Innovation Nation: Evidence from Broadening Access to Ph.D. Training in the U.S.

[Not For Circulation]

Francisca Antman, Kirk Doran, Xuechao Qian, and Bruce A. Weinberg

Abstract

The United States rose to become a global leader in scientific research in the early 20th Century. Matching the universe of ProQuest Ph.D. recipients to the full count decennial US Censuses (1850-1940), this paper investigates how access to Ph.D. training developed research manpower in the US during this critical period. Evidence from our event study design suggests that opening one more Ph.D. program during the peak ages of graduate study induces nearly 3 more Ph.D. recipients per 1 million people born in that state. Furthermore, the expansion of Ph.D. programs improved access for minority, immigrant, rural, and lower socioeconomic status families. We also 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.

1. Introduction

Rising inequality in the U.S. and increased global competition for innovative and technological dominance both point to a need for the U.S. to expand access to Ph.D. training. This paper explores what we can learn from the historical experience in which the United States entered the 20th Century as a research laggard, but, through a series of broad expansions in access to education, especially research education, rose to become a global leader in scientific research and innovation. This transformation was made possible by radically altering the institutions providing research education and engaging a far wider range of the population in research (Geiger, 1986). We look back to that period to identify lessons on the potential for broadening access to research education today.

This paper investigates the impacts of expanding research education institutions on the propensity to obtain a Ph.D. degree, the location and field of study among Ph.D. recipients. Furthermore, we estimate the extent to which broadening access to research education affects underserved communities that have the potential to generate advances and technologies of the greatest benefit to underrepresented and economically challenged groups.

We generate a wealth of population-scale data on U.S. Ph.D. recipients over much of the 20th Century by linking historical U.S. Censuses (1850-1940) with population-scale 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 (Abramitzky et al., 2020) and a historical census record linkage by IPUMS), 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, socioeconomic background, parental occupations, sibship characteristics as well as gender, race, immigrant status, birth cohort and other defining traits. Research studies have only begun to link demographic changes with contributions to science, and these have been confined to specific episodes in history and only for particular groups (Borjas and Doran 2012; Moser et al., 2014). Our unprecedented data resources allow us to trace out the broadest expansion in scientific communities over the longest period of time, giving us a unique opportunity to observe the effects of this expansion.

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 programs and the interactions with diverse population groups, as evidence suggests diversity is associated with economic performance (Alesina and La Ferrara, 2005), firm outcomes (Østergaard et al., 2011) and innovation ideas (Hofstra et al., 2020). 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 preeminent data sets allow us to determine when Ph.D. programs begin and enable us to link the timing of program openings with

individual outcomes. We also provide a series of "placebo" checks to bolster our identification assumptions and pinpoint causal relationships. Our event study estimates indicate that opening one more Ph.D. program in a state during the peak ages of graduate study induces nearly 3 more people born in that state to obtain a Ph.D. per million people born in that state. Furthermore, the expansion of Ph.D. programs improved access for minority, immigrant, and lower socioeconomic status families. We also 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. Furthermore, we leverage exogenous variation brought about by major institutional expansions, such as HBCUs, Co-education colleges, Land Grant universities, as well as historical emerging of private universities to investigate the main driving (policy) force of the research institution expansion.

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 and Giuliano, 2014), secondary (Goldin and Katz 2009), and college education (Bound and Turner 2002; Chetty et al. 2017, and Hoxby and Avery 2014). Ours is the first to focus on the expansion of research education. 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. 2019) 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. 2018; Azoulay et al. 2019; Berkes and Nencka 2020). It also makes a substantial contribution to the literature examining the human impact of research universities on the economy (Kantor and Whalley 2014).

In the future work, we will also study the effect of research education on innovation to contribute to the literature on the rate of inventive activity (Khan and Sokoloff 1993; Lamoreaux and Sokoloff 2003). 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.

2. Data

2.1 Data Linkage

The enabling resource for this project is a set of population scale links (shown in Figure 1) between the historical U.S. Censuses and population-scale data on doctoral dissertations. To the best of our knowledge this database is the first longitudinal dataset in which individual innovators can be linked back to their characteristics, including those of their families of origin, and ultimately forward to their research achievements. Here we detail each data component in turn.

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.[1] All the full count U.S. Censuses have been linked across the whole period from 1850 to 1940 by the Census Linking Project (Abramitzky et al., 2020) at the individual level and a second link is being produced by the IPUMS Multigenerational Longitudinal Panel. The comprehensive information in the longitudinally-linked historical U.S. Census enables us to match the U.S. population with Ph.D. recipients from

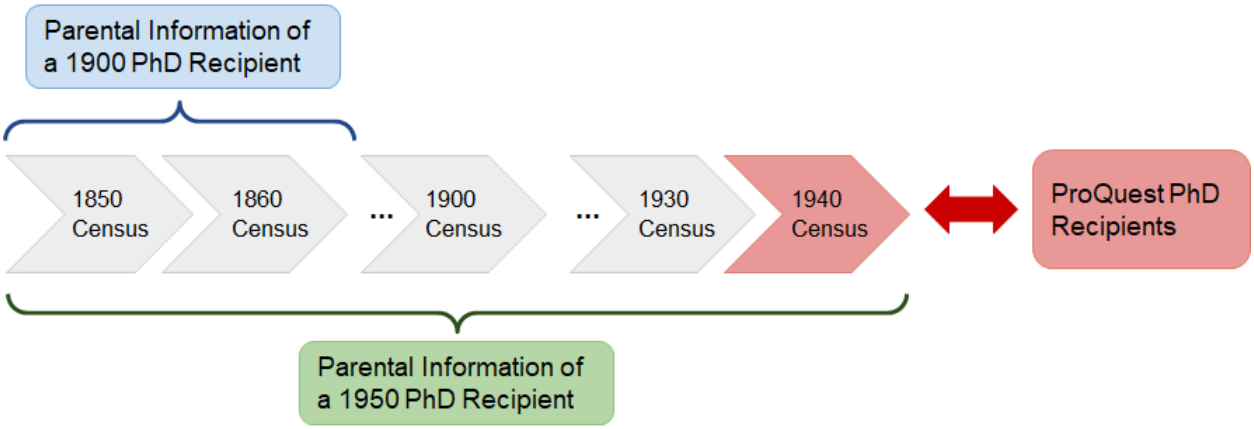
[1] The 1890 Census no longer exists. We expect the 1950 Census to be available toward the end of our study and will pilot adding those data.

U.S. institutions and innovators with patents registered in the U.S. (and also, at some later date, with authors of research publications). As illustrated in Figure 1, by using the longitudinal links, we will not only be able to document who these researchers and innovators are using their demographic characteristics, such as race, gender, nativity, etc. from the Census data, but look backwards in the Census to identify the kinds of families they grow up from with the help of the Census household records including parental socioeconomic status.

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, research subject, institution information (a school identifier and name), and the degree information of the recipient (the degree and date). As discussed below, we focus on people who received their Ph.D.s before 1960 to maximize our ability to match recipients to their Census records.

Figure 1. Linking approach.



Linking Approach

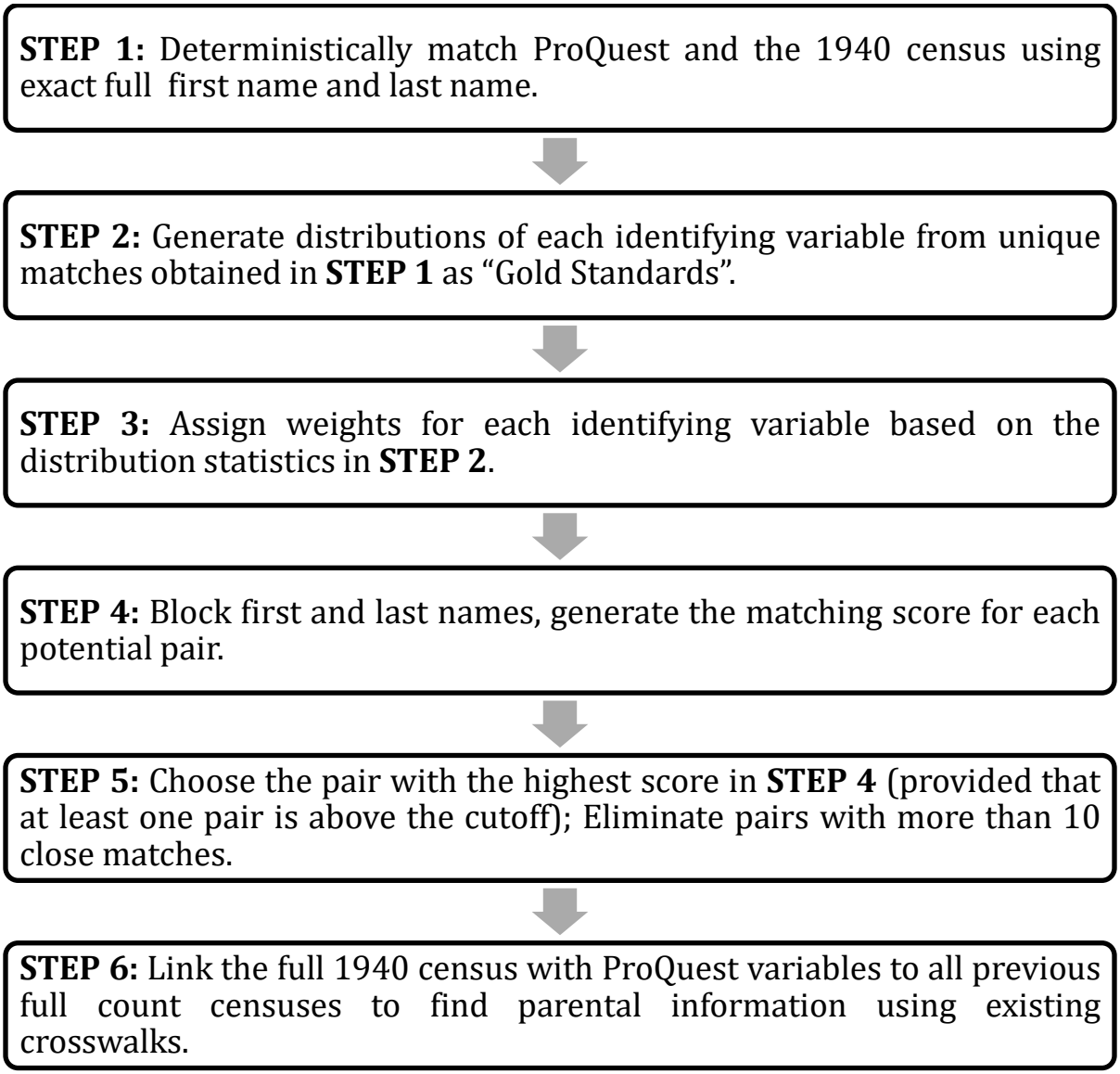
Given that the 1940 Census is the most recent public full count Census and the first Census containing educational attainment information, our linking work starts with matching ProQuest Ph.D. recipients to the 1940 Census.[2] Our matching strategy is outlined in Figure 2. Step 1 uses unique, exact matches 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 (e.g. education, implied age of degree receipt based on the ProQuest degree date, etc.) among likely true matches. 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 at the 1940 Census in our resulting dataset. Unlike our gold standard, we do not require a unique match. Thus, for common pairings of first and last names there will be multiple individuals in each dataset who exactly match on first and last name. To obtain unique matches, we generate a match score for each potential matched pair from the ProQuest and Census based on the probabilistically weighted set of identifying characteristics and the highest-score pair is taken as the best match if and only if the match score is above a predetermined cutoff. When there are multiple candidate matches to one Ph.D. recipient with the same highest score, we keep all the potential matches if and only if the total number of candidates is below 10.[3] In order to compare the Ph.D. recipients

[2] Initially, we will focus on people who receive Ph.D.s in and before 1960 because high rates of false positive matches are likely when matching people who receive Ph.D.s after 1960. This is due to the fact that people who receive Ph.D.s after 1960 are likely to be very young as of the 1940 Census. The total number of Ph.D. recipients awarded in and before 1960 in the ProQuest is 133,002. We will explore adding additional years (e.g. Ph.D.s through 1970) when the 1950 full count Census becomes available after 2022.

[3] When there are N candidate matches to one Ph.D. recipient, a weight as 1/N is assigned to all the candidates for all empirical estimates.

to (the overwhelming majority of) the population that did not receive Ph.D.s, we retain observations from the 1940 Census records that did not match the ProQuest dataset.

Figure 2: The flow chart of the probabilistic linking algorithm



Using this procedure, we matched ProQuest to the 1940 Census. We obtained 102,642matches 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 matching rate is 66%. As shown in Appendix Figure 1, 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), and degrades for people who received Ph.D.s in the very early 20th or late 19th centuries, and also during World War I. Appendix Figure 1 also shows a similar match rate across fields.

Table 1 reports the distribution of characteristics in the 1940 Census and the distribution of characteristics among the people in ProQuest, who we matched to the 1940 Census. These data indicate that the Ph.D. recipients are considerably more likely to be men, white, and to live in a city in 1940. They are slightly more likely to have immigrants for one or both of their parents.

