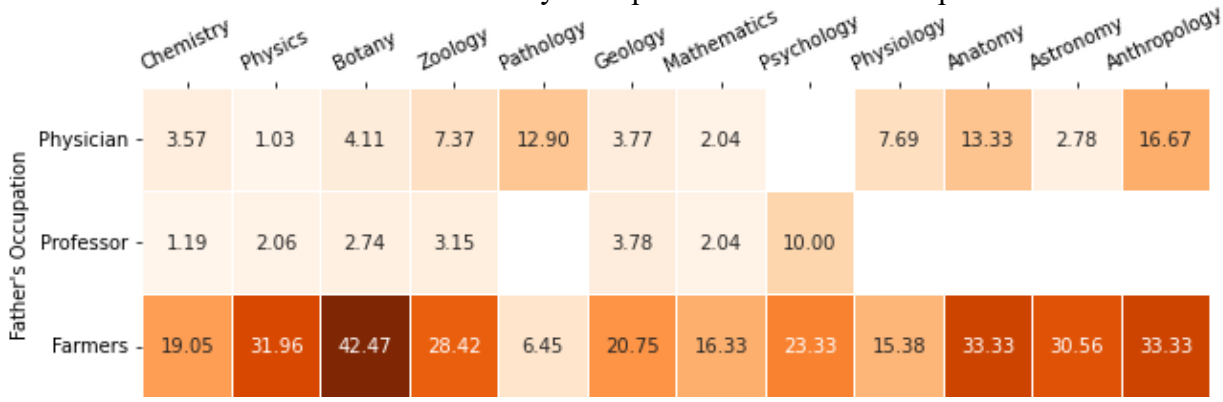
Notes: To investigate the quality of the census matching, we compare the distribution of stars (Panel A) and other scientists (Panel B) for scientists whom we have matched with the 1880 census to the same distributions for other, unmatched scientists. Data include 4,687 scientists whom we observe in the census of 1900 and 3,104 US-born male scientists born before 1900 whom we cannot match with the census. Scientists born after 1882 (to the right of the red dashed line) were minors in 1900. 15% of the matched and 16% of the unmatched scientists are stars.

