From the holdings of the Department of Special Collections , Princeton University Library, Princeton, NJ

These documents are either in the public domain, or can only be used for educational and research purposes (Fair use) as per U.S. Copyright law (text below). By accessing this file, all users agree that their use falls in the public domain or within fair use as defined by the copyrightlaw. We do not charge any permission or usage fees for the publication of images of material inour collections, including those provided by Princeton University Library, via our website, catalog, or directly from staff.For more information regarding copyright, credit, and citation please review the information onthe Copyright, Credit and Citation Guidelines page of our website.
https://library.princeton.edu/special-collections/policies/copyright-credit-and-citation-guidelines. If you have any questions, including how to order higher-quality images from the collection, contact Special Collections Public Services via our Ask Us! Form. https://library.princeton.edu/special-collections/ask-us

---

Doctoral Dissertations, 2011-2026 Economics

2023

Essays on Economic History, Immigration, and Human Capital

Hao, Stephanie

Princeton, NJ : Princeton University

---

http://arks.princeton.edu/ark:/88435/dsp01xs55mg34f

U.S. Copyright law: The copyright law of the United States (Title 17, §108, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be used for any purpose other than private study, scholarship, or research. If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of fair use that user may be liable for copyright infringement. This institution reserves the right to refuse to accept a copying order if, in its judgment, fulfillment of the order would involve a violation of copyright law. https://www.copyright.gov/title17/title17.pdf

Essays on Economic History, Immigration, and Human Capital

Stephanie Hao

A Dissertation
Presented to the Faculty
of Princeton University
in Candidacy for the Degree
of Doctor of Philosophy

Recommended for Acceptance
by the Department of
Economics
Advisers: Leah Boustan and Ilyana Kuziemko

May 2023

Abstract

This dissertation consists of three chapters on topics in labor economics and economic history.

Chapter 1 documents the shifting relationship between birth order and educational attainment throughout U.S. history. I compile a new dataset of siblings spanning over one hundred years. After controlling for family fixed effects, I find that older siblings in the 19th century were less likely to be in school and completed fewer years of education than their younger siblings. This positive birth-order effect has gradually decreased, switching sign for cohorts born in the 1930s and remaining negative until the modern day. The decline of child labor in the U.S. in the early 1900s may have contributed to this trend and may explain the different birth-order effects observed in high- versus low-income countries today.

Chapter 2 explores the effect of high-skilled immigration on U.S. innovation. I compare changes in patenting in county-research field cells with different levels of exposure to the 1992 Chinese Student Protection Act, which exogenously increased the number of high-skilled immigrants in the U.S. Cells with a 1-percentage-point higher exposure to the CSPA experienced a 3 percent increase in patent applications post-1992. Increases in private sector non-academic patenting drive the results. The number of new inventors also increased in more-exposed cells. These findings contribute to the policy debate on the effects of extending immigration opportunities to high-skilled foreign workers in the U.S. today.

Chapter 3, co-authored with Ran Abramitzky, Leah Boustan, and Katherine Eriksson, investigates whether the perceived cultural assimilation of immigrants affects their labor market outcomes. We document that, in the early 20th century, children of immigrants who were given more-foreign first names completed fewer years of schooling, earned less, and married less assimilated spouses. However, we find few differences in the adult outcomes of brothers with more and less foreign-sounding first

names. This pattern suggests that the negative association between ethnic names and adult outcomes in this era does not stem from discrimination on the basis of first names but instead reflects household differences associated with cultural assimilation. We cannot rule out discrimination on the basis of other ethnic cues.

Acknowledgements

I would first like to thank my advisers, Leah Boustan and Ilyana Kuziemko, for their invaluable guidance and support. Their knowledge and experience, patient and enthusiastic mentorship, and thoughtful and incisive suggestions have been instrumental in shaping not only the essays in this dissertation but also my growth as an economist. I am incredibly grateful for the time, advice, and encouragement they have given me throughout the PhD.

I have also benefited greatly from comments and feedback from other Princeton faculty members, including Janet Currie, Ellora Derenoncourt, Hank Farber, David Lee, and Alex Mas. I am grateful for the support provided by the staff of the Industrial Relations Section – Valerie Ching, Indiana Gaffney, Jeannie Moore, Gene Piccola, and Patti Tracey – and by Laura Hedden. I gratefully acknowledge the financial support provided by the Graduate School and by the Industrial Relations Section. I also thank my professors at UC San Diego, Julie Cullen and Ross Starr, for introducing me to economic research and encouraging me to apply to graduate school.

I thank my classmates and friends at Princeton University. I especially thank Dorisz Albrecht for being there since day one, for teaching me new things, and for always making me laugh. I thank Chris Szerman for her wisdom, thoughtfulness, and generosity. I thank my peers in the Industrial Relations Section, including Christine Cai, Jiwon Choi, Rachel Fung, Carl Lieberman, Chris Mills, and Carolyn Tsao, for their insightful comments and conversation. I thank my PhD cohort for their friendship over the past six years.

Finally, I am grateful to my family for their unconditional love and support. My parents, my sister, Cynthia, and my grandparents have believed in me and encouraged me every step of the way – thank you for everything.

For my grandparents.

Contents

Abstract ii

Acknowledgements iv

List of tables ix

List of figures xii

1 Birth Order and Educational Attainment Throughout the 20th Century **1**

1.1 Introduction 1

1.2 Data 5

1.2.1 Data using school attendance as outcome variable 6

1.2.2 Data using years of education as outcome variable 8

1.3 Empirical strategy 12

1.4 Results 13

1.5 Mechanisms 15

1.5.1 Decreasing family size 15

1.5.2 Decline in prevalence of child labor 17

1.6 A model of intra-household allocation of schooling 21

1.7 Conclusion 26

1.8 Tables and figures 28

Appendices **36**

1.A Appendix tables and figures 36

1.B Data appendix . . . 51
1.B.1 Data using school attendance as outcome variable . . . 51
1.B.2 Data using years of education as outcome variable . . . 53

2 Green Card Eligibility of High-Skilled Workers and U.S. Innovation 58
2.1 Introduction . . . 58
2.2 Background on the Chinese Student Protection Act . . . 62
2.3 Data . . . 64
2.3.1 U.S. patent data . . . 65
2.3.2 Patent applications . . . 65
2.3.3 Determining CSPA exposure . . . 66
2.3.4 Characteristics of final dataset . . . 69
2.4 Empirical strategy . . . 70
2.5 Results . . . 72
2.5.1 Robustness . . . 74
2.6 Mechanisms . . . 75
2.6.1 Academic vs. non-academic patenting . . . 75
2.6.2 Extensive vs. intensive margin . . . 76
2.7 Conclusion . . . 78
2.8 Tables and figures . . . 80

Appendices 92
2.A Appendix tables and figures . . . 92
2.B Data appendix . . . 110

3 Discrimination and the Returns to Cultural Assimilation in the Age of Mass Migration 112
3.1 Introduction . . . 112
3.2 Data and definitions . . . 115

3.2.1 Measuring the foreignness of given names 115
3.2.2 Creating a linked Census sample 116
3.3 Foreign names and outcomes in the 1940 labor and marriage markets 117
3.4 Conclusion 121
3.5 Tables and figures 122