	Mean	S.D.	Mean	S.D.
	1940 population born in US (N= 120,033,290)		Matched Ph.D. recipients (N=102,642)	
Ph.D. recipient	0.0009	0.0292		
Male	0.50	0.50	0.88	0.32
White	0.89	0.31	0.95	0.21
Urban residence	0.53	0.50	0.70	0.46
One foreign born parent	0.05	0.21	0.07	0.26
Two foreign born parents	0.05	0.21	0.07	0.26

2.2 Quantifying the Expansion of and Changes in U.S. Research Education

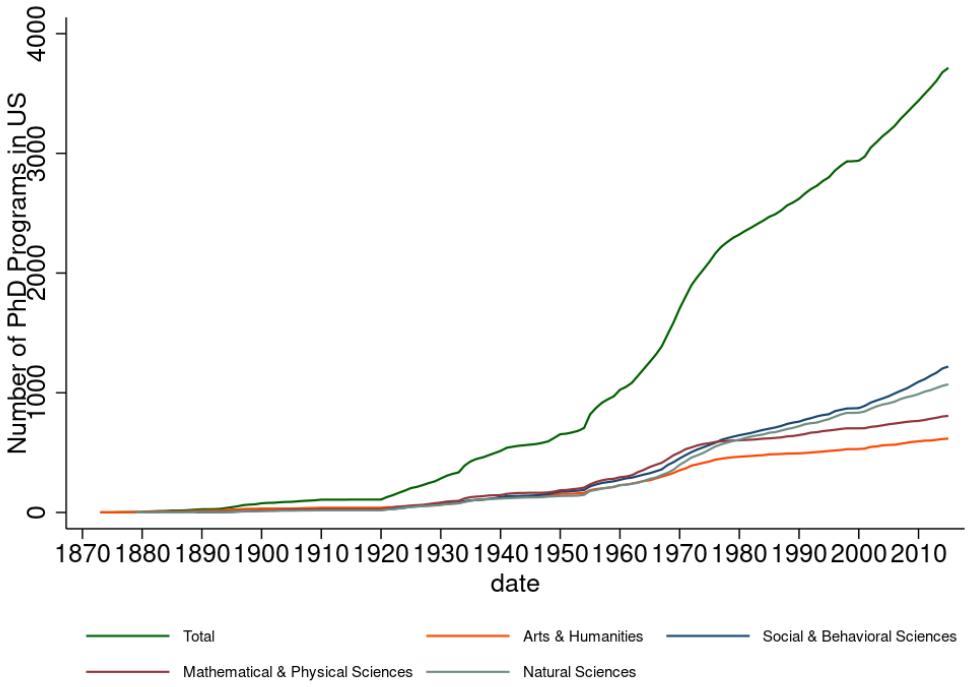
We begin by using ProQuest to trace the trajectory of American research education by documenting the establishment of new programs in existing fields, the emergence of entirely new fields for programs, and the growth in degrees awarded. Here we focus on research fields, institution types (e.g. public / private and new / old), and geography.

As indicated, ProQuest provides population-level data on individual Ph.D.s awarded in the U.S. from 1861 onward. We are able to not only observe the total number of Ph.D. recipients in each year since 1861, but also see the awarding institution and the research field categorized at a very fine level for each dissertation. We use these data to infer the year in which a Ph.D. program in each field at each university was actively awarding degrees and reached any given scale.[4] While we could associate the first Ph.D. degree(s) in a particular field awarded by each university as the time that university had a Ph.D. program that was actively awarding degrees in that field, we choose to be conservative and impose a threshold size and only recognize the opening of a Ph.D. program once it has awarded three degrees in its history or per year.

[4] Because we do not observe matriculation in our data, everything is dated from around graduation. As an alternative would be to subtract a fixed number of years to reflect the length of time required to complete a PhD from the graduation years to proxy for the beginning of PhD programs.

Figure 2. Trends in the U.S. Ph.D. Programs.

A. Cumulative number of Ph.D. programs opened in US (total and by major field)



B. Share of cumulative Ph.D. programs opened in US by major field.

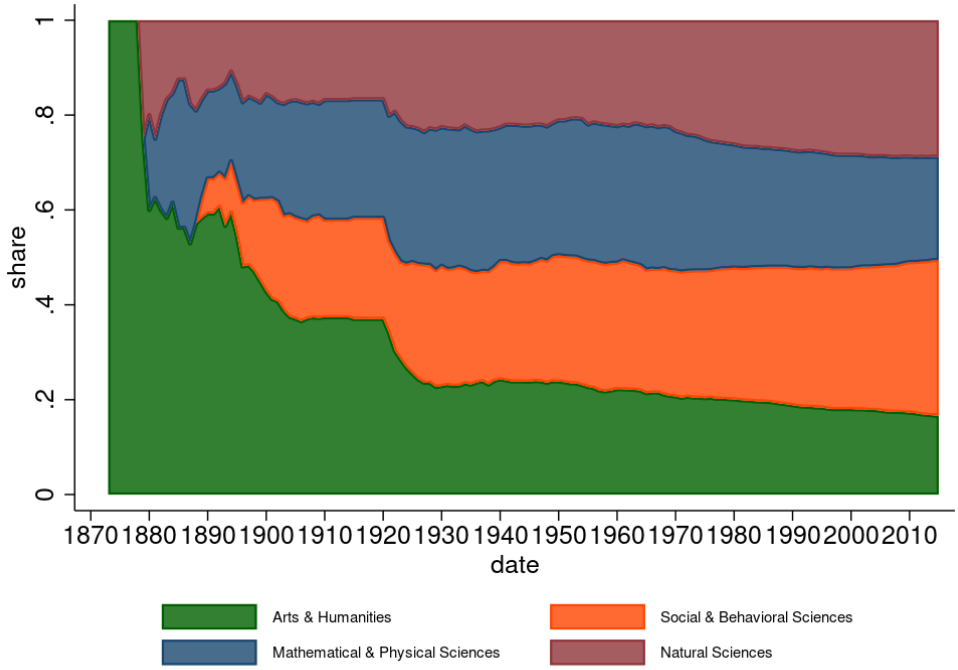
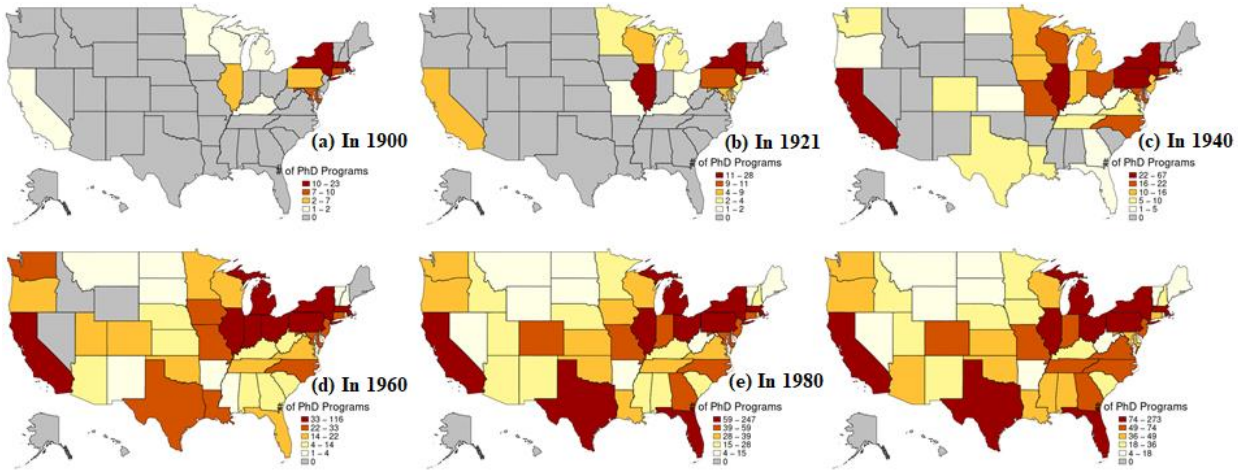


Figure 2 provides basic trends in the number of Ph.D. programs that have been opened in the U.S. Panel A shows the enormous growth of programs. Panel B shows the dramatic shift from the Arts and Humanities, which dominated Ph.D. education in the late 19th Century, toward other fields. In recent years, the Social and Behavioral Sciences are the single largest area of Ph.D. studies while the Arts and Humanities are smallest, behind the Natural Sciences and Mathematical and Physical Sciences.

Figure 3. Geographic spread of Ph.D. programs.



As the number of Ph.D. programs increased, they spread out geographically. As shown in Figure 3, in 1900, virtually all Ph.D. programs in the U.S. were in the Mid Atlantic (including Massachusetts), Midwest, or California. By 1940, Ph.D. programs were operating in most states other than the parts of the Mountain region and South, while by 1980, they were operating in all states. Thus, the expansion of programs, involved a geographic spreading out into new locations, and a wider range of disciplinary foci.

3. Empirical Analysis

3.1 The Effects of Program Expansion on Access

Having laid out the underlying source of variation, our first regression strategy is a nonlinear model the probability that people obtained a Ph.D. as a functionof 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. (i.e. appears in ProQuest) to the number of programs in their state of birth and nearby states. We use an event study design, which involves including leads and lags of the number of programs to leverage the timing of opening of programs.

Our basic specification is:

$$\boldsymbol{P(Ph.D._i) = \Phi(\beta_0 + \beta_{Birth,InState} N_{Birth,InState,i} +}$$

$$\boldsymbol{\beta_{10,InState} N_{10,InState,i} + \ldots \beta_{70,InState} N_{70,InState,i} + \beta_{Birth,Nearby} N_{Birth,Nearby,i} +}$$

$$\boldsymbol{\beta_{10,Nearby,i} N_{10,Nearby,i} + \ldots \beta_{70,Nearby} N_{70,Nearby,i}} + X_i\Gamma + \eta_s + \lambda_c + \varepsilon_i) \qquad (1)$$

Here $\boldsymbol{Ph.D._i}$is a dummy variable equal to 1 if individual i receives a Ph.D. degree in her life; $N_{Birth,InState,i}$, $N_{10,InState,i}$,... $N_{70,InState,i}$is a set of measures describing the number of Ph.D. programs in the state in which *i* was born starting in the year in which they were born to age 70 (at 10 year intervals); $N_{Birth,Nearby,i}$, $N_{10,Nearby,i}$,... $N_{70,Nearby,i}$is a comparable set of measures for the states near *i*'s state of birth (e.g. within 200 miles), but excluding the state of birth itself. X_i is a set of individual characteristics; η_s is a set of fixed effects (dummy variables) for the state of birth; and λ_c is a set of fixed effects (dummy variables) for the year of birth.[5] The total number of nation-wide programs at each point in a person's life are captured by the cohort fixed effects, so one should interpret the coefficients of interest as the effects of having more programs in the state of birth or near the state of birth relative to a given number of nation-wide programs. State of birth

[5] The number of programs in the rest of the U.S. at each age is subsumed in λ_c, so it is not included explicitly (intuitively, the total number of programs in the U.S. at each age is the same for everyone in a given cohort, so we are estimating the "extra" effect of having programs nearby relative to anywhere in the U.S.)

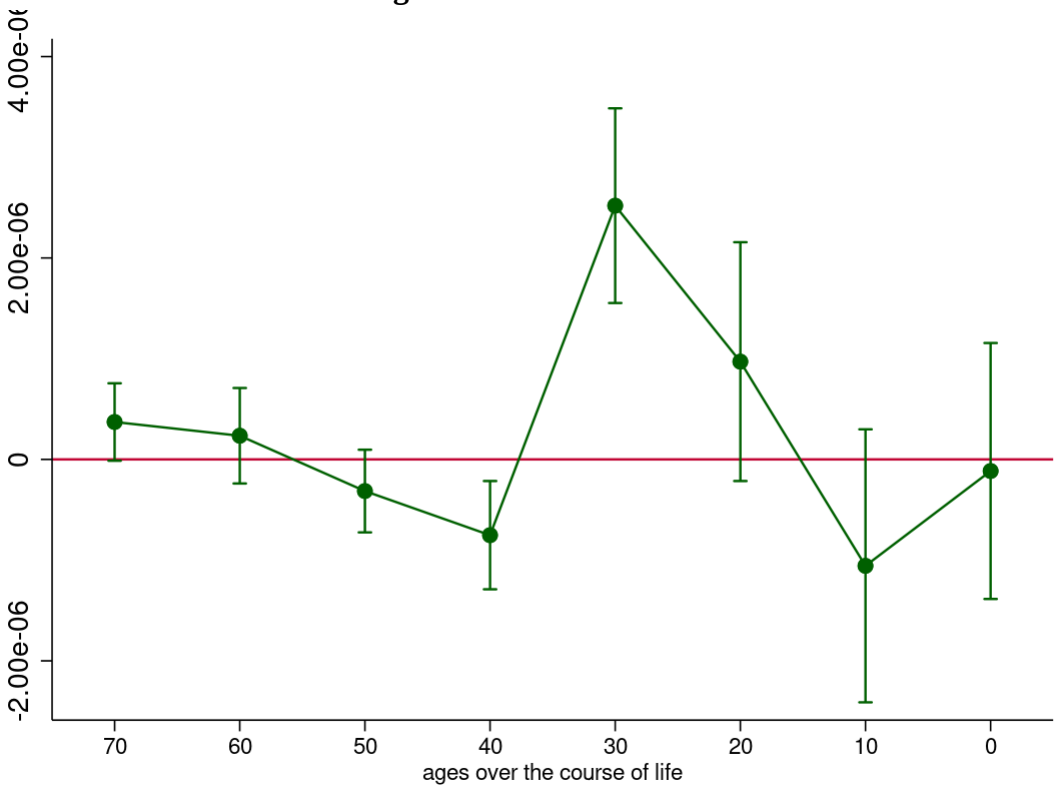
fixed effects control for time-invariant underlying differences in the propensity of obtaining a Ph.D. across states. Here and below, where we have aggregate-level independent variables (e.g. variables that are common to everyone born in a given state in a given year), we cluster our standard errors to account for the possibility that observation errors may be correlated within clusters.