FIGURE A6 – THE CHILDREN OF FARMERS AND PHYSICIANS ACROSS DISCIPLINES

Panel A: Share of Scientists by Discipline and Fathers' Occupation

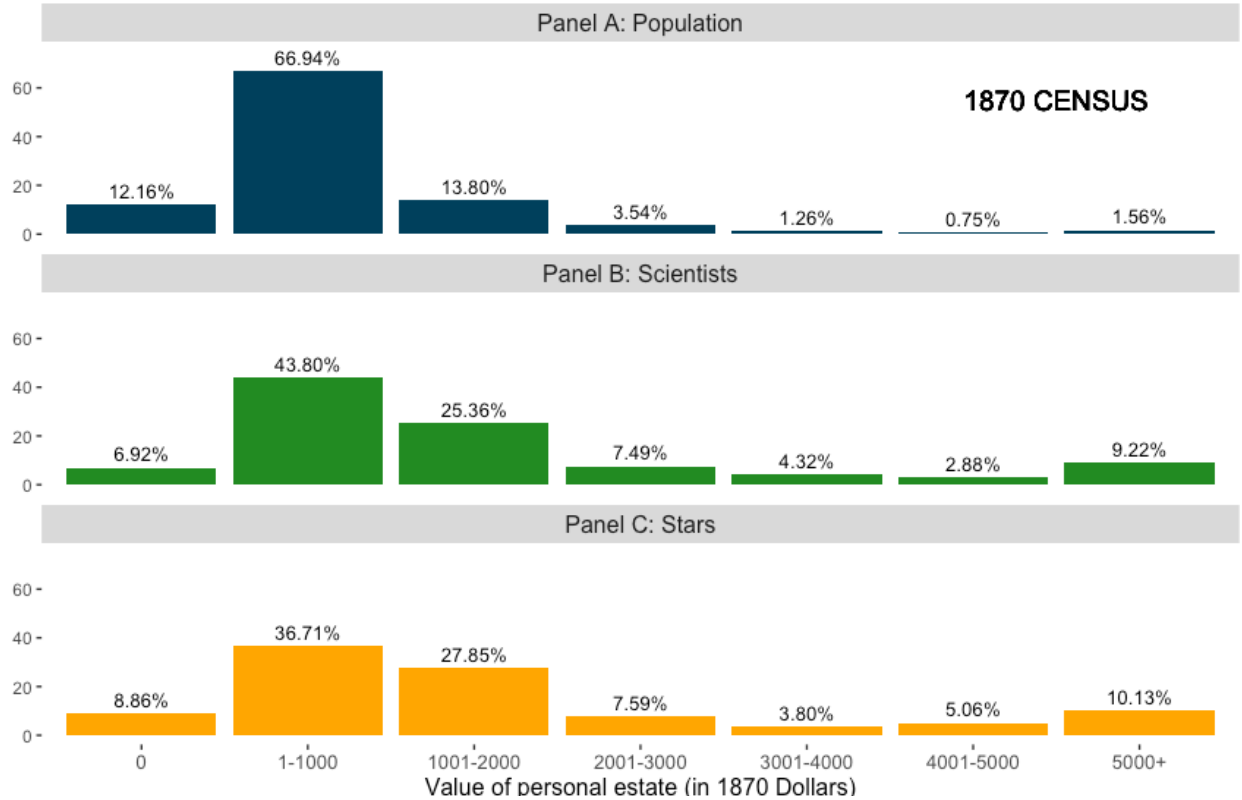


Panel B: Share of Stars by Discipline and Fathers' Occupation



Notes: To investigate whether boys are more likely to become scientists (Panel A) or stars (Panel B) in disciplines that are adjacent to their fathers' work, we plot the share of scientists (stars) in each discipline whose father was a physician, a professor, a farmer. Darker shades signify higher shares. Disciplines are arranged by their size (measured by the number of scientists in that discipline). Data include 4,067 scientists whom we match with at least one census wave when they are minors.