Appendices **124**

3.A Appendix tables and figures 124

Bibliography **151**

List of Tables

1.1 Summary statistics 29
1.2 Summary statistics by probability of child labor 30
1.A.1 Coefficients in birth-order literature 37
1.A.2 Summary statistics for 1870 to 1940 Census sample 38
1.A.3 Summary statistics for 1900 and 1910 Census subsample matched to 1940 Census 39
1.A.4 Probit regression results for probability of labor force participation – Pooled Census CHSURV sample 40
1.A.5 Results by family size and individual birth-order position – Pooled NLS and WLS sample 41

2.1 NBER technology categories and subcategories 80
2.2 Summary statistics by county-NBER subcategory cell 81
2.3 Effect of the CSPA on patent applications 81
2.4 Patent applications for non-academic vs. academic assignees 82
2.5 Results for extensive vs. intensive margin 82
2.A.1 Top 15 most frequent non-immigrant visa classes of CSPA LPR recipients 92
2.A.2 Top 15 most frequent classes of admission of PRC green card recipients between 1972-2000 93
2.A.3 Most common surnames on patent applications between 1984-1988 by ethnicity 94

2.A.4 Distribution of inventors across NBER categories by ethnicity . . . 95
2.A.5 PRC inventors and patent applications by NBER subcategory, 1984-1988 . . . 96
2.A.6 PRC inventors and patent applications by county, 1984-1988 . . . 96
2.A.7 Results using share of other pre-period ethnic inventors as exposure measure . . . 97
2.A.8 Robustness to alternative definitions of outcome variable . . . 97
2.A.9 Results for placebo CSPAs in pre-period . . . 98

3.1 Effect of given name foreignness on labor and marriage market outcomes . . . 122
3.A.1 Comparing mean characteristics of matched and unmatched second-generation men . . . 125
3.A.2 Effect of given name foreignness on labor and marriage market outcomes: 5% of full 1940 population . . . 126
3.A.3 Effect of F-indices calculated at birth and at age 20 on labor and marriage market outcomes . . . 127
3.A.4 Effect of given name foreignness on labor and marriage market outcomes: men with names below median frequency . . . 128
3.A.5 Effect of given name foreignness on labor and marriage market outcomes: men with names at or above median frequency . . . 129
3.A.6 Effect of given name foreignness on probability of self-employment 130
3.A.7 Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 3-year band, age difference within 5-year band . . . 131
3.A.8 Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 3-year band, age difference within 3-year band . . . 132

3.A.9 Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 5-year band, age difference within 5-year band 133
3.A.10 Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 5-year band, age difference within 3-year band 134
3.A.11 Effect of given name foreignness on labor and marriage market outcomes: EM matches 135
3.A.12 Effect of given name foreignness on labor and marriage market outcomes: ABE-conservative matches restricted to states used in EM sample 136

List of Figures

1.1 Composition of sample by birth cohort of eldest child (NLS and WLS) 31
1.2 Effect of birth order on educational attainment over time 32
1.3 Birth-order effects on high school completion and college attendance 33
1.4 Main results reweighted by family size distribution 34
1.5 Heterogeneity analysis 35
1.A.1 School attendance and labor force participation results using 1870 to 1940 Censuses 42
1.A.2 Years of education results for matched Census subsample 43
1.A.3 Share of 10- to 17-year-old children in 1870 to 1940 Censuses in school 44
1.A.4 Unweighted results for Figure 1.5 45
1.A.5 Heterogeneity analysis 46
1.A.5 Heterogeneity analysis (cont.) 47
1.A.6 Share of 10- to 17-year-old children in 1870 to 1940 Censuses in labor force 48
1.A.7 Sample split by propensity for child labor, separately by region . . 49
1.A.8 Composition of sample by family size (NLS and WLS) 50

2.1 Number of immigrants admitted to the U.S. 83
2.2 CSPA recipients by class of non-immigrant entry 84
2.3 Variation in pre-period share of PRC inventors by county and NBER category 85
2.4 Raw means of patent applications per cell 86

2.5 Effect of CSPA on log patent applications 87
2.6 Robustness checks 88
2.7 Effect of other immigrant groups on patent applications 89
2.8 Effects of CSPA on non-academic vs. academic patenting 90
2.9 Effect of CSPA on inventors 91
2.A.1 Share of foreign S&E PhD students with plans to stay in U.S. . . . 99
2.A.2 Share of PRC-born individuals with > 16 years of education in ACS who are naturalized citizens 100
2.A.3 Raw means of patent applications per cell over time, by NBER category 101
2.A.4 Raw means of patent applications per cell over time, by county . . 102
2.A.5 Correlation between research fields of pre-period and CSPA PRC PhD students 103
2.A.6 Correlation between NBER technology subcategories of pre- and post-CSPA PRC inventors 104
2.A.7 Geographic variation in pre-period PRC inventor share 105
2.A.8 Main results by year 106
2.A.9 Robustness to alternative treatments of outcome variable 107
2.A.10 Results for placebo CSPAs in pre-period 108
2.A.11 Patents per inventor 109

3.1 Effect of a 20-point shift in F-index 123
3.A.1 Results using F-index calculated from 1895-1905 birth cohorts . . . 137
3.A.2 Results for F-index on additional labor market outcomes 138
3.A.3 Results using country-specific F-indices 139
3.A.4 Results for weighted samples 141
3.A.5 Results using more conservative linking algorithms 142
3.A.6 Results by country: sample matched using 3-year uniqueness bands, 5-year age bands 143

3.A.7 Results by country: sample matched using 3-year uniqueness bands, 3-year age bands 145
3.A.8 Results by country: sample matched using 5-year uniqueness bands, 5-year age bands 147
3.A.9 Results by country: sample matched using 5-year uniqueness bands, 3-year age bands 149

Chapter 1

Birth Order and Educational Attainment Throughout the 20th Century[1]

1.1 Introduction

Childhood environment plays an important role in determining adult outcomes. Within families, birth order is one factor that may lead to different outcomes even between siblings: children born early or late in the birth order may have different ways of forming human capital, experience different financial constraints during childhood, or receive different investments of time or resources from their parents.

Social scientists have long studied how birth order may affect educational attainment. Two standard theories put forth by psychologists to explain why earlier-born children may receive more education than their younger siblings include the confluence hypothesis (Zajonc, 1976) and the resource-dilution hypothesis (Blake, 1981).

[1]I am incredibly grateful to Leah Boustan and Ilyana Kuziemko for their guidance and support in this project. I also thank Dorisz Albrecht, Jiwon Choi, Ellora Derenoncourt, Hank Farber, David Lee, Alex Mas, Christiane Szerman, Carolyn Tsao, and participants in the Industrial Relations Section Labor Lunch for helpful comments. All errors are my own.

The confluence hypothesis posits that earlier-born children perform better relative to their younger siblings because they grow up in an environment with higher average intellectual development, due to a greater share of household members being adults. The resource-dilution hypothesis proposes that older children benefit from undivided parental attention in their early developmental years.

Using data from high-income countries, economists have found empirical results consistent with these theoretical explanations for a negative association between birth order and education. Black et al. (2005), exploiting a rich panel dataset covering the entire population of Norway, were the first to use family fixed effects to estimate the effect of birth order on educational attainment separately from the effect of family size.[2] They find that children earlier in the birth order attain more years of schooling than their younger siblings, a result that has been corroborated in later studies using modern data in other high-income countries (De Haan, 2010, and Kim, 2020, in the U.S.; Barclay, 2018, in Sweden; Booth and Kee, 2006, in the U.K.; and Härkönen, 2014, in Germany, among others).