If access to programs increases the probability of completing a Ph.D., we would expect the number of programs in a state around the time people turn 30 (and hence would be receiving a Ph.D.) to be strongly related to the probability of getting a Ph.D. We might see some effects at younger ages (e.g. if having more Ph.D. programs in a state encourages more people to go to college and then, once they are there, to continue to graduate school. In our setup, parallel pretrends take the form of a weak relationship between the number of programs and later ages (i.e. after people have completed their education). As indicated, 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.

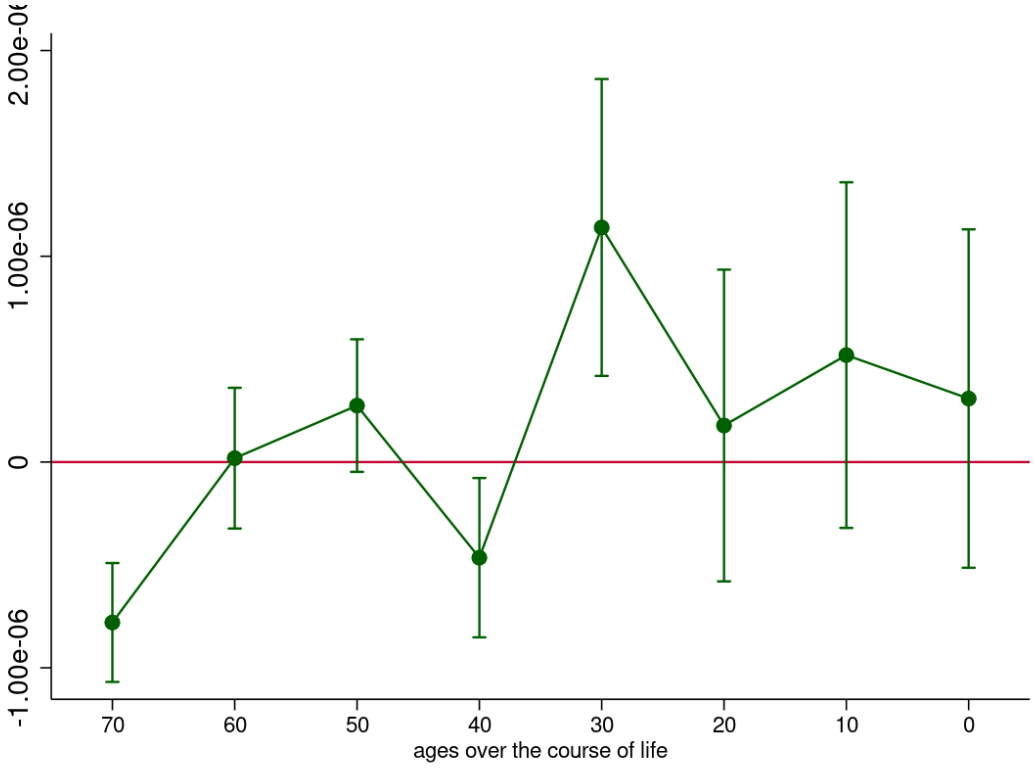
The estimates are shown in Figure 4. Panel A indicates a very strong relationship between the number of programs in the state of birth at the time that people turn 30 and the probability of obtaining a Ph.D. There is a relatively large effect at age 20 as well, but this is not statistically significant. The estimates at other ages, especially at older ages are quite small and remarkably flat. The estimates for age 30 indicate that the addition of one additional Ph.D. program at age 30 induces 2 people per million to obtain a Ph.D.

Figure 4. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree

A. Programs in the state of birth



B. Programs in nearby states



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 for programs in the individual's own state. These estimates are reassuring from the perspective of causality in that we expect nearby 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 that the relationship 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 programs rather than unobserved factors that drive the prevalence of programs and completion.

3.2 Heterogeneity Effects by Individual Characteristics

We can also identify the individual characteristics that increase or reduce sensitivity to the number of programs. To do this, we can augment Equation (1) by including interactions between the number of Ph.D. programs and individual demographics, such as gender, race, nativity, rural residence in childhood, and parental socioeconomic measures (i.e. X_i), to study how the impacts of expanded Ph.D. study opportunities vary with population groups and whether the access is more equalized thanks to the increased research training opportunity. We show estimates after adding the interactions only with one of the five characteristics and also a version with all the individual demographics in Figure 5.

Although somewhat noisy, the estimates in Figure 5 provide a mixed picture of how access to programs affected Ph.D. attainment. On the one hand, the number of in-state programs at age 30 increases the probability that men obtain a Ph.D. relative to women, enhancing existing disparities. However, Figure 5 shows suggestive evidence that increasing the number of in-state programs at age 30 increases Ph.D. attainment among non-white population and people with one

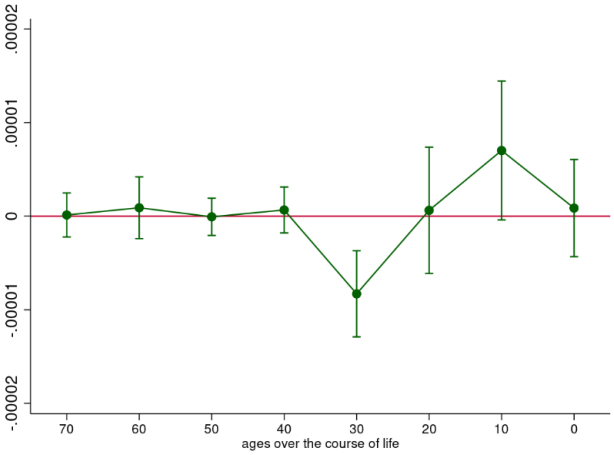
or two foreign born parents than the rest of the population.6 While obtaining a Ph.D. may be more challenging for people with foreign born parents, they are somewhat more likely to obtain a Ph.D. Panel A.5 and B.5 also show that in-state programs reduce disparities - the number of in-state programs has larger effect on people who grew up from rural areas and whose parents have lower socio-economic status (as measured using a Duncan index). Thus, increased access to Ph.D. training may be reducing inequality in some ways. It is noteworthy that the pre-trends tend to be quite flat for most of the series.

Figure 5. Effect of number of in state programs by individual characteristics.

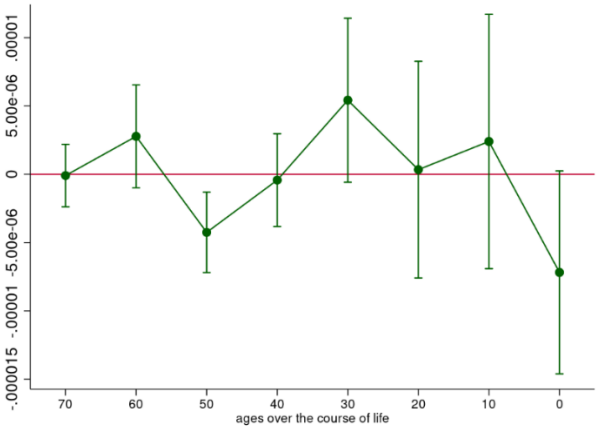
[6] Note that we do not study people who are themselves foreign born because they would be less likely to appear in historic censuses.

Panel A: Single interaction with each characteristic

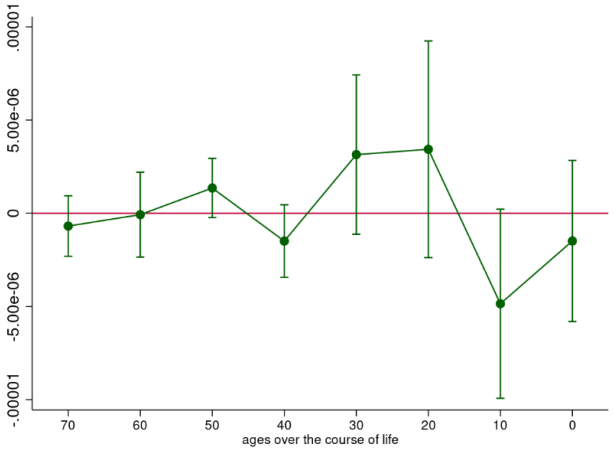
A1. Interaction with female



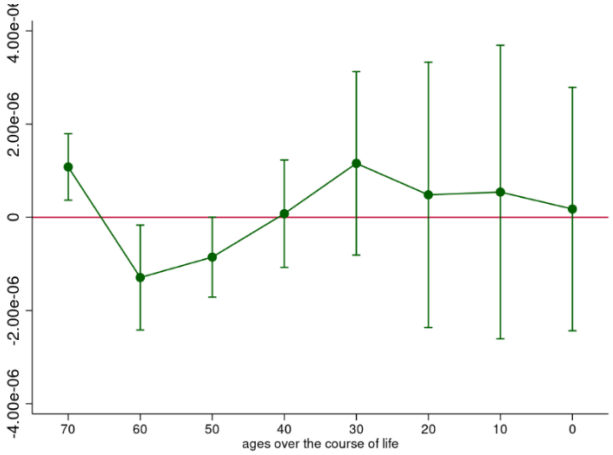
A2. Interaction with non-white



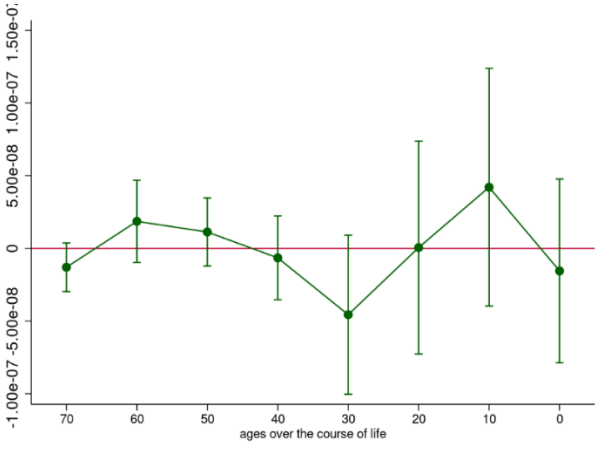
A3. Interaction with foreign parent(s)