FIGURE A7 – DIFFERENCES IN PERSONAL WEALTH AMONG FARMERS

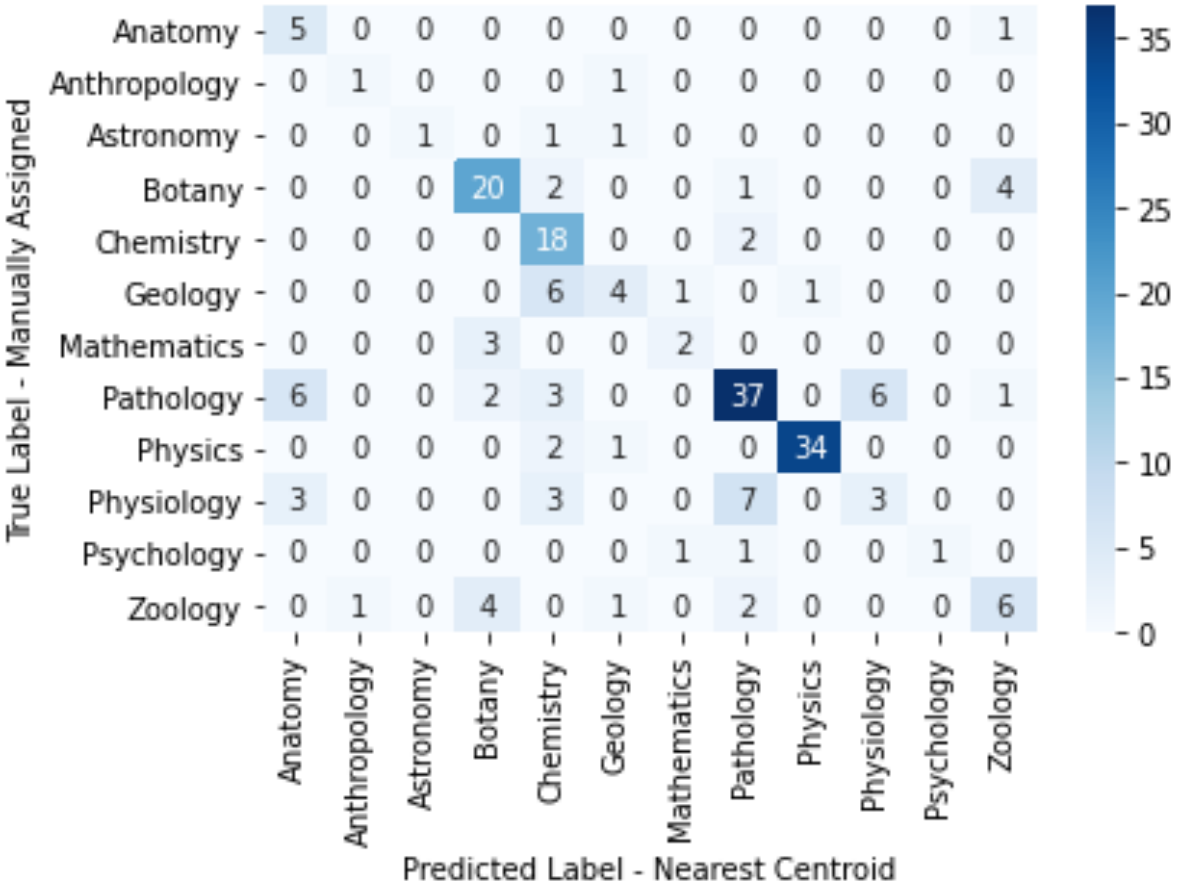


Notes: A large share of our scientists and the population were the children of farmers and farm laborer. Among scientists who were children in the census of 1870, for example, 31% were the children of farmers, and 29% of star scientists, compared with 46% of boys of the same age in the population (Figure 2). To investigate differences in SES within this large group of farmers, we collect data on personal wealth from the 1870 census. The personal wealth of farmers ranged between 0 and $999, 997 in 1870 (roughly than 23 million measured by its relative share in GDP per capita, MeasuringWorth.com 2024).

Appendix B: Investigating the Performance of the Nearest Centroid Algorithm

We check the precision of the nearest centroid algorithm by manually assigning 200 scientists to a discipline and constructing a confusion matrix to count true matches, as well as false positives and negatives. In this matrix, manually assigned disciplines are true label (rows), and disciplines predicted by nearest centroid algorithm are predicted labels (columns). Values on the diagonal report true matches. For example, 20 scientists are assigned to the discipline of botany manually and by the algorithm; these 20 entries are a true match. 131 of 200 predicted disciplines are a true match, and mismatches are predominantly in related disciplines. For instance, two scientists are assigned to botany manually but to chemistry by the algorithm and four are assigned to botany by the algorithm but to zoology manually.

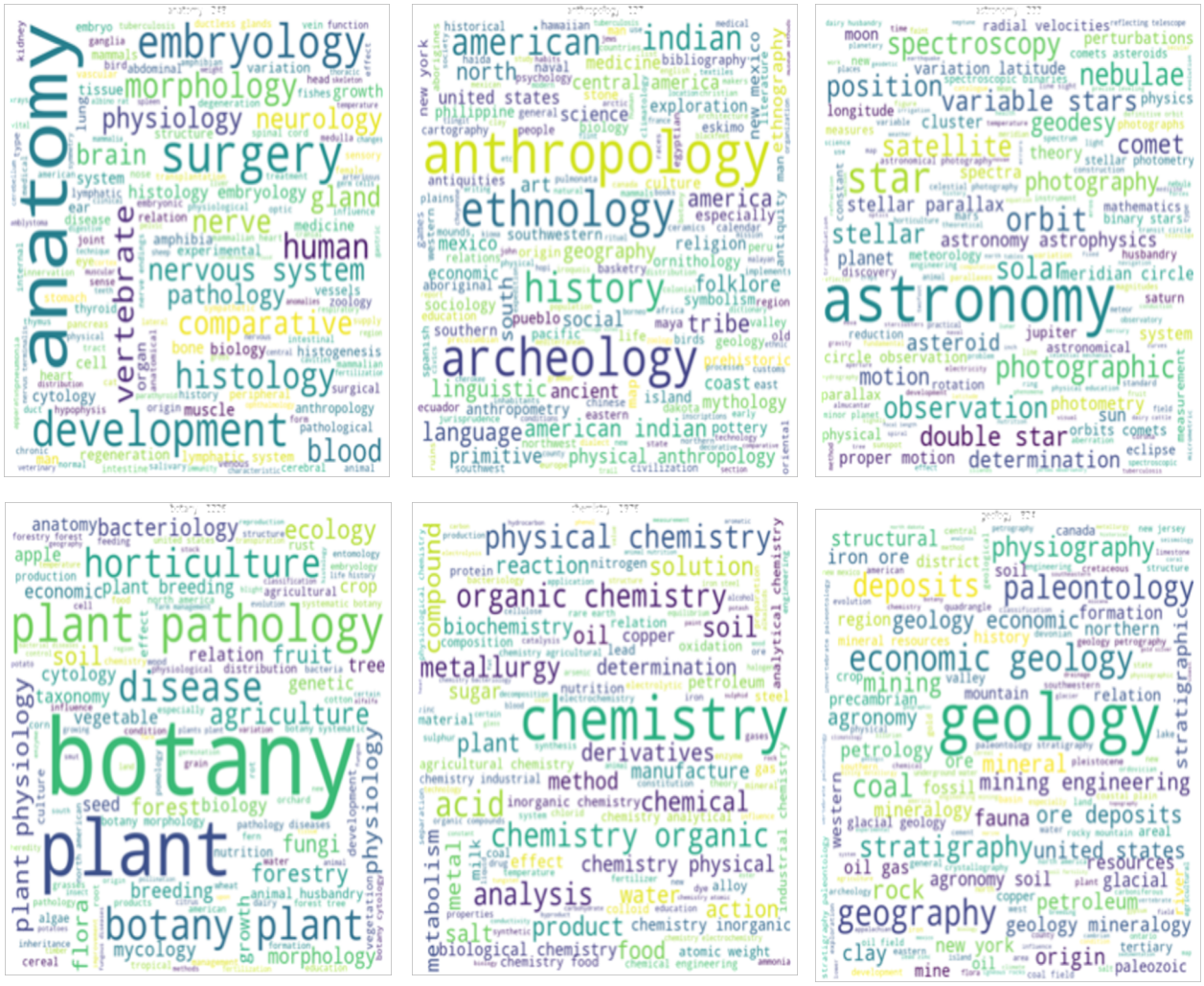
Figure B1 - Confusion Matrix
Comparing True (Manually Assigned) Labels with Predicted Labels