However, while the effects of birth order on education in high-income countries are consistent across countries, studies set in lower-income countries yield more heterogeneous results. Ejrnaes and Pörtner (2004), Emerson and Souza (2008), De Haan et al. (2014), and Weng et al. (2019) find that later-born children receive more education compared to their older siblings in the Philippines, Brazil, Ecuador, and China respectively; Tenikue and Verheyden (2010) find different results across 12 African countries; and Esposito et al. (2020) that older siblings receive more schooling in Mexico. Overall, however, many studies find that the relationship between birth order and educational attainment in lower-income countries is opposite that in high-income countries.[3]

[2]Although early studies (Lindert, 1977; Hauser and Sewell, 1985; Hanushek, 1992) find differing effects of birth order on education, these studies used small datasets and lacked power to sufficiently control for family size, parental background, and cohort effects.

[3]I provide a summary of the literature in Appendix Table 1.A.1.

What can explain these differences? In this paper, I turn to U.S. history to shed light on how and why different birth-order effects may emerge across different contexts.

The primary contribution of this paper is the documentation of trends in birth-order effects in the U.S. over time. While a few other papers have investigated the intergenerational effects of birth order (Kim, 2020; Havari and Savegnago, 2022), or birth order differences between two or three cohorts (Barclay, 2018; Handy and Shester, 2022), this paper is, to the best of my knowledge, the first to study how birth-order effects change over the course of a country's economic development.

To examine long-term trends in birth order and education, I compile a new dataset of siblings covering nearly 150 years, including data from the 1900 and 1910 complete-count Censuses, six cohorts of the National Longitudinal Survey (NLS), and the Wisconsin Longitudinal Study (WLS). A secondary contribution of this paper is the inclusion of Census data and the NLS Original Cohorts in my analysis: these data sources allow me to investigate birth-order effects during late 19th and early 20th century U.S., a setting that has not previously been studied in the birth-order literature.

I find that the relationship between birth order and educational attainment in the U.S. has shifted from positive to negative over the 20th century. In the late 1800s and early 1900s, birth-order effects were similar to those in low-income countries today: children born later in the birth order were significantly more likely to be attending school than their older siblings (around 0.02 percentage points more likely per successive birth-order rank); they also completed around 0.2 more years of education.[4] However, this positive relationship between birth-order rank and years of education diminishes over time: there is no significant difference in educational attainment by birth order for cohorts born in the 1930s, and a negative relationship emerges for

[4]While aggregate rates of schooling are increasing over this time period as well, this result holds when controlling flexibly for year of birth.

cohorts born after the 1930s. The negative relationship persists into the modern day: for families with an eldest child born in the 1970s and beyond, I find that each successive birth-order rank is associated with 0.2 to 0.3 fewer years of schooling, consistent with magnitudes found in other high-income countries today.

I then explore potential mechanisms behind the observed decrease in birth-order effects over time. Prior research (e.g., Ejrnaes and Pörtner, 2004; Emerson and Souza, 2008; De Haan et al., 2014) has suggested that the presence of child labor may explain the negative birth-order effects in lower-income countries today: earlier-born children may have left school earlier to work to support their younger siblings. Likewise, I provide suggestive evidence that the trend I observe is consistent with declines in child labor in the U.S. over time. I find that children in families with higher propensities for child labor experienced the largest decrease in birth-order effects. Effects for children in families with lower propensities for child labor, however, have remained relatively constant over time and are comparable those observed in the U.S. and other rich countries today. I also show that decreases in family size over time do not explain this trend.

Finally, I outline a simple model that formalizes the relationship between birth order and educational attainment in the presence of child labor. Parents value both consumption and their children's human capital and will send their child to school if the marginal utility of more schooling (via increases in the child's human capital) is greater than that of more consumption (through receiving wages from child labor). I assume that, due to a more beneficial early childhood environment, earlier-born children form human capital more productively and thus provide parents with a higher return to each year of schooling. In high-income countries today, where the effective wage of child labor is zero (due to, e.g., child labor legislation and compulsory schooling laws), earlier-born children will attain more education because of their higher "ability" in accumulating human capital through schooling. However, when child

labor is legal, e.g. in lower-income countries today, earlier-born children may also receive a higher wage for child labor than their younger siblings, and parents may choose to send their older children to work instead. Thus, the absence or presence of child labor may explain the negative birth-order effects in high-income countries and the ambiguous birth-order effects found in lower-income countries today.

This paper proceeds as follows. I describe the data and sample construction in Section 1.2 and outline my empirical strategy in Section 1.3. I present results in Section 1.4. I discuss potential mechanisms in Section 1.5 and suggest a simple model that may explain my empirical results in Section 1.6. Section 1.7 concludes.

1.2 Data

Determining the effect of birth order on education requires controlling for family size, age or cohort effects, and parent and background characteristics. Family size could be confounded with birth order if, for example, children in larger families (which will mechanically include children with higher birth orders) receive less education. Age could be a confounder if children born in later years (and with mechanically higher birth orders) receive more education. Previous studies in high-income countries have exploited large panel datasets to identify the effects of birth order using a family fixed effects strategy; De Haan et al. (2014), Oliveira (2018), and Weng et al. (2019), among others, apply this methodology to lower-income countries.

I construct a sample spanning the 1883 to 1995 birth cohorts using three main sources of data: the complete-count 1900 and 1910 population Censuses; the National Longitudinal Surveys (NLS), conducted between 1966 and 2018; and the Wisconsin Longitudinal Study (WLS), conducted between 1957 and 2011. Previous birth-order studies set in the U.S. have also used data from the WLS (Hauser and Sewell, 1985; De Haan et al., 2010; Kim, 2020) and later cohorts of the NLS (Hotz and Pantano,

2015; Lehmann et al., 2018); however, I am, to the best of my knowledge, the first to use data from the Census or the NLS Original Cohorts. I provide an overview of the data sources and sample construction below and more detailed documentation in Appendix 1.B.

1.2.1 Data using school attendance as outcome variable

1900 and 1910 complete-count Censuses

I use data from the 1900 and 1910 digitized complete-count U.S. Censuses for the earliest cohorts in my analysis. While studies set in high-income countries primarily use years of education (Black et al., 2005; De Haan et al., 2010; Barclay, 2015) or college completion (Handy and Shester, 2022) as an outcome variable, many studies set in lower-income countries, especially those that examine birth-order effects on child labor, use school attendance and labor force participation instead (Emerson and Souza, 2008; De Haan et al., 2014; and Esposito et al., 2020).

The Census does not ask respondents about completed years of schooling until 1940. I therefore follow the literature for lower-income countries and use school attendance as an outcome variable.[5] I classify a child as "in school" if they report that they have attended school in a time period specified by the Census; this time period is defined as "within the past year" in the 1900 Census (when respondents were asked on June 1st) and "since September 1" of the previous year in the 1910 Census (when respondents were asked on April 15th).[6]

[5]Although data on completed years of education is not available until 1940, I am still able to use years of education as an outcome variable for a select subsample of my Census sample by matching individuals from their childhood Census to the 1940 Census using the Census Linking Project (Abramitzky, Boustan, and Rashid, 2020). However, doing so restricts my sample to men (and hence families with at least two sons) only; furthermore, matched individuals are much more likely to be white, more highly educated, have foreign-born parents, and live in non-Southern non-rural areas than the general U.S. population (see summary statistics for the matched subsample in Appendix Table 1.A.3). I show results using years of education as an outcome variable for the matched subsample in Appendix Figure 1.A.2.

