Innovation Nation: The Impact of Broadening Access to Doctoral Education on Ph.D. and Patenting Outcomes*

Francisca M. Antman† Kirk B. Doran‡ Xuechao Qian§
Bruce A. Weinberg¶

July 2026

Abstract

This paper explores how broadening access to doctoral education impacts Ph.D. and patenting outcomes. Matching the universe of Ph.D. recipients to the full count U.S. Censuses (1850-1940), an event study research design shows that the expansion in Ph.D. programs in the early 20th century increased doctorates and patentors. The expansion increased Ph.D. access for underrepresented groups, including minorities, children of immigrants, rural households, and lower socioeconomic status families; however, men gained more relative to women. Additional evidence suggests spillovers that improve patent quality and promote more inventors, including among non-doctorates.

Keywords: Ph.D. education expansion, Diversity, Patenting, Innovation
JEL Codes: I23, I24, O31

*We are grateful to Ran Abramitzky, Daniel Fetter, Donna Ginther, Lawrence Katz, Petra Moser, Heidi Williams, seminar participants at The Ohio State University, Stanford University, University of Colorado Boulder, Harvard University, University of California Merced, University of California Berkeley Goldman School of Public Policy, University of Houston, University of Toronto, Cornell University, University of Carlos III of Madrid, University of Glasgow, Autonomous University of Madrid, University of Mannheim, Virginia Tech, Tilburg University, University of Illinois Urbana-Champaign, and participants of the 2022 Southern Economic Association Annual Meeting, 2023 NBER Investments in Early Career Scientists: Promoting Diversity in Science Careers, 2024 Society of Labor Economists Annual Conference, and 2023 Virtual Economic History Seminar. We also gratefully acknowledge data on coeducational institutions from Lawrence Katz.

The authors declare that they have no relevant or material financial interests that relate to the research described in this paper.

†Department of Economics, The University of Colorado Boulder; IZA Institute for the Study of Labor, francisca.antman@colorado.edu

‡Department of Economics, University of Notre Dame, IZA Institute for the Study of Labor, kdoran@nd.edu

§Department of Economics, Ball State University, xuechao.qian@bsu.edu

¶Department of Economics, The Ohio State University; IZA Institute for the Study of Labor; National Bureau of Economic Research, weinberg.27@osu.edu

1 Introduction

Persistently low rates of participation in the U.S. innovation ecosystem suggest that increasing access to higher education could serve as a mechanism to increase economic opportunities for historically marginalized groups (Fry et al., 2021). At the same time, greater global competition for innovative and technological dominance also highlights the need for the U.S. to increase access to research training as broadly as possible. Although both of these motivations make a strong case for increasing doctoral research training in the U.S., it is not obvious whether a transformational expansion in Ph.D. education would necessarily lead to greater participation or diversity in the research labor force or greater innovative outcomes. For instance, increased take-up of educational opportunities in one area might simply displace them in another, and those populations who already had access to training opportunities may be the primary beneficiaries of increased access to new ones. This paper explores this question by examining the historical experience in which the United States rose to research preeminence in the early 20th century through a series of broad expansions in access to education, especially research education. This transformation was made possible by radically altering the institutions providing research education and engaging a far wider range of people in research (Geiger, 1986).

Specifically, this paper investigates the impacts of expanding research education institutions on the propensity to obtain a Ph.D. degree and the location and field of study among Ph.D. recipients. Furthermore, we estimate the extent to which broadening access to research education affects underserved populations that have the potential to generate more unique advances and technologies, including those that may be particularly beneficial for underrepresented and economically disadvantaged groups (Dossi, 2024). We find that access to research education was an important determinant of minority Ph.D. attainment. Linking research education to innovation productivity, this paper also identifies potential spillover effects of research education expansion, which increase the likelihood of innovation and enhance patent qualities, among inventors without Ph.D. degrees.

We generate a wealth of population-level data on U.S. Ph.D. recipients over much of the 20th century by linking historical U.S. Censuses (1850-1940) with population-level data on Ph.D.s awarded by U.S. institutions from ProQuest's Dissertation and Theses Database. Specifically, we match ProQuest data on dissertations to the longitudinally-linked U.S. Censuses from 1850-1940 (using valuable existing code generated by Abramitzky et al. (2020)), so that we can "look back" to earlier Censuses when Ph.D. recipients were children to characterize their family backgrounds. This allows us to identify the impact of the expansion of research education on the types of individuals who obtained Ph.D.s in terms of geography, parental socioeconomic status, as well as gender, race, and parental immigrant status. Research has only begun to

link demographic changes with contributions to science, and these have been limited to specific episodes in history and only for particular groups (Borjas & Doran, 2012; Moser et al., 2014; Moser & San, 2020; Hager et al., 2024). In this paper, we construct a broad population-level dataset on inventors, comprised of those who are registered in the European Patent Office's PATSTAT database, whom we match to historical U.S. Censuses, and also to the ProQuest dataset of doctoral recipients. Our unprecedented data resources allow us to observe scientific communities over the longest period of time, giving us a unique opportunity to trace out the effects of the expansion of research training institutions on the broadest possible innovation outcomes.

We first document the historical changes in U.S. research education to lay the groundwork for our empirical analysis. Geiger (1986) emphasizes that, as the 19th Century was coming to an end, research institutions in the U.S. were primarily focused on educating the sons of the elite in the humanities at a relatively small number of traditional institutions, but that over the course of the 20th century there was a tremendous expansion of institutions due to the founding of Land Grant Universities as well as institutions founded by the ultra-wealthy (e.g., Carnegie and Mellon, Chicago, Stanford, etc.). As a consequence, the locus of research education shifted from the Northeast to the Midwest, South, and West and, at the same time, the substance of education shifted toward more practical subjects, including science, engineering, and agriculture.

We then focus on identifying causal impacts of accessing research education overall and for different population subgroups, in line with the literature suggesting that diversity can be associated with better economic performance (Alesina & Ferrara, 2005), firm outcomes (Østergaard et al., 2011) and innovative ideas (Hofstra et al., 2020). Specifically, we use an event study empirical strategy, leveraging the timing of opening of Ph.D. programs relative to the age of pursuing Ph.D. education and the geographic variation in the opening of new Ph.D. programs. Our population-level data sets allow us to determine when Ph.D. programs begin and to link the timing of program openings with individual outcomes. This approach enables us to simultaneously provide a series of "placebo" checks showing that the expansion of Ph.D. programs later in life is not associated with Ph.D. attainment, which typically occurs in the mid-30s, thereby bolstering our identification strategy and allowing us to pinpoint causal relationships.

Our identifying assumption is that Ph.D. program expansions by state over time are exogenous, conditional on the control variables in our regression model, which also includes birth year fixed effects and state of birth fixed effects. These controls should capture any unobserved differences that might drive program openings which are common to all states in any given year or common to all cohorts in any given state, such as those which might reflect factors influencing demand for doctoral education. Nevertheless, we perform various robustness checks to address any remaining concerns about time-varying

endogenous program openings. These include adding controls for the number of colleges at the birth state-year level, and adding linear times trends by birth-region interacted with cohort and birth-census division interacted with cohort.

Our event study estimates indicate that one additional Ph.D. program in the birth state during a person's peak ages of graduate study induces nearly 2 more people born in that state to obtain a Ph.D. per million people. The impact of opening one additional nearby, out-of-birth-state Ph.D. program is about half as large, resulting in 1 more person per million people earning a Ph.D. degree. Overall, we estimate that the expansion in the number of U.S. Ph.D. programs between 1875 and 1960, both in the birth state and in nearby states, generated approximately 128 additional Ph.D. recipients per million people or 17 thousand Ph.D.s. This effect represents a 15% increase in the share of Ph.D. recipients relative to that observed in the 1940 U.S. population.

Furthermore, we find that the expansion of Ph.D. programs improved access for minorities, children of immigrants, and lower socioeconomic status families, relative to their counterparts. This suggests that the expansion of Ph.D. programs can play an important role in diversifying pathways to a larger research workforce. At the same time, we find that the nationwide expansion of Ph.D. programs resulted in greater gains for men relative to women.

In addition, we leverage exogenous variation brought about by major institutional expansions, such as Land Grant universities, co-education colleges, as well as the historical emergence of private universities, to investigate the main (policy) force driving the expansion in research training. Focusing on these policy changes also helps to address potential concerns that the results we observe are driven by endogenous changes in Ph.D. program openings. This analysis suggests that a primary policy driver of our results is the expansion in land grant institutions, which is reassuring as it is likely to be exogenously driven by federal policies. Furthermore, we find that the location and field of study of the Ph.D. program openings help to determine the institution and research area of doctoral recipients. These findings also reassure us that Ph.D. program openings and degree attainment are causally linked by region, institution type, and field of study.

In addition to establishing the impact of Ph.D. program expansion on Ph.D. outcomes, we also connect it with direct measures of innovation, including the quantity and quality of patents. Our estimates predict about 90 additional inventors per one million people due to increases in the number of Ph.D. programs in an individual's birth state and surrounding states within a 250-mile radius between 1875 and 1960, measured at age 30. This is equivalent to approximately 6.8% of inventors in the 1940 U.S. population. We also find strong evidence that Ph.D. program openings had a spillover effect on inventors without Ph.D. degrees, a result that may be explained by local agglomeration effects spurred by the opening of Ph.D. programs. Moreover, research educational programs

directly promoted innovation within fields, by stimulating patenting which relied more heavily on doctoral research training.

Our work represents a significant extension of the literature on the impacts of broadening participation in education, which until now has focused on primary (Clay et al., 2012; Card & Giuliano, 2014), secondary (Goldin & Katz, 1997), and college education (Bound & Turner, 2002; Chetty et al., 2017; Hoxby & Avery, 2013). Ours is one of the first to focus on the expansion of research education (MacLeod & Urquiola, 2021). By estimating the impacts of Ph.D. program expansion on women's representation among Ph.D. recipients, this paper also makes a noteworthy contribution to the burgeoning literature on women in science and the barriers they have faced (Kim & Moser, 2021; Iaria et al., 2024). Recent work has also drawn attention to institutions dedicated to serving historically underrepresented groups, such as Historically Black Colleges and Universities (HBCUs), which have been found to be positively associated with graduation rates (Gordon et al., 2021) and future faculty careers for African Americans (Perna, 2001). Our work connects closely with other studies that have found that barriers in access and exposure directly affect the resulting stock of scientists and innovators and the representation of historically underrepresented groups in the scientific community (Aghion et al., 2017; Bell et al., 2019; Azoulay et al., 2021; Berkes & Nencka, 2024). It also makes a substantial contribution to the literature examining the human impact of research universities on the economy (Kantor & Whalley, 2014).

This work also contributes to the broader literature on the impact of research and graduate education on innovation and the rate of inventive activity (Khan & Sokoloff, 1993; Lamoreaux & Sokoloff, 1999; Aghion et al., 2009; Toivanen & Väänänen, 2016). While this period in American history stands out for the ascension of the U.S. into a research and technology superpower, relatively little is known about the link between the expansion in access to graduate education over this period and the resulting growth in ideas and innovation; our work makes possible further contributions in this area.

The paper proceeds as follows. Section 2 describes the data sources and linking methods, reviews summary statistics, and describes the expansion in Ph.D. programs in the U.S. based on our data. Section 3 reviews the empirical strategy and event study methodology used to examine the impact of the expansion of Ph.D. programs on innovation outcomes. Section 4 presents the results and goes through a series of robustness checks. Section 5 concludes.

2 Data

2.1 Data Sources

We construct the first longitudinal dataset in which individual Ph.D. recipients and innovators are linked back to their characteristics and families of origin. Our historical linkage is built upon three major datasets: 1) the full count U.S. Census microdata from 1850 to 1940, 2) ProQuest Dissertation and Theses database, and 3) the European Patent Office's PATSTAT database.[1]

Historical Censuses The complete count U.S. Census microdata (1850-1940) is publicly available from the Integrated Public Use Micro Samples (IPUMS) USA. We use a version that includes full names provided by the National Bureau of Economic Research. In this study, the 1940 Census, which is the first decennial census to collect information on educational attainment for everyone, is the baseline population. Each individual census record in 1940 includes birth place, birth year, gender, race, ethnicity, birth place of each parent, urban residential status, and current occupation. It is then linked with all previous full count U.S. Censuses from 1850 to 1930 based on the Census Linking Project (Abramitzky et al., 2020) at the individual level. The comprehensive information in the longitudinally-linked historical U.S. Censuses enables us to match the U.S. population with the population-level Ph.D. recipients from U.S. institutions, and innovators with patents registered in the U.S.

ProQuest The ProQuest Dissertation & Theses data is the world's largest curated collection of dissertations and theses, and the official offsite dissertation repository for the U.S. Library of Congress. It contains dissertations published from 1861 and provides the comprehensive information of each dissertation, including a unique ProQuest identification number, author name, title, research fields, institution information (a school identifier and name), and the degree information of the recipient (the degree and date awarded). We only use Ph.D. degree recipients from the ProQuest and focus on people who were awarded their degrees before 1960 to maximize our ability to match recipients to their Census records.

PATSTAT Our innovator-level data come from the European Patent Office's PATSTAT database, which has the detailed information on each patent application since 1899, including inventor's full name, year of application, International Patent Classification (IPC) code, and number of citations. We then aggregate patent application level information to the individual innovator level to measure individual patenting outcomes. The main innovator level outcome measures for each individual are: 1) an indicator for ever having patented, 2) total number of lifetime patent applications granted

[1]Most of the 1890 Census records were destroyed by a fire in 1921. The preliminary version of the full count 1950 Census was released in 2024 pending further data cleaning at IPUMS.

per innovator, 3) total number of lifetime patent citations, and 4) average number of citations per granted patent.

2.2 Data Linkage

We implement a set of population scale links between the historical U.S. Censuses and population-level data on doctoral dissertations and innovators. Figure 1 illustrates our approach to match across the three datasets. By using the longitudinal links, we are able to document who these researchers and innovators are using their demographic characteristics, including birth place, birth year, gender, race, ethnicity, birthplace of each parent, current urban residential status, and current occupation, from the Census data. We are also able to look backwards in the Census to identify the kinds of families they grew up in with the help of the Census household records including the place where they lived as children (e.g., urban / rural) and parental socioeconomic status.

(Figure 1 here)

Census-ProQuest Given that the 1940 Census is the most recent, clean, public, full-count Census and the first Census containing educational attainment information, our Census-ProQuest linking work starts with matching ProQuest Ph.D. recipients to the 1940 Census. Our matching strategy is outlined in Appendix B. Step 1 uses unique, exact matches based on full first and full last names, which are likely to be very high quality, to obtain a "gold standard" dataset. In step 2, we use this gold standard dataset to estimate the distribution of other informative linking variables among likely true matches. We rely on three identifying variables in our baseline linkage: 1) highest education grade, 2) implied age of degree receipt based on the ProQuest degree date, and 3) geographical distance between the coordinates (latitude and longitude) of the Ph.D.-granting institution and the coordinates of the centroid of each person's current residential county.[2] [3] In step 3, we establish probabilistic matching rules by using the distribution of a set of selected identifying characteristics in the gold standard data set produced in Step 1. To reduce false matches, we block on the most important identifying variables, the first and last name of the dissertation author, meaning that we require an exact match of full first name and full last name for all matches between ProQuest and the 1940 Census in our resulting dataset.

[2]The county-level centroid longitudes and latitudes come from the U.S. Gazetteer Files (2017): https://www.census.gov/geographies/reference-files/time-series/geo/gazetteer-files.2017.html#list-tab-264479560.

[3]To ensure that the relationship that we estimate between programs opening in a state at the time that people would be enrolling is not an artifact of our data construction, we relax our linking restrictions by dropping either the implied age of degree receipt or the geographical distance between current residential location and Ph.D. education institution to regenerate the linked sample for robustness checks.

Some might be concerned that this matching procedure might induce selection patterns into our sample of women, since the matching procedure relies on blocking on first and last name and women are more likely to change their names over their lifetimes. To address this, we graph the female share of Ph.D. recipients in the ProQuest-Census matched sample alongside the share of female Ph.D. recipients obtained solely from the ProQuest data set over this period, where gender is identified algorithmically based on first name in ProQuest.[4] As Figure A.1 shows, the share of female Ph.D. recipients observed over time is very similar across data sources, suggesting that the matching algorithm does a relatively good job of matching women in our sample and minimizing the threat of any underlying selection patterns across gender. Moreover, the inclusion of women in our data set allows us to investigate whether there is a gendered pattern of results and extend the contribution of this paper to highlight factors impacting the representation of women in scientific fields, an area in which there is relatively little historical research, in part because population-level data are needed for studying women given that they are heavily underrepresented.

Unlike our gold standard data set made up of unique matches across data sources, in our full sample, there will be multiple individuals from each data source who exactly match on first and last name. To obtain unique matches, in step 4, we generate a match score for each matched pair from the ProQuest and Census based on the probabilistically weighted set of identifying characteristics. The highest-score pair is taken as the best match if and only if the match score is above a predetermined cutoff.[5] In order to compare Ph.D. recipients to (the overwhelming majority of) the population who did not receive Ph.D.s, we retain observations from the 1940 Census records that did not match to the ProQuest dataset. Lastly, we use existing crosswalks (Abramitzky et al., 2020) to link the full 1940 Census with ProQuest variables to all previous full count Censuses to find parental and family information.

Using this procedure, we obtain 102,639 matches between ProQuest Ph.D. recipients and the 1940 Census with matching scores above the predetermined cutoff. Out of all the Ph.D. recipients who graduated from U.S. institutions through 1960, our match rate is 66%, meaning we are able to match two-thirds of the population of Ph.D. recipients in ProQuest to a U.S. Census record. As shown in Figure B.4, the match rate is reasonably flat and at or above 60% for most of the period. It increases somewhat for people who were young in 1940 (i.e., who got Ph.D.s in the 1950s and 1960s). Figure B.4 also shows that

[4]We use the Python packages: gender-guesser and ethnicolr to impute gender and race/ethnicity of each Ph.D. recipient in ProQuest based on author names, in keeping with other work - for example, Hofstra et al. (2020).

[5]Appendix B presents the distribution of each identifying variable among unique matches and lists the probabilistic match scores. In the rare case when there are multiple candidates in the 1940 Census matched to one Ph.D. recipient with the same highest match score, we keep all the potential matches if and only if the total number of candidates is below 10. This approach reduces precision but increases recall.

the match rate is largely uniform across fields, suggesting that the matching procedure does not introduce selection bias.

Census-PATSTAT We merge PATSTAT patent data into the full count U.S. Census to capture the demographics and family characteristics of all individuals who have at least one patent granted in their lifetime, using the approach in Davison et al. (2023). Our matching procedure only uses unique PATSTAT inventor names and 1940 Census names, which are available for about 30% of the 1940 U.S. population, following the ABE approach (Abramitzky et al., 2012, 2014, 2019). We obtain exact matches between the cleaned first, middle, and last names. In accordance with the age distribution of great innovators (Jones, 2010), for each innovator, we restrict the age of last invention to be no earlier than 19 years old and no later than 85 years old, and the age of first invention to be no later than age 50 and also no earlier than age 10. In total, we link 159,709 inventors from PATSTAT to the 1940 Census. Similar to our Census-ProQuest linked data, our linked inventors are longitudinally linked to previous Censuses to obtain family background and demographic characteristics.