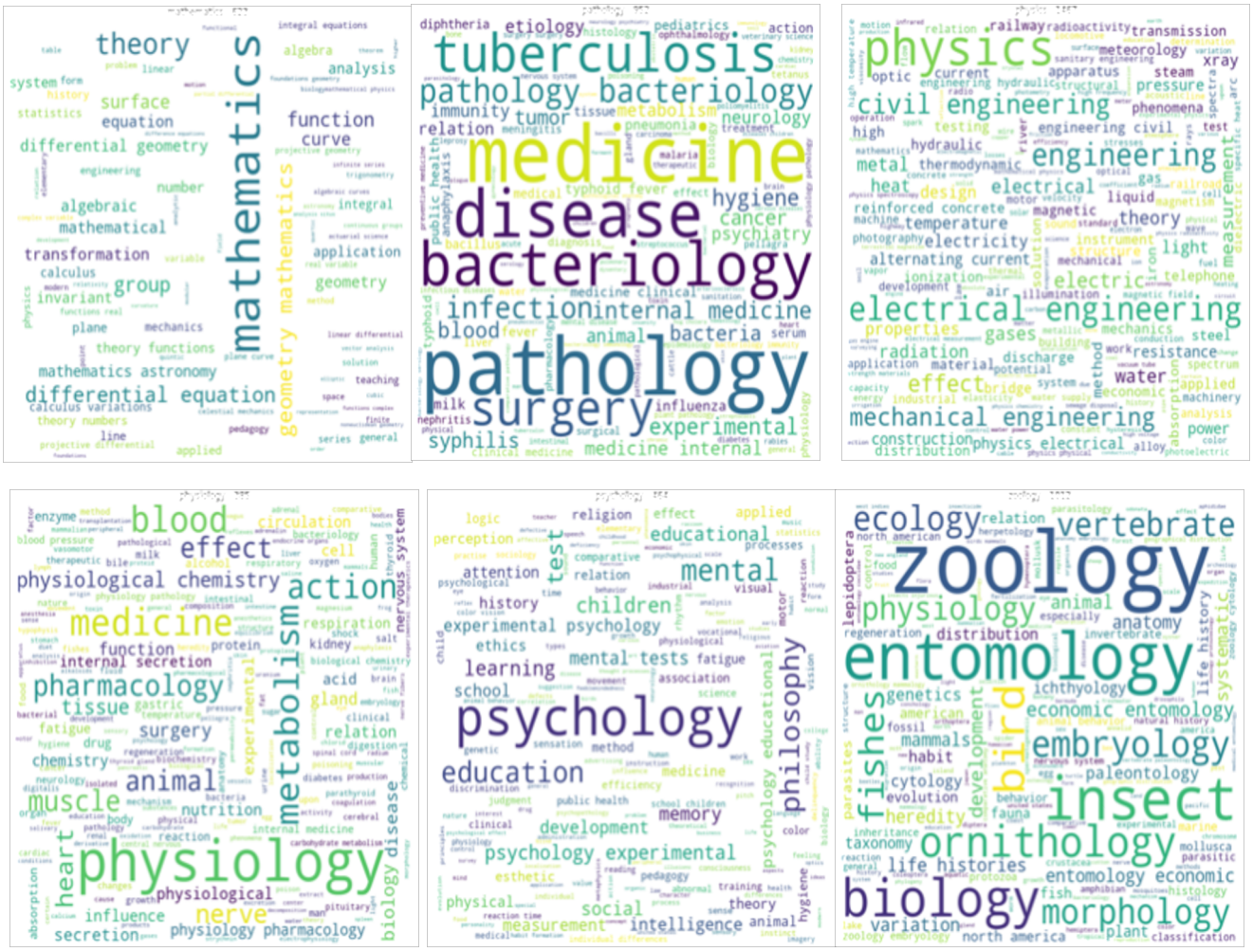
Notes: This matrix records true matches (on the diagonal), as well as false positives and false negatives for a hand-matched random sample of 200 scientists.