A4. Interaction with rural residence in childhood

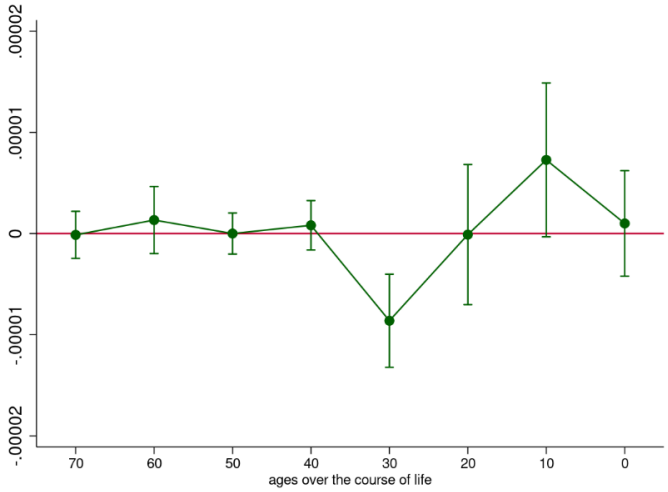


A5. Interaction with Duncan socioeconomic index

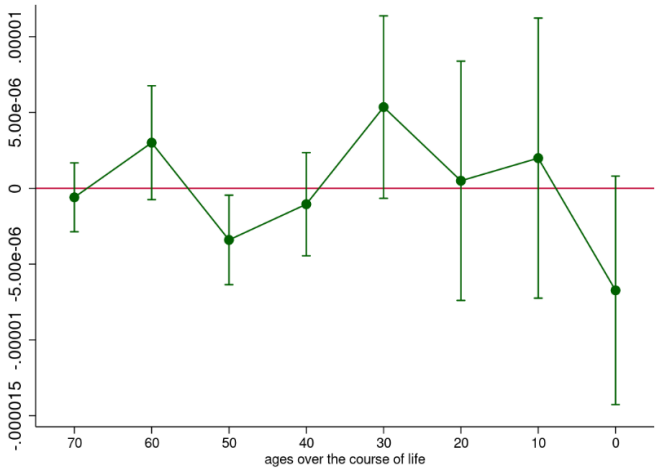


Panel B: A full set of interactions with all characteristics

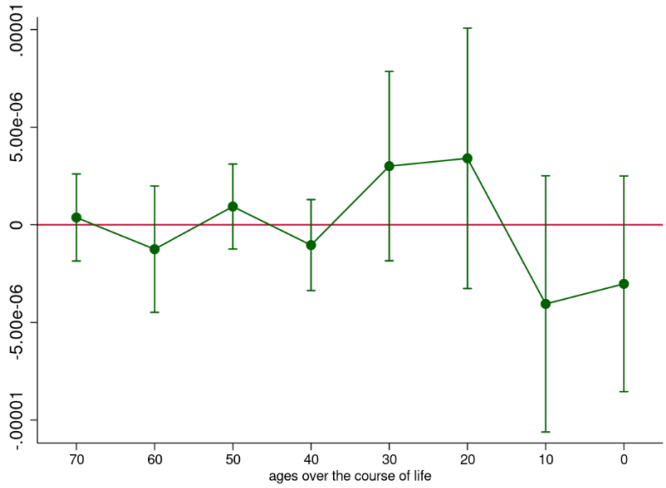
B1. Interaction with female



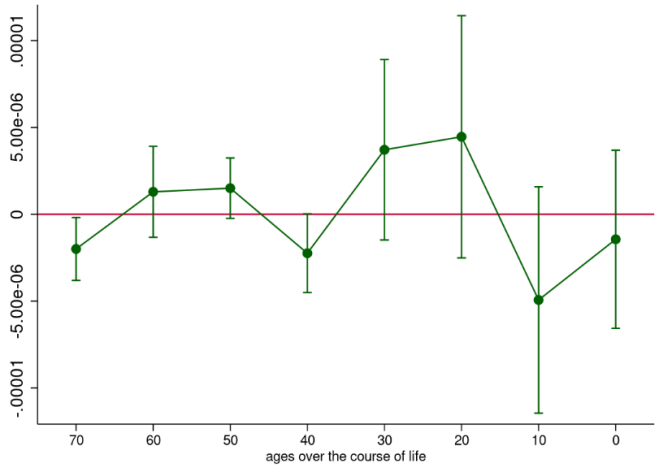
B2. Interaction with non-white



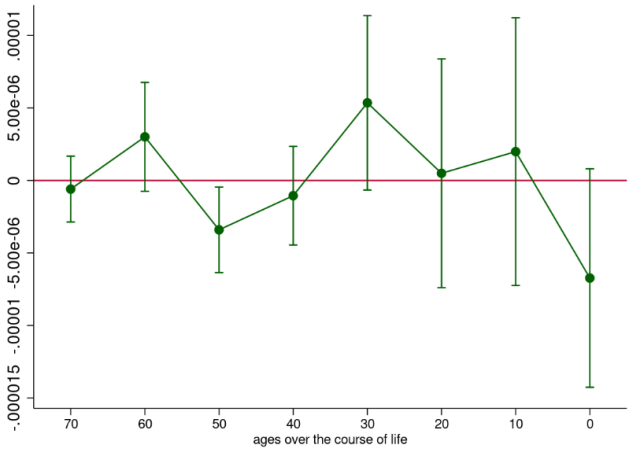
B3. Interaction with one foreign parent



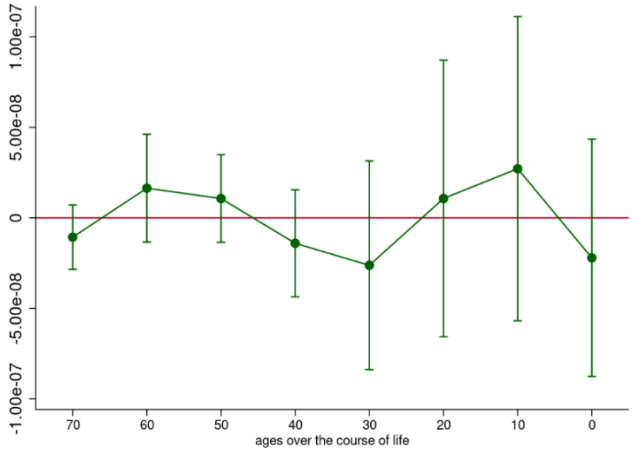
B4. Interaction with two foreign parents



B5. Interaction with rural residence in childhood



B6. Interaction with Duncan socioeconomic index



3.3 The Effects of Program Expansion on Choices of Ph.D. Recipients

Our next analysis focuses on the choices made by Ph.D. recipients, such as where to pursue their degree, what type of institution to attend (e.g. private or public), and which field to study. This analysis is important substantively and provides another set of checks on causality. For instance, if an increase in the number of programs in a state increases enrollments, one would expect to see some increase in nearby states, but the greatest increase at in-state institutions. Similarly, an increase in the number of programs in nearby states might increase the number of people obtaining Ph.D.s in state, but would be expected to have a larger effect on people obtaining Ph.D.s in nearby states. We only use the sub-sample of Ph.D. recipients for this part of analysis to show the relative effects across different choices conditional on receiving a Ph.D. degree.

For this analysis, we will use a Multinomial Logit Model with the set of equations below as an example. Focusing on the in state / nearby state / distant state case, we estimate the probability that someone who obtains a Ph.D. does so in state (in equation (2A)) or in a nearby state (in equation (2B)) relative to a distant (far away) state. Formally,

$$ln\frac{P(Y_i = InState)}{P(Y_i = FarAway)} = \beta_{\boldsymbol{Birth,InState\rightarrow InState}}\boldsymbol{N_{Birth,InState,i}} + \dots \boldsymbol{\beta_{70,InState\rightarrow InState}}\boldsymbol{N_{70,InState,i}}$$
$$+\beta_{\boldsymbol{Birth,Nearby\rightarrow InState}}\boldsymbol{N_{Birth,Nearby,i}} + \dots \boldsymbol{\beta_{70,Nearby\rightarrow InState}}\boldsymbol{N_{70,Nearby,i}}$$
$$+X_i\Gamma_{InState} + \eta_{InState,s} + \lambda_{InState,c} + \varepsilon_{InState,i} \qquad (2A)$$

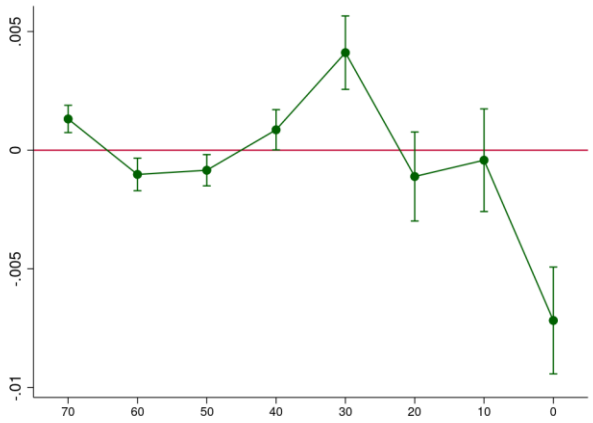
$$ln\frac{P(Y_i = Nearby)}{P(Y_i = FarAway)} = \beta_{\boldsymbol{Birth,InState\rightarrow Nearby}}\boldsymbol{N_{Birth,InState,i}} + \dots \boldsymbol{\beta_{70,InState\rightarrow Nearby}}\boldsymbol{N_{70,InState,i}}$$
$$+\beta_{\boldsymbol{Birth,Nearby\rightarrow Nearby}}\boldsymbol{N_{Birth,Nearby,i}} + \dots \boldsymbol{\beta_{70,Nearby\rightarrow Nearby}}\boldsymbol{N_{70,Nearby,i}}$$
$$+X_i\Gamma_{Nearby} + \eta_{Nearby,s} + \lambda_{Nearby,c} + \varepsilon_{Nearby,i} \qquad (2B)$$

As above, $\boldsymbol{N_{Age,InStatel,i}}$ gives the number of programs in *i*'s state of birth at the time that *i* turned age *Age* and $\boldsymbol{N_{Age,Nearby,i}}$ gives the corresponding number in nearby states. As above, we include a set of individual characteristics, X_i; a set of fixed effects (dummy variables) for the state of birth,η_s; and a set of fixed effects (dummy variables) for the year of birth, λ_c. Unlike equation (1), here we estimate a set of models, one for each outcome, for example obtaining a degree in the birth state and obtaining a degree in a nearby state, where both are relative to obtaining a degree in a distant state. As a consequence, each variable can have a different effect on each outcome. As an example, $\boldsymbol{N_{30,InState,i}}$ is expected to have a strong effect on Ph.D. receipt in state (i.e. $\boldsymbol{\widehat{\beta}_{Birth,InState\rightarrow InState} > 0}$), but it may actually reduce the probability of obtaining a degree in a nearby state ($\boldsymbol{\widehat{\beta}_{Birth,InState\rightarrow Nearby} < 0}$). Indeed, a negative coefficient would be reassuring in terms of having estimated a causal effect).

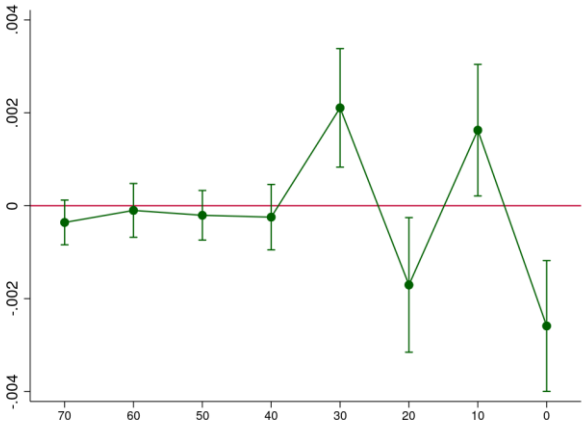
Figure 6 shows that an increase in the number of in-state programs increases the probability of obtaining a Ph.D. in-state and plots the marginal effects in each panel. Panel A and Panel B suggest that expanding Ph.D. programs in state and nearby states both increase the probability of getting a Ph.D. in home states. However, Panel B shows that the effect of an increase in the number of programs in nearby states is roughly one half the size of that of instate expansion. Panel C and Panel D also show that opening Ph.D. programs in nearby states generate the largest impact on the probability of studying in a Ph.D. program in nearby states. Panel E and Panel F further verify that increases of the number of Ph.D. programs in home states or nearby states are negatively related to the probability of obtaining a Ph.D. degree in states far away from home areas. Therefore, Figure 6 suggests the significance of geographical distance in accessing Ph.D. training.

Figure 6. Effects on location of Ph.D. attainment

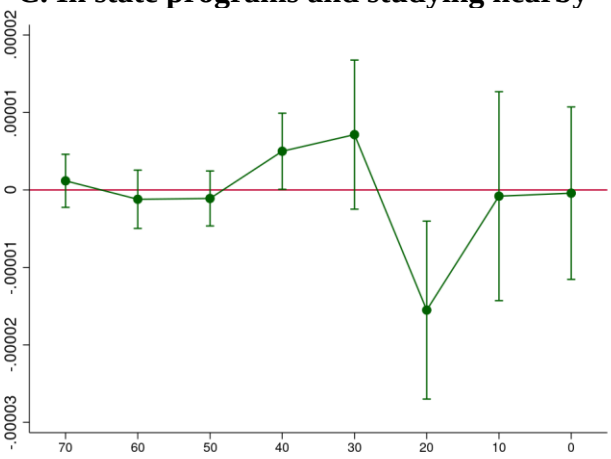
A. In state programs and studying in state



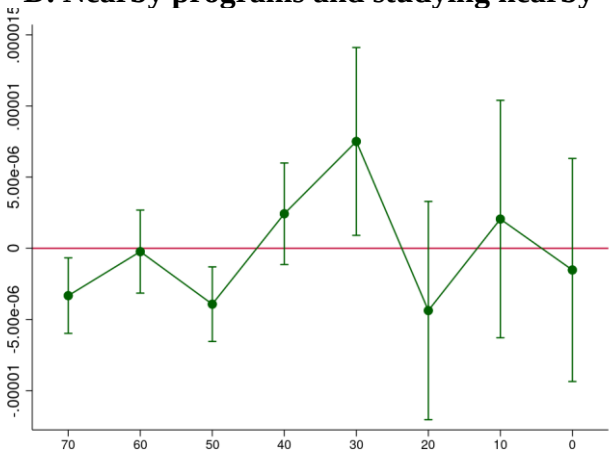
B. Nearby programs and studying in state



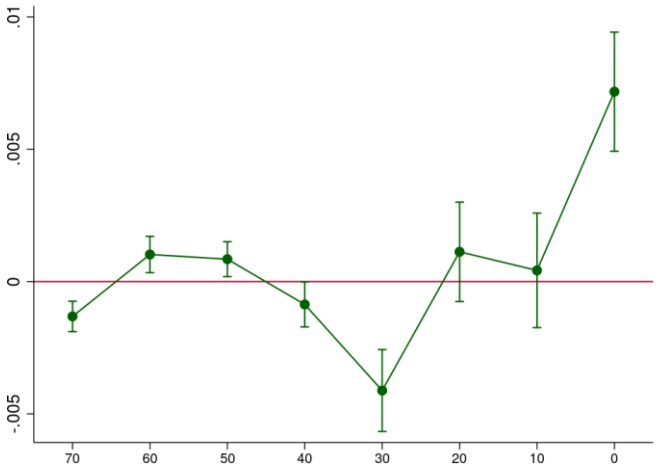
C. In state programs and studying nearby