2.3 Data Description

2.3.1 Doctoral Research Training Programs in the U.S.

(Figure 2 here)

We first describe the historical evolution of Ph.D. programs using the universe of dissertation data in ProQuest from 1861 to 2015 to trace the trajectory of American research education.[6] Ph.D. programs are defined at the institution and fine academic field level, which can be further aggregated into four major fields using the "tree" structure of the ProQuest field classification system: Arts & Humanities, Social & Behavioral Sciences, Mathematical & Physical Sciences, and Natural Sciences.[7] [8]

Figure 2 presents the historical trends in the total number of Ph.D. programs in the U.S. overall and by categories (on a log scale for readability). In all three panels, the solid, thick black line shows that the number of doctoral training programs was under 5 before 1880 but grew exponentially to 77 by 1900 and to 1025 by 1960. Figure 2(a) shows that the plurality of Ph.D. programs before 1920 were in the Arts & Humanities, which was surpassed gradually by the other three areas: Social & Behavioral Sciences, Mathematical & Physical Sciences, and Natural Sciences. We note that doctoral programs existed in

[6]This description only uses the ProQuest data, as it does not require any link to the U.S. Census to describe Ph.D. program openings.

[7]The list of fine fields and their respective grouped major fields is in Table A.1.

[8]To reduce measurement error, we first identify the initial year of a Ph.D. program as the year in which the institution first awarded three Ph.D.s in one cohort in one field. Our focus on when degrees are awarded means that our estimates should be interpreted in terms of the time / age at which people complete rather than enroll in programs.

Europe at this time, so the expansion in the U.S. represented the adoption of an existing practice and was not due to a shock to knowledge. By 2010, the most programs were in the social sciences and natural sciences, both with a relatively similar number of programs, followed by math and physical sciences programs and then the arts and humanities.

Figure 2(b) presents the number of doctoral training programs for various types of institutions, mainly based on whom they serve.[9] Comparing trends in the number of Ph.D. programs offered in "Women-granting" or "Minority-granting" institutions with those of Ph.D. programs in co-educational institutions, Figure 2(b) suggests that graduate education in the U.S. was opened to minorities essentially from the start and to women from the beginning of the 20th Century, considerably earlier than widespread undergraduate education opportunities were available for these groups. Additionally, the establishment of Land-Grant universities also contributed to the increase in doctoral research training programs substantially over the past century in the U.S.

Figure 2(c) divides Ph.D. programs based on whether the institution is private or public using information from the Integrated Postsecondary Education Data System (IPEDS). It shows that doctoral research training programs were exclusively offered in private institutions until the end of the 19th century. However, the number of Ph.D. programs in public institutions, many of which are land grant institutions, expanded more rapidly starting around the 1920s, eventually constituting the majority of U.S. doctoral training programs. Overall, Figure 2 demonstrates heterogeneous sources of Ph.D. program expansion, especially in the first half of the 20th century, with the rise of natural sciences, social & behavioral sciences, and public institutions.

(Figure 3 here)

As the number of Ph.D. programs increased, they spread out geographically. Figure 3 shows how Ph.D. programs expanded across the country geographically through the 20th century. At the end of the 19th century, states with doctoral research training programs were almost entirely in the Mid Atlantic region, Massachusetts, the Midwest, or California, while the vast majority of states did not have any Ph.D. programs at all. By 1940, Ph.D. programs were operating in most states other than some states in the Mountain region and South, while by 1980, they were operating in all states. Thus, much of the spread was due to geographic expansions to areas that previously had few programs.

[9] In this period, not all institutions granted degrees to women and non-Whites. Once a Ph.D.-granting institution grants a female or non-White student a Ph.D. degree, we define this institution as a "Women-granting" or "Minority-granting" institution. We use the Python packages: gender-guesser and ethnicolr to impute gender and race/ethnicity of each Ph.D. recipient in ProQuest based on author names. These algorithms tend to have a reasonably high precision but low recall rate for imputing minorities, defined here as Hispanics, non-Hispanic Blacks, non-Hispanic Asians.

2.3.2 Summary Statistics

(Table 1 here)

Table 1 summarizes major statistics for the entire 1940 U.S. population, our linked Ph.D. sample, and our linked inventor sample. In the version of the full count U.S. Censuses to which we have access, the 1940 population of the U.S. is 120,033,290. Among the entire U.S. population, 50% are male, 89% are White, 53% live in urban areas, and 9% have either or both parents born in a foreign country (2nd-generation immigrants). We successfully link 102,639 Ph.D. recipients to the Census, implying a population rate of Ph.D. receipt of 855 per 1 million people (0.000855). Within the sample of Ph.D. recipients, 88% are male, 95% are White, 70% live in urban areas in 1940, and 14% are 2nd-generation immigrants. We also matched 159,709 inventors to the Census. This match implies a population rate of patent inventorship of 0.00133 (1,330 per 1 million), which is higher than the rate of Ph.D. receipt. Among the linked inventors, 93% are male, 97% are White, 62% live in urban areas in 1940, and 11% are 2nd-generation immigrants.

Figure 4 plots trends in the share of each demographic group among all ProQuest Ph.D. recipients relative to the 1940 US population share. It suggests that, relative to the U.S. population as a whole, the share of women, non-Whites, and 2nd-generation immigrants increased substantially among doctoral degree recipients over the course of the early 20th century, however, the share of women appears to have increased markedly between 1910-1915, but then stabilized at a relatively low level. Moreover, the representation of rural and low-SES groups among Ph.D. recipients remained relatively flat, and the rural representation shows some decline. Overall, these descriptive graphs point to the possibility of gains in educational attainment for underrepresented groups that coincided with the expansion in Ph.D. programs, although the benefits may have varied across groups. Our econometric analysis relates changes in Ph.D. attainment overall and for these groups to Ph.D. program expansion over time using the individually reported demographic data to track gains for underrepresented groups.

(Figure 4 here)

Returning to Table 1, we link 159,709 patent inventors, of whom 1% have a doctoral degree, 93% are male, 97% are White, 62% live in urban areas in 1940, and 11% are 2nd-generation immigrants. As with Ph.D. recipients, it appears that inventors were more likely to be from privileged social groups (i.e., male, White, and urban), compared to the U.S. population, although it is noteworthy that they are also more likely to be children of immigrants. Table 2 summarizes the individual level patenting outcomes among the 159,709 inventors. On average, each inventor has four patents granted and 22 citations per application. The average number of citations per patent is about 6.[10]

[10]We exclude extreme large (top 1%) values of the total number of patents and number of citations

(Table 2 here)

3 Empirical Strategy

Our empirical strategy is to estimate a discrete choice model of the probability that people obtained a Ph.D. or became an inventor as a function of the expansion of graduate education programs nationally from 1900 through 1960. At its core, our approach is to relate the probability that any given individual in the Censuses receives a Ph.D. in the U.S., as given by whether they appear in ProQuest, to the number of Ph.D. programs in their state of birth and nearby out-of-state programs over the course of their life. We use an event study design, which includes leads and lags of the number of Ph.D. programs to describe the timing of Ph.D. program openings using the following regression model:

$$Outcome_{isc} = \beta_0 + \sum_{k=0}^{7} \beta_{k,InState} N_{i,k,InState(s)} + \sum_{k=0}^{7} \beta_{k,Nearby} N_{i,k,Nearby(s)} + X_{isc}\Upsilon + \mu_s + \lambda_c + \epsilon_{isc} \quad (1)$$

where $Outcome_{isc}$ is an outcome for individual i from birth state s, born in cohort c. Our outcomes include an indicator for ever receiving a Ph.D. degree, PhD_{isc}, and an indicator for ever receiving a patent, $Inno_{isc}$. $N_{i,k,InState(s)}, k = 0, 1, 2, ..., 7$ are the number of Ph.D. programs in individual i's birth state, from her birth to age 70 (at 10 year intervals). $N_{i,k,Nearby(s)}, k = 0, 1, 2, ..., 7$ is an analogous set of measures indicating the number of Ph.D. programs within 250 miles of individual i's birth state but excluding programs in the state of birth itself, which we refer to as "nearby programs."[11] The total number of Ph.D. programs nationally at each point in a person's life is absorbed by the cohort fixed effects, λ_c, so one should interpret the coefficients of interest ($\beta_{k,InState}$ and $\beta_{k,Nearby}$), for which we report the marginal effects, as the effects of having one more program in the state of birth or near the state of birth conditional on the number of programs nationwide. X_{isc} is a set of individual characteristics, including gender, race, urban residence, and indicators for parental immigrant status. In addition to birth year (or cohort) fixed effects λ_c, we control for state of birth fixed effects μ_s. The state of birth fixed effects control for time-invariant underlying differences in the propensity to obtain a Ph.D. or the propensity to become an inventor across states. Standard errors are clustered at the birth-state-cohort level.

If access to programs increases the probability of completing a Ph.D. or obtaining

per inventor in the analysis.

[11] The distance is calculated based on the longitudes and latitudes of the centroid of the state of birth and the longitudes and latitudes of each institution/university. The state-level centroid longitudes and latitudes come from Google: https://developers.google.com/public-data/docs/canonical/states_csv. Unfortunately, birth state is the only birth place information that is available for the whole population in the full count censuses. Linking individuals to their childhood county across censuses reduces our sample by 51%.

a patent, we would expect the number of programs in a state to have the strongest impact around the ages when people would be most likely to finish a Ph.D.[12] Given that we estimate the impacts in 10-year intervals, we expect this will be around the time people turn 30, which is approximately the peak age of Ph.D. receipt among the unique matches in our linked Census-ProQuest data. We might see some effects of Ph.D. program expansion at younger ages. For example, having more Ph.D. programs in a state might encourage more people to go to college or, once they are there, to continue to graduate school, however, we do not expect it to capture access. Also, we do not expect to see impacts at older ages.

Indeed, a finding of impacts in older ages would likely suggest that other correlated effects are driving the results, and thus including Ph.D. program openings at older ages in the specification serves as a placebo test to see whether Ph.D. program openings might be correlated with other unobserved variables that might serve to increase the probably of earning a Ph.D.

4 Results

4.1 Baseline Effects of Ph.D. Program Expansion on Ph.D. Attainment

Figure 5 shows the estimated coefficients (marginal effects) on the number of Ph.D. programs in the birth state ($\beta_{k,InState}$) in Panel (a) and in nearby, non-birth state institutions (within 250 miles of the birth state centroid) ($\beta_{k,Nearby}$) in Panel (b) from birth to age 70, at 10 year intervals. Panel (a) indicates a very strong relationship between the number of Ph.D. programs in the state of birth at the time that people turn 30 and the probability of obtaining a Ph.D. The estimates at other ages, especially at older ages, are quite small and flat, suggesting unobserved variables are unlikely to be driving the results. The weak results at older ages are reassuring in terms of causality. The weak effects at younger ages indicates that access at the relevant age is a sufficient statistic to capture the role of programs.

Figure 5 Panel (b) shows that the number of programs in neighboring states at age 30 is also related to the probability of obtaining a Ph.D., but the estimated marginal effect is roughly half that of programs in the individual's own birth state. These estimates are reassuring from the perspective of causality in that we expect nearby, non-birth-states to be broadly similar economically and socially to a person's own state, but less important as a determinant of obtaining a Ph.D. Thus, our estimates showing that the relationship

[12] As noted above, our ProQuest data allows us to identify programs according to the time that people are receiving degrees, not the time at which they matriculate, so it is natural to look according to the time that people would naturally receive degrees.

between the number of in-state Ph.D. programs at age 30 is a stronger predictor of the probability of getting a Ph.D. suggests that the effects are being driven by access to research training programs rather than unobserved factors that drive the prevalence of programs and completion, which are likely to be similar across states in a region. Specifically, the estimates at age 30 (reported in Column (1), Table A.2) indicate that one additional Ph.D. program in the birth state around the graduate education completion age induces about 2 more people per million to obtain a Ph.D. degree, while the addition of one program in nearby, non-birth-state institutions at age 30 generates nearly 1 additional Ph.D. recipient per million in the state's population.

We use the main estimates of the effect of Ph.D. programs available during a person's graduate education years, both in their birth state and in nearby, non-birth-states, to estimate the overall impact of Ph.D. education expansion between 1875 and 1960 on the U.S. population. In 1875, the average number of Ph.D. programs per person at age 30 was approximately 0.011 in the birth state and 0.162 in surrounding states within a 250-mile radius. By 1960, for individuals reaching the age for Ph.D. education, the average number of programs in the birth state had increased to 37.092, while programs in nearby, non-birth-state institutions rose to 80.716. As a result, the total effect of the increased number of Ph.D. programs, both in birth states and neighboring states, corresponds to about 128 additional Ph.D. recipients per one million population (0.000128), equivalent to nearly 17 thousand Ph.D. degrees. [13] This overall impact accounts for approximately a 15% increase, relative to the share of Ph.D. receipt among the 1940 U.S. population–around 855 people per million.

(Figure 5 here)

4.2 Heterogeneous Effects by Individual Characteristics

From a public resource perspective, it is reassuring that the expansion in Ph.D. programs over this period substantially increased overall Ph.D. attainment. At the same time, it is reasonable to investigate which groups responded the most to this expansion in access to Ph.D. training. To do this, we estimate how individual and parental characteristics mediate the response to the number of Ph.D. program openings over this period. To measure parental socioeconomic status, we cross-link the 1940 U.S. Census to earlier Censuses and keep the observation that is closest to the year in which the individual was 10 years old.[14] With the resulting sample, we are able to control for urban residence

[13] To calculate the total effect, we multiply the coefficient for in-state programs from Column (1), Table A.2, by the increase in the number of in-state programs from 1875 to 1960: $1.5 \times 10^{-6} \times 37.081$. A similar calculation is performed for nearby state programs: $0.9 \times 10^{-6} \times 80.553$. The total effect is then obtained by summing these two components, yielding approximately 0.000128.

[14] Unfortunately, the total linked sample size drops by 51% as many people cannot be linked back to censuses during their childhood.

and parental socioeconomic status both around age 10, in addition to the individual characteristics in the main specification noted above (i.e., gender, race, and indicators of parental immigrant status). We then augment Equation (1) by including interactions between the number of Ph.D. programs and these demographics controls. Doing so allows us to estimate how the expansion of Ph.D. programs impacted the probability of earning a Ph.D. degree for various demographic groups and, more generally, whether the expansion of Ph.D. programs resulted in improved access and greater equality of opportunity to research training for underrepresented groups.

(Table 3 here)

We report estimates that include interactions with all individual demographics—gender, race, nativity, and family characteristics from the linked census, including urban residential location and parental socioeconomic status measures around age 10, in Table 3. For socioeconomic background, we have five measures of socioeconomic status in the censuses. These are the occupational income score, the Duncan socioeconomic index, the Siegel occupational prestige score, the occupational earnings score, and the occupational education score. In each column of Table 3, we include a different measure of socioeconomic status interacted with the number of programs.

The coefficients on the interaction terms with socioeconomic status are consistent across Columns (1) to (5). The coefficient on the interaction term between female and the number of in-state Ph.D. programs at age 30 is negative, indicating that Ph.D. program expansions had a negative relative impact on the probability of earning a Ph.D. for women compared to men, thus increasing existing gender disparities. Given the small number of women earning Ph.D.s at this time, the ones who did may have been significantly motivated, so that institution location was not as important a factor in women's decisions, compared with those of men. Table 3 also shows that increasing the number of in-state programs at age 30 has a positive relative impact on the probability of Ph.D. attainment among non-Whites compared to Whites and among people with immigrant parents relative to children of U.S. natives. Table 3 further shows that the number of in-state programs has a positive relative effect on the probability of Ph.D. attainment for people who grew up in rural areas compared with those from urban areas (statistically significant in Columns (3) and (5)) and whose parents have lower socio-economic status (statistically significant in Columns (2) and (4)). Thus, increased access to Ph.D. training appears to have reduced inequality in several ways, unlocking innovative potential from underrepresented population groups.

To explore whether the overall effect on subgroups was positive, we estimate the impacts separately for subgroups of women, men, minority, White, 2nd-gen immigrant, and native groups and report the results in panel A of Table A.3. The results can be noisy

due to smaller population proportions, but indicate that some groups clearly benefited from the program expansion overall (men, minorities, White subgroups, and children of U.S. natives). Results for women and second-generation immigrants are also positive, but not statistically significant. Thus, the greater effect of program access on men relative to women in Table 3 indicates only that men's Ph.D. attainment is increasing relative to women and not that women's Ph.D. attainment is declining.[15]

4.3 Heterogeneous Effects by Institution Type

To investigate the major driving forces of the effect of Ph.D. program expansions, we estimate separate effects by institution type: Land Grant universities, institutions that are co-educational (whether they were initially co-educational colleges or converted to co-educational colleges), whether the university grants Ph.D.s to women or non-White students, and private versus public universities. We also look by research fields: STEM versus non-STEM Ph.D. programs. In addition, we analyze results separately for female and non-White samples to see whether underrepresented groups in particular are affected differentially by these types of Ph.D. program expansions.

(Table 4 here)

Table 4 shows that while Land Grant universities contribute substantially to increasing the population probability of earning a Ph.D. degree, the increase in the number of co-educational institutions was not the main driving force in expanding access for women. Perhaps because they are far more prevalent, we find that universities that have ever granted a Ph.D. to a woman or minority have more statistically significant effects on increasing Ph.D. receipt. We also find that the opening of Ph.D. programs in public institutions had slightly larger effects on the probability of earning a Ph.D. in the full sample and for non-Whites.

Similarly, the opening of non-STEM programs have a larger impact on the probability of earning a Ph.D. compared with STEM programs. While we find the number of Ph.D. programs in most sub-categories does not differentially affect women, programs in Land Grant universities, co-educational institutions, institutions that have granted degrees to either women or minorities, and public universities generate larger effects on the probability of earning a Ph.D. among non-Whites. Only the number of Ph.D. programs in private universities generates a relatively large and statistically significant effect on Ph.D. receipt among women.

[15] We also estimate the effects for subgroups defined by urban or rural residential location and parental socioeconomic status around age 10 in Table A.4. Since the linked Census sample is reduced by around 50%, most of the coefficients on the number of in-state programs at age 30 are not statistically significant. High versus low parental SES is defined using the median of the average score across the five socioeconomic measures.

Overall, Table 4 suggests that establishing Land Grant and public universities and expanding the body of Ph.D. recipients to both female and minority students contributes substantially to expanding the population receiving advanced research degrees. The positive coefficients of interest in columns (2) and (3) suggest a possible mechanism underlying these drivers could be the increased numbers of Ph.D. students who are women and minorities at these types of institutions.

4.4 Effects by Region, Field, and Type of Institution

We extend our analysis to examine links by region, field, and type of institution. Note that this also constitutes a way to probe the causal relationship between Ph.D. program expansions and individual level doctoral degree attainment. For instance, if our estimates were causally linked, we would expect the number of Ph.D programs in nearby, non-birth-states (within 250 miles) to have a larger effect on the probability of obtaining a Ph.D. degree in nearby, non-birth-states, relative to doctoral degrees earned either in the birth state or farther away. Table A.5 presents the effects of Ph.D. programs at each geographical distance on the probability of earning a Ph.D. from institutions in the respective area. Column (1) shows that, compared to programs in states within 250 miles of a person's birth state, the number of Ph.D. programs in birth states has almost twice the effect on the probability of earning a Ph.D. within the person's birth state. Meanwhile, the probability of earning a Ph.D. degree in a nearby state is solely determined by the number of programs in nearby, non-birth-states. Moreover, the probability of earning a Ph.D. in a state that is more than 250 miles from the birth state is decreasing in the number of programs within 250 miles of the birth state.