FIGURE B2 - WORD CLOUDS FOR 12 DISCIPLINES

PANEL A: ANATOMY, ANTHROPOLOGY, ASTRONOMY, BOTANY, CHEMISTRY, AND GEOLOGY



PANEL B: MATHEMATICS, PATHOLOGY, PHYSICS, PHYSIOLOGY, PSYCHOLOGY, AND ZOOLOGY



Notes: Word clouds display frequent words that describe the research of scientists across 12 disciplines. Disciplines are observable for stars in Visher (1947), but not for other scientists. To assign the remaining (non-star) scientists to disciplines, we use the text that describes the research of stars to create a training data set for a nearest-neighbor algorithm. This algorithm assigns each of the remaining scientists uniquely to one of the 12 disciplines. We plot word clouds to check the quality of the algorithm.

APPENDIX C: MATCHING SCIENTISTS WITH RECORDS IN THE CENSUS OF 1870, 1880, AND 1900

To link scientists in the MoS (1921) with individual census records in 1870, 1880 and 1900, we implement a machine-learning approach proposed by Feigenbaum (2016). The idea is to 1) build and manually match a smaller training sample of candidate record-census matches; 2) estimate a matching score function based on the textual characteristics of each candidate match pair; 3) tune hyper-parameters to select valid matches among those with the highest matching scores. Following this process, we use the matching score function and the tuned hyper-parameters to find census matches for scientists.

Specifically, we match 7,791 US-born male scientists in the MoS (1921) with 16.5 million US-born entries in the census of 1870, 21.8 million in 1880, and 33 million in 1900, respectively. Before matching scientists with the census, we exclude 1,408 scientists whose birthplace is unknown or in a foreign country (because we cannot observe them in the US census as children) and 355 women (because women used to change their name upon marriage, which makes it difficult to match them algorithmically). We also drop 104 scientists with missing data on birth years and missing data on first and last names, because we need this information to match scientists with the census. We then run pre-processing and cleaning scripts to prepare the reduced sample of scientists for matching.

In the first step of the matching process, we create a manually matched sample of scientist-census candidate pairs. For a random subset of scientists in the MoS (1921), we select all candidate matches in the US census who are born in the same state within 3 years of the scientist's birth year and whose first and last names are within a 0.2 Jaro-Winkler threshold for string distances of the scientist's name (implementing the same threshold as Feigenbaum 2016). Specifically, we match 2,000 scientists with the census of 1880 (when the median scientist in our data was a child), and 1,000 scientists each with the 1870 and 1900 census. This process matches 95% of scientists with at least one census record for each census wave.

In the second step, we use matches for 1,000 scientists with the 1880 census to train the algorithm, and the remaining samples to test the out-of-sample performance of the matching algorithm. Tests of out-of-sample performance across waves show that the characteristics of successful candidate scientist-census match pairs are stable across census waves. Therefore, we train the model using data for the 1880 census, when we observe the largest number of scientists as children.

In the third step, we use the manually classified data for 1,000 scientists, matched with the 1880 census, to estimate a matching score function that returns a score between 0 and 1 for each candidate scientist-census pair. Using these data, we estimate a probit model of the probability that a candidate scientist-census pair is a match as a function of the textual characteristics of the two records and the characteristics of the census matching pool for each MoS record. Values near 0 indicate that the scientist and census record are different, while values close to 1 indicate similarity. To select variables for the matching score function, we measure similarities in names using Jaro-Winkler and Soundex distances and indicators for a match between the first or last letter of the first, middle, and last name, as in Feigenbaum (2016). To measure distances in birth years, we use dummies for births that are one, two, or three years apart. Additional matching variables include an indicator for an exact name match, the total number of exact name matches in the sample, and the number (in levels and squared) of census candidates for each MoS record. Table C1 reports these probit estimates, which we use to calculate a matching score for each scientist-census pair and select the census candidate observation with the highest matching score.