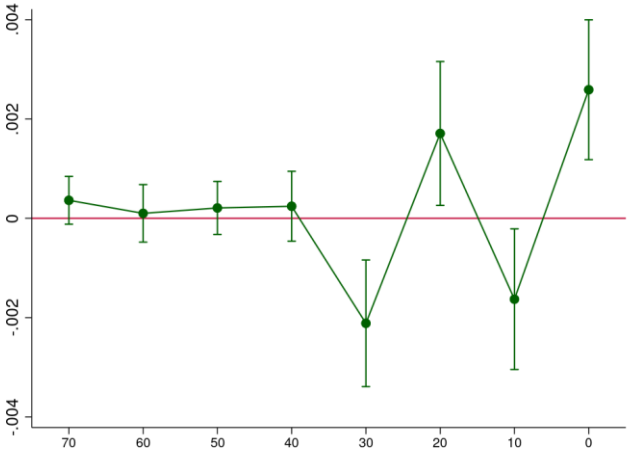
D. Nearby programs and studying nearby



E. In state programs and studying far Away



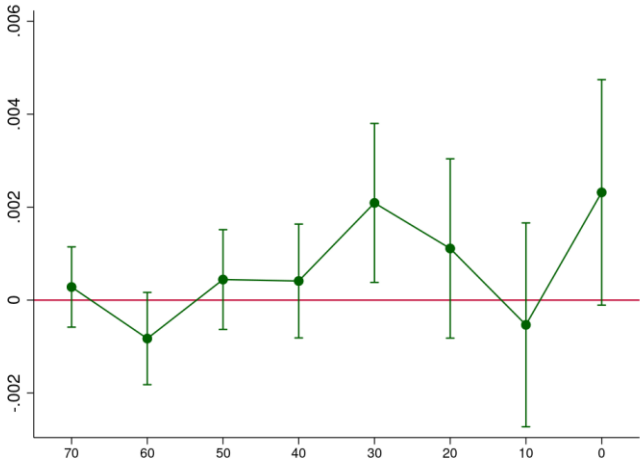
F. Nearby programs and studying far Away



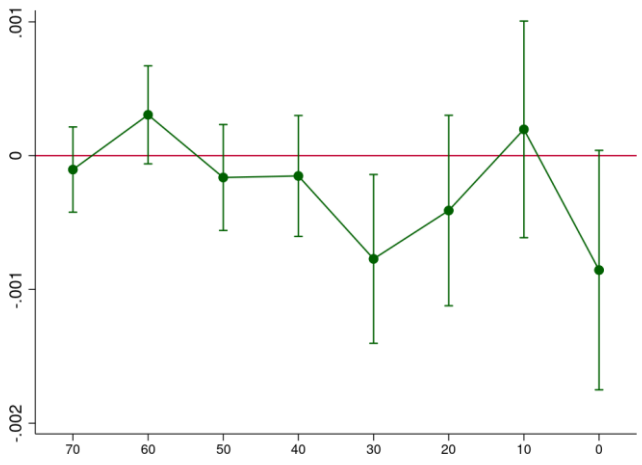
We estimate similar models by research field, so that we can see how field choices change with the fields of newly opened Ph.D. programs. Figure 7 suggests that the expansion of a specific field is positively correlated with the probability of choosing to study in that field.

Figure 7. Effects on earning a Ph.D. degree in Arts and Humanities

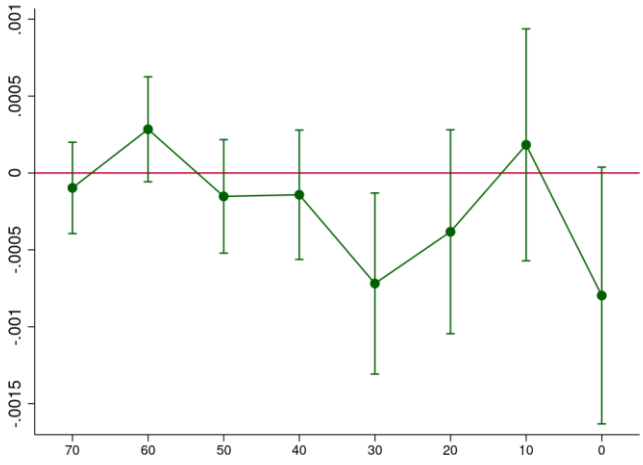
A. Arts and Humanities programs in state



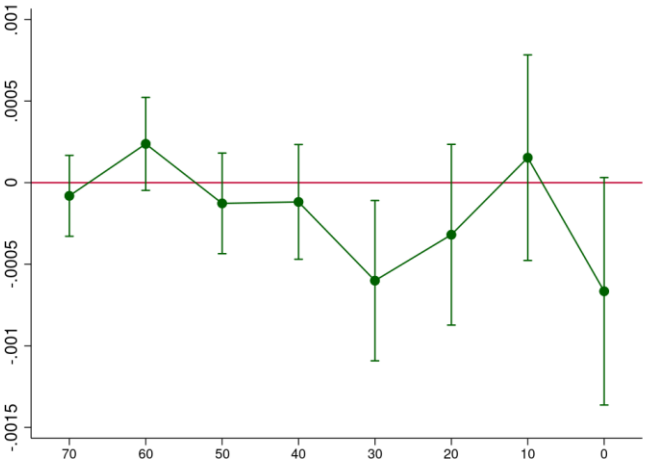
B. Social and Behavioral Sciences (Business) programs in state



C. Mathematical and Physical Sciences programs in state



D. Natural Sciences programs in state

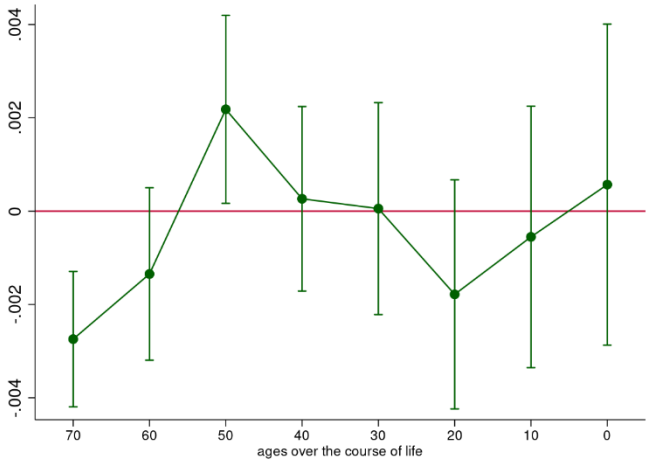


We also estimate the Multinomial Logit Model to show how the choice of attending a private versus public research program is affected by the provision of research training programs at private and public institutions. Figure 8 shows that, among all the Ph.D. recipients, the

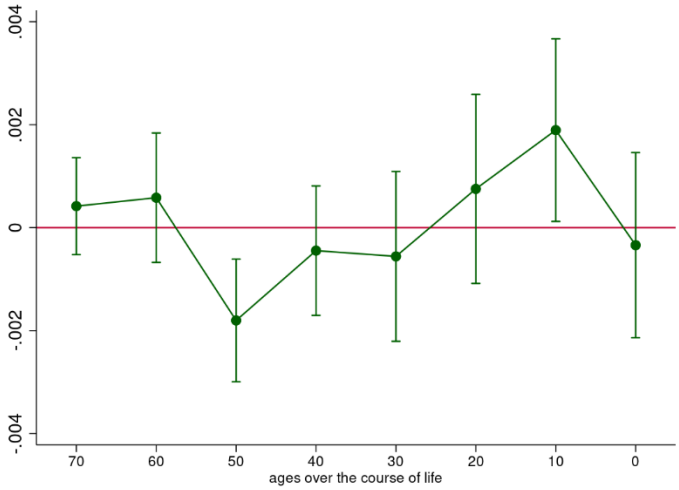
probability of attending a public program is only associated with the number of public research programs in birth state. These estimates provide useful placebo checks (for instance an increase in the number of programs at, for instance, public schools should not be causally associated with an increase in degrees at private schools.)

Figure 8. Effects on earning a Ph.D. degree from public research programs

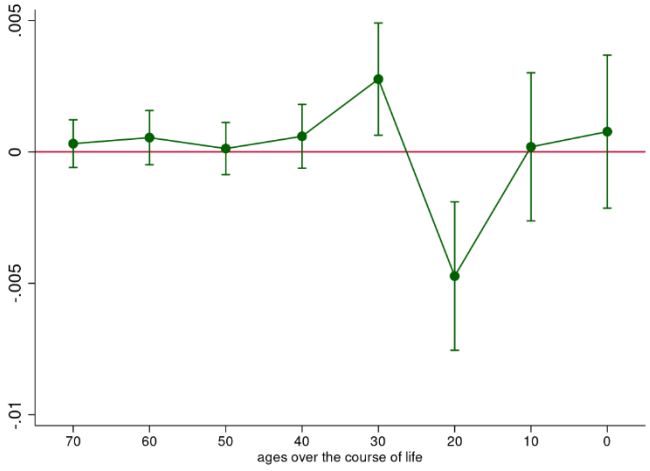
A. Private programs in state



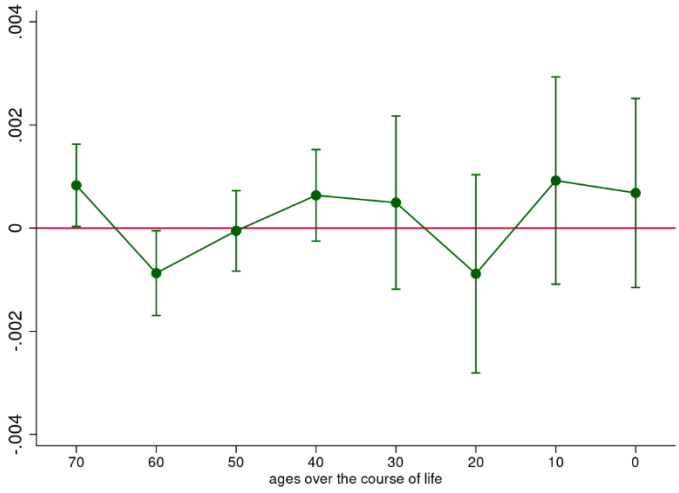
B. Private programs in nearby states



C. Public programs in state



D. Public programs in nearby states



3.4 Robustness Checks

There are a number of modeling decisions that we have made that may affect our results. First, we rely heavily on the linkage of names in ProQuest to the Census and use additional features to refine name-based matches. In particular, we use distance between the institution where people received their PhDs to where we locate them in the Census to link people between the two datasets. This may generate a spurious relationship between where people are born and where they receive their PhDs. To address this possibility, we replicate the main analysis using a sample with updated matching criteria removing distance as a variable to refine matches. These results, which are substantially unchanged, are reported in Figure A.2.

There may be trends in the propensity to obtain a PhD that are related to the opening of PhD programs stemming from demand shocks or some third source of causality. To address this concern, we control for census division-specific cohort trends. Thee estimates are reported in Figure A.3. Lastly, we recognize that the opening of PhD programs is related to the expansion of colleges in general. To address this possibility, we have collected data on the number of colleges in each state. Figure A.4 reports results that include the number of colleges in each person's state of birth as a control. All told, our main estimates are robust to matching methods and control variables.

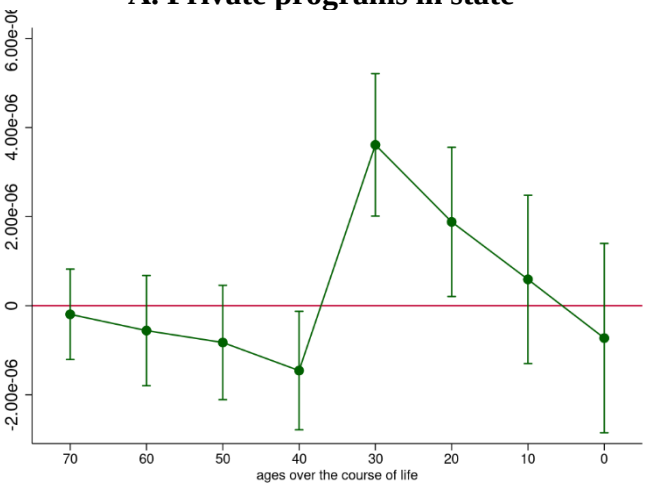
While the population rate of earning a Ph.D. degree is generated using the ProQuest and Census matched sample, we also use some census generated outcomes to validate the main estimates. Figure A.5 shows that, among all the adults above age 40, the number of programs in birth state around age 30 is also highly correlated with the probability of completing some post-college education and the chance of working in a professional occupation. This evidence using census outcomes is consistent with the main estimates in Figure 4.

3.5 Source of the Effect of Ph.D. Program Expansion

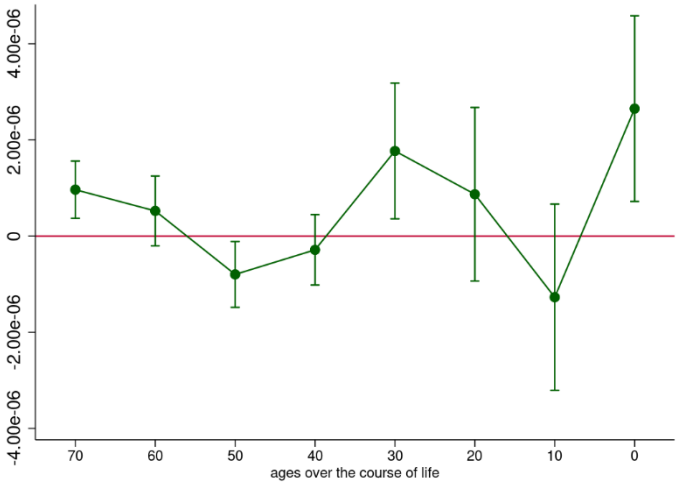
To investigate the main driving force of the estimated effect of Ph.D. program, we decompose the number of Ph.D. within each life stage and each geographic range by funding source: private versus public, and by the institutional policy establishing universities: Land Grant, HBCU, and Co-educational switch. Figure 9 presents the estimated effect of both opening private and public universities on population probability in obtaining a Ph.D. degree. The horse-race result suggests that while both type of institutions have effect on the chance of earning a Ph.D. degree, the larger impact comes from private universities. This finding is consistent with the finding that the number of newly opened private universities is about five times the number of public universities during early 20th century (Goldin and Katz, 1999).