Table A.6 divides the number of Ph.D. programs further based on whether the institution is public or private, and also distinguishes Ph.D. receipt from public and private universities. The results show that the probability of earning a Ph.D. from a public institution increases with the number of Ph.D. programs in both in-state public and private institutions, while the effect is substantially larger for public programs. The probability of earning a Ph.D. degree from a private institution is solely related to the number of Ph.D. programs at private institutions in nearby, non-birth-states.

We also study how the number of programs in four broad fields relates to the subjects in which people obtain degrees. Table A.7 shows that the probability of earning a Ph.D. in Arts & Humanities is solely determined by the number of Ph.D. programs in the Arts & Humanities in the birth state. The probability of obtaining a Ph.D. degree in Natural Sciences is strongly related to the number of Natural Sciences Ph.D. programs in the

state of birth. [16]

4.5 Robustness Checks

4.5.1 Endogeneity and Unobserved Variables

One concern with the analysis above is that Ph.D. program openings may be endogenous, being driven by demand forces or other economic changes, and that these unobserved variables drive our results. For instance, some may be concerned that the opening of Ph.D. programs could be closely related to the expansion of higher education more generally, making it important to distinguish between these two forces. To address this possibility, we first point to Figure A.2 showing national trends in the number of research institutions and in undergraduate colleges which indicates that Ph.D. program openings accelerate after 1920 in a way that college openings do not. This suggests that Ph.D. program expansion is likely to be independent of college expansion. More broadly, some may be concerned that Ph.D. program expansion is driven by other demand-side forces that may also affect the probability of earning a Ph.D.

To address these concerns, we present three robustness checks in Table A.2. For comparison, Column (1) presents our baseline estimates discussed above. Column (2) includes controls for the number of colleges analogous to our measures of the number of Ph.D. programs (i.e., controls for the number of colleges in each person's birth state and in states within 250 miles of the birth state from birth to age 70, at 10 year intervals). Columns (3) and (4) include richer time trends to account for possible changes in local economic conditions that might be related to Ph.D. program expansions and the probability of earning a Ph.D. Specifically, Column (3) controls for birth-region specific linear cohort trends and Column (4) controls for birth-division specific linear cohort trends. All told, our main estimates on the coefficient of interest, measuring the effect of the number of programs at age 30 in the birth state, are robust to the inclusion of these controls. The estimates for programs in nearby, non-birth-states are somewhat more sensitive, but broadly consistent with our main estimates. Overall, these results indicate that our main results are not likely to be driven by unobserved variables or other economic forces.

4.5.2 Robustness to Matching Algorithm

Another concern relates to our matched sample, and whether some elements of the algorithm used to match the ProQuest and Census data might generate a relationship between the number of Ph.D. programs and the probability of earning a Ph.D. degree.

[16] It is also positively related to Arts and Humanities programs in the birth state but negatively related to Behavioral and Social Science and Mathematical and Physical Science programs, suggesting some collinearity.

We address this concern in two ways. First, we use self-reported data on educational attainment from the Census to confirm our results are consistent with those from alternative data sources. Second, we redo our match after dropping some variables from the matching criteria to check the consistency of our results.

We begin by generating an indicator variable for having some graduate education (i.e., above four-year college) using completed education grades in the 1940 Census, and check to see if our results are similar. Panel (a) in Figure A.3 shows that, among all adults above age 40, the number of programs in the birth state at age 30 is a strong predictor of the probability of self-reporting having completed some post-college education in the Census. Moreover, the pattern of estimates is similar to the main results in Figure 5, indicating that the positive relationship between Ph.D. completion and the number of Ph.D. programs in the state of birth peaks at age 30. Thus, the evidence using Census outcomes is consistent with the main results using ProQuest.

We also address the specifics of our matching criteria, which rely on matching names in ProQuest to the Census and refining the matches using other features. In particular, we use the age when an individual completed Ph.D. study and the distance between the institution where people received their Ph.D.s (in ProQuest) and where we observe them to be located in the Census to generate a matching score and to select the most likely match for people who cannot be uniquely matched to the Census based solely on their name. To address any possible bias based on using these variables to refine our matches, we reproduce the main results using a matched sample that does not include age at degree in the match criteria and separately present results for a matched sample where distance to institution is not included in the match criteria. These results are presented in Panel (b) and Panel (c) of Figure A.3, respectively. Both sets of estimates are substantially unchanged from the baseline estimates in Figure 5. Thus, the particulars of our matching algorithm do not appear to drive our results.

4.6 Effects of Ph.D. Program Expansion on Patenting

4.6.1 Overall Patenting Results

Our extended objective is to estimate the effect of doctoral education on the quantity and quality of innovation itself in addition to the effects on the STEM workforce. To begin to understand the effects of doctoral education on innovation, we examine how the expansion of Ph.D. programs impacts patenting. There are two channels by which Ph.D. programs might increase patenting. First, people who are induced by the presence of a program to earn a Ph.D. may generate more patents than they otherwise would—a "direct" effect of Ph.D. programs operating via Ph.D. receipt. Second, the presence of a Ph.D. program may have a "spillover" effect by providing inspiration, knowledge, or access to networks and resources for others who do not obtain a Ph.D. themselves (e.g., undergraduates

or employees at the same institutions, people who begin but do not complete graduate programs, or participation of others in the surrounding community). To investigate these hypotheses, we return to Equation (1) and re-estimate it with patenting outcomes using the same event study framework. To exclude the potential association with the expansion of higher education more generally, we further control for the total number of colleges in each person's birth state and in states within 250 miles of the birth state from birth to age 70 (at 10 year intervals).

(Figure 6 here)

Figure 6 plots all the coefficients for the number of Ph.D. programs in birth states.[17] Consistent with Figure 5, the number of Ph.D. programs in the birth state around the time they were 30 years old is associated with a substantial increase in the probability of ever being an inventor. Estimates in Table A.8 relate patenting outcomes among people born in a state to the number of Ph.D. programs in that state and non-birth-states within 250 miles at age 30. Column (1) highlights that the presence of one additional Ph.D. program in the birth state around the graduate education completion age induces about 2 more people per million to obtain patents, while the effect of adding one program in nearby, non-birth-states (within 250 miles) at age 30 is positive but much smaller in magnitude and not statistically significant.

Utilizing the estimates of the effect of Ph.D. programs open during a person's graduate education years, both in their birth state and in nearby, non-birth states, on the probability of becoming an inventor, along with changes in the number of programs over time, allows us to recover an estimate of the overall impact of Ph.D. education expansion between 1875 and 1960 on innovation in the U.S. Based on the average number of Ph.D. programs per person at age 30 in the birth state and surrounding states within a 250-mile radius in 1875 and 1960, we estimate that the total effect of the increased number of Ph.D. programs, both in birth states and neighboring states, corresponds to about 90 more inventors per one million people (0.000090), equivalent to nearly 12 thousand inventors.[18] This overall impact accounts for approximately 6.8% of inventors among the 1940 U.S. population–around 1,331 people per million. While this effect may seem somewhat large, note that it includes both the direct effect coming from Ph.D. program receipt on patenting as well as spillover effects (which may be derived from agglomeration impacts, for example) in the locations surrounding universities where Ph.D. programs are expanding. At the same time, we note that the impact on patenting is smaller than the impact on Ph.D. program receipt, which may also relate to a low level of transferrable

[17] Coefficients for the number of nearby, non-birth-state Ph.D. programs are plotted in Figure A.4.

[18] To calculate the total effect, we multiply the coefficient for in-state programs from Column (1), Table A.8, by the increase in the number of in-state programs from 1875 to 1960: $2 \times 10^{-6} \times 37.081$. A similar calculation is performed for nearby state programs: $0.2 \times 10^{-6} \times 80.553$. The total effect is then obtained by summing these two components, yielding approximately 0.000090.

scientific knowledge between universities and the private sector. This is consistent with recent research suggesting there may be low levels of transfer of scientific knowledge across public and private sectors due to labor market frictions among the academic and private research labor force (Arora et al., 2023).

In particular, Column (2) suggests a null effect of Ph.D. program expansion on increasing patenting activities among Ph.D. recipients and Column (3) shows a positive effect on being an inventor conditional on not receiving a Ph.D. This is consistent with other evidence from this historical period when patent innovation may have required more rudimentary knowledge and skills not directly conferred by Ph.D. study. Thus, it appears that spillover effects from the availability of Ph.D. programs are larger than the direct effects.

The remaining columns in Table A.8 focus on the quantity and quality of patents. Column (4) shows a positive but insignificant effect on the number of patents ever granted. Columns (5) and (6) show positive effects of the number of in-state Ph.D. programs at age 30 on the total number of citations and the average number of citations per patent, respectively. Thus, Table A.8 provides evidence of strong spillover effects of the expansion of research education on patenting activities, particularly by increasing the number of inventors among non-Ph.D. recipients and improving the average quality of innovations.

(Table 5 here)

4.6.2 Results for High and Low Research Intensity Patents

To strengthen the link between research education and innovation, we investigate the relationship between academic area of Ph.D. training and patent types. To do this, we first classify all patents into low and high research intensity.[19] Among inventors with low-research intensity patents, 0.8% have earned Ph.D. degrees, while among those with high-research intensity patents, 3% have earned Ph.D. degrees. Therefore, among Ph.D. recipients, we expect stronger effects of Ph.D. programs on patenting in high-research intensity areas.

Table 5 reports estimates of the impact of Ph.D. program openings in the birth state at the peak ages of graduate study in four academic areas (Arts & Humanities, Social Behavioral Sciences, Mathematical Physical Sciences, and Natural Sciences) on the probability of patenting, the number of patents granted, the number of citations, and the average number of citations per patent in low- and high-research intensity areas.

[19]Specifically, based on the CPC classification: https://www.uspto.gov/web/patents/classification/cpc/html/cpc.html, we define as low research intensity, patents in "Human necessities", "Performing operations; Transporting", "Fixed constructions", and "Mechanical engineering; Lighting; Heating; Weapons; Blasting". We classify as high research intensity, patents in the other categories. These are "Chemistry; Metallurgy", "Textiles; Paper", "Physics", "Electricity", and "New technological developments; Cross-sectional technologies spanning over several sections of the IPC".

Among the four academic field areas, we find that the number of Ph.D. programs in Arts & Humanities, which is likely least relevant to innovation, is negatively related to all patenting outcomes. This is reassuring in that it suggests that access to any Ph.D. program *per se* is not associated with greater patenting. At the same time, Table 5 also suggests that the number of Ph.D. programs in Mathematical & Physical Sciences, which includes chemistry and physics, has a larger positive effect on the probability of patenting in high-research intensity areas than in low-research intensity areas. Table 5 also suggests that the number of Ph.D. programs in Mathematical and Physical Sciences has a relatively larger positive impact on patent quality measures in the high-research intensity patenting categories, as suggested by the larger estimates in the latter groups across Columns (5) through (8). Overall, these results suggest that Ph.D. programs have a spillover effect on inventors without Ph.D. degrees. They also promote innovation within fields, as evidenced by patenting in areas that rely more heavily on research training.

4.6.3 Patenting Heterogeneity by Demographic Group

One important contribution of this paper is to examine the impacts of Ph.D. program expansion on outcomes by demographic group. Thus, just as we did for the main results on Ph.D. receipt, we also explore whether Ph.D. program expansion affected patenting outcomes differentially for underrepresented individuals.

First, we separately estimate equation (1) by demographic subgroup for all outcomes in Table A.3 using only the 1940 Census. Second, we estimate the overall impacts for heterogeneous groups based on childhood characteristics using the linked-census sample (Table A.4 Panels (B) to (E)). Note that the latter estimates are based on a much smaller sample as the linked-census sample covers less than 50% of the 1940 national population. We find that men, White individuals, and children of U.S. natives are more likely to ever patent when there are more programs open at age 30 (Table A.3, Panel (B)). Although Table A.3 shows no statistically significant effects for some underrepresented groups, such as minorities (non-Whites) and second-generation immigrants, their sample sizes are also very small. As indicated, estimates for subgroups based on early childhood / parental characteristics are reported in Table A.4, which shows that individuals who lived in urban areas around age 10 and who had both low- and high-SES parents are statistically more likely to patent if they are exposed to Ph.D. program expansions at age 30 (Panel (B)). Moreover, Table A.4 does not suggest a statistically significant negative effect on any demographic group.

To explore relative effects across the majority and underrepresented groups along the extensive margin, we turn to regressions interacting our main Ph.D. program independent variables of interest with individual and family characteristics. While Table A.9 and Table A.10 generally show weak effects on the probability of ever patenting for individuals from

various underrepresented subgroups relative to the omitted majority groups, there is some indication that doctoral programs increase patenting for people from lower income backgrounds. In particular, the estimates on the interactions between in-state Ph.D. programs and Parental Occupational Income Score (Table A.10, Column 1) as well as Parental Occupational Earnings Score (Table A.10, Column 7), respectively, suggest that there may be a reduction in the gap between high- versus low-SES families in the probability of patenting in high research-intensity fields.

Table A.11 through Table A.16 investigate the relative impacts of Ph.D. program openings by demographic subgroups on the remaining patenting outcomes which capture the intensive margin and patent quality, including lifetime number of patents, lifetime number of citations, lifetime average citations per patent—both overall and in high and low research intensity areas. In general, results from these tables show a negative relative impact of Ph.D. program openings on innovative outcomes for women and individuals who grew up in rural areas, as well those who come from lower socioeconomic backgrounds, relative to the comparable majority group (i.e., men, and those from urban and higher socioeconomic backgrounds). While we also found a negative impact of Ph.D. program openings on the probability of women receiving a Ph.D. compared with men, the relative disadvantage for rural and lower SES backgrounds in these patenting outcomes contrasts with the results on Ph.D. attainment and may indicate that the spillover effects driving the patenting innovation outcomes disadvantage these already less advantaged groups.

Together, the results from Table A.10, Table A.12, Table A.14, and Table A.16 show that the potentially widened gaps between women and men, rural and urban residential locations around age 10, and low- and high-SES families are predominantly driven by low research-intensity patent fields, where Ph.D. receipt is less likely among inventors. In general, magnitudes of the relatively negative impacts for women are also higher in the low research intensity patent areas. Table A.12, Table A.14, and Table A.16 also show that individuals with high-SES parents are more positively affected by exposure to the number of Ph.D. programs at age 30, but only in low research-intensity fields. This difference between low- and high-research-intensity areas may reflect the fact that patenting in low-research-intensity fields is more often driven by spillover effects outside the direct channels of Ph.D. education, which are also less likely to benefit underrepresented groups, such as women and children from low-SES families.

Overall, our results exploring the heterogeneous impacts of Ph.D. program openings on patenting outcomes suggest that there are some similarities with the results for Ph.D. receipt and some differences. Our results on the extensive margin of patenting suggest reduced disparities in the probability of patenting when there are more Ph.D. program openings. This may be reflective of the direct channel of Ph.D. program openings on patenting operating through Ph.D. receipt, which would be consistent with

the results observed for high versus low research intensity patents, given that Ph.D. receipt is more likely in high research areas. At the same time, our results examining heterogeneous impacts on the intensive margin of patenting outcomes are suggestive of widening disparities, which may be linked to the spillover effects of Ph.D. program openings which operate outside of Ph.D. receipt.

5 Conclusion

We demonstrate that the expansion of research education in the United States at the beginning of the 20th century was not merely a scaling up of the existing enterprise but a significant transformation in terms of geography, subject matter, institutions, and participation. Moreover, opening programs increased and broadened participation in research. Specifically, we find that opening one additional Ph.D. program during the peak ages of graduate study results in nearly two more Ph.D. recipients per one million residents in a state. We also find evidence suggesting that the expansion of U.S. research education significantly improved access for non-White individuals, children of immigrants, rural residents, and individuals from lower socioeconomic status families, although men appear to have gained more from the expansion in Ph.D. programs than women. These effects are primarily driven by Land Grant universities, public institutions, and programs granting degrees to women and minorities, as well as emerging STEM fields.

Our study also provides potential evidence that expanding research education promotes innovation activities beyond academia. We present evidence of a spillover effect of opening new Ph.D. programs on promoting new inventors and improving the quality of patents. Additionally, we find a direct impact of Ph.D. program openings on high-research patenting among doctoral degree recipients, thus establishing a close link between research training and innovation. Moreover, the heterogeneous impacts of Ph.D. program openings on patenting outcomes suggest the possibility of widening disparities in patenting outcomes between majority and underrepresented groups, and our results indicate these differences may be more acute in the low research areas. Since Ph.D. receipt is less likely in low research intensity areas, this suggests spillover effects outside the direct channels of Ph.D. education may be less likely to benefit underrepresented groups. While this paper breaks new ground in examining the impacts of research training on the U.S. population, further research should explore the underlying mechanisms behind the heterogeneous impacts on Ph.D. versus patenting outcomes to better understand why some groups benefit more than others.

References

Abramitzky, R., Boustan, L., & Eriksson, K. (2019). To the New World and Back Again: Return Migrants in the Age of Mass Migration. *ILR Review*, *72*(2), 300–322. Publisher: SAGE Publications Inc.
URL https://doi.org/10.1177/0019793917726981

Abramitzky, R., Boustan, L., Eriksson, K., Pérez, S., & Rashid, M. (2020). Census linking project: Version 2.0 [dataset].
URL https://censuslinkingproject.org

Abramitzky, R., Boustan, L. P., & Eriksson, K. (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.
URL https://www.aeaweb.org/articles?id=10.1257/aer.102.5.1832

Abramitzky, R., Boustan, L. P., & Eriksson, K. (2014). A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration. *Journal of Political Economy*, *122*(3), 467–506. Publisher: The University of Chicago Press.
URL https://www.journals.uchicago.edu/doi/full/10.1086/675805

Aghion, P., Akcigit, U., Hyytinen, A., & Toivanen, O. (2017). The Social Origins of Inventors.
URL https://www.nber.org/papers/w24110

Aghion, P., Boustan, L., Hoxby, C., & Vandenbussche, J. (2009). The causal impact of education on economic growth: evidence from us. *Brookings papers on economic activity*, *1*(1), 1–73.