[6]Note that there may be some measurement error in school attendance and educational attainment reported in the 1900 to 1940 Censuses; see Goldin (1998) for a discussion on the accuracy of

One challenge of using cross-sectional data is correctly assigning birth order within a family. I first determine each child's position in the birth order by ranking all children in the same family (i.e., all individuals who are classified as a child of the household head, regardless of age) by birthyear. To increase the accuracy of the birth-order imputation, I restrict the sample to children whose mothers are at least 40 years old and have likely finished having children. I then exclude families with multiple children born in the same year, as I cannot determine birth order in these cases.

I am only able to calculate birth-order rank for children currently living in a household. Thus, to account for children who may no longer be living with their parents, I use the CHSURV variable, which reports the number of children born to each ever-married woman who are still alive on Census day. For my preferred sample, I use families where the number of children currently living in the household is less than or equal to the total number of surviving children born to the mother.[7] If the number of children in the household is less than the number of surviving children born to the mother, I assume that the missing children are older and have already left their parents' household and adjust the birth-order rank for observed children accordingly. The combination of the mother's age restriction and the CHSURV restriction allows me to determine family size and the birth-order position of each individual with reasonable accuracy.

While full-count data is available for all Censuses between 1850 to 1940, women were only asked about their total number of children surviving in 1900 and 1910. Thus, my preferred sample includes children from the 1900 and 1910 Censuses only.[8]

these variables and Appendix Figure 1.A.3 for the percentage of 10- to 17-year-olds in school across Censuses.

[7]For the families where number of children in the household is larger than the mother's number of children surviving, I assume some are adopted or stepchildren. I drop these families for consistency, as I can only observe education for full siblings or biological children for most of my modern data. This restriction drops around 5 percent of children from the sample.

[8]I can also run my analysis using the 1870 to 1940 Censuses, which allows me to observe trends over a longer time horizon but uses a more imprecise measure of birth order. Summary statistics for

After having determined birth order for all children within a household, I use in my analysis all 10- to 17-year-olds still living in their childhood household at the time of each Census. I exclude children younger than 10 because they were less likely to be in the labor force. I also exclude children older than 17 because most older teenagers have already left their childhood household; thus, any who are still observable in my sample may be subject to selection bias. To allow for family fixed effects, I restrict my sample to only families who have two or more children in the age range.[9] Finally, because I focus on changes in the U.S. over time, I restrict my sample to families where all children were born in the U.S. and hence are likely to have similar exposure to the American education system.[10] After these restrictions, my sample consists of 1,136,536 observations from 439,398 families in the 1900 Census and 1,290,044 observations from 499,885 families in 1910.

1.2.2 Data using years of education as outcome variable

Unlike in the cross-sectional Census data, I can observe individuals in the NLS and WLS even after they have left their childhood household. The NLS and WLS thus offer two advantages. First, I can calculate birth order more accurately, as I have information on age and birthyear for all children in a family rather than only for those living in their parents' household in a particular Census year. Second, I can observe longer-run outcomes such as years of completed education, rather than school

this sample are presented in Appendix Table 1.A.2. Results are shown in Appendix Figure 1.A.1. I do not include the 1850 and 1860 Censuses, as labor force data in those years were collected only for free males and free persons aged 15 or older respectively, or the 1890 Census, as the records were damaged in a fire.

[9] I include all resident children of the household head when calculating birth order, even if they are not necessarily in my sample. For example, in a three-child family, a fifteen-year-old and a thirteen-year-old living at home with their eighteen-year-old sibling would have birth order ranks of 2 and 3 respectively.

[10] For example, some children born abroad may have attended school in countries where the mandatory attendance age is different from the U.S.

attendance, which only measures educational attainment at a point in time.[11] I pool the NLS and WLS data and use years of completed education as my main outcome variable for the remainder of my sample (individuals born between 1906 and 1995).

I provide a brief overview of each survey used below and additional detail in Appendix 1.B.

NLS Original Cohorts

The NLS Original Cohorts comprise two cohorts of men from the National Longitudinal Survey of Older and Young Men (NLSM) and two cohorts of women from the National Longitudinal Survey of Mature and Young Women (NLSW). Respondents were asked about the age, sex, and years of schooling of all siblings in three of the four Original Cohorts: the NLSM Young Men, the NLSW Mature Women, and the NLSW Young Women. Unlike the other cohorts, respondents in the NLSM Older Men cohort were not asked about their siblings; they were asked, however, about the educational attainment of each of their children (including nonresident children). Thus, I include the children of the NLSM Older Men in my analysis rather than the respondents themselves. Respondents in my sample from the Original Cohorts were born between 1923 and 1954, first interviewed between 1966 and 1968, and asked about siblings or children ten years after their initial interview. Of the 20,487 original respondents, I drop 5,524 due to nonresponse and 894 due to their being only children (or, for the NLSM Older Men, childless).

NLSY79 and NLSY79 Child and Young Adult

The National Longitudinal Survey of Youth 1979 (NLSY79) consists of 12,686 respondents born between 1957 and 1965 (aged 14 to 22 when first interviewed in

[11] School attendance also provides less information on educational outcomes in the late 20th century than it does in late 19th: by 1940, nearly 90 percent of teenagers living in their childhood household complete high school and only 4 percent are in the labor force.

1979). In 1993, respondents were asked about their highest grade completed and the sex, age, and highest grade completed of each sibling. Of the original respondents from 1979, I drop 3,690 due to nonresponse in 1993 and 331 due to their being only children.

I also use the National Longitudinal Survey of Youth 1979 Child and Young Adult (NLSY79 CYA) surveys, which record information for every child born to a woman in the original NLSY79 respondent group. The NLSY79 CYA began data collection in 1986 and is ongoing as of 2022; the latest survey wave I use is 2018. The full sample includes 11,545 children from 4,941 mothers from the NLSY79; I drop 1,178 individuals who were only children at the time of the 2018 survey. While questions about education in my other sources of modern data are retrospective, the NLSY79 CYA collects information about each child's education every two years.[12]

Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study is a study of of 10,317 respondents who graduated from Wisconsin high schools in 1957. Survey data were collected in various years between 1957 and 2011. In 1975, respondents were asked to provide a full sibling roster (including the respondents themselves), with age, sex, and the highest grade of schooling for each sibling. Of the original respondents from 1957, I drop 754 due to nonresponse in 1975 and 646 due to their being only children.

Sample construction and restrictions

To construct my modern sample, I first pool data from the NLS and WLS and keep families with at least two children.[13] To accurately determine birth order within a family, I keep only families with birthyears or ages given for all children.[14] I further

[12]The wording and coding of this question changed after 2012; I document how I recode post-2012 responses in Appendix 1.B.

[13]For multiple-respondent families in the NLS and WLS, I count each family only once.

[14]I topcode birth order at 7 for the 1,174 families with more than seven children.

restrict the sample to families where all children were born in the U.S. I drop families with children born the same year and families with more than 25 years between the birthyears of the oldest and youngest children. I keep only families for whom I am able to observe completed education for at least two siblings, topcoding years of schooling at 17. Finally, I keep only individuals who are 23 or older at the time of the survey and thus have likely finished their education. These restrictions yield a preferred sample of 69,734 individuals from 19,745 families.