Figure 9. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree by funding source of universities

A. Private programs in state



B. Public programs in state

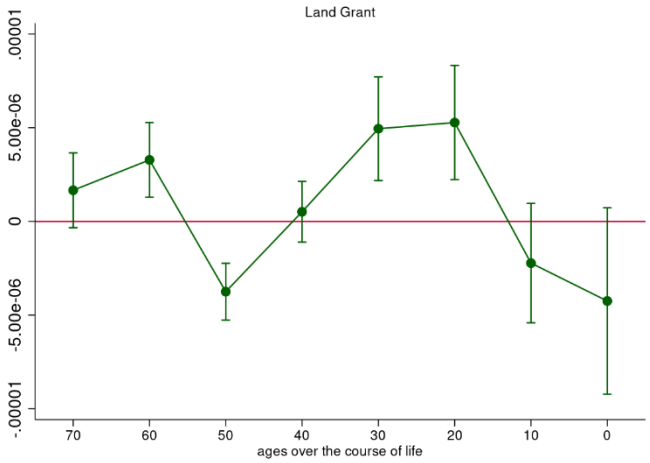


We also compare the effects of the number of research programs opened in universities that are established or reformed by different institutional policies. We mainly focus on three changes: Land Grant, HBCU, and Co-education. Figure 10 presents the horse-raced effects of the number of Ph.D. programs in Land Grant, HBCU, and Co-education universities opened in birth

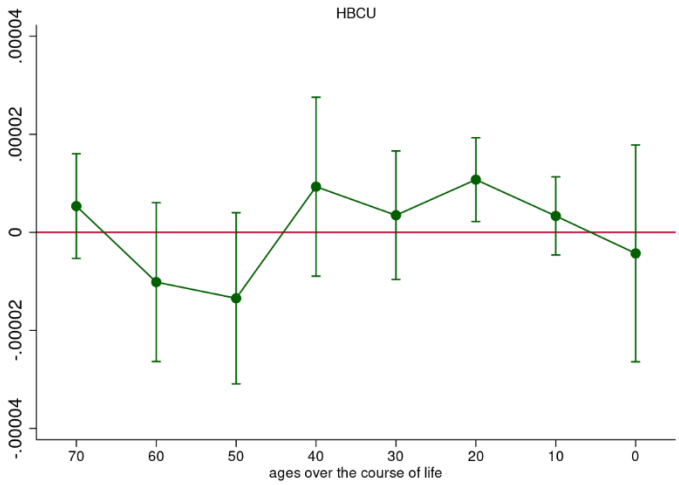
state. It suggests that neither HBCU opening nor Co-education switching contributes to the effect that increasing population probability in obtaining Ph.D. degrees, and only the Ph.D. programs opened in Land Grant universities significantly increases the average chance of getting a Ph.D. degree. Figure 11 further confirms that the university level total number of Ph.D. programs and the total number of Ph.D. students per program both increases after the Land Grant universities are opened.

Figure 10. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree by policy of institutional changes

A. Land Grant



B. HBCU



C. Co-edu

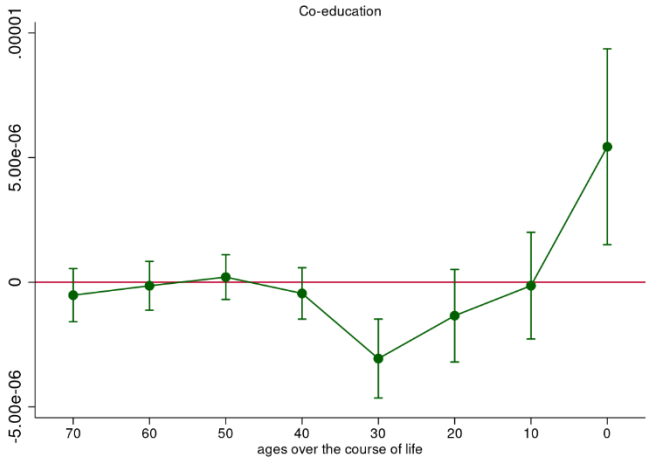
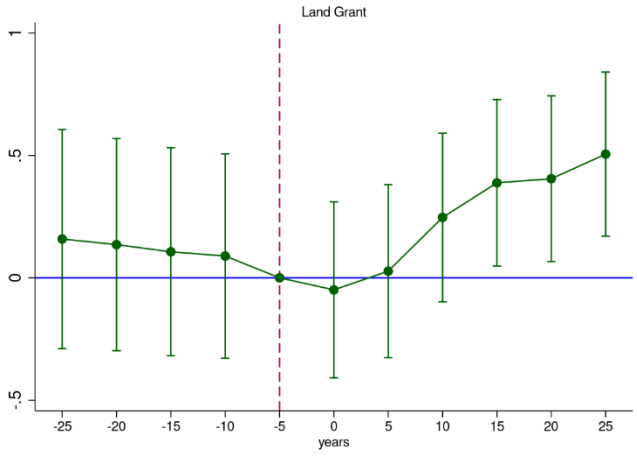
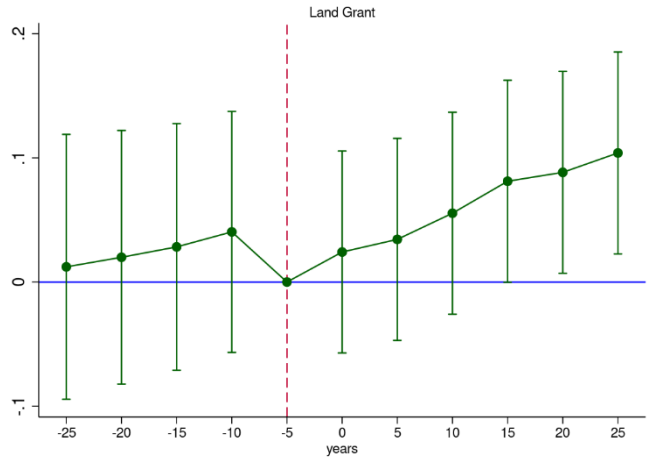


Figure 11. Event study estimates of one education institutional change on log total number of research programs per university and log total number of Ph.D. students

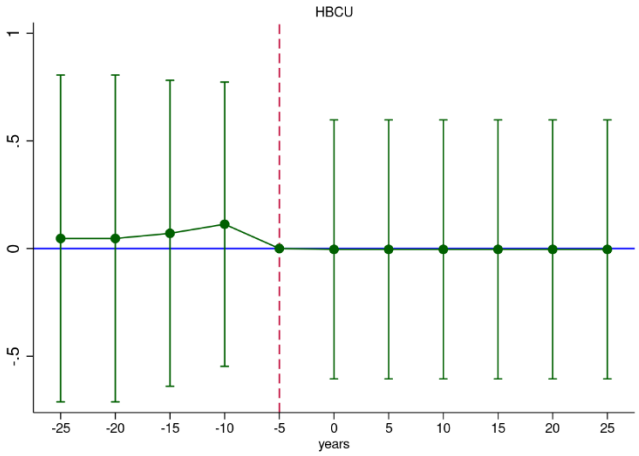
A. Land Grant: # of Ph.D. programs



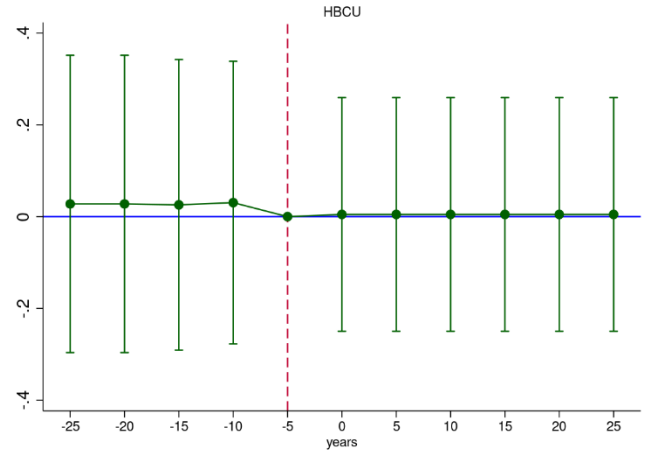
B. Land Grant: # of Ph.D. students per program



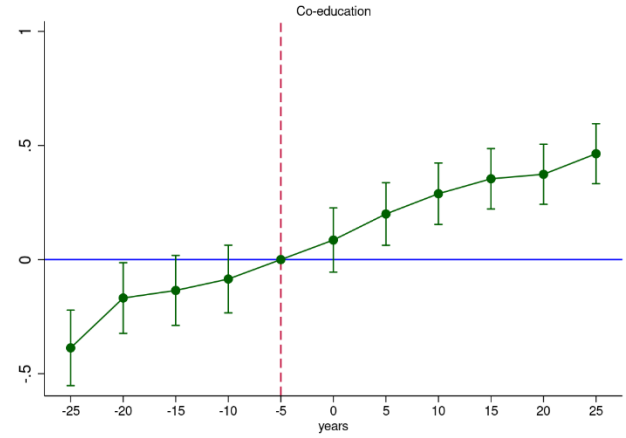
C. HBCU: # of Ph.D. programs



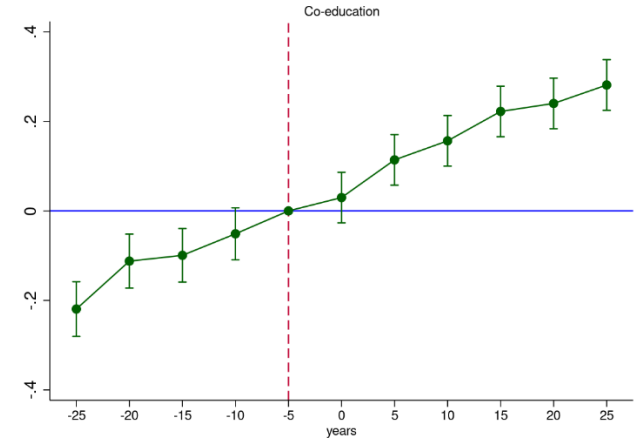
D. HBCU: # of Ph.D. students per program



E. Co-edu: # of Ph.D. programs



F. Co-edu: # of Ph.D. students per program



4. Conclusion and Future Work

We have shown 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 massive transformation in terms of geography, subject matter, institutions and, in many ways, participation. To quantify the effect of program expansion on participation, we find that the opening of a new Ph.D. program generated 2 additional Ph.D. recipients per 1 million residents in a state. One important question relates to whether increasing innovation in one location reflects an increase in overall innovation or whether it crowds out innovation elsewhere. Our finding that new programs differed along many dimensions from existing ones suggests that there may well have been limited crowding out.

Our ultimate objective is to estimate the effect of increased numbers of programs on the quantity and quality of innovation itself. To do this, we return to equation (1) and re-estimate it using linear or count models with the number or quality of patents awarded to people as the dependent variable using the same event study framework. Here we will make use of the availability of direct links from inventors on US patents to respondents to the historical censuses (Doran and Yoon, 2020). These estimates will indicate whether the presence of more Ph.D. programs in the state in which someone was born increases innovation among people who are likely exposed to them at formative ages. As above, we can include interaction terms between the presence of Ph.D. programs and individual characteristics to explore what types of people are induced to patent more by the presence of Ph.D. programs. And, as above, the event study framework will allow us to use timing to shed light on causal effects.

There are two channels by which Ph.D. programs might increase patenting. First, people who are induced by the presence of a program to obtain a Ph.D. may obtain more patents than they

otherwise would, a "direct" effect of Ph.D. receipt. Second, the presence of a Ph.D. program may have a "spillover" effect by providing inspiration, knowledge, or access to resources for people (e.g. undergraduates at the same institutions) who do not obtain a Ph.D. themselves. While our design does not allow us to distinguish these effects, we can at least perform some back-of-the-envelope calculations to get a general sense of the relative importance of the two channels. Specifically, we will estimate how many patents Ph.D. recipients receive and combine these estimates with the causal effect of program access on Ph.D. receipt to get a rough sense of the direct effect. We can then quantify the spillover effect by observing the patents of people "exposed" to a new Ph.D. program (geographically and temporally) who were not themselves the right age to be at risk of earning a Ph.D.

A similar set of analyses can be conducted using the occupation and industry data in the Censuses. We can estimate the probability of being in an innovative occupation and/or industry. Here we will explore using the industry as an analogue of the field or technology class (i.e. people in innovative careers in agriculture to identify people who are innovating in agriculture). We can obtain some mapping between industries and technologies using the industries of patent inventors.

References

Abramitzky, Ran, Leah Boustan and Myera Rashid. Census Linking Project: Version 1.0 [dataset]. 2020.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas. "The Rise of American Ingenuity: Innovation and Inventors of the Golden Age," NBER Working Paper N. 23047, 2017.