Alesina, A., & Ferrara, E. L. (2005). Ethnic Diversity and Economic Performance. *Journal of Economic Literature*, *43*(3), 762–800.
URL https://www.aeaweb.org/articles?id=10.1257/002205105774431243

Arora, A., Belenzon, S., Cioaca, L., Sheer, L., & Zhang, H. (2023). The Effect of Public Science on Corporate R&D.
URL https://papers.ssrn.com/abstract=4645071

Azoulay, P., Greenblatt, W. H., & Heggeness, M. L. (2021). Long-term effects from early exposure to research: Evidence from the NIH "Yellow Berets". *Research Policy*, *50*(9), 104332.
URL https://www.sciencedirect.com/science/article/pii/S0048733321001311

Bell, A., Chetty, R., Jaravel, X., Petkova, N., & Van Reenen, J. (2019). Who Becomes an Inventor in America? The Importance of Exposure to Innovation*. *The Quarterly Journal of Economics*, *134*(2), 647–713.
URL https://doi.org/10.1093/qje/qjy028

Berkes, E., & Nencka, P. (2024). Knowledge Access: The Effects of Carnegie Libraries on Innovation. *The Review of Economics and Statistics*, (pp. 1–45).
URL https://doi.org/10.1162/rest_a_01446

Borjas, G. J., & Doran, K. B. (2012). The Collapse of the Soviet Union and the Productivity of American Mathematicians*. *The Quarterly Journal of Economics*, *127*(3), 1143–1203.
URL https://doi.org/10.1093/qje/qjs015

Bound, J., & Turner, S. (2002). Going to War and Going to College: Did World War II and the G.I. Bill Increase Educational Attainment for Returning Veterans? *Journal of Labor Economics*, *20*(4), 784–815. Publisher: The University of Chicago Press.
URL https://www.journals.uchicago.edu/doi/full/10.1086/342012

Card, D., & Giuliano, L. (2014). Does Gifted Education Work? For Which Students?
URL https://www.nber.org/papers/w20453

Chetty, R., Friedman, J. N., Saez, E., Turner, N., & Yagan, D. (2017). Mobility Report Cards: The Role of Colleges in Intergenerational Mobility.
URL https://www.nber.org/papers/w23618

Clay, K., Lingwall, J., & Stephens, J., Melvin (2012). Do Schooling Laws Matter? Evidence from the Introduction of Compulsory Attendance Laws in the United States.
URL https://www.nber.org/papers/w18477

Davison, C., Doran, K., & Yoon, C. (2023). Compulsory innovation: How education forms inventors.
URL https://colindavison.github.io/working_papers/compulsory_innovation/

Dossi, G. (2024). Race and Science.
URL https://www.dropbox.com/scl/fi/cbe2tu5f2us9wde3wl1i9/JMP_Dossi.pdf?dl=0&rlkey=jce2ug3kq3mfjgaxr0drvwnq6

Fry, R., Kennedy, B., & Funk, C. (2021). STEM Jobs See Uneven Progress in Increasing Gender, Racial and Ethnic Diversity.
URL https://www.pewresearch.org/social-trends/2021/04/01/stem-jobs-see-uneven-progress-in-increasing-gender-racial-and-ethnic-diversity/

Geiger, R. L. (1986). *To Advance Knowledge: The Growth of American Research Universities, 1900-1940*. Oxford University Press. Google-Books-ID: 2PhnqzZR5HIC.

Goldin, C., & Katz, L. F. (1997). Why the United States Led in Education: Lessons from Secondary School Expansion, 1910 to 1940.
URL https://www.nber.org/papers/w6144

Gordon, E. K., Hawley, Z. B., Kobler, R. C., & Rork, J. C. (2021). The Paradox of HBCU Graduation Rates. *Research in Higher Education*, *62*(3), 332–358.
URL https://doi.org/10.1007/s11162-020-09598-5

Hager, S., Schwarz, C., & Waldinger, F. (2024). Measuring Science: Performance Metrics and the Allocation of Talent. *American Economic Review*, *114*(12), 4052–4090.
URL https://www.aeaweb.org/articles?id=10.1257/aer.20230515

Hofstra, B., Kulkarni, V. V., Munoz-Najar Galvez, S., He, B., Jurafsky, D., & McFarland, D. A. (2020). The Diversity–Innovation Paradox in Science. *Proceedings of the National Academy of Sciences*, *117*(17), 9284–9291. Publisher: Proceedings of the National Academy of Sciences.
URL https://www.pnas.org/doi/abs/10.1073/pnas.1915378117

Hoxby, C., & Avery, C. (2013). The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students. *Brookings Papers on Economic Activity*, *44*(1 (Spring)), 1–65. Publisher: Economic Studies Program, The Brookings Institution.
URL https://ideas.repec.org//a/bin/bpeajo/v44y2013i2013-01p1-65.html

Iaria, A., Schwarz, C., & Waldinger, F. (2024). Gender Gaps in Academia: Global Evidence Over the Twentieth Century.
URL https://papers.ssrn.com/abstract=4150221

Jones, B. F. (2010). Age and Great Invention. *The Review of Economics and Statistics*, *92*(1), 1–14.
URL https://doi.org/10.1162/rest.2009.11724

Kantor, S., & Whalley, A. (2014). Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks. *The Review of Economics and Statistics*, *96*(1), 171–188. Publisher: The MIT Press.
URL https://www.jstor.org/stable/43554921

Khan, B. Z., & Sokoloff, K. L. (1993). "Schemes of Practical Utility": Entrepreneurship and Innovation Among "Great Inventors" in the United States, 1790–1865. *The Journal of Economic History*, *53*(2), 289–307.
URL https://www.cambridge.org/core/journals/journal-of-economic-

history/article/schemes-of-practical-utility-entrepreneurship-and-innovation-among-great-inventors-in-the-united-states-17901865/4AA08392C755E038FE042E64893DF563#

Kim, S. D., & Moser, P. (2021). Women in Science. Lessons from the Baby Boom. URL https://www.nber.org/papers/w29436

Lamoreaux, N. R., & Sokoloff, K. L. (1999). Inventive Activity and the Market for Technology in the United States, 1840-1920. URL https://www.nber.org/papers/w7107

MacLeod, W. B., & Urquiola, M. (2021). Why Does the United States Have the Best Research Universities? Incentives, Resources, and Virtuous Circles. *Journal of Economic Perspectives*, *35*(1), 185–206. URL https://www.aeaweb.org/articles?id=10.1257/jep.35.1.185

Moser, P., & San, S. (2020). Immigration, Science, and Invention. Lessons from the Quota Acts. URL https://papers.ssrn.com/abstract=3558718

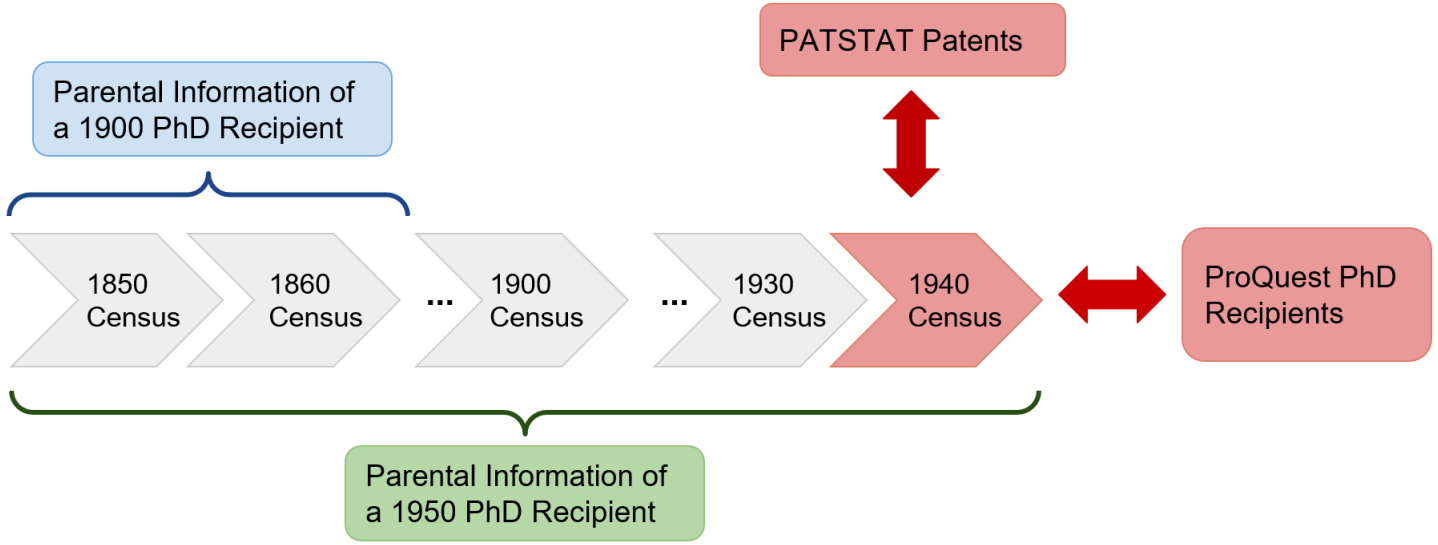
Moser, P., Voena, A., & Waldinger, F. (2014). German Jewish Émigrés and US Invention. *American Economic Review*, *104*(10), 3222–3255. URL https://www.aeaweb.org/articles?id=10.1257/aer.104.10.3222

Perna, L. W. (2001). The Contribution of Historically Black Colleges and Universities to the Preparation of African Americans for Faculty Careers. *Research in Higher Education*, *42*(3), 267–294. URL https://doi.org/10.1023/A:1018869922415

Toivanen, O., & Väänänen, L. (2016). Education and Invention. *The Review of Economics and Statistics*, *98*(2), 382–396. Publisher: The MIT Press. URL https://www.jstor.org/stable/43830355

Østergaard, C. R., Timmermans, B., & Kristinsson, K. (2011). Does a different view create something new? The effect of employee diversity on innovation. *Research Policy*, *40*(3), 500–509. URL https://www.sciencedirect.com/science/article/pii/S0048733310002398

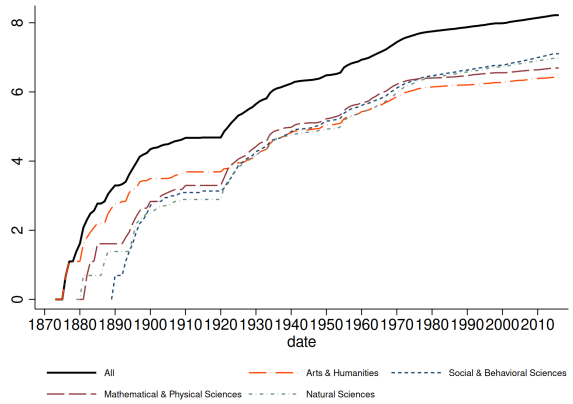
Figure 1: Data linkage schematic



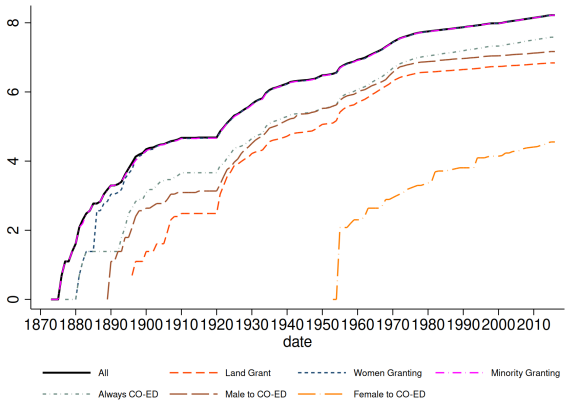
Note: This figure illustrates how we link the full count U.S. censuses with ProQuest Ph.D. recipients and PATSTAT patents.

Figure 2: Trends in the number Ph.D. programs in the U.S.

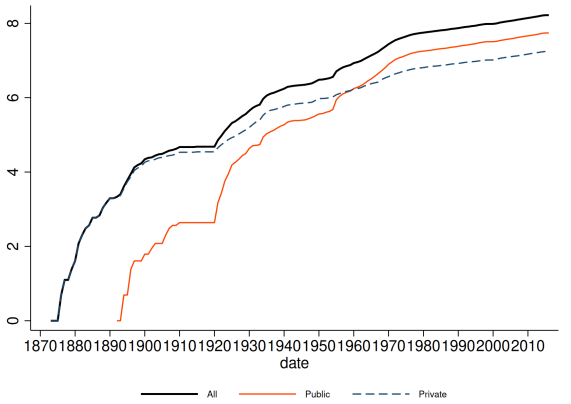
(a) By academic field



(b) By educational institutions/policies



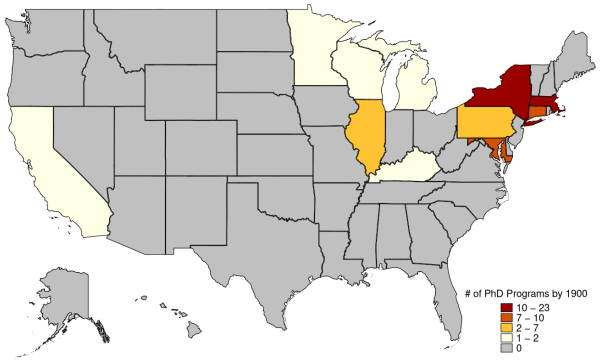
(c) By public VS private



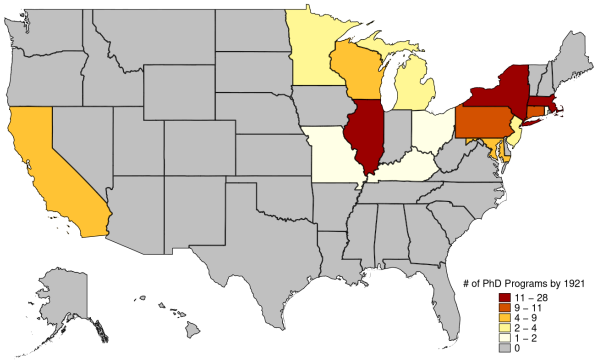
Note: Data come from the ProQuest Dissertation & Theses data. This figure plots the log number of Ph.D. programs in the U.S. overall and by sub-categories.

Figure 3: Distribution of Ph.D. programs across U.S. states over time

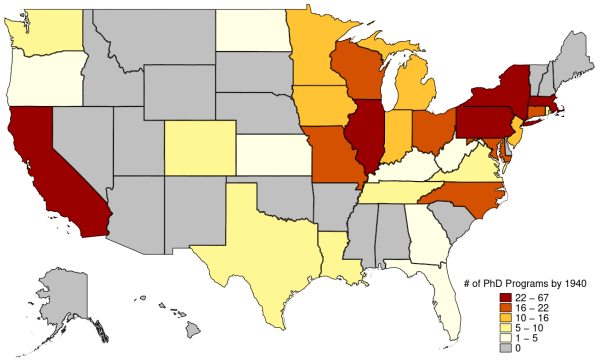
(a) By 1900



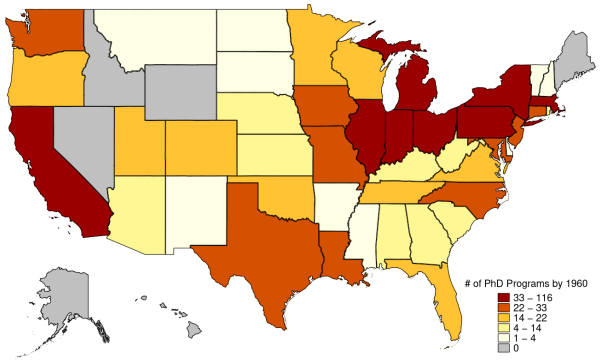
(b) By 1921



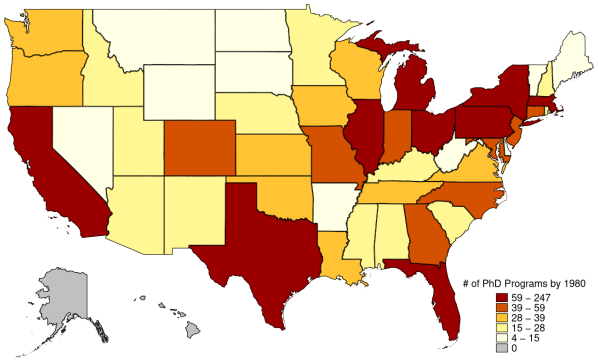
(c) By 1940



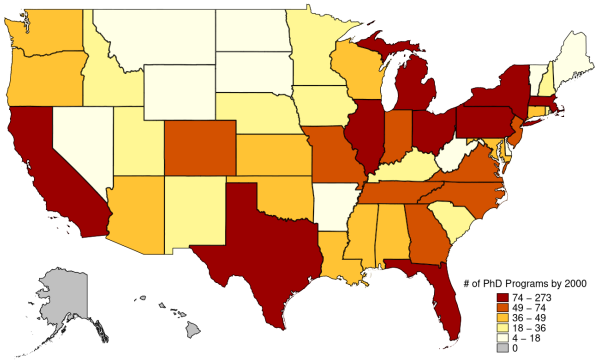
(d) By 1960



(e) By 1980

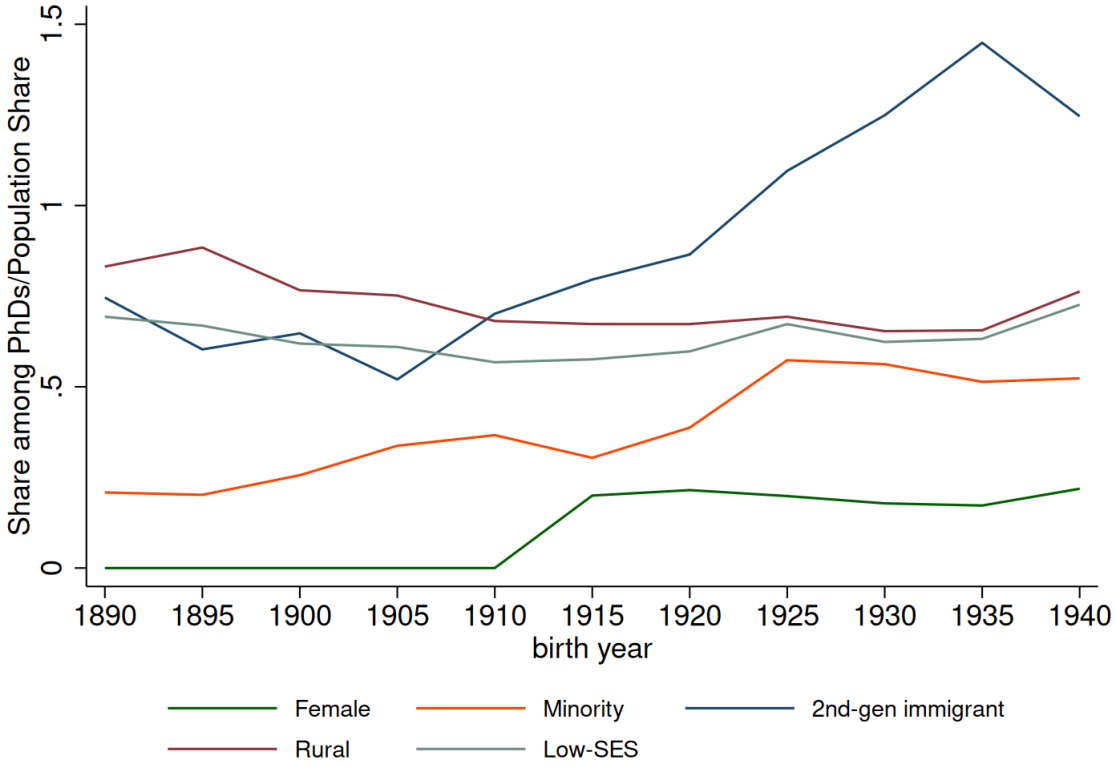


(f) By 2000



Note: Data come from the ProQuest Dissertation & Theses data. This figure presents the number of Ph.D. programs in each U.S. state by year.