My preferred sample thus comprises data from the 1900 and 1910 Censuses, the NLS, and the WLS and spans the 20th century. Summary statistics are presented in Table 1.1 separately by data source. The oldest individuals in the sample were born in 1883 (children from the 1900 Census), while the youngest individuals were born in 1995 (children of the NLSY79 cohort). Reassuringly, the summary statistics by data source reflect trends in the U.S. over time: individuals in earlier data sources (e.g. the Census and the NLS Mature Women) had on average less education, more siblings, and were more likely to live in a rural area than later data sources (e.g. NLSY79 CYA). The NLS did, however, oversample certain populations such as Black Americans, especially in later surveys. Thus, to adjust for any sampling biases, I weight all regressions using the NLS or WLS to match the U.S. native-born population by race (White, Black, or other), urban status, farm status, geographic region (South or non-South), and whether any parent was foreign-born, for a total of 48 cells per birthyear.[15]

[15] Weights are calculated separately for each birthyear, based on population counts from the Census in which individuals born that year were aged ten or younger (e.g., weights for individuals born in 1929 were calculated using all native-born 9-year-old children in the 1920 Census). For the WLS, because I cannot observe race or immigration status, and because all individuals in the data lived in the same state, I weight by urban status and farm status to match Wisconsin only. For the NLSY79 CYA, because I do not observe farm or immigration status, I weight by race, urban status, and geographic region only.

1.3 Empirical strategy

Following the birth-order literature, I use the household fixed-effects methodology introduced in Black et al. (2005). For my Census sample, I estimate:

$$Y_{ij} = \alpha + \beta_1 \cdot \mathbb{1}[\text{Census} = 1900] \cdot BirthOrder_{ij} + \beta_2 \cdot \mathbb{1}[\text{Census} = 1910] \cdot BirthOrder_{ij} + \gamma_i + \delta_i + \zeta_j + \epsilon_{ij}. \quad (1.1)$$

where Y_{ij} is an indicator for whether individual i in household j is attending school. $BirthOrder_{ij}$ is the birth order of individual i in household j; γ_i and δ_i are birthyear and sex fixed effects respectively; and ζ_j are childhood household fixed effects. Standard errors are clustered at the household level.

The $\hat{\beta}$ coefficients show the effect of being born one position later in the birth order on the probability of attending school. The birthyear fixed effects control for trends in educational attainment over time: if levels of school attendance are rising and children later in the birth order mechanically belong to later birth cohorts, $\hat{\beta}$ may be biased upward. The inclusion of birthyear fixed effects requires that I observe at least two people born in different years for every birth-order rank (i.e., two eldest children with different birthyears, two second-born children with different birthyears, etc.). The household fixed effects absorb all time-invariant characteristics between families that may affect educational attainment as well as birth order, such as family size, family income, parental education, and mother's age at first birth; for example, $\hat{\beta}$ may be biased downward if larger families (and thus individuals with higher birth orders) tend to have lower educational attainment. In order to identify the effect of birth order using household fixed effects, I must have at least two observations in each household; furthermore, siblings must be unequally spaced so that birth order and birthyear are not perfectly correlated.

To examine changes in birth-order effects on years of education over time, I estimate a version of equation (1.1) using the pooled NLS and WLS data. I now interact the birth-order variable with indicators for whether the oldest child in a household was born in one of 31 birth-cohort bins, where each bin comprises a three-year interval between 1900 and 1990. The composition of each three-year bin by data source is shown in Figure 1.1. Specifically, I estimate:

$$Y_{ij} = \alpha + \sum_{c=0}^{30} \beta_c \cdot \mathbb{1}_j^c \left[1900 + 3c \leq y_j < 1900 + 3\left(c+1\right)\right] \cdot BirthOrder_{ij} + \gamma_i + \delta_i + \zeta_j + \epsilon_{ij}. \quad (1.2)$$

Here, Y_{ij} denotes individual i's completed years of schooling instead of school attendance, and y_j is the birthyear of the eldest child in household j. I furthermore weight each observation as described in the previous section to reflect the demographics of the full U.S. native-born population for each birthyear.[16] Again, the coefficients of interest are the $\hat{\beta}$s, which represent how the effect of birth order on completed years of schooling changes over time.

1.4 Results

I find that birth-order effects have decreased – and flipped sign – from positive (younger siblings receive more education) to negative (older siblings receive more education) in the U.S. over the 20th century.

The $\hat{\beta}$ coefficients from equations (1.1) and (1.2) are presented in Figure 1.2.[17] For children in my Census sample (the 1883 and 1893 cohorts), I find positive and

[16] Unweighted results are presented in Appendix Figure 1.A.4. The results are robust to the exclusion of weights.

[17] The figure does not show coefficients for cohorts with fewer than 100 observations (1900, 1903, and 1969); however, they are included in the regressions. There are no observations in the 1966 cohort.

significant birth-order effects on school attendance. Because the definition of the outcome variables changes slightly between Censuses, I do not interpret the difference in levels of $\hat{\beta}$s between 1900 and 1910 as necessarily indicative of trends in birth-order effects over this time period. Nevertheless, children in the Census were on average around 1.5 to 2 percentage points more likely to attend school for each successive rank in the birth order, relative to a base attendance rate of around 0.8 for the eldest child; these magnitudes are consistent with those found in low-income countries today.[18]

Furthermore, this positive relationship between birth order and education in the earliest cohorts persists when I use years of education as the outcome variable. The coefficients from equation (1.2) are positive and significant for children with eldest siblings belonging to the 1906 to 1918 birth cohorts: younger children in these families attain significantly more years of schooling than their older siblings.

However, the key result I find, and the novel contribution this paper makes to the literature, is the clear downward trend in the relationship between birth order and years of education that emerges across time. While older siblings have significantly worse educational outcomes than their younger siblings in both the Census and the earliest NLS cohorts, older siblings in the 1924-1930 cohorts do not do significantly worse, and by the 1930s, this positive relationship between birth order and years of education has become significant and negative. By the late 1950s, later-born children completed between 0.1 and 0.2 fewer years of education per successive birth rank than their older siblings, an estimate that is in line with studies set in modern-day high-income countries such as Denmark, Norway, and the U.K. This negative relationship between birth order and education persists until the last cohorts I observe in my dataset, those with eldest siblings born after 1990.

[18] Appendix Figure 1.A.1 presents the $\hat{\beta}$ coefficients for equation (1.1) using an extended Census sample (from the 1880 to 1940 Censuses) with both school attendance and labor force participation as outcome variables. The results are robust to this longer time period. Note that due to differences in how the outcome variables were defined across Censuses, changes in levels of $\hat{\beta}$ do not necessarily reflect trends in birth-order effects over time.

Reassuringly, I find that this trend and its timing is consistent across data sources: panel (b) of Figure 1.2 presents the estimates separately by data source. Younger siblings from the earlier cohorts in both the NLS Mature Women and the WLS complete more years of education than their older siblings, but younger siblings in the latest data source, the NLSY79 CYA, complete fewer years of education in all cohorts.