Since not all of these observations are true matches, we further tune the matching procedure. For instance, the highest matching score may be very low, if the candidate pool for a specific scientist does not contain strong matching candidates. Additionally, if there are multiple

candidates with similar characteristics and very high matching scores, we may not have enough information to confidently choose a final match.

We tune two hyper-parameters: b_1, the minimum score to be considered a match, and b_2, the minimum distance between the highest and second-to-highest candidate match (Table C2). These hyper-parameters are chosen to maximize a combination of the True Positive Rate and the Positive Predictive Value. Intuitively,

TPR = (True Positives) / (True Positives + False Negatives) measures recall, and
PPV = (True Positives) / (True Positives + False Positives) measures precision.

We choose b_1 and b_1 to maximize the sum of TPR and PPV, conditional on reaching a minimum 90% in-sample PPV. This leads to a lower match rate compared to an unconstrained maximization procedure but provides us with better quality matches. Table C2 reports the tuned hyper-parameters.

Finally, we assess out-of-sample performance by deploying the estimated probit model and the tuned hyper-parameters to identify matches in the remaining manually matched samples (Table C3). We compare the algorithm matches with hand matches to identify True Positives, False Positives, True Negatives, and False Negatives, and to calculate the TPR and PPV for each census wave. ranges between 80.7% and 86.7%, whereas precision ranges between 83.3 and 88.2%. Recall is highest for the 1880 census wave. Precision increases over time, reaching the highest value for the 1900 census.

TABLE C1 –PROBIT ESTIMATES

First and last name match	1.4332***
	(-0.193)
First name distance, Jaro-Winkler	-6.1839***
	(-1.519)
Last name distance, Jaro-Winkler	-12.1736***
	(1.529)
Absolute value difference in year of birth is 1	-0.8953***
	(0.086)
Absolute value difference in year of birth is 2	-2.2709***
	(0.209)
Absolute value difference in year of birth is 3	-2.0093***
	(0.177)
First name Soundex match	0.1553*
	(0.09)
Last name Soundex match	0.8639***
	(0.201)
Hits	-0.0239***
	(0.002)
Hits-squared	5.179E-05***
	(4.68E-06)
More than one match for first and last name	-2.1872***
	(0.267)
First letter of first name matches	0.3766
	(0.489)
First letter of last name matches	-0.3862
	(0.293)
Last letter of first name matches	-0.1414
	(0.152)
Last letter of last name matches	-0.0533
	(0.161)
Middle initial matches (if there is a middle initial)	1.1884***
	(0.137)
Constant	-0.0964
	(0.677)
Observations	25,159
Log Likelihood	-629.78

Notes: Estimates of a Probit model of the probability a candidate scientist-census match pair is a match. Matching variables include Jaro-Winkler and Soundex distances and indicators for a match between the first or last letter of the first, middle, and last name; dummies for births that are one, two or three years apart; an indicator for an exact name match, the total number of exact name matches in the sample, and the number (in levels and squared) of census candidates for each MoS record.