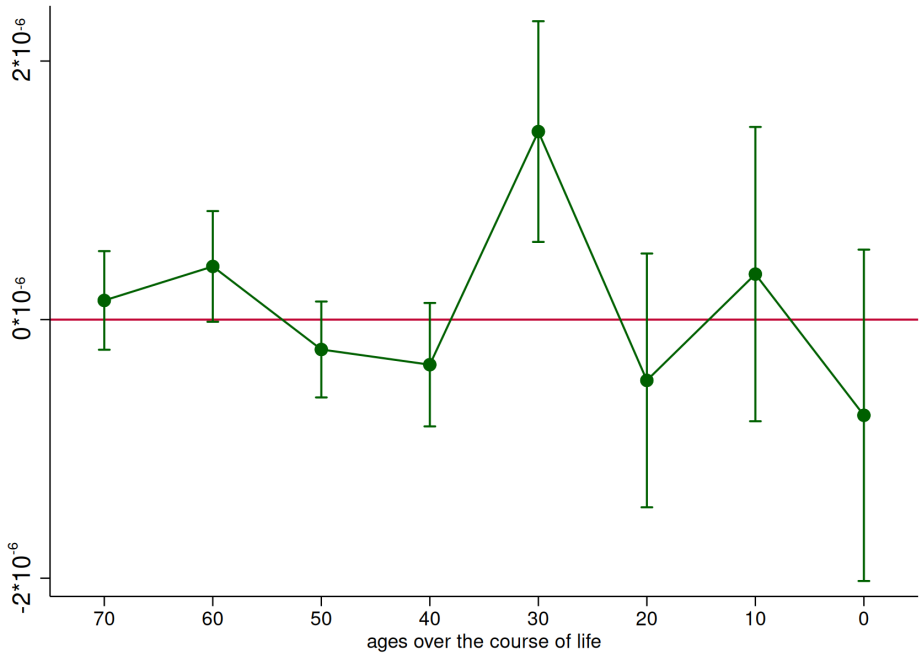
Figure 4: Representation of groups among Ph.D. recipients



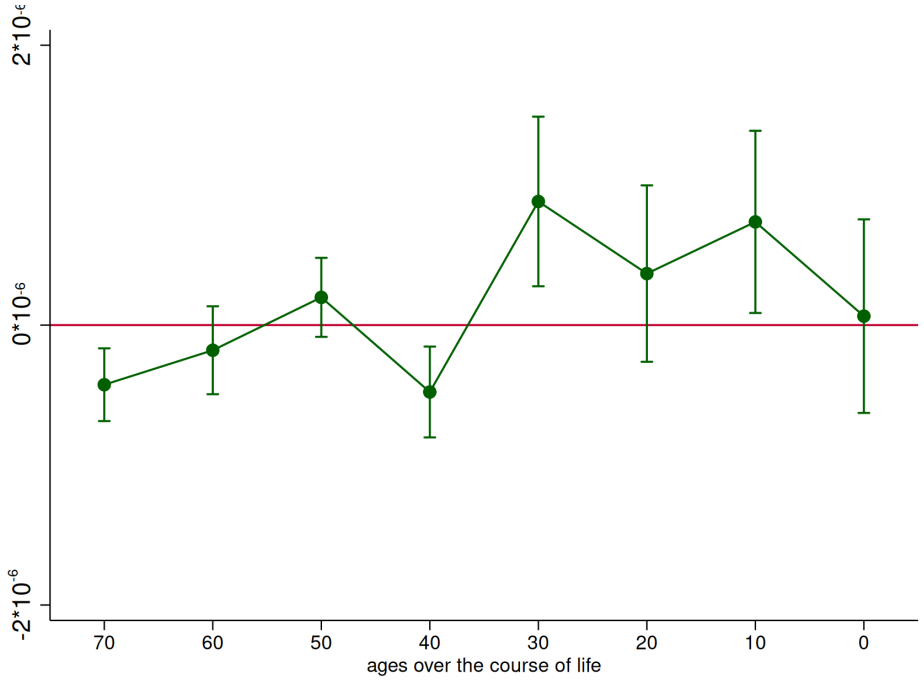
Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This figure plots the share of each demographic group among all the ProQuest Ph.D. recipients relative to the 1940 US population share.

Figure 5: Effect of the number of Ph.D. programs on Ph.D. receipt

(a) Ph.D. programs in birth state

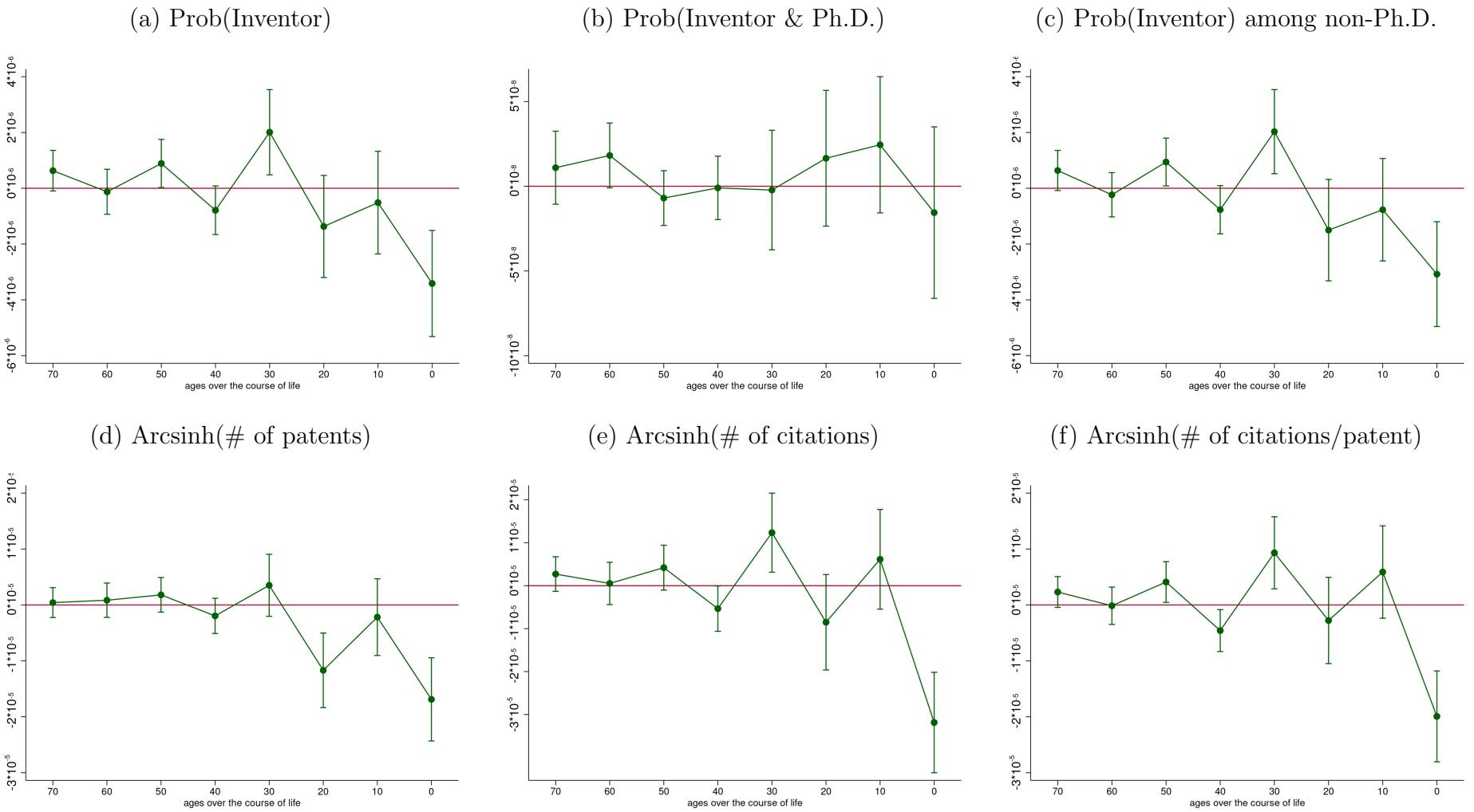


(b) Ph.D. programs in nearby states



Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This figure plots the estimated coefficients (marginal effects) on Ph.D. receipt of the number of Ph.D. programs in the birth state (Panel (a)) and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth, (Panel (b)) from birth to age 70 at 10 year intervals. All coefficients in Panel (a) and Panel (b) are estimated in one equation. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level and presented with 95% confidence intervals.

Figure 6: Effects of Ph.D. programs on patent outcomes: In-state programs



Note: Data come from the linked Census-PATSTAT innovator sample. This figure plots the estimated coefficients (marginal effects) on patent outcomes of the number of Ph.D. programs in the birth state from birth to age 70 (at 10 year intervals). We control for individual characteristics, including gender, race, urban residence, and nativity indicators, the total number of colleges in the birth state and in nearby states other than the state of birth (i.e., within 250 miles) from birth to age 70 (at 10 year intervals), birth state fixed effects, and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level and presented with 95% confidence intervals.

Table 1: Summary demographic statistics

	Mean	S.D.	Mean	S.D.	Mean	S.D.
	1940 population born in US (N=120,033,290)		Matched Ph.D. recipients (N=102,639)		Matched inventors (N=159,709)	
Ph.D. recipient	0.00086	0.02923			0.01	0.11
Inventor	0.00133	0.03645	0.02	0.14		
Male	0.50	0.50	0.88	0.32	0.93	0.26
White	0.89	0.31	0.95	0.21	0.97	0.19
Urban residence	0.53	0.50	0.70	0.46	0.62	0.49
2nd-gen immigrant	0.09	0.29	0.14	0.35	0.11	0.31

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample and the linked Census-PATSTAT innovator sample. This table summarizes demographic characteristics for the full count U.S. population in 1940, linked Ph.D. recipients, and linked inventors.

Table 2: Summary statistics of patenting outcomes

	Mean	S.D.	Min	Max
# of patents	3.70	4.74	1	39
# of citations	21.82	40.24	0	346
# of citations per patent	5.60	7.25	0	291

Note: Data come from the linked Census-PATSTAT innovator sample. This table summarizes individual level patenting outcomes among our linked inventors. We trim extreme outcomes by excluding the top 1%.

Table 3: Effect of Ph.D. programs on Ph.D. receipt by demographic groups

	(1)	(2)	(3)	(4)	(5)
	Prob(earning a Ph.D. in life)				
# of in-state programs at age 30*Female	-4.5**	-4.3**	-4.4**	-4.5**	-4.4**
	(2.2)	(2.2)	(2.2)	(2.3)	(2.2)
# of in-state programs at age 30*Minority	9.0***	8.6***	9.5***	8.4***	8.9***
	(3.0)	(2.9)	(2.9)	(3.1)	(2.9)
# of in-state programs at age 30*2nd-gen immigrant	5.4***	4.8**	5.2***	5.5***	5.4***
	(1.9)	(1.9)	(1.9)	(2.0)	(1.9)
# of in-state programs at age 30*Rural childhood neighborhood	2.0	1.9	2.4*	1.1	2.2*
	(1.3)	(1.2)	(1.3)	(1.3)	(1.2)
# of in-state programs at age 30*Parental Occupational Income Score	-7.4				
	(5.0)				
# of in-state programs at age 30*Parental Duncan Socioeconomic Index		-4.7*			
		(2.6)			
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)			-4.6		
			(5.0)		
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis				-5.6**	
				(2.2)	
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis					-1.0
					(2.5)
Observations	56,155,926	56,155,926	56,155,926	56,118,273	56,118,273

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table 4: Effects of Ph.D. programs on Ph.D. receipt by type of program

	(1) Full sample	(2) Female	(3) Non-White
# of in-state programs at age 30: Land Grant	2.2**	-0.5	3.0*
	(0.9)	(0.9)	(1.7)
# of in-state programs at age 30: others	1.2**	0.3	1.8*
	(0.5)	(0.5)	(0.9)
# of in-state programs at age 30: CO-ED colleges	0.8*	-0.2	2.1**
	(0.5)	(0.5)	(0.9)
# of in-state programs at age 30: others	1.6**	0.7	2.4
	(0.8)	(0.8)	(1.5)
# of in-state programs at age 30: institutions granting females Ph.D.s	1.6***	0.6	1.5*
	(0.5)	(0.4)	(0.8)
# of in-state programs at age 30: others	-1.4	-0.5	0.6
	(1.9)	(1.6)	(2.8)
# of in-state programs at age 30: institutions granting minority Ph.D.s	0.9*	0.2	1.9**
	(0.5)	(0.4)	(0.8)
# of in-state programs at age 30: others	2.5	1.4	3.1
	(2.1)	(1.9)	(4.0)
# of in-state programs at age 30: public	2.0***	-0.5	2.1**
	(0.6)	(0.6)	(1.0)
# of in-state programs at age 30: private	1.6***	1.2**	2.0
	(0.6)	(0.6)	(1.3)
# of in-state programs at age 30: STEM	-0.2	0.2	2.1
	(0.9)	(0.8)	(1.6)
# of in-state programs at age 30: non-STEM	2.8***	0.7	2.0
	(1.0)	(0.9)	(1.6)
Observations	115,004,798	57,526,406	12,495,126

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state for various types of Ph.D. programs. Column (1) uses the full 1940 population sample, while Columns (2) and (3) are specifically for women and non-Whites. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

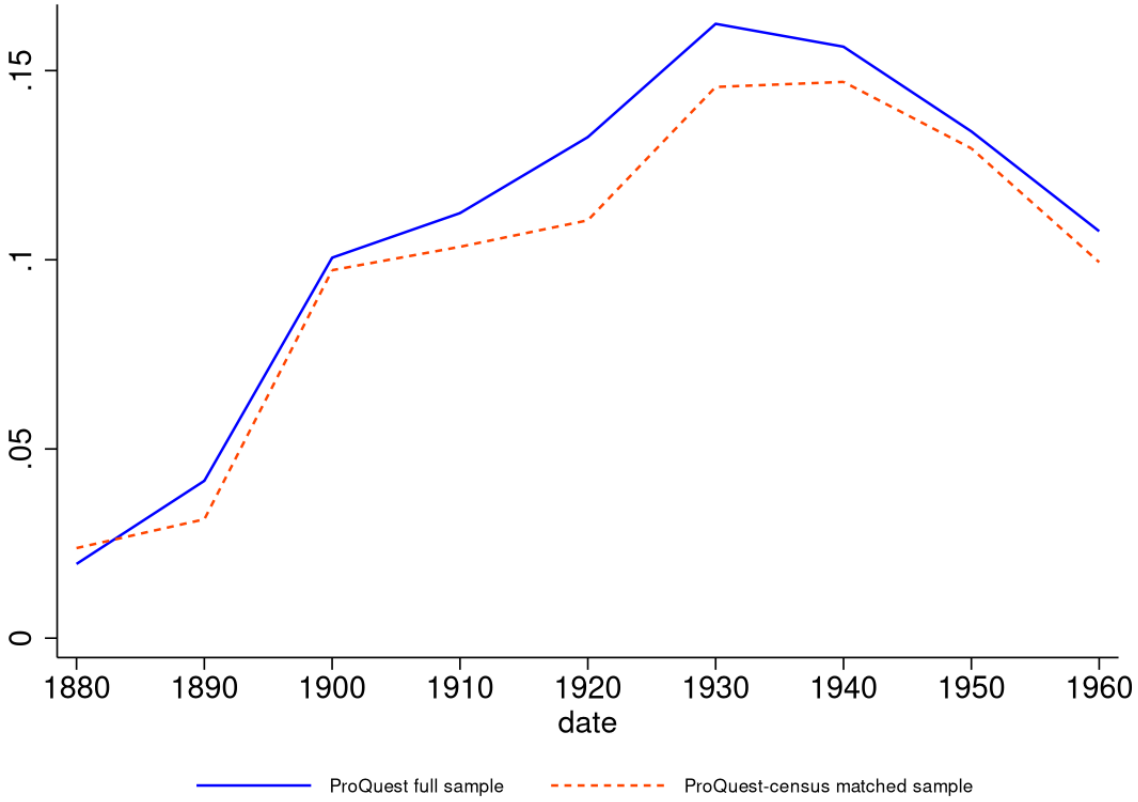
Table 5: Effects of Ph.D. programs on patent outcomes among Ph.D. recipients by patent classification

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Prob(Inventor)		Arcsinh(# of patents)		Arcsinh(# of citations)		Arcsinh(# of citations/patent)	
	Low research	High research	Low research	High research	Low research	High research	Low research	High research
# of in-state programs at age 30: Arts and Humanities	-0.0005173	-0.0009277*	-0.0014511*	-0.0023161	-0.0029460*	-0.0027872	-0.0019627	-0.0015773
	(0.0003456)	(0.0004796)	(0.0007687)	(0.0014294)	(0.0015645)	(0.0023832)	(0.0012182)	(0.0014942)
# of in-state programs at age 30: Social and Behavioral Sciences	0.0001517	-0.0006093	0.0007161	-0.0008538	0.0011069	-0.0018957	0.0006216	-0.0014856
	(0.0003548)	(0.0004699)	(0.0007265)	(0.0013343)	(0.0015233)	(0.0022191)	(0.0011962)	(0.0013956)
# of in-state programs at age 30: Mathematical and Physical Sciences	0.0004727*	0.0008907**	0.0007864	0.0005919	0.0009083	0.0012156	0.0004528	0.0011724
	(0.0002693)	(0.0003745)	(0.0006007)	(0.0010950)	(0.0011895)	(0.0018239)	(0.0009027)	(0.0011351)
# of in-state programs at age 30: Natural Sciences	-0.0000154	-0.0000334	0.0002924	0.0011764	0.0012768	0.0017565	0.0009505	0.0007323
	(0.0003692)	(0.0004977)	(0.0008251)	(0.0013621)	(0.0016187)	(0.0023329)	(0.0012199)	(0.0014808)
Observations	100,724	100,338	101,266	101,266	101,266	101,266	101,266	101,266
R-squared	-	-	0.0048	0.0073	0.0045	0.0073	0.0044	0.0073

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state separately for four academic areas: Arts Humanities, Social Behavioral Sciences, Mathematical Physical Sciences, and Natural Science. It only uses the ProQuest-census linked sample. The outcomes are the probability of ever patenting, lifetime patents, lifetime citations, and average number of citations per patent. (For the latter three, we use an in inverse hyperbolic sine transformation.) We split patents into two groups. Low research patent classifications are: "Human necessities", "Performing operations; Transporting", "Fixed constructions", and "Mechanical engineering; Lighting; Heating; Weapons; Blasting." Among inventors in these classifications, 0.8% have earned Ph.D. degrees. High research classifications are: "Chemistry; Metallurgy", "Textiles; Paper", "Physics", "Electricity", and "New technological developments; Cross-sectional technologies spanning over several sections of the IPC." 3% of the inventors in these classifications have earned Ph.D. degrees. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, the number of colleges in the birth state and nearby states other than the state of birth from birth to age 70 (at 10 year intervals), birth state fixed effects, and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Appendix A