To determine what levels of education may be driving these patterns, I re-estimate equation (1.2) for my modern sample using high school completion (at least 12 years of education) and college attendance (at least 13 years of education) as outcome variables and present results in Figure 1.3. I find that the positive birth-order effects in the earliest cohorts are driven primarily by younger siblings completing high school. As high school completion becomes more common in the U.S. over time, the coefficients grow smaller in magnitude, until they are no longer statistically different from zero after the 1930s. The negative birth-order effects I observe for the later cohorts, on the other hand, are driven primarily by college attendance: older siblings in later cohorts are more likely to receive some college education than their younger siblings.

1.5 Mechanisms

1.5.1 Decreasing family size

What might have caused this increase in educational attainment of older siblings relative to younger over time? I first investigate whether decreases in family size may explain the decline in birth-order effects shown in Figure 1.2. Previous studies in modern-day high-income countries (e.g., Black et al., 2005) find that birth-order effects are more negative for smaller families than for large families. Thus, if larger families make up the majority of earlier cohorts in my sample and smaller families

make up the majority of later cohorts, this shift in sample composition may yield a mechanical decline in observed birth-order effects over time.

I first show the number and percentage of households in my sample across cohorts from each family size in Appendix Figure 1.A.8. Family sizes in my sample have indeed shifted over time, reflective of trends in the general population: individuals from families with five or more children comprise over 60 percent of my sample in cohorts before the 1930s and less than 10 percent in cohorts after 1980.

Furthermore, I also find that smaller families do experience larger birth-order effects: I run equation (1.2) using my pooled NLS and WLS sample separately by family size and present results in panel (a) of Appendix Table 1.A.5; panel (b) shows results using individual birth-order position dummies instead of a single *BirthOrder* variable. Younger siblings in two-child families receive around 0.3 fewer years of schooling than their older siblings, while younger siblings in six-child families receive only 0.05 fewer years per successive birth-order position. Thus, it is possible that the decreasing trend in birth-order effects is driven by changes in family size over time.

In order to more directly test this hypothesis, I reweight my sample such that the family size remains constant over time; that is, I choose one cohort in my sample and then reweight families every other cohort to reflect the distribution of family sizes in the chosen cohort. Figure 1.4 presents the results of this reweighting exercise, where my chosen cohorts are 1924 and 1954. If decreasing family size is the main mechanism through which the relationship between birth order and education changes over time, I would expect to see flat trends for coefficients from the reweighted regressions. I find, however, that the decline in birth-order effects persists after reweighting. Thus, I argue that decreases in family size do not drive the trend I observe in Figure 1.2.

1.5.2 Decline in prevalence of child labor

The presence or absence of child labor may also explain why birth-order effects may differ between high- and low-income countries. Edmonds (2006), Emerson and Souza (2008), De Haan et al. (2014) find that older children in lower-income countries are more likely to be in the labor force than their younger siblings. Furthermore, even within lower-income countries, birth-order effects may differ between families: De Haan et al. (2014) and Esposito et al. (2020) find that older siblings in richer families in Ecuador and Mexico, respectively, receive more schooling, while those in poorer families receive less. Thus, it is possible that while older siblings may benefit from their environment in early childhood, in the face of financial constraints (and absent any child labor or compulsory schooling legislation), parents may also decide to send them to work earlier to support the family.

The cohorts studied in this paper provide an attractive setting to examine the effect of child labor on birth-order effects, as the U.S. experienced large changes in both child labor and compulsory schooling in the early 20th century. Appendix Figure 1.A.6 shows rates of labor force participation, which I define as whether an individual reports any occupation, for children aged 10 to 17 who are classified as a child of the household head in the 1880 to 1940 Censuses.[19] Labor force participation for children in this age range has declined from 26 percent for children in the 1880 Census (the 1863-1870 birth cohorts) to 5 percent for children in the 1940 Census (the 1923-1930 birth cohorts), with a marked decrease starting in 1920 (the 1903-1910 birth cohorts).

Could the decline in child labor have contributed to decreasing birth-order effects? If the returns to labor market participation (i.e., wages in childhood) were greater than the returns to schooling (i.e., increases in future human capital) for older children in poorer families, it is possible that the positive birth-order effects observed during

[19] All individuals aged 10 and older were asked about occupation in the 1880 to 1930 Censuses. However, occupation was only asked for individuals aged 14 and above in the 1940 Census; thus, all children younger than 14 are classified as "not in the labor force" for this year.

the early 20th century are primarily driven by these families. Then, as wages for child labor decrease over time, older siblings in these families may be able to pursue more education, and birth-order effects for these families should decline as well.

Many papers in the child labor literature exploit the differential timing of child labor and compulsory schooling laws across states in the early 20th century to examine the causal effect of legislation on educational attainment. However, the efficacy and "bite" of these laws has been widely debated (see Lleras-Muney, 2002; Goldin and Katz, 2011; Stephens and Yang, 2014). In this paper, I take a different approach to determine how the overall *trend* of declining child labor may have contributed to declining birth-order effects, rather than the effect of the laws themselves. I first calculate the propensity for child labor by estimating the following probit regression for my pooled Census sample:

$$\Pr(LaborForce_i = 1 \mid X_i) = \Phi(X_i'\beta) \tag{1.3}$$

where I define $LaborForce$ as an indicator for whether a child in my Census sample reported any occupation when asked. All respondents were asked about occupation in the 1900 Census, and respondents aged 10 or older were asked in the 1910 Census; thus, I should be able to observe labor force participation for my entire Census sample as all individuals in the sample are between 10 and 17 years old. X_i is a vector of covariates that includes the total number of siblings (including the respondent); indicators for Black and other race; and indicators for urban status, farm status, foreign-born parent, and residency in the South.

Using the coefficients from the probit regression, I then predict the propensity for child labor for each child in my full sample. That is, given each child's observable characteristics, I calculate the probability that they would have been in the labor force *if they had lived in the early 20th century.*

The coefficients from equation (1.3) are reported in Appendix Table 1.A.4. Column 1 of Appendix Table 1.A.4 reports results from the full Census sample, while columns 2 and 3 report results separately for Southern vs. non-Southern states. Because I observe notable differences in the coefficients between the South and non-South, likely due to regional differences in demographics and types of work available, I predict each child's propensity for child labor using columns 2 and 3.[20]

Note that the observable characteristics included on the right-hand side of the probit regression, and hence the predicted propensities for child labor, should be the same across children in the same family. Thus, I am able to split families in my full sample into those with above- and below-median propensities for child labor. Summary statistics for the sample split into above- and below-median propensities are provided in Table 1.2.

I then re-estimate equations (1.1) and (1.2) after interacting the birth order variable with an indicator for above-median propensity for child labor and including region by birthyear fixed effects.[21] If high wages for child labor in the late 19th century did contribute to the positive birth-order effects observed for the earliest cohorts in Figure 1.2, I would expect the decrease in birth-order effects to be concentrated in families above the median; that is, older siblings in more financially constrained families are to able attain progressively more schooling over time. Specifically, I estimate:

$$Y_{i,j} = \alpha + \beta_1 \cdot \mathbb{1}[\text{Census} = 1900] \cdot \mathbb{1}[X_j = 1] \cdot BirthOrder_{i,j} + \beta_2 \cdot \mathbb{1}[\text{Census} = 1910] \cdot \mathbb{1}[X_j = 1] \cdot BirthOrder_{i,j} + \gamma_i + \delta_i + \zeta_j + \epsilon_{i,j}. \quad (1.4)$$

[20] For example, residency in an urban area is positively and significantly associated with being in the labor force in the North, Midwest, and West, while it is negatively associated with child labor in the South, suggesting that child labor seemed to be primarily concentrated in manufacturing in the non-South and agriculture in the South. The large positive and significant coefficients on Black and farm residency provide further suggestive evidence on the agricultural nature of work in the South.