Aghion, Philippe, Ufuk Akcigit, Ari Hyytinen, and Otto Toivanen, "The SocialOrigins of Inventors," NBER Working Paper No. 24110, 2017.

Alesina, Alberto, and Eliana La Ferrara. "Ethnic diversity and economic performance." *Journal of Economic Literature* 43.3 (2005): 762-800.

Azoulay, Pierre, Wesley H. Greenblat, and Misty L. Heggeness. 'Long-term Effects from Early Exposure to Research: Evidence from the NIH "Yellow Berets"' NBER Working Paper No. 26069, 2019.

Bell, Alex, Raj Chetty, Xavier Jaravel, Neviana Petkova, and John Van Reenen. "Who Becomes an Inventor in America? The Importance of Exposure to Innovation." *The Quarterly Journal of Economics,* 134(2): 647-713.

Berkes, Enrico and Peter Nencka, "Knowledge Access: The Effects of Carnegie Librarieson Innovation," Working Paper, Ohio State University. 2020.

Borjas, George J. and Kirk B. Doran. "The Collapse of the Soviet Union and the Productivity of American Mathematicians." *The Quarterly Journal of Economics,* 1143-1203.

Bound, John and Sarah Turner. 2002. "Going to War and Going to College: Did World War II and the GI Bill Increase Education Attainment for Returning Veterans?" *Journal of Labor Economics,* 20(4): 784-815.

Catherine Buffington, Benjamin Cerf, Christina Jones, Bruce A. Weinberg. STEM Training and Early Career Outcomes of Female and Male Graduate Students: Evidence from UMETRICS Data linked to the 2010 Census. 2016. *American Economic Review* 106 (No. 5, May): 333-38.

Card, David, and Laura Giuliano, "Does Gifted Education Work? For Which Stu-dents?," NBER Working Paper no. 20453, 2014.

Chetty, Raj, John Friedman, Emmanuel Saez, Nicholas Turner, and Danny Yagan,"Mobility Report Cards: The Role of Colleges in Intergenerational Mobility,"NBER Working Paper no. 23618, 2017.

Clay, Karen, Jeff Lingwall, and Melvin Stephens, Jr. 2012. "Do Schooling Laws Matter? Evidence from the Introduction of Compulsory Attendance Laws in the United States." NBER Working Paper No. 18477.

Cook, Lisa and Chaleampong Kongchareon. “The Idea Gap in Pink and Black,” NBER Working Paper No. 16331, September 2010

Doran, Kirk and Chungeun Yoon. 2020 “Immigration and Invention: Evidence from the Quota Acts.” Unpublished manuscript.

Geiger, Roger L. *To advance knowledge: The growth of American research universities, 1900-1940*. Routledge, 2017.

Gordon, Ethan K., Zackary B. Hawley, Ryan Carrasco Kobler, and Jonathan C. Rork. 2019. “The Paradox of HBCU Graduation Rates.” *Research in Higher Education*.

Goldin, Claudia and Lawrence F. Katz. “Why the United States Led in Education: Lessons from Secondary School Expansion, 1910 to 1940.” In: *Human Capital and Institutions*, ed. D. Eltis, F. Lewis, and K. Sokoloff. Cambridge University Press.

Goldin, Claudia, and Lawrence F. Katz. "The shaping of higher education: The formative years in the United States, 1890 to 1940." *Journal of Economic Perspectives* 13, no. 1 (1999): 37-62.

Hofstra, B., Kulkarni, V. V., Galvez, S. M. N., 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.

Hoxby, Caroline and Christopher Avery, "The Missing "One-Offs": The Hidden Supply of High-Achieving, Low-Income Students," *Brookings Papers on Economic Activity*, 2014.

Husbands Fealing, Kaye, Yufeng Lai, and Samuel L. Myers Jr. “Pathways vs. pipelines to broadening participation in the STEM workforce” *Journal of Women and Minorities in Science and Engineering* 21(4), 271-93.

Khan, B. Zorina, and Kenneth L. Sokoloff, “Schemes of Practical Utility: En-trepreneurship and Innovation among Great Inventors in the United States,1790–1865,”Journal of Economic History, 53 (1993), 289–307.

Lamoreaux, Naomi R. and Kenneth L. Sokoloff, “Understanding the Regional Dimensions of Technological Change: Inventive Activity in the United States during the Late Nineteenth and Early Twentieth Centuries,” *Economic History Yearbook* (Moscow, 2003), 539-65.

Marschke, Gerald, Allison Nunez, Bruce A. Weinberg, Huifeng Yu. “Last Place? The Intersection of Ethnicity, Gender, and Race in Biomedical Authorship.” 2018. *American Economic Review Papers and Proceedings* 108 (No. 5, May): 222-27.

Mora, Marie T. 2019. “Best Practices in Mentoring Underrepresented Minority Women in Economics: An Introduction.” *AEA CSWEP Newsletter*, Spring 2019, pp. 1, 3.

Mora, Marie T. 2018. “Voices from the Field: Diversity in Academic Administration.” *The Minority Report*, Issue 10, pp. 12-14.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German Jewish Émigrés and US Invention." *American Economic Review*, 104(10): 3222-3255.

Myers, Kyle. "The Elasticity of Science." *American Economic Journal: Applied Economics* 12, no. 4 (October 2020): 103–134.

Østergaard, Christian R., Bram Timmermans, and Kari Kristinsson. "Does a different view create something new? The effect of employee diversity on innovation." *Research policy* 40.3 (2011): 500-509.

Perna, Laura 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): 267-94.

Universities-National Bureau Committee for Economic Research and Committee on Economic Growth of the Social Science Research Council. *The rate and direction of inventive activity: Economic and social factors*. Princeton University Press, 1962.

Weinberg, Bruce. A., Jason Owen-Smith, Rebecca Rosen, Lou Schwarz, Barbara McFadden Allen, Roy Weiss, Julia Lane. "Science funding and short-term economic activity." 2014. *Science, 344*(6179), 41-43.

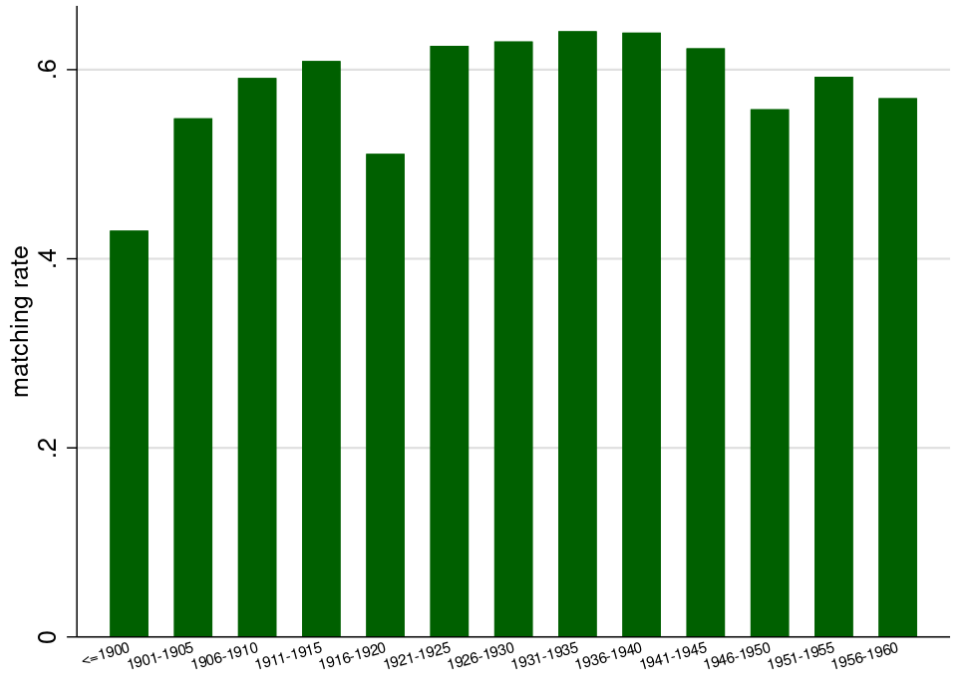
Whalley, Alex and Shawn Kantor. 2014. "Knowledge Spillovers from Research Universities: Evidence from Endowment Value Shocks." *Review of Economics and Statistics*, 96(1): 171-188.

Zolas, Nik, Nathan Goldschlag, Ronald Jarmin, Paula Stephan, Jason Owen-Smith, Rebecca Rosen, Bruce A. Weinberg, and Julia I. Lane. "Wrapping it up in a person: Examining employment and earnings outcomes for Ph. D. recipients." 2015. *Science, 350*(6266), 1367-1371. https://doi.org/10.1126/science.aac5949

Appendix

Figure A.1. Matching Rates of ProQuest to Census

A. By Degree Year



B. By Field

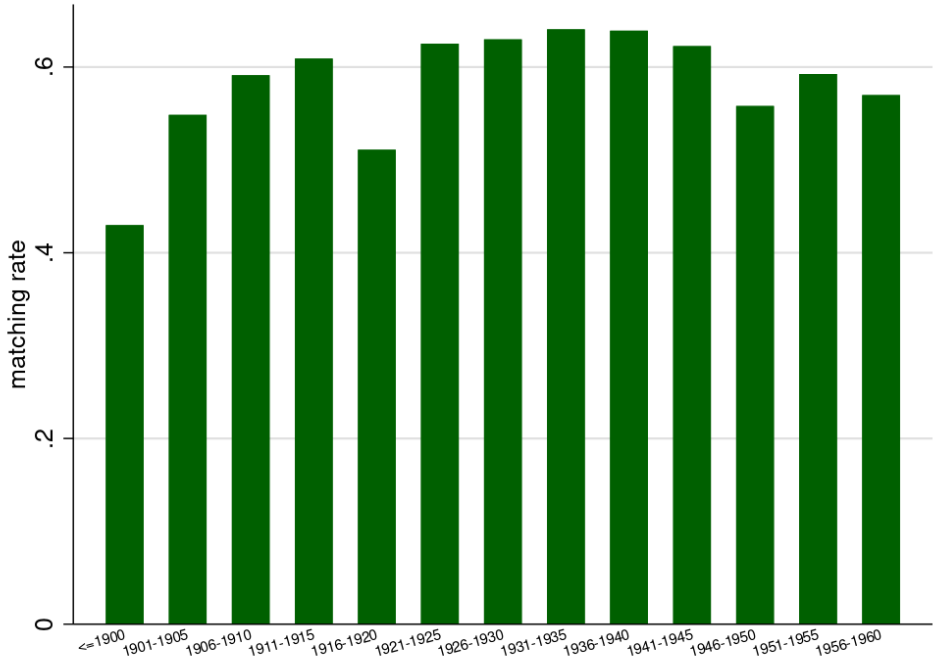
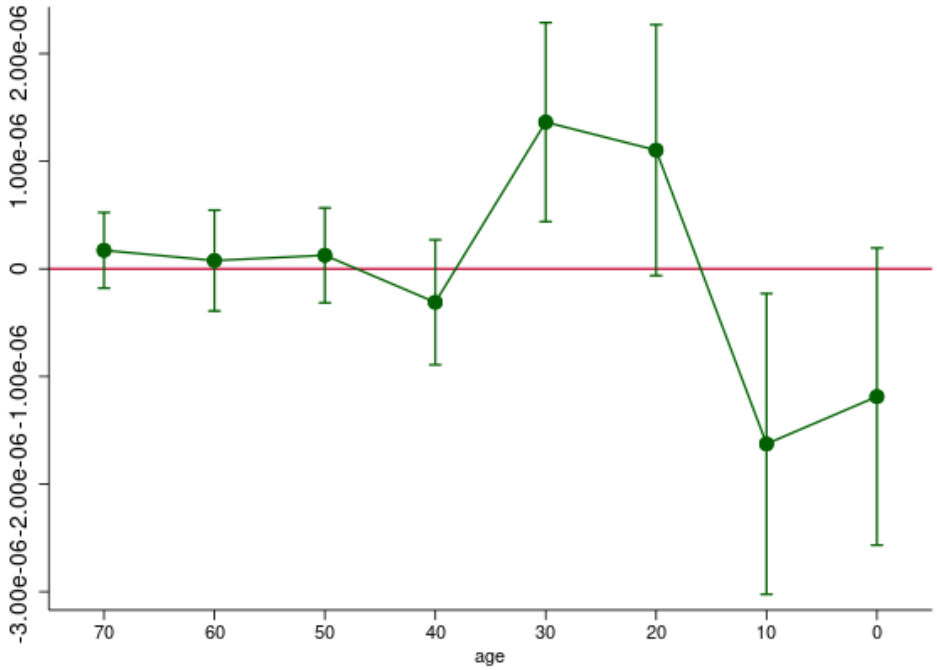


Figure A.2. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree (revised matching)