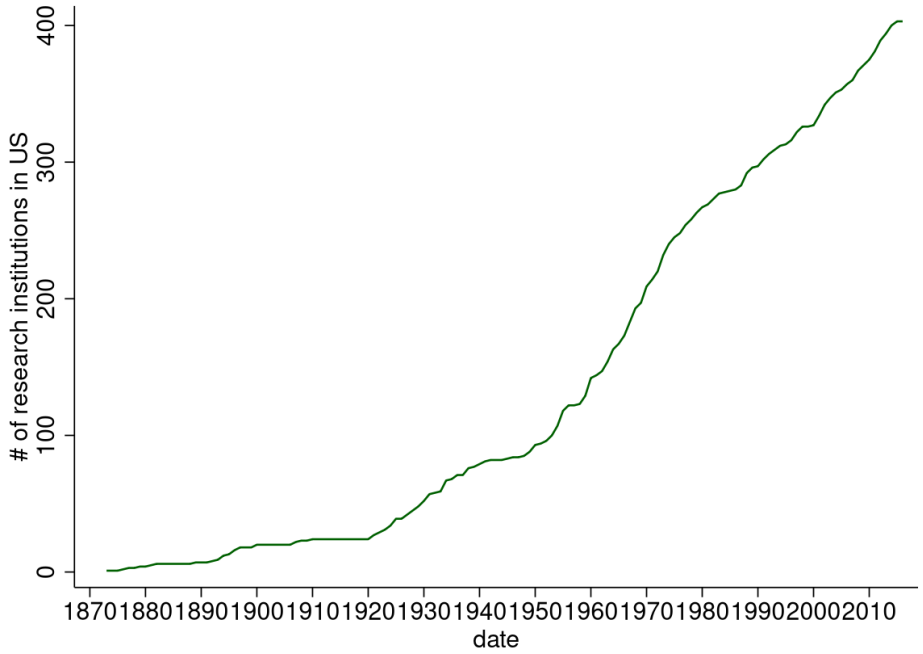
Figure A.1: Trends in the share of female Ph.D. recipients



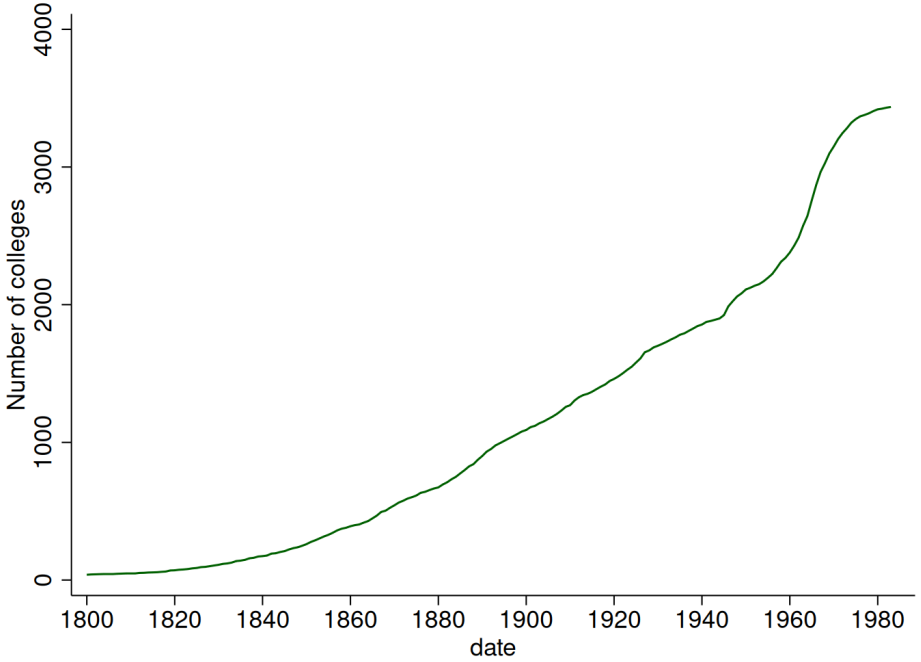
Note: Data come from the ProQuest Dissertation & Theses full sample and the linked Census-ProQuest Ph.D. recipient sample. This figure plots trends in the share of female Ph.D. recipients using both the ProQuest full sample (using names to impute genders) and in the linked Census-ProQuest Ph.D. recipient sample.

Figure A.2: Trends in the number of educational institutions in the U.S.

(a) Research universities



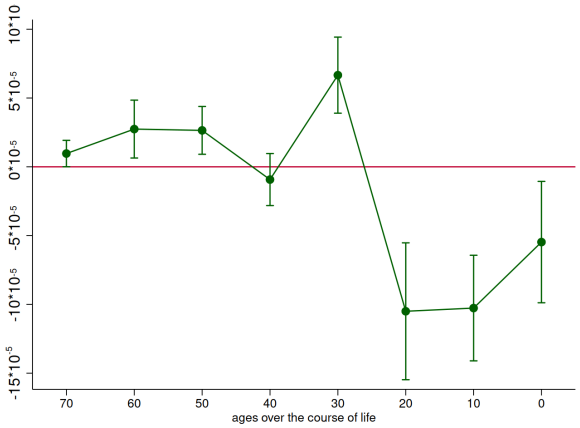
(b) Colleges



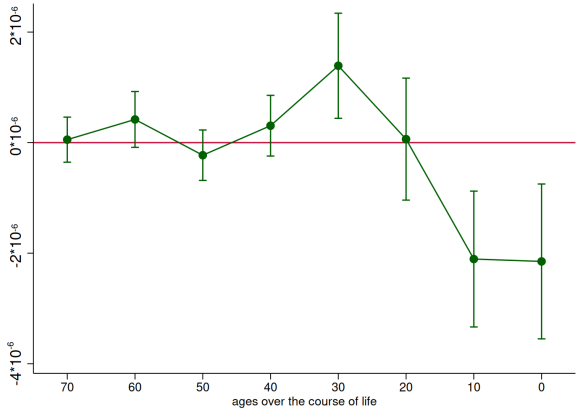
Note: Data come from the ProQuest Dissertation & Theses data and the Integrated Postsecondary Education Data System. This figure plots trends in the national number of research institutions (institutions offering Ph.D. programs) in Panel (a) and colleges in Panel (b).

Figure A.3: Effects of Ph.D. programs on Ph.D. receipt: alternative specifications and samples

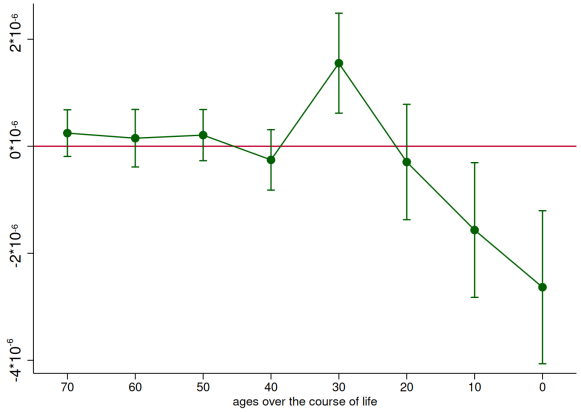
(a) Earning a graduate degree (1940 census above age 40)



(b) Earning a Ph.D. (without matching current distance to university)

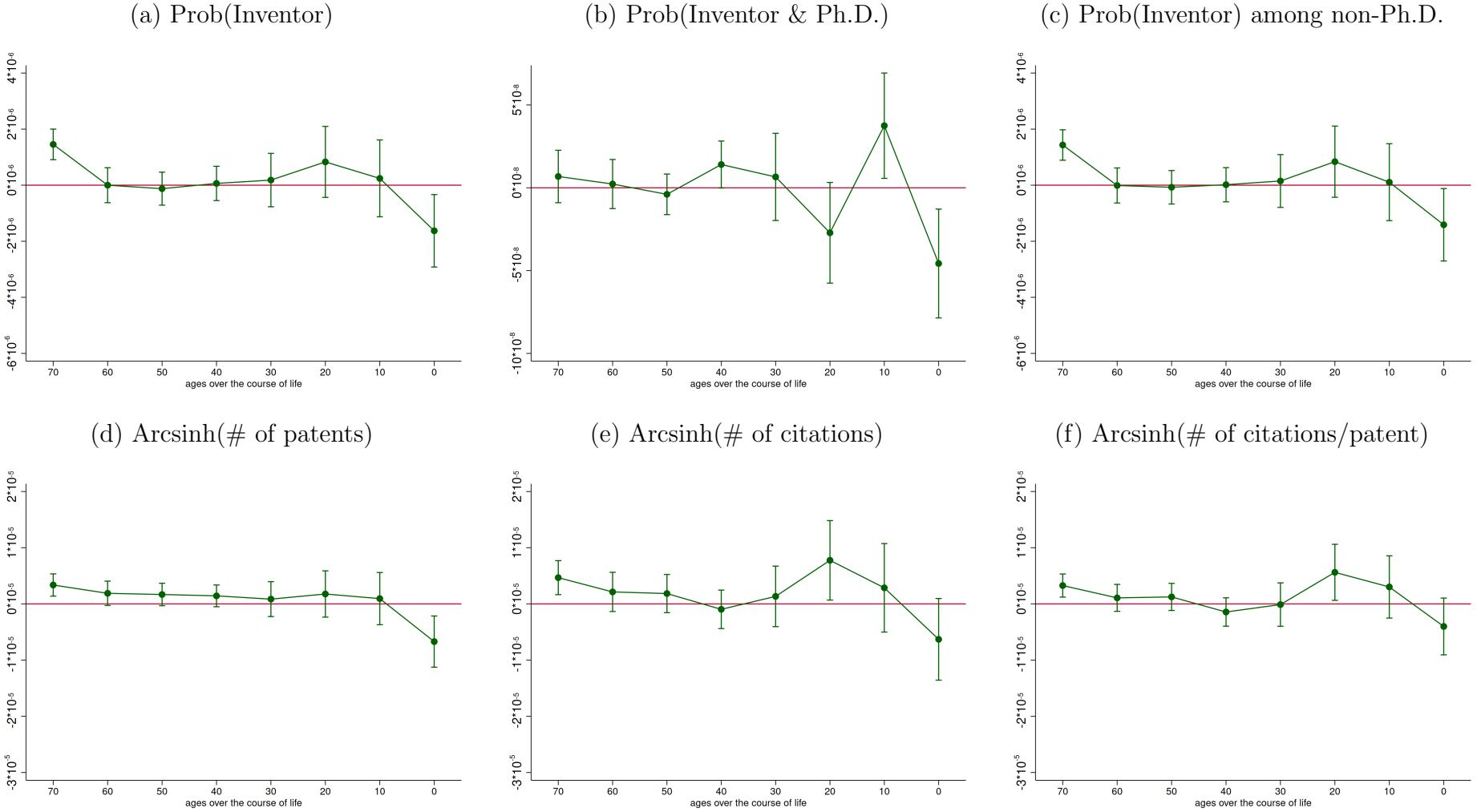


(c) Earning a Ph.D. (without matching degree age)



Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This figure plots estimated coefficients (marginal effects) on Ph.D. receipt of the number of Ph.D. programs in the birth state from birth to age 70 at 10 year intervals. In Panel (a), the outcome is a dummy for having some post-college education as measured in the 1940 census. We modify the variables we use to link ProQuest to the Census by dropping distance from the current residence to the degree university (in Panel (b)) and by dropping implied age at Ph.D. (in Panel (c)). Individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects are controlled. Standard errors are clustered at the birth-state-cohort level and presented with 95% confidence intervals.

Figure A.4: Effects of Ph.D. programs on patent outcomes: Nearby, out-of-state programs



Note: Data come from the linked Census-PATSTAT innovator sample. This figure plots the estimated coefficients (marginal effects) on patent outcomes of the number of Ph.D. programs in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth, from birth to age 70 at 10 year intervals. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, the total number of colleges in the birth state and in nearby states other than the state of birth (i.e., within 250 miles) from birth to age 70 (at 10 year intervals), birth state fixed effects, and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level and presented with 95% confidence intervals.

Table A.1: ProQuest field crosswalk

Fine Field	Aggregated Major
Language and Literature	Arts & Humanities
Law and Legal Studies	Arts & Humanities
Fine and Performing Arts	Arts & Humanities
Philosophy and Religion	Arts & Humanities
History	Arts & Humanities
Social Sciences	Social & Behavioral Sciences
Education	Social & Behavioral Sciences
Communication and Information Sciences	Social & Behavioral Sciences
Area, Ethnic, and Gender Studies	Social & Behavioral Sciences
Business	Social & Behavioral Sciences
Behavioral Sciences	Social & Behavioral Sciences
Mathematics	Mathematical & Physical Sciences
Engineering	Mathematical & Physical Sciences
Physical Sciences	Mathematical & Physical Sciences
Chemistry	Mathematical & Physical Sciences
Agriculture	Natural Sciences
Ecosystem Sciences	Natural Sciences
GeoSciences	Natural Sciences
Architecture	Natural Sciences
Environmental Sciences	Natural Sciences
Health and Medical Sciences	Natural Sciences
Biological Sciences	Natural Sciences

Note: Data come from the ProQuest Dissertation & Theses data. This table lists the ProQuest fine fields and the respective grouped aggregated majors.

Table A.2: Effects of Ph.D. programs on Ph.D. receipt

	(1)	(2)	(3)	(4)
	Prob(earning a Ph.D. in life)			
# of in-state programs at age 30	1.5***	1.3***	1.7***	1.6***
	(0.4)	(0.5)	(0.4)	(0.4)
# of nearby state programs within 250 miles at age 30	0.9***	1.4***	0.4	0.6*
	(0.3)	(0.3)	(0.3)	(0.3)
Number of colleges		Y		
Birth Region Trend			Y	
Birth Division Trend				Y
Observations	115,004,798	115,004,798	115,004,798	115,004,798

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) on Ph.D. receipt of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects in all columns. Column (1) reports the baseline estimates. Additionally, we control for the total number of colleges in birth state and nearby states other than the state of birth from birth to age 70 (at 10 year intervals) in Column (2), birth-region cohort trends in Column (3), and birth-division cohort trends in Column (4). All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.3: Effects of Ph.D. programs on Ph.D. attainment and patent outcomes by demographic groups, using only the 1940 census

	(1) Female	(2) Male	(3) Minority	(4) White	(5) 2nd-gen immigrant	(6) Native
			A: Prob(earning a Ph.D. in life)			
Mean	198.0	1,515.0	381.0	913.0	1,310.0	808.0
# of in-state programs at age 30	0.5	3.7***	1.9**	1.5***	2.1	1.0**
	(0.4)	(1.2)	(0.8)	(0.5)	(1.9)	(0.4)
# of nearby state programs within 250 miles at age 30	0.1	2.5***	-0.7	1.2***	1.0	1.0***
	(0.3)	(0.9)	(0.4)	(0.4)	(1.4)	(0.3)
Observations	57,526,406	57,478,392	12,495,126	102,424,531	11,182,775	103,793,129
			B: Prob(Inventor)			
Mean	188.0	2,478.0	436.0	1,439.0	1,495.0	1,314.0
# of in-state programs at age 30	0.5	6.7***	0.6	2.4***	2.1	1.7**
	(0.6)	(2.5)	(1.8)	(0.9)	(2.1)	(0.8
Observations	57,526,328	57,476,049	12,579,780	102,422,185	11,211,339	103,791,038
			C: Arcsinh(# of patents)			
Mean	281.0	4,117.0	693.0	2,377.0	2,503.0	2,163.0
# of in-state programs at age 30	0.3	7.4	1.0	5.8*	9.7	2.1
	(1.0)	(5.5)	(4.6)	(3.0)	(7.1)	(2.8)
Observations	57,526,328	57,476,049	12,580,192	102,422,185	11,211,339	103,791,038
R-squared	0.0000	0.0005	0.0002	0.0011	0.0010	0.0011
			D: Arcsinh(# of citations)			
Mean	456.0	6,663.0	1,165.0	3,842.0	4,708.0	3,432.0
# of in-state programs at age 30	0.1	26.1***	1.5	13.8***	18.1	10.3**
	(1.7)	(9.2)	(8.5)	(5.1)	(13.1)	(4.5)
Observations	57,526,328	57,476,049	12,580,192	102,422,185	11,211,339	103,791,038
R-squared	0.0000	0.0003	0.0002	0.0009	0.0010	0.0009
			E: Arcsinh(# of citations/patent)			
Mean	337.0	4,651.0	843.0	2,689.0	3,421.0	2,392.0
# of in-state programs at age 30	0.2	19.6***	1.1	9.8***	14.0	7.8**
	(1.3)	(6.4)	(6.2)	(3.6)	(9.4)	(3.1)
Observations	57,526,328	57,476,049	12,580,192	102,422,185	11,211,339	103,791,038
R-squared	0.0000	0.0003	0.0002	0.0009	0.0011	0.0009

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample and the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth. The outcome variable is the probability of earning a Ph.D. degree (Panel (A)), the probability of ever patenting (Panel (B)), the lifetime number of patents (Panel (C)), the lifetime number of citations (Panel (D)), and the lifetime average number of citations per patent (Panel (E)) (all using in inverse hyperbolic sine transformation in Panels (C) to (E)). We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. We also control for the number of colleges in the birth state and in nearby states other than the state of birth for all patent outcomes. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.4: Effects of Ph.D. programs on Ph.D. attainment and patent outcomes by family characteristics, using the census-linked sample

	(1) Rural childhood	(2) Urban childhood	(3) Low parental SES	(4) High parental SES
		A: Prob(earning a Ph.D. in life)		
Mean	823.0	1,661.0	715.0	1,588.0
# of in-state programs at age 30	0.1	0.5	0.9	0.2
	(0.8)	(1.0)	(0.9)	(0.9)
# of nearby state programs within 250 miles at age 30	-0.4	1.8**	-0.9*	1.9***
	(0.5)	(0.8)	(0.5)	(0.7)
Observations	30,168,220	25,987,706	26,128,852	30,027,074
		B: Prob(Inventor)		
Mean	1,609.0	1,996.0	1,445.0	2,048.0
# of in-state programs at age 30	0.5	4.2**	2.7*	3.1*
	(1.6)	(1.7)	(1.6)	(1.6)
Observations	30,167,461	25,986,698	26,128,325	30,025,834
		C: Arcsinh(# of patents)		
Mean	2,610.0	3,366.0	2,337.0	3,431.0
# of in-state programs at age 30	2.4	-1.8	-7.8	3.9
	(9.6)	(5.3)	(5.6)	(6.9)
Observations	30,167,461	25,986,698	26,128,325	30,025,834
R-squared	0.0011192	0.0016838	0.0009524	0.0016946
		D: Arcsinh(# of citations)		
Mean	4,295.0	5,992.0	3,788.0	6,060.0
# of in-state programs at age 30	-7.4	9.1	4.6	1.0
	(9.1)	(11.7)	(8.6)	(11.7)
Observations	30,167,461	25,986,698	26,128,325	30,025,834
R-squared	0.0007880	0.0013345	0.0006203	0.0013209
		E: Arcsinh(# of citations/patent)		
Mean	3,053.0	4,275.0	2,684.0	4,326.0
# of in-state programs at age 30	-4.8	7.8	3.4	1.6
	(6.4)	(8.0)	(6.0)	(8.0)
Observations	30,167,461	25,986,698	26,128,325	30,025,834
R-squared	0.0007427	0.0012694	0.0005757	0.0012511