[21] Due to data limitations, I increase the size of the cohort bins in equation (1.2) to six years; e.g., families in the 1903 cohort have an eldest child born between 1903 and 1908.

and

$$Y_{i,j} = \alpha + \sum_{c=1}^{16} \beta_c \cdot \mathbb{1}_j^c \left[1897 + 6c \leq y_j < 1897 + 6c + 6\right] \cdot \mathbb{1}[X_j = 1] \cdot BirthOrder_{i,j} + \gamma_i + \delta_i + \zeta_j + \epsilon_{i,j}. \quad (1.5)$$

where $X_j = 1$ if the family has an above-median propensity for child labor and $X_j = 0$ otherwise, and where $Y_{i,j}$ represents school attendance in equation (1.4) and highest grade completed in equation (1.5).

The coefficients, split by propensity for child labor, are plotted in panel (a) of Figure 1.5. I find that birth-order effects on school attendance and years of education are significantly more positive for families with above-median propensities for child labor in the earliest cohorts. Thus, in families that would have most benefited from higher wages for children, later-born children did indeed attain more education than their older siblings. This result is consistent with the story that older siblings were more likely to enter the labor force earlier to help support the family. On the other hand, birth-order effects in families with below-median propensities for child labor are closer to zero and remain relatively constant over time: that is, the relationship between birth order and education in richer families in the past looks more similar to that observed in the U.S. today.

However, as wages for child labor decrease over time (as a result of, e.g., stricter legislation, increasing returns to schooling, or decreasing demand by employers for child labor), birth-order effects on education for families with above-median propensities for child labor experience a steeper decline as well. By the 1940s, birth-order effects for families above the median have converged to those for families below the median and look similar to those for families today.[22]

[22] In Appendix Figure 1.A.7, I present results separately for Southern and non-Southern families. I find larger initial differences between above- and below-median families, as well as a larger decrease for above-median families, in the non-South. This heterogeneity may reflect the higher propensity

Additionally, the NLS and WLS also include father's education, which may provide a more direct proxy for socioeconomic status and propensity for child labor. I thus re-estimate equation (1.2) where I interact *BirthOrder* with an indicator for whether the father of each family was a high school graduate. The $\hat{\beta}$ coefficients split by father's education are presented in panel (b) of Figure 1.5 and provide further evidence to corroborate the results in panel (a): families with less-educated fathers experienced more pronounced positive birth-order effects in earlier cohorts and a more dramatic decline, while birth-order effects in families with more-educated fathers remain negative or not significantly different from 0 and relatively flat over time.

Again, this finding seems to suggest that older siblings in families with lower socioeconomic statuses may have left school earlier to support their families when children were able to work for higher wages. Furthermore, like in panel (a), birth-order effects for both groups have converged by the latter half of the 20th century, when child labor was far less common. Taken together, panels (a) and (b) from Figure 1.5 provide suggestive evidence that child labor may contribute to the difference in birth-order effects found in high- and low-income countries.

1.6 A model of intra-household allocation of schooling

To formalize the relationship between birth order and educational attainment in the presence of child labor, I present a simple model of intra-household allocation of schooling, drawing from those developed by Becker (1981), Emerson and Souza (2002 WP), and Edmonds (2006). The model illustrates how child labor may result in

for child labor in Southern families overall; furthermore, child labor and compulsory schooling laws were targeted at manufacturing and mining, which was more prevalent in the non-South, and often did not apply to agricultural work, which was more prevalent in the South.

differences in education by birth order even without differences in parental preferences between children.

Suppose that households are composed of one adult and two children. I assume fertility is exogenous and that the number of children in the household has been previously determined; doing so allows the model to be extended to households with N children. Let each child be denoted by $i = 1, 2$, where i is the child's rank in the birth order; i.e., the eldest child has $i = 1$.

There are $N = 2$ periods in the model. The adult values consumption c_i and their child's human capital h_i and makes only one decision in each period: in period i, the adult can choose whether child i attends school (s_i) or works (l_i) or both, where the child has time endowment T and can only spend time either in school or at work ($T = s_i + l_i$). For simplicity, I normalize $T = 1$.

If the child attends school, he increases his human capital h_i, where h_i is increasing and concave in schooling s_i. That is, we can define h_i as a human capital accumulation function $h_i = h_i(s_i)$ for $i = 1, 2$, where $\frac{\partial h_i(s_i)}{\partial s_i} > 0$ and $\frac{\partial^2 h_i(s_i)}{\partial {s_i}^2} < 0$ for all s_i. I assume that parents have no preferences between children, so that h_1 and h_2 are perfect substitutes. It is important to note that h_i may be different for each child even if parents do not explicitly prefer one child over the other. In particular, h_i can reflect differences in ability or returns to schooling between children by birth order that may arise due to an early childhood environment with successively lower average intellectual development (Zajonc, 1976) or increasingly divided parental attention (Blake, 1981) as more children are born into the family.

If the child works, he earns wages in childhood w_i that increase the household's consumption c_i in period i. I assume wages in childhood w_i are not dependent on schooling s_i or on the human capital acquired from schooling h_i. This assumption is reasonable as most children in the 19th- and 20th-century U.S. (and in lower-income countries today) worked in low-skill industries that did not require much

formal education (e.g., mining, textile and other manufacturing, agriculture, and domestic work).

I assume that the adult spends all his time at work, and that $w_{a,1}$ and $w_{a,2}$ are the wage rates of the adult in periods 1 and 2. For simplicity, I assume that $w_{a,1}$ and $w_{a,2}$ have been previously determined, although they may be a function of the adult's human capital that depends on how much schooling she received in her own childhood. Wages of the adult and of the children are perfect substitutes in each period.

We can now fully characterize the adult's optimization problem. The adult chooses the child's school attendance s_i in each period to maximize:

$$U(c_1, c_2, h_1, h_2) \tag{1.6}$$

subject to the budget constraints:

$$c_1 \leq w_{a,1} + (1 - s_1)w_1 \tag{1.7}$$

for period 1 and

$$c_2 \leq w_{a,2} + (1 - s_2)w_2 \tag{1.8}$$

for period 2, where U is strictly concave and $s_i + l_i = 1$.