TABLE C2 – CENSUS MATCHING: HYPERPARAMETERS

Hyperparameters	
b_1	b_2
0.1651	1.7405

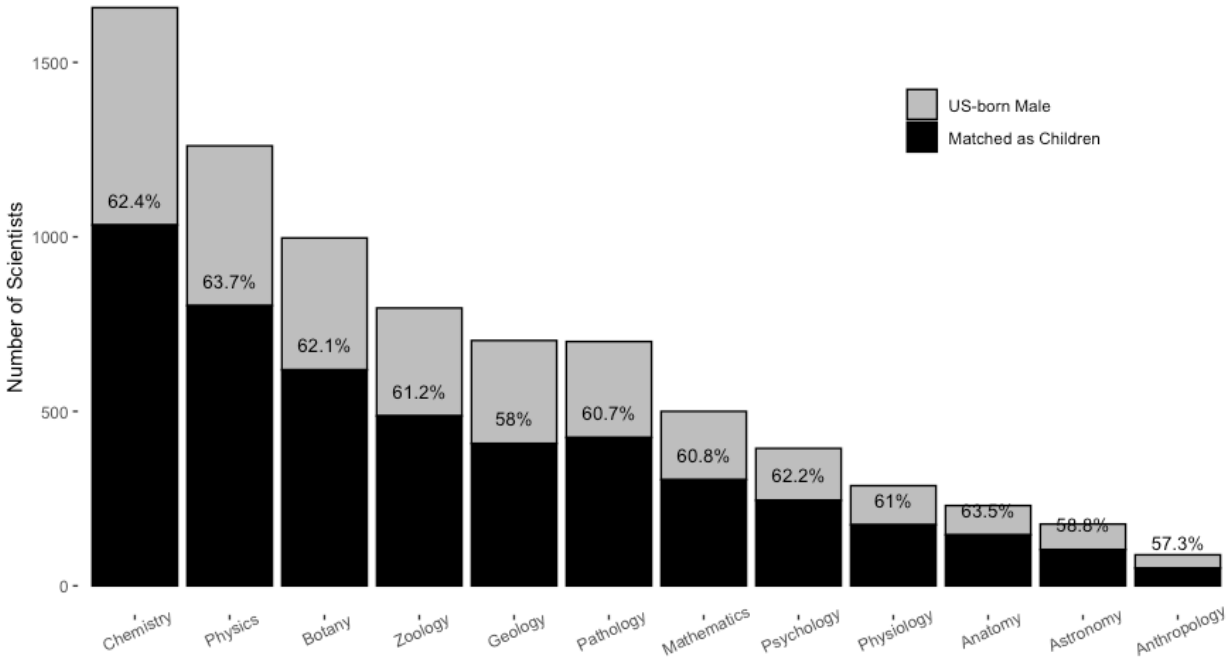
Notes: Hyperparameters of the census matching model, tuned to maximize a combination of the *True Positive Rate* and the *Positive Predictive Value*, conditional on achieving a Positive Predicted Value $\geq 90\%$. b_1 is the minimum score to be considered a match; b_2 is the minimum distance between the highest and second-to-highest candidate match.

TABLE C3 – CENSUS MATCHING: OUT-OF-SAMPLE PERFORMANCE

	Out of Sample Performance	
	Recall	Precision
1870	80.7%	83.3%
1880	86.7%	85.6%
1900	85.4%	88.2%

Notes: Out-of-sample performance of the matching algorithm, by census wave. Recall is the share of correctly identified matches over the set of possible matches, True Positive Rate = True Positives/ (True Positives + False Negatives)). Precision is defined as the share of correctly identified matches over the number of scientist-census pairs classified as matches, Positive Predicted Value= True Positives/ (True Positives + False Positives).

FIGURE C1 – MATCH RATE ACROSS 12 DISCIPLINES



Notes: The distribution of 4,067 scientists, whom we observe in their childhood home across 12 disciplines. The black area represents the number of matched scientists; the number above the black bar is the share of matched scientists among all scientists in that discipline.

APPENDIX D: COEFFICIENTS AS DIFFERENCES IN ODDS OR PROBABILITIES

Interpreting Coefficients as a Difference in Odds

Our preferred interpretation of the coefficients from the logistic regression is in terms of a change in odds. Coefficients are reported as the natural logarithm (ln) of odds, and we convert them to non-logged odds using the exponential function $e^{\widehat{\beta}}$. Converted in this way, estimates capture the odds ratio of being a star for scientists from high-SES families compared with other, low-SES scientists. We calculate and report the change in odds subtracting 1 from the odds ratio.

$$\%\,\Delta \text{ in odds} = e^{\widehat{\beta}} - 1$$

Interpreting Coefficients as a Difference in Probabilities

Converting estimates into the corresponding change in probabilities depends on estimates for other coefficients (besides $\hat{\beta}$) as well as the values of other control variables. To address this issue, we calculate the implied changes in probabilities for a baseline scientist of average age (46) in the largest discipline (chemistry) and the 1880 census wave (the wave in which we observe the largest number of scientists as children (Table D1). We calculate the change in probability at this baseline:

$$\%\,\Delta \text{ in probability} = \frac{P(star|SES_{high\ SES}=1)}{P(star|SES_{high\ SES}=0)} - 1\ ,$$

where:

$$P(star|SES_{high\ SES}=1) = \frac{e^{(\hat{\alpha}+\hat{\beta}+\hat{X})}}{1+e^{(\hat{\alpha}+\hat{\beta}+\hat{X})}} \text{ and } P(star|SES_{high\ SES}=0) = \frac{e^{(\hat{\alpha}+\hat{X})}}{1+e^{(\hat{\alpha}+\hat{X})}}$$

and $\hat{X}$ is the vector representing the dot product of remaining coefficients and its corresponding baseline values reported in Table D1.