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample and the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth. The outcome variable is the probability of earning a Ph.D. degree (Panel (A)), the probability of ever patenting (Panel (B)), the lifetime number of patents (Panel (C)), the lifetime number of citations (Panel (D)), and the lifetime average number of citations per patent (Panel (E)) (all using in inverse hyperbolic sine transformation in Panels (C) to (E)). We control for individual characteristics, including gender, race, nativity, and urban residence around age 10 indicators, parental SES index (mean of five socioeconomic status scores in census: Occupational Income Score, Duncan Socioeconomic Index, Occupational Prestige Score, Occupational Earnings Score, and Occupational Education Score), birth state fixed effects, and birth year fixed effects. We also control for the number of colleges in the birth state and in nearby states other than the state of birth for all patent outcomes. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.5: Effects of Ph.D. programs on Ph.D. location

	(1) Prob(earning a Ph.D. in birth state)	(2) Prob(earning a Ph.D. in nearby states other than the state of birth within 250 miles)	(3) Prob(earning a Ph.D. in out states over 250 miles)
# of in-state programs at age 30	0.7***	0.1	0.2
	(0.2)	(0.1)	(0.3)
# of nearby state programs within 250 miles at age 30	0.4***	0.3***	-0.7***
	(0.1)	(0.1)	(0.2)
Observations	113,175,230	92,184,221	115,004,798

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth. The outcome variable is the probability of earning a Ph.D. degree in the birth state (Column (1)), nearby states other than the state of birth (Column (2)), and in a state that is over 250 miles from the birth state (Column (3)). We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.6: Effect of Ph.D. programs on earning a Ph.D. from a public or private institution

	(1) Prob(earning a Ph.D. from a public institution)	(2) Prob(earning a Ph.D. from a private institution)
# of in-state programs at age 30: public	1.3***	-0.4
	(0.3)	(0.5)
# of nearby state programs within 250 miles at age 30: public	0.2	-0.5
	(0.2)	(0.4)
# of in-state programs at age 30: private	0.8**	0.4
	(0.3)	(0.4)
# of nearby state programs within 250 miles at age 30: private	-0.3	1.2***
	(0.3)	(0.3)
Observations	115,004,798	115,004,798

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth, separately for public and private institutions. The outcome variable is the probability of ever earning a Ph.D. degree from a public (Column (1)) and private (Column (2)) institution. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.7: Effects of Ph.D. programs on Ph.D. research fields

	(1) Prob(earning a Ph.D. in Arts and Humanities)	(2) Prob(earning a Ph.D. in Social and Behavioral Sciences)	(3) Prob(earning a Ph.D. in Mathematical and Physical Sciences)	(4) Prob(earning a Ph.D. in Natural Sciences)
# of in-state programs at age 30: Arts and Humanities	1.8*	1.2	0.3	2.4***
	(1.0)	(1.0)	(0.6)	(0.8)
# of in-state programs at age 30: Social and Behavioral Sciences	-0.9	0.3	-0.2	-2.3***
	(0.9)	(1.0)	(0.6)	(0.7)
# of in-state programs at age 30: Mathematical and Physical Sciences 30	-0.7	-1.0	-0.5	-1.5***
	(0.7)	(0.7)	(0.4)	(0.6)
# of in-state programs at age 30: Natural Sciences	0.9	1.4	0.5	2.1***
	(0.9)	(1.0)	(0.6)	(0.8)
Observations	115,004,505	115,004,505	115,004,505	115,004,505

Note: Data come from the linked Census-ProQuest Ph.D. recipient sample. This table reports the estimated coefficients (marginal effects) of the number of Ph.D. programs at age 30 in the birth state separately for four academic areas: Arts & Humanities, Social & Behavioral Sciences, Mathematical & Physical Sciences, and Natural Sciences. The outcome variables are the probability of earning a Ph.D. degree ever in each academic area. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, birth state fixed effects, and birth year fixed effects. All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.8: Effects of Ph.D. programs on patent outcomes

	(1) Prob(Inventor)	(2) Prob(Inventor & Ph.D.)	(3) Prob(Inventor) among non-Ph.D.	(4) Arcsinh(# of patents)	(5) Arcsinh(# of citations)	(6) Arcsinh(# of citations/patent)
# of in-state programs at age 30	2.0***	-0.0	2.0***	3.5	12.3***	9.3***
	(0.8)	(0.0)	(0.8)	(2.8)	(4.7)	(3.3)
# of nearby state programs within 250 miles at age 30	0.2	0.0	0.1	0.9	1.3	-0.1
	(0.5)	(0.0)	(0.5)	(1.6)	(2.7)	(2.0)
Observations	115,002,377	111,597,554	114,901,111	115,002,377	115,002,377	115,002,377
R-squared	-	-	-	0.0010	0.0009	0.0009

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) on patent outcomes of the number of Ph.D. programs at age 30 in the birth state and in the nearby states, defined as states other than the state of birth with centroids within 250 miles of that of the state of birth. We control for individual characteristics, including gender, race, urban residence, and nativity indicators, the number of colleges in the birth state and in nearby states other than the state of birth (defined as those within 250 miles) from birth to age 70 (at 10 year intervals), birth state fixed effects, and birth year fixed effects. The individual outcomes from Columns (1) to (3) are indicators for ever patenting, ever patenting and earning a Ph.D. degree, and ever patenting among non-Ph.D.s. Columns (4) to (6) report estimates for the lifetime number of patents, the lifetime number of citations, and the average number of citations per patent (all using in inverse hyperbolic sine transformation). All coefficients and standard errors have been multiplied by 10^6. Standard errors are clustered at the birth-state-cohort level. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.9: Effect of Ph.D. programs on patenting by demographic groups

	(1)	(2)	(3)	(4)	(5)
			Prob(Inventor)		
# of in-state programs at age 30*Female	-4.6	-4.7	-4.8	-4.5	-4.7
	(4.6)	(4.5)	(4.6)	(4.6)	(4.6)
# of in-state programs at age 30*Minority	0.1	0.7	1.5	-0.7	0.7
	(5.3)	(5.2)	(5.2)	(5.4)	(5.3)
# of in-state programs at age 30*2nd-gen immigrant	1.6	1.5	1.9	1.5	2.0
	(2.9)	(2.9)	(2.9)	(2.9)	(2.9)
# of in-state programs at age 30*Rural childhood neighborhood	-3.2	-2.7	-2.1	-4.4**	-2.6
	(2.0)	(2.0)	(2.0)	(2.1)	(2.0)
# of in-state programs at age 30*Parental Occupational Income Score	-7.3				
	(8.4)				
# of in-state programs at age 30*Parental Duncan Socioeconomic Index		-0.9			
		(4.0)			
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)			3.0		
			(7.6)		
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis				-6.4*	
				(3.7)	
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis					0.2
					(4.2)
Observations	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the probability of ever patenting. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.10: Effect of Ph.D. programs on patenting by demographic groups: High vs. low research intensity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Prob(Inventor)									
	High research	Low research	High research	Low research	High research	Low research	High research	Low research	High research	Low research
# of in-state programs at age 30*Female	-1.2	-4.6	-1.3	-4.6	-1.3	-4.7	-1.2	-4.6	-1.3	-4.6
	(2.8)	(3.5)	(2.7)	(3.4)	(2.7)	(3.5)	(2.8)	(3.5)	(2.8)	(3.5)
# of in-state programs at age 30*Minority	0.8	-2.9	1.2	-2.6	1.8	-2.2	0.3	-3.4	1.3	-2.6
	(3.2)	(4.3)	(3.1)	(4.3)	(3.1)	(4.3)	(3.3)	(4.4)	(3.2)	(4.3)
# of in-state programs at age 30*2nd-gen immigrant	-0.7	3.3	-0.8	3.2	-0.6	3.5	-0.7	3.2	-0.7	3.6
	(1.6)	(2.3)	(1.6)	(2.3)	(1.6)	(2.3)	(1.7)	(2.3)	(1.6)	(2.3)
# of in-state programs at age 30*Rural childhood neighborhood	-1.8	-0.8	-1.5	-0.6	-1.0	-0.4	-2.4*	-1.7	-1.3	-0.6
	(1.2)	(1.7)	(1.2)	(1.7)	(1.2)	(1.7)	(1.3)	(1.8)	(1.2)	(1.7)
# of in-state programs at age 30*Parental Occupational Income Score	-8.6*	0.2								
	(4.7)	(7.1)								
# of in-state programs at age 30*Parental Duncan Socioeconomic Index			-2.2	1.1						
			(2.3)	(3.4)						
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)					0.2	4.3				
					(4.3)	(6.2)				
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis							-4.9**	-2.7		
							(2.1)	(3.1)		
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis									-1.5	2.9
									(2.3)	(3.5)
Observations	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507	56,116,507	56,116,507

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the probability of ever patenting in fields that are classified in Figure 5. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.11: Effect of Ph.D. programs on lifetime number of patents by demographic groups

	(1)	(2)	(3)	(4)	(5)
	Arcsinh(# of patents)				
# of in-state programs at age 30*Female	-26.4***	-26.6***	-26.4***	-26.6***	-26.8***
	(9.6)	(9.6)	(9.6)	(9.7)	(9.7)
# of in-state programs at age 30*Minority	-9.2	-6.9	-5.7	-12.0	-9.3
	(7.4)	(7.3)	(7.1)	(7.6)	(7.3)
# of in-state programs at age 30*2nd-gen immigrant	1.6	2.5	3.8	0.9	2.8
	(8.7)	(8.7)	(8.8)	(8.7)	(8.8)
# of in-state programs at age 30*Rural childhood neighborhood	-12.5**	-11.4*	-13.6**	-16.1***	-13.1**
	(6.0)	(5.9)	(6.1)	(6.0)	(5.9)
# of in-state programs at age 30*Parental Occupational Income Score	32.3				
	(34.5)				
# of in-state programs at age 30*Parental Duncan Socioeconomic Index		31.6*			
		(16.4)			
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)			77.9***		
			(28.3)		
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis				-2.5	
				(11.5)	
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis					47.0**
					(23.2)
Observations	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507
	0.0015	0.0016	0.0015	0.0014	0.0015

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime number of patents. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.12: Effect of Ph.D. programs on lifetime number of patents by demographic groups: High vs. low research intensity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Arcsinh(# of patents)									
	High research	Low research	High research	Low research	High research	Low research	High research	Low research	High research	Low research
# of in-state programs at age 30*Female	-11.1**	-18.1***	-11.3**	-18.1***	-11.4**	-17.8***	-11.1**	-18.3***	-11.4**	-18.3***
	(5.1)	(6.6)	(5.1)	(6.6)	(5.1)	(6.6)	(5.1)	(6.6)	(5.1)	(6.6)
# of in-state programs at age 30*Minority	-4.1	-7.0	-2.7	-5.9	-1.0	-5.9	-5.0	-9.4*	-3.5	-7.4
	(4.6)	(5.3)	(4.7)	(5.2)	(4.6)	(5.1)	(4.8)	(5.4)	(4.6)	(5.2)
# of in-state programs at age 30*2nd-gen immigrant	-2.1	4.2	-1.9	4.9	-1.2	5.7	-2.5	3.5	-2.0	5.1
	(5.7)	(6.3)	(5.7)	(6.3)	(5.7)	(6.3)	(5.7)	(6.3)	(5.7)	(6.3)
# of in-state programs at age 30*Rural childhood neighborhood	-7.2*	-6.1	-6.4*	-5.7	-6.0	-8.3*	-8.5**	-9.1*	-6.8*	-7.1
	(3.9)	(4.6)	(3.7)	(4.6)	(3.7)	(4.6)	(4.0)	(4.7)	(3.6)	(4.5)
# of in-state programs at age 30*Parental Occupational Income Score	-3.9	35.3								
	(21.5)	(26.8)								
# of in-state programs at age 30*Parental Duncan Socioeconomic Index			5.2	27.3**						
			(10.8)	(12.6)						
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)					33.6*	52.7***				
					(18.4)	(19.9)				
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis							-5.0	-0.1		
							(7.2)	(8.7)		
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis									11.0	41.7**
									(15.9)	(16.6)
Observations	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507	56,116,507	56,116,507
R-squared	0.0005	0.0014	0.0006	0.0014	0.0006	0.0014	0.0005	0.0013	0.0006	0.0014

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime number of patents in fields that are classified in Figure 5. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.13: Effect of Ph.D. programs on lifetime number of citations by demographic groups

	(1)	(2)	(3)	(4)	(5)
	Arcsinh(# of citations)				
# of in-state programs at age 30*Female	-43.2**	-44.0**	-44.1**	-43.0**	-44.2**
	(19.3)	(19.3)	(19.3)	(19.3)	(19.4)
# of in-state programs at age 30*Minority	-13.4	-9.5	-6.9	-17.7	-13.5
	(14.2)	(14.0)	(13.6)	(14.6)	(14.0)
# of in-state programs at age 30*2nd-gen immigrant	2.4	3.6	5.6	1.5	4.6
	(16.6)	(16.6)	(16.7)	(16.6)	(16.7)
# of in-state programs at age 30*Rural childhood neighborhood	-21.5*	-19.5*	-22.1**	-27.1**	-22.2**
	(11.0)	(10.8)	(11.2)	(11.1)	(10.8)
# of in-state programs at age 30*Parental Occupational Income Score	52.0				
	(63.6)				
# of in-state programs at age 30*Parental Duncan Socioeconomic Index		49.1			
		(29.9)			
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)			120.1**		
			(52.1)		
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis				-2.5	
				(20.9)	
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis					76.2*
					(44.5)
Observations	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507
R-squared	0.0011	0.0012	0.0012	0.0011	0.0012

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime number of citations. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.14: Effect of Ph.D. programs on lifetime number of citations by demographic groups: High vs. low research intensity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Arcsinh(# of citations)									
	High research	Low research	High research	Low research	High research	Low research	High research	Low research	High research	Low research
# of in-state programs at age 30*Female	-15.2	-32.9**	-15.7	-33.3**	-16.0	-33.2**	-15.0	-33.0**	-15.8	-33.6**
	(10.1)	(13.8)	(10.1)	(13.8)	(10.1)	(13.8)	(10.1)	(13.8)	(10.2)	(13.8)
# of in-state programs at age 30*Minority	-8.9	-10.4	-6.0	-9.3	-3.0	-8.4	-9.3	-14.8	-7.5	-11.7
	(9.2)	(10.7)	(9.2)	(10.5)	(9.1)	(10.2)	(9.4)	(10.8)	(9.2)	(10.4)
# of in-state programs at age 30*2nd-gen immigrant	-7.8	15.6	-7.6	16.4	-6.5	17.9	-8.1	14.4	-7.5	17.1
	(10.9)	(12.4)	(11.0)	(12.3)	(11.1)	(12.4)	(10.9)	(12.4)	(11.1)	(12.4)
# of in-state programs at age 30*Rural childhood neighborhood	-10.2	-12.3	-8.3	-12.2	-7.2	-15.3*	-11.1	-17.9**	-9.0	-14.2
	(7.3)	(8.8)	(7.1)	(8.8)	(7.0)	(8.9)	(7.5)	(9.0)	(6.9)	(8.7)
# of in-state programs at age 30*Parental Occupational Income Score	-17.5	67.9								
	(42.0)	(51.1)								
# of in-state programs at age 30*Parental Duncan Socioeconomic Index			5.6	43.1*						
			(20.5)	(23.6)						
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)					46.1	91.0**				
					(35.6)	(38.5)				
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis							-6.1	1.7		
							(13.9)	(16.7)		
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis									12.0	73.1**
									(31.3)	(32.4)
Observations	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507	56,116,507	56,116,507
R-squared	0.0005	0.0009	0.0006	0.0010	0.0005	0.0009	0.0005	0.0009	0.0006	0.0009

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime number of citations in fields that are classified in Figure 5. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.15: Effect of Ph.D. programs on lifetime average number of citations per patent by demographic groups

	(1)	(2)	(3)	(4)	(5)
	Arcsinh(# of citations/patent)				
# of in-state programs at age 30*Female	-28.1**	-28.8**	-29.0**	-27.8*	-28.9**
	(14.2)	(14.3)	(14.3)	(14.2)	(14.3)
# of in-state programs at age 30*Minority	-9.3	-6.5	-5.1	-11.9	-9.3
	(10.4)	(10.3)	(10.0)	(10.7)	(10.3)
# of in-state programs at age 30*2nd-gen immigrant	1.8	2.5	3.6	1.4	3.5
	(12.0)	(12.0)	(12.1)	(12.0)	(12.1)
# of in-state programs at age 30*Rural childhood neighborhood	-15.5**	-13.9*	-15.4*	-18.9**	-15.8**
	(7.9)	(7.8)	(8.1)	(8.0)	(7.9)
# of in-state programs at age 30*Parental Occupational Income Score	32.7				
	(45.4)				
# of in-state programs at age 30*Parental Duncan Socioeconomic Index		32.6			
		(21.6)			
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)			72.3*		
			(37.2)		
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis				-0.8	
				(15.0)	
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis					49.1
					(32.2)
Observations	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507
	0.0011	0.0011	0.0011	0.0010	0.0011

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime average number of citations per patent. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Table A.16: Effect of Ph.D. programs on lifetime average number of citations per patent by demographic groups: High vs. low research intensity

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
	Arcsinh(# of citations/patent)									
	High research	Low research	High research	Low research	High research	Low research	High research	Low research	High research	Low research
# of in-state programs at age 30*Female	-8.6	-23.6**	-9.0	-24.0**	-9.3	-24.1**	-8.5	-23.6**	-9.1	-24.3**
	(7.5)	(10.7)	(7.6)	(10.7)	(7.6)	(10.7)	(7.5)	(10.7)	(7.6)	(10.7)
# of in-state programs at age 30*Minority	-7.4	-8.1	-4.9	-7.5	-3.2	-6.8	-7.3	-11.0	-6.1	-9.3
	(6.8)	(8.4)	(6.8)	(8.2)	(6.7)	(8.0)	(7.0)	(8.4)	(6.9)	(8.1)
# of in-state programs at age 30*2nd-gen immigrant	-7.1	14.0	-6.9	14.6	-6.3	15.6*	-7.1	13.3	-6.7	15.3
	(8.1)	(9.3)	(8.1)	(9.3)	(8.2)	(9.4)	(8.1)	(9.3)	(8.2)	(9.4)
# of in-state programs at age 30*Rural childhood neighborhood	-6.7	-8.8	-4.9	-8.8	-4.2	-10.7	-7.0	-12.3*	-5.5	-10.2
	(5.4)	(6.7)	(5.2)	(6.6)	(5.1)	(6.8)	(5.5)	(6.8)	(5.2)	(6.7)
# of in-state programs at age 30*Parental Occupational Income Score	-18.3	52.0								
	(30.9)	(37.8)								
# of in-state programs at age 30*Parental Duncan Socioeconomic Index			3.5	30.6*						
			(14.8)	(17.6)						
# of in-state programs at age 30*Parental Occupational Prestige Score (Siegel)					24.1	64.3**				
					(25.7)	(28.7)				
# of in-state programs at age 30*Parental Occupational Earnings Score, 1950 basis							-4.8	4.3		
							(10.1)	(12.6)		
# of in-state programs at age 30*Parental Occupational Education Score, 1950 basis									5.7	52.6**
									(22.8)	(24.1)
Observations	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,154,159	56,116,507	56,116,507	56,116,507	56,116,507
R-squared	0.0005	0.0008	0.0006	0.0009	0.0005	0.0008	0.0005	0.0008	0.0006	0.0008

Note: Data come from the linked Census-PATSTAT innovator sample. This table reports the estimated coefficients (marginal effects) of the interactions terms between the number of Ph.D. programs at age 30 in the birth state and a set of individual and family characteristics. The outcome is the lifetime average number of citations per patent in fields that are classified in Figure 5. Individual characteristics are gender, race, urban residence around age 10, and nativity indicators, along with a measure of parental socioeconomic status. We control for the number of colleges in the birth state and in nearby states other than the state of birth, birth state fixed effects and birth year fixed effects. Standard errors are clustered at the birth-state-cohort level. All coefficients and standard errors have been multiplied by 10^6. Significant level at ***p<0.01, **p<0.05, *p<0.1.