A. Programs in the state of birth



B. Programs in nearby states

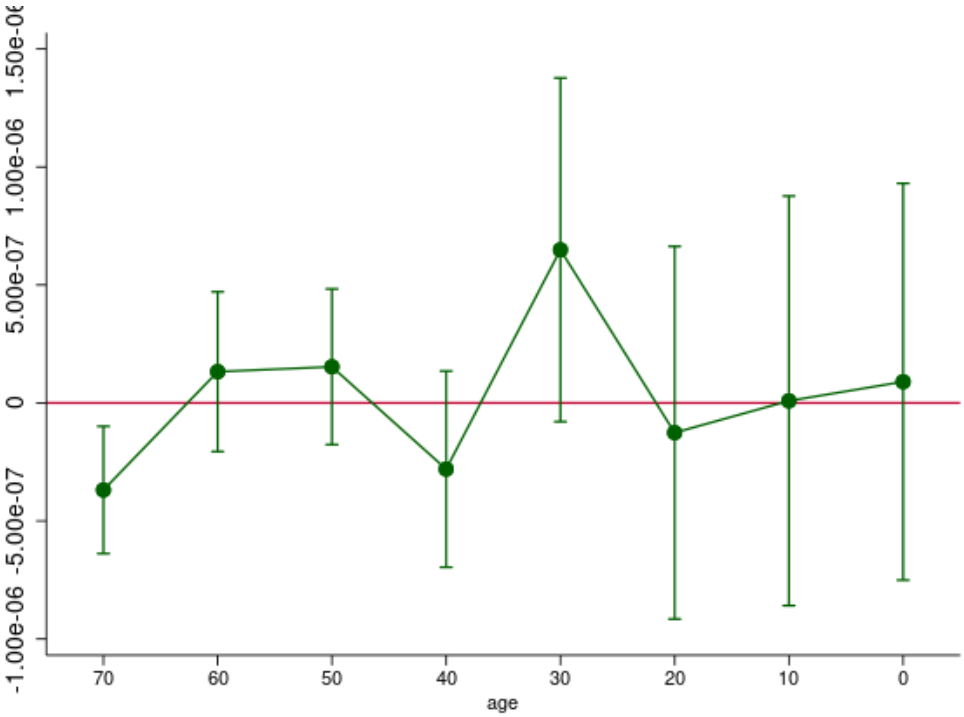
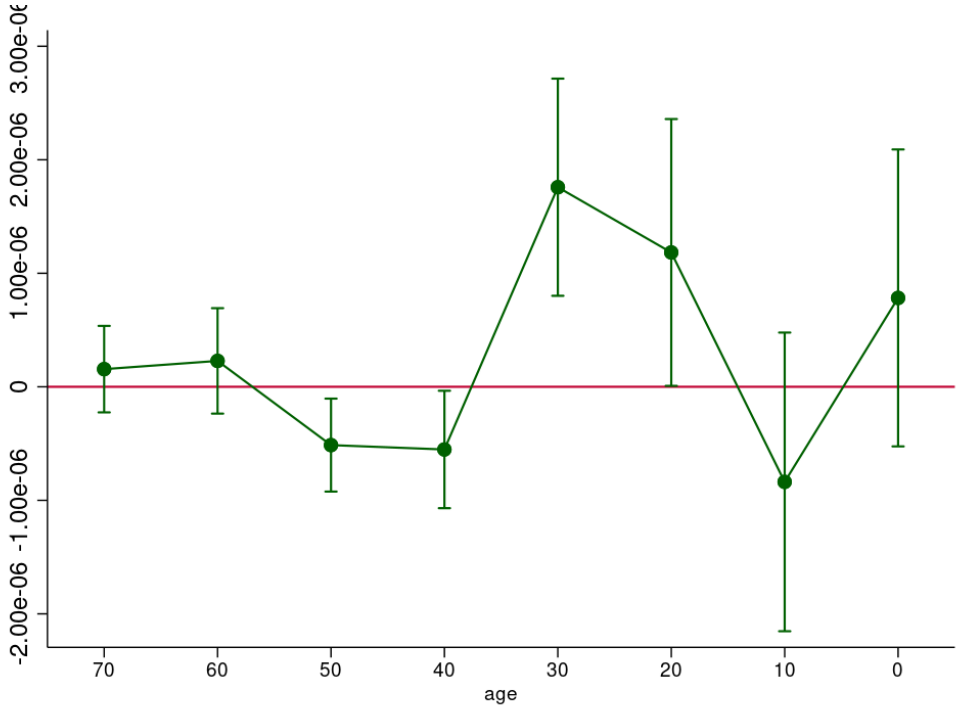


Figure A.3. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree (controlling division-cohort trends)

A. Programs in the state of birth



B. Programs in nearby states

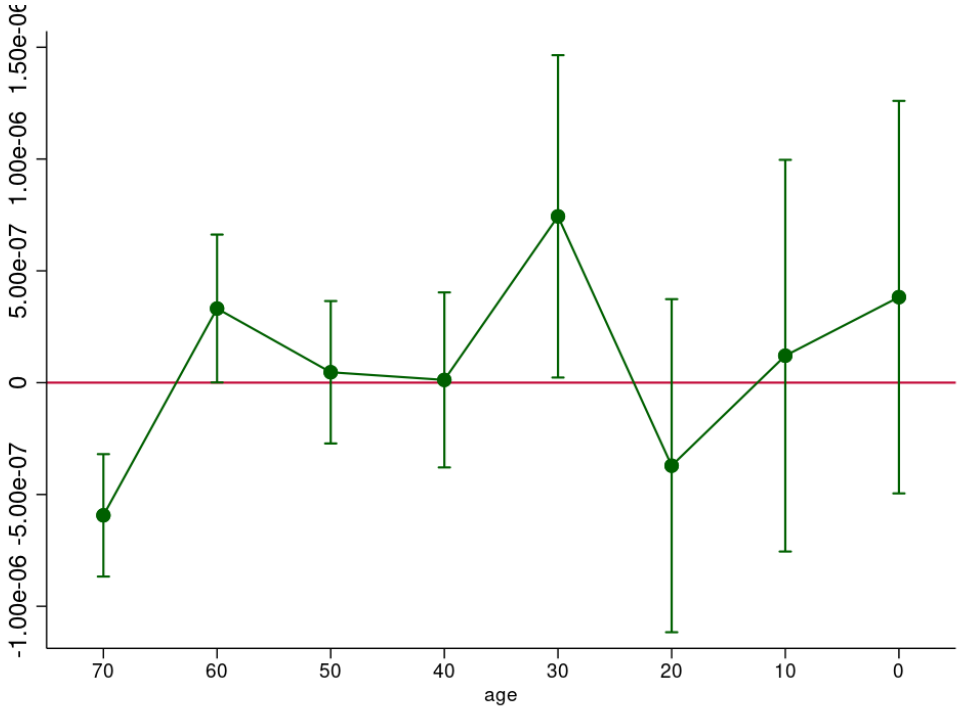
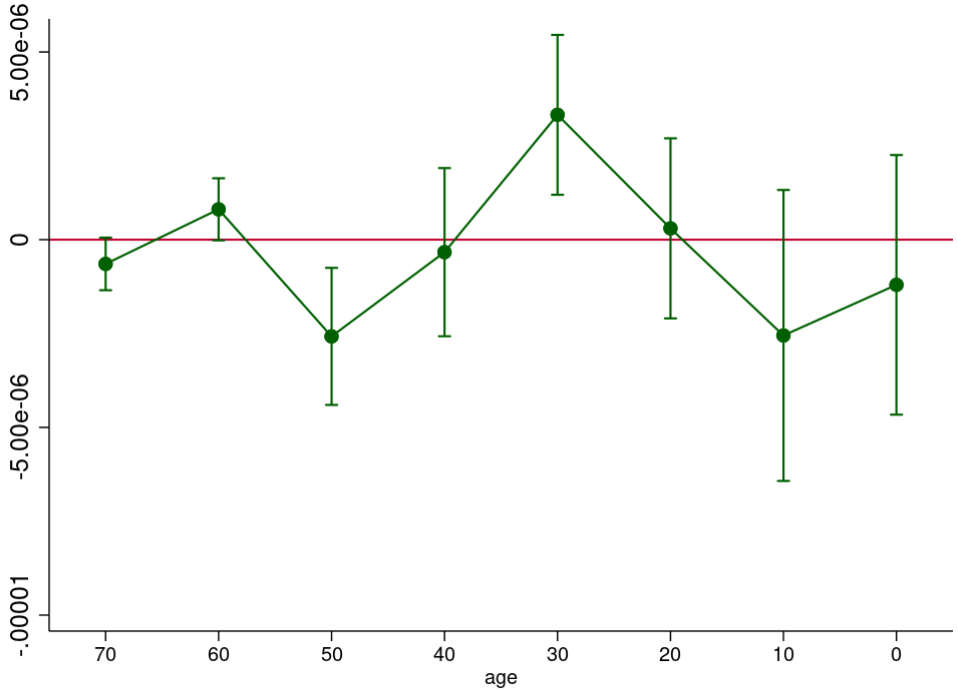


Figure A.4. Marginal effects of the number of Ph.D. programs on the probability of ever earning a Ph.D. degree (controlling the number of colleges)

A. Programs in the state of birth



B. Programs in nearby states

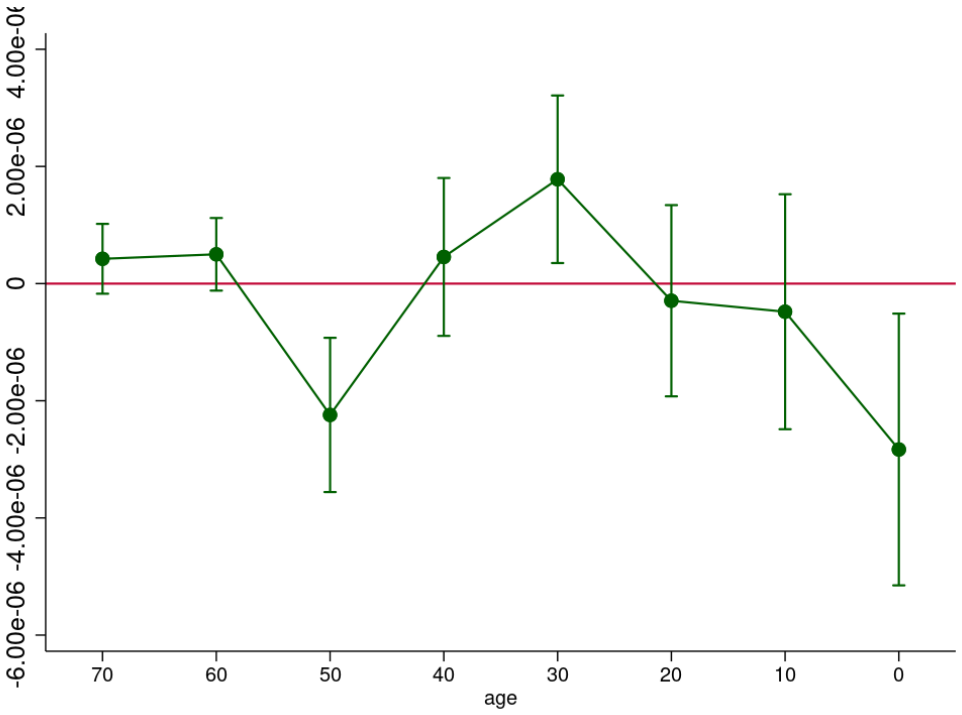
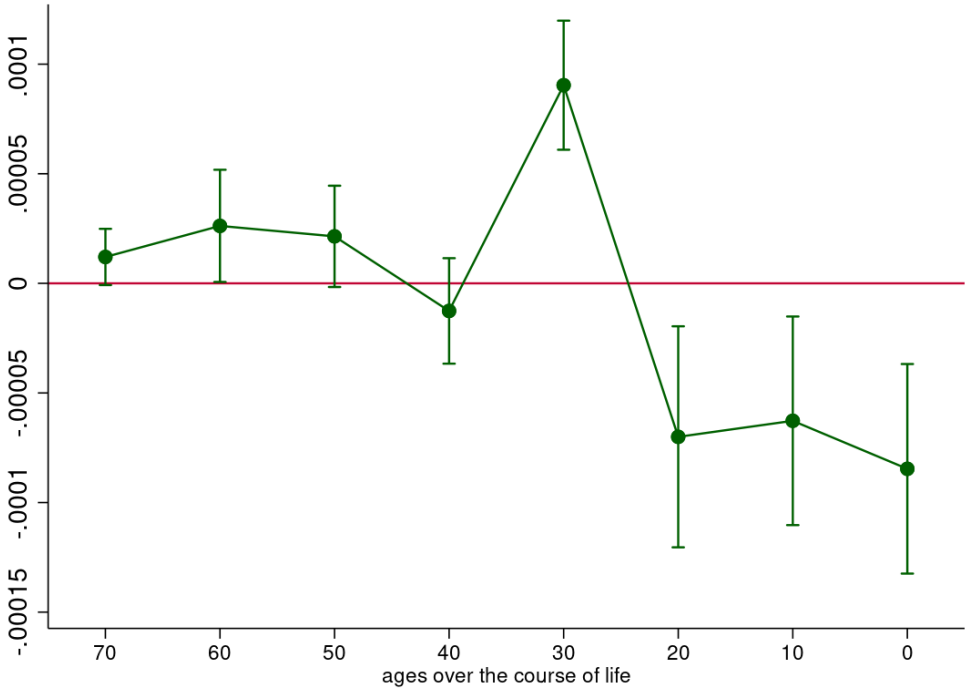


Figure A.5. Marginal effects of the number of Ph.D. programs on census outcomes among all adults older than 40 years old in 1940

A. Completing post-college education



B. Working in a professional job