Change in odds and probabilities reported in Table 3, column 1 are calculated as below:

$$\%\,\Delta \text{ in odds} = \exp(0.327) - 1 = 0.3866 = 38.66\%$$

$$\%\,\Delta \text{ in probability} = \frac{\frac{e^{(-4.223+46*0.044+0.3269)}}{1+e^{(-4.223+46*0.044+0.3269)}}}{\frac{e^{(-4.223+46*0.044)}}{1+e^{(-4.223+46*0.044)}}} - 1 = 0.3346 = 33.46\%$$

Since the overall share of star in our sample is 13%, the change in odds overestimate the change in probability by a small margin (the odds ratio approximates the true probability ratio when the probability of the outcome is small, less than 10%).

Marginal effects of citations

Besides interpreting the main coefficient of High SES, we are also interested in quantifying the marginal effect of citations on the chance of being star in equation (4):

$$\%\,\Delta\, in\ odds\ (star \mid \Delta citation) = e^{\theta_2 \text{asinh}\,(\Delta cit)} - 1$$

Since we are interested in the marginal change at the baseline (where the mean citation is 4), $x > 2$, and or $x > 2$, $\text{asinh}(x) \approx \ln(2x) = \ln(x) + \ln(2)$, so

$$\%\,\Delta\, in\ odds\ (star \mid \Delta citation) \approx e^{\theta_2 \Delta \ln\,(2*cit)} - 1$$

For a 10% increase in citation at the baseline (from 4 citations to 4.4 citations):

$$\Delta \ln(2 * cit) = ln(2 * 1.1 * cit) - ln(2 * cit) = \frac{ln(2.2*cit)}{ln(2*cit)} = ln(1.1) = 0.0953.$$

$$\%\,\Delta\, in\ odds\ (star \mid \Delta citation) \approx e^{\theta_2 + 0.0953} - 1$$

For a 100% increase in citation at the baseline (from 4 citations to 8 citations):

$$ln(2 * 2 * cit) - ln(2 * cit) = \frac{ln(4*cit)}{ln(2*cit)} = ln(2) = 0.693.$$

$$\%\,\Delta\, in\ odds\ (star \mid \Delta citation) \approx e^{\theta_2 + 0.693} - 1$$

This approximation assumes the marginal effect of citations to be identical for scientists who grew up with *high SES* and scientists who grew up with *low SES*. Supporting this assumption, Table D2 shows that differences are negligible, especially when the change in citation is small.

TABLE D1 – BASELINE VALUES OF CONTROL VARIABLES

Variable	Baseline
Continuous	
Age	46
asinh(pub)	2.0947
asinh(cit)	2.0947
Number of siblings	2
Fixed effect	
Census Year	1880
Discipline	Chemistry
Dichotomous	
Elite undergrad	0
Elite grad	0
First-born	0
Foreign-born parents	0

Notes: This table shows the baseline values we have used to interpret logistic coefficients as the difference in the probability that a high SES scientist is becoming a star (compared with a low SES scientist). To interpret these coefficients, we set the baseline value of the continuous variables (age, *asinh*(pub), *asinh*(cit), number of siblings) at the mean of the pooled data (across all census waves) for scientists whom we observe as minors in at least one census wave and whose fathers' occupation is known. The mean values of a*sinh*(pub) and a*sinh*(cit) above are equivalent to the mean of 4 publications and 4 citations. We choose the census wave with the largest number of observations (1880) and the discipline with the largest number of scientists (chemistry) as the baseline. For dichotomous variables, we set the baseline at zero: no elite degrees, not a first-born, and no foreign-born parents.

TABLE D2 – MARGINAL EFFECT OF CITATIONS AT THE BASELINE

	Low SES	High SES	All
Baseline	0.0686	0.0924	
$\Delta citation$ = 10%	0.0718	0.0965	
$\% \Delta in Prob\ (star \mid \Delta citation)$	4.66%	4.44%	
$\% \Delta in Odds\ (star \mid \Delta citation)$			4.28%
Baseline	0.0686	0.0924	
$\Delta citation$ = 100%	0.0950	0.1207	
$\% \Delta in Prob\ (star \mid \Delta citation)$	38.48%	30.63%	
$\% \Delta in Odds\ (star \mid \Delta citation)$			35.67%

Notes: This table uses estimates from Table 6, column 1 to estimate how a 10% and 100% change in citation changes the odds and the probability of being a star. A 10% increase in citation at the baseline is equivalent to an increase from 4.0 to 4.4 citations or an increase from 2.0947 $\text{asinh}(cit)$ to 2.1874 $\text{asinh}(cit)$. A 100% increase in citation at the baseline is equivalent to an increase from 4 to 8 citations or an increase from 2.0947 $\text{asinh}(cit)$ to 2.7765 $\text{asinh}(cit)$.