Appendix B

We briefly describe our matching algorithm linking the universe of ProQuest Ph.D. recipients from all US institutions to the full count 1940 Census in this section. Our method is a probabilistic matching approach based on exact matches of first and last names. Our method is novel and effective in linking US census data to datasets without any demographic information, such as ProQuest dissertations.

Step 1 We start by obtaining all the unique and exact matches based on full first and full last names in both data sets. As name pairs that are uniquely matched are very likely to be high quality matches, we define this sample as our "gold standard" dataset. We utilize this "gold standard" dataset to develop our probabilistic approach in the following steps.

Step 2 We estimate the distributions of three key identifying variables that can be observed in the "gold standard" dataset and are highly related to whether someone earned a Ph.D. degree. The highest grade of education attended or completed is provided in the 1940 census. The implied age of Ph.D. receipt is calculated for each possible match based on the ProQuest degree date and birth year in 1940 census. The geographical distance is measured based on the distance between the location of the Ph.D. institution and the centroid of each person's current residential county.

Step 3 For each value in the distribution of identifying variables in Step 2, we calculate the ratio between the share of observations in the uniquely matched "gold standard" data set and the analogous share in the 1940 US population. For each identifying variable, we use this ratio between the distribution of unique "gold standard" matches and the US population distribution — the relative ratio (RR) distribution—to determine the weights on a potential match given its value of each identifying variable.[20] We begin by assigning an arbitrary numerical weight to the peak value of the distribution of each RR distribution of each identifying variable. We then back out the weights for the remaining values of that variable based on the relative size of each value compared with the peak value in the RR distribution of that variable. For example, both Panel (a) and Panel (b) in Figure B.1 show that 7 years of college is the peak value in the RR distribution of highest education grade completed; therefore, we set a potential match with 7 years of college to have the maximum weight in cohorts born before 1940 and 1940-1950. As seen in Panel (a), the values of 5, 6, 7, and 8 years of college in the RR distribution are 68, 31, 137, and 61, respectively. This means that in the uniquely matched "gold standard" RR distribution, a person with 7 years of college appears about 2.6 times more often than a person with a grade that is two grades off, where we smooth the distribution by

[20]We conduct this exercise for people who are awarded Ph.D. degrees before 1940, 1940 to 1950, and 1950 to 1960 separately to address potential cohort differences. For example, highest education degree could be less distinctive for people who got Ph.D. degrees after 1950 as they are relatively young in 1940, as Panel (c) of Figure B.1 suggests.

taking the average across adjacent value (i.e., 137/[(68+31+61)/3]). Thus, the weight for a potential match with 7 years of college is assigned as 8 and the weight for 5 years of college is assigned as 3 (approximately 2.6 times smaller than 8). [21]

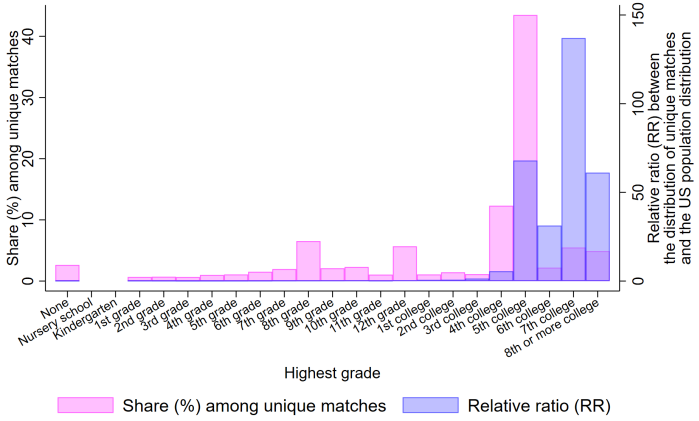
Step 4 After assigning weights based on the RR distributions for each identifying variable for each cohort, we obtain all the relative weights in Table B.1. To reduce false matches, we block on the first and last names of the dissertation author, requiring an exact match of full first name and full last name for all matches between ProQuest and the 1940 Census in our resulting dataset. Then, for each pair with an exact first and last full name match, we sum the weights of all identifying variables to determine the matching score and drop all observations beneath a cutoff score. The cutoff score is the sum of all minimum positive weights for each identifying variable.

Step 5 To further increase the match precision, we also assign weights based on additional identifying variables, for example, whether there is a match of middle full name or initial, name frequency in 1940 census, and population size of current residential county, as indicated in Table B.2. Among all the matched pairs above the cutoff score with the same first and last name, we rank the total match score and pick the pair with the highest score. In the rare case when there are multiple matched pairs with the same highest score, we keep all the potential matches if and only if there are 10 or fewer total candidates with the highest score. The overall matching rate is presented in Figure B.4, with an average rate around 60%.

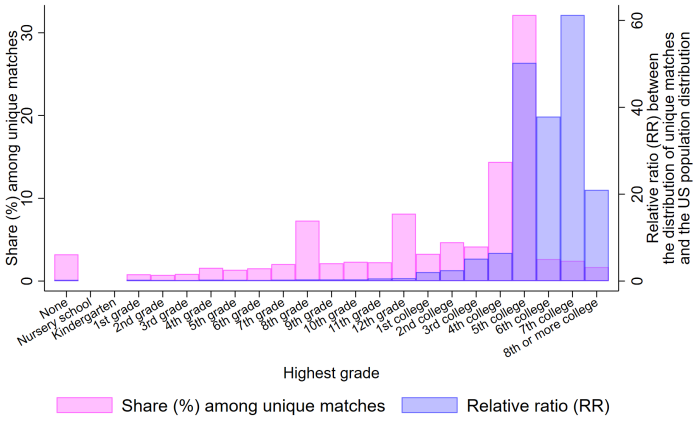
[21] When assigning weights to the peak values, we require them to be relatively similar across all identifying variables to ensure comparable importance.

Figure B.1: Identifying variable - Highest education grade

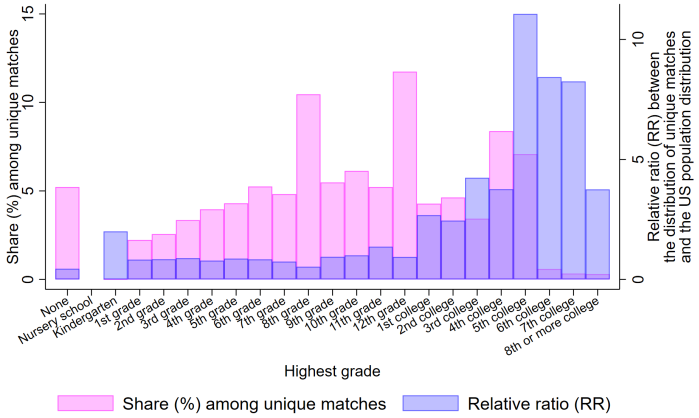
(a) Ph.D. recipient cohort before 1940



(b) Ph.D. recipient cohort 1940 to 1950



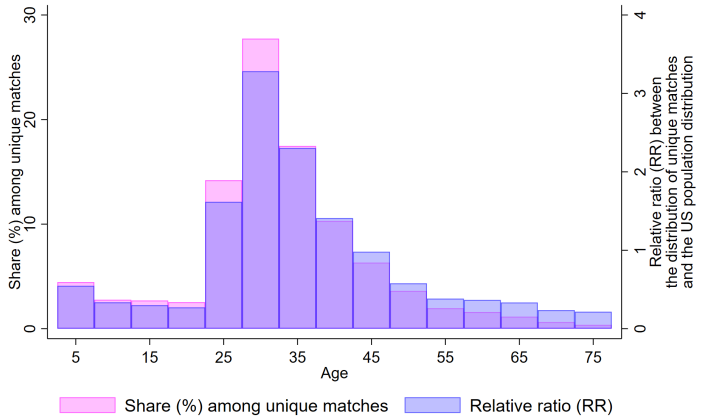
(c) Ph.D. recipient cohort 1950 to 1960



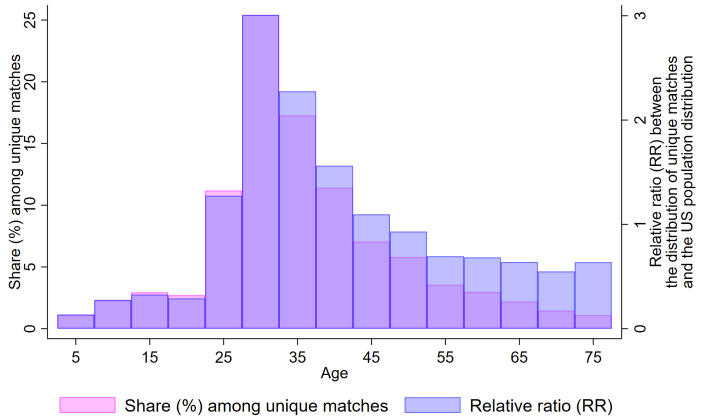
Note: This figure plots the percentage share of people with each level of highest education grade in the "gold standard" dataset and the ratio between the share of highest education grade among unique matches and US population education share in 1940.

Figure B.2: Identifying variable - Implied age of degree receipt

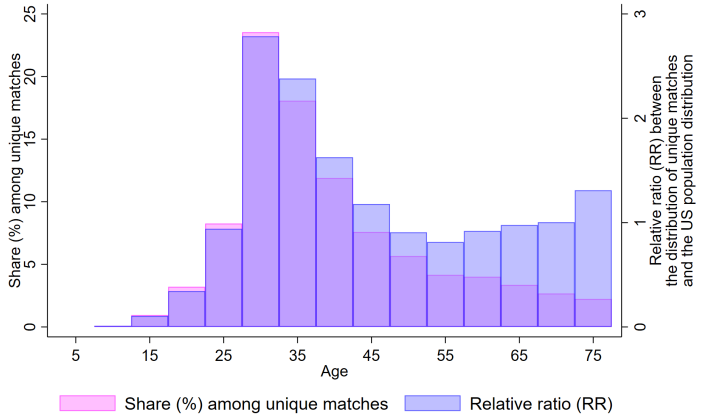
(a) Ph.D. recipient cohort before 1940



(b) Ph.D. recipient cohort 1940 to 1950



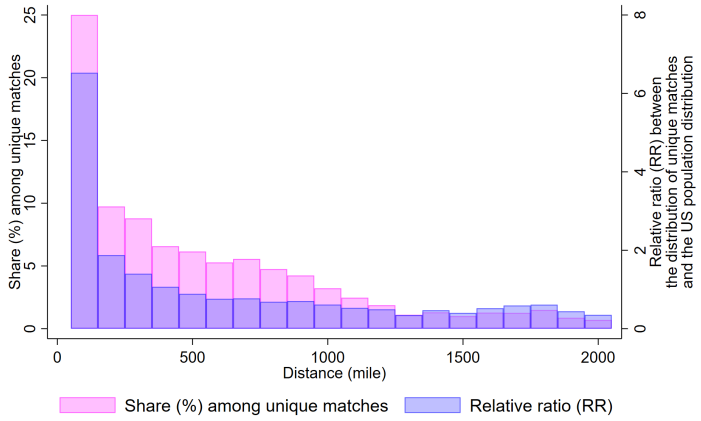
(c) Ph.D. recipient cohort 1950 to 1960



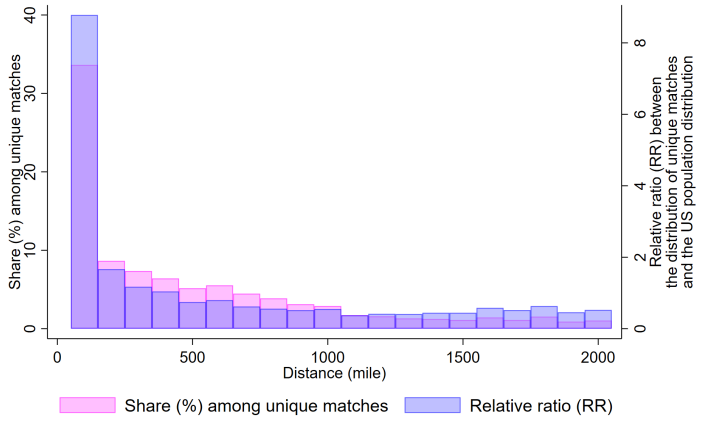
Note: This figure plots the percentage share of people earning Ph.D. degree in each age group in the "gold standard" dataset and the ratio between the share of Ph.D. age among unique matches and US population age share in 1940.

Figure B.3: Identifying variable - Distance between birth state centroid and Ph.D. institution location

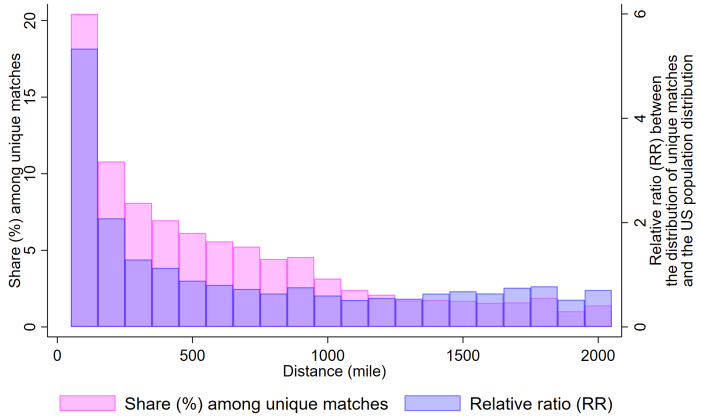
(a) Ph.D. recipient cohort before 1940



(b) Ph.D. recipient cohort 1940 to 1950

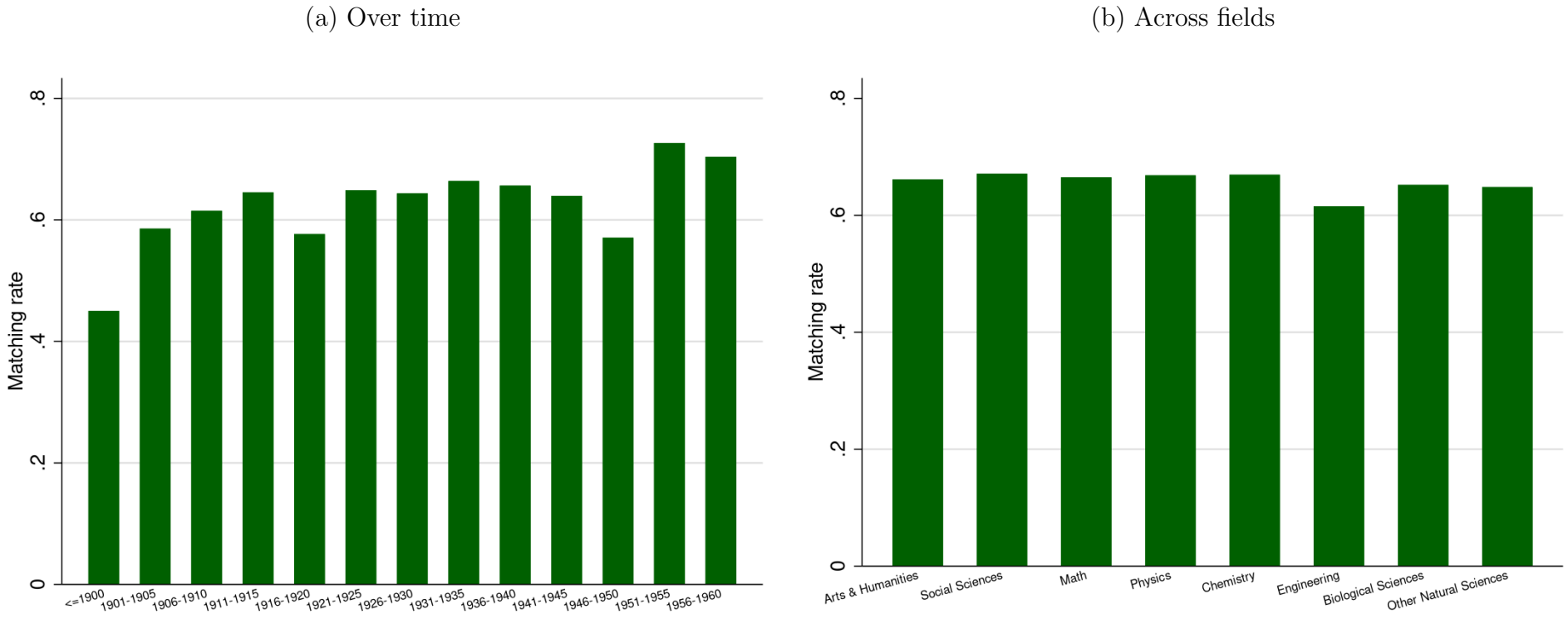


(c) Ph.D. recipient cohort 1950 to 1960



Note: This figure plots the percentage share of geographical distance between attended Ph.D. institution and birth state centroid in the "gold standard" dataset and the ratio between the share of institution-birthplace distance among unique matches and the distribution of distance between randomly pairing a county centroid to an research institution.

Figure B.4: Match rate of ProQuest to the Census by year and field



Note: This figure plots the match rate of ProQuest to the Census as a share of ProQuest Ph.D. recipients from U.S. research institutions. Panel (a) shows the matching rate over time. Panel (b) shows the matching rate by academic field.

Table B.1: Weights of each value of each identifying variable

	Earlier than 1940	1940-1950	1950-1960
Highest education grade			
7th college	8	8	
8th college	3	4.8	
6th college	3	4.8	
5th college	3	4.8	
4th college	0.3	0.7	
3th college	-12	0.7	
others	-12	-12	
Longitude			
within 100 mile	7	7	7
100-200	2	1	2.8
200-500	1	0.7	1.4
500-1000	0.7	0.4	1
others	-2	-2	-2
Latitude			
within 100 mile	7	7	7
100-200	2	1	2.8
200-500	1	0.7	1.4
500-1000	0.7	0.4	1
others	-2	-2	-2
Implied degree age			
25-35	6	6	6
20-25/35-40	3	3	3
40-50	1.5	2.3	2.4
others	-12	-12	-12
Cutoff	3.2	3.8	4.4

Note: This table lists the assigned weight for each value of each identifying variable within each cohort.

Table B.2: Weights of additional identifying variable

Middle name	
exact full name match	8
mid initial match(one full name+one initial)	4
either one missing	0
only mid initial match, rest not	-3
mid initial not match	-9
full middle not match	-16
Uncommon name	
Very uncommon (once)	3
Relatively uncommon (1-10 times)	0.3
Less uncommon (10-100 times)	0.05
Small county	
Very small (<50,000)	1.8
Relatively small (50,000-100,000)	0.6
Moderately small (100,000-200,000)	0.3
Less small (200,000<500,000)	0.1

Note: This table lists the assigned weight for middle name, frequency of names in 1940 census, and population size of the current residential county.