The FOCs for each child i are:

$$\frac{\partial U}{\partial s_i} : \frac{\partial U}{\partial c_i}\frac{\partial c_i}{\partial s_i} + \frac{\partial U}{\partial h_i}\frac{\partial h_i}{\partial s_i} = 0$$

The parent will thus allocate schooling and child labor such that the ratios of the marginal utility of consumption to the marginal utility of the child's marginal gain

in human capital are equal for both children:

$$\frac{(\partial U/\partial c_1)(\partial c_1/\partial s_1)}{(\partial U/\partial h_1)(\partial h_1/\partial s_1)} = \frac{(\partial U/\partial c_2)(\partial c_2/\partial s_2)}{(\partial U/\partial h_2)(\partial h_2/\partial s_2)} \tag{1.9}$$

We can see that $\frac{\partial c_i}{\partial s_i} = -w_i$; that is, increasing the child's school attendance will result in a decrease in consumption equal to the wage the child could have earned. We can also see that, because parents do not have specific preferences over their children, $\frac{\partial U}{\partial h_1} = \frac{\partial U}{\partial h_2}$ for all h. Thus the previous equation simplifies to:

$$\frac{(\partial U/\partial c_1)w_1}{(\partial h_1/\partial s_1)} = \frac{(\partial U/\partial c_2)w_2}{(\partial h_2/\partial s_2)} \tag{1.10}$$

The model can first show how differences in s_1 and s_2 (i.e., the birth-order effect on schooling) may arise *within* a family, even in a setting without explicit parental preferences between children. We can see that differences in s_1 and s_2 will depend on differences in the adult's wage in each period, $w_{a,1}$ and $w_{a,2}$; differences in the first and second child's wage, w_1 and w_2; and differences between their marginal gain in human capital $\frac{\partial h_1}{\partial s_1}$ and $\frac{\partial h_2}{\partial s_2}$.

I first assume that $w_{a,1} < w_{a,2}$ (that is, the adults have an age-earnings profile such that the budget constraint is always looser for the second child) and that there is no borrowing between periods. I also assume that $\frac{\partial h_1}{\partial s_1} > \frac{\partial h_2}{\partial s_2}$ for all values of s; i.e., the older child's human capital will increase more than that of the younger child from the same increase in schooling. This difference in the human capital function reflects the older child's higher ability as a result of a more beneficial early childhood environment.

Then, the model generates the following predictions:

1. The lower the wage in childhood, the less children will work. The marginal gain in human capital will outweigh the wage from child labor, and the parent will decide to send the child to school.

2. The higher the adult's wage, the less children will work. This follows from the concavity of U (i.e., $\frac{\partial U}{\partial c_i} > 0$ and $\frac{\partial^2 U}{\partial {c_i}^2} < 0$). That is, if w_a is high, the marginal gain in utility from consumption from the child's wage will be small.

3. If children's wages are the same ($w_1 = w_2$) and the first child has higher ability, the first child will receive more schooling. This follows from the concavity of h_i.

4. If the wage of the first child is greater than the wage of the second ($w_1 > w_2$) and the first child has higher ability, the birth-order effect will be ambiguous.

In high-income countries, where adult wages are high and children's wages are low (effectively zero), we can see that neither child will work and that, within a family, older siblings will receive more schooling. The negative birth-order effects found in numerous papers set in high-income countries lend to support to this case.

In low-income countries, where adult wages may be lower in the first period, and the wage for the first child may be higher than that for the second, birth-order effects will be ambiguous. Indeed, empirical results across papers set in low-income countries with a higher prevalence of child labor are less consistent, suggesting a setting similar to this.

Furthermore, the model may explain the change from positive to negative birth-order effects over time, i.e. the difference in birth-order effects *between* families in the past and in the present. In particular, as demand for child labor decreases over time (as a result of, e.g., stricter legislation), wages for child labor, w_i, will also decrease. In countries that are in the process of transitioning away from child labor, the difference in wages between the first and second child, $(w_1 - w_2)$, will be large; younger siblings will receive a lower wage in childhood than their older siblings and thus attend more school. However, as eventually wages for child labor w_1 and w_2 both go to zero, and hence $(w_1 - w_2) \to 0$, parents in the present will allocate more schooling to the child

with a higher ability relative to parents in the past. Thus, older children will attain more schooling relative to their younger siblings over time.

Finally, if the wages of the adult during period 1 (when parents are deciding whether to send the older child to work) increase over time, older children will work less as well. This may be the case if wages for adults are rising due to greater economic development, or if, due to the proliferation of birth control, rising rates of college attendance and female labor force participation, people delay having children until they are older. Thus, increasing economic development could lower child labor and generate a switch from positive to negative birth-order effects on education over time, even without changes in child labor legislation.

1.7 Conclusion

This paper documents a shift in the effect of birth order on educational attainment in the U.S. from the late 19th to the late 20th century using a newly compiled dataset comprising siblings from the 1900 and 1910 Census, the National Longitudinal Surveys, and the Wisconsin Longitudinal Study.

First, I find that birth-order effects on both school attendance and years of education were positive in the late 19th century: older children during this time period were less likely to be in school and completed fewer years of schooling than their younger siblings. This positive relationship between birth order and schooling is similar to those found in low-income countries today.

The main result of this paper is the striking decline in birth-order effects on years of education in the first half of the 20th century. Families with eldest children born after the 1930s exhibited negative birth-order effects, and this negative relationship persists until the present day. By the 1990s, children attained around 0.2 fewer years

of schooling per successive birth-order rank, consistent with magnitudes found in high-income countries today.

To my knowledge, this is one of the first papers to examine changes in birth-order effects over a long time period. While some previous papers have compared birth-order effects between two different cohorts (Barclay, 2015; Handy and Shester, 2022), or examined whether birth-order effects may be intergenerational (Kim, 2020; Havari and Savegnago, 2022), I document how the relationship between birth order and educational attainment has changed over multiple birth cohorts spanning more than one hundred years.

Finally, I explore whether decreases in family size or declines in child labor in the U.S. in the early 20th century could have contributed to these decreasing birth-order effects. I provide suggestive evidence that child labor may be one channel through which differences in birth-order effects on educational attainment have emerged between high- and low-income countries. I find that families with above-median propensities for child labor experienced the largest positive birth-order effects on education in the late 19th century and a more dramatic decline in birth-order effects over time; in contrast, families with below-median propensities for child labor exhibit smaller or even negative birth-order effects and less change over time, similar to effects observed in the U.S. today. Thus, child labor (and its decline in the U.S. over the 20th century) might explain part of this observed trend and may be one reason birth-order effects differ between high- and low-income countries today.

1.8 Tables and figures

Table 1.1: Summary statistics

	1900 Census	1910 Census	NLS Mature Women	WLS	NLS Older Men (children)	NLS Young Men	NLS Young Women	NLSY 79	NLSY 79 (children)
Background									
Year of birth	1886.57	1896.48	1929.49	1939.65	1942.86	1945.15	1946.90	1959.54	1984.98
Male	0.51	0.51	0.38	0.50	0.47	0.67	0.37	0.50	0.51
White	0.92	0.93	.	.	.	.	.	.	.
Black	0.08	0.07	0.18	.	0.26	0.23	0.23	0.30	0.34
Childhood									
Total # of children in household	4.28	4.13	4.93	4.42	3.90	4.43	4.64	4.75	3.15
Total # of surviving children born to mother	4.94	4.78	.	.	.	.	.	.	.
Father's age in childhood Census	46.95	47.26	.	.	.	.	.	.	.
Mother's age in childhood Census	43.91	43.86	.	.	.	.	.	.	.
Father's highest grade completed	.	.	8.22	10.19	9.02	8.73	10.01	10.88	.
Mother's highest grade completed	.	.	8.82	10.56	9.95	10.08	10.26	10.88	12.88
Urban	0.33	0.41	0.65	0.62	0.58	0.69	0.71	0.78	0.75
Farm	0.45	0.40	0.28	0.20	0.12	0.21	0.16	0.05	.
South	0.30	0.28	0.39	.	0.39	0.40	0.40	0.36	0.41
Foreign-born parent	0.30	0.33	0.13	.	0.04	0.05	0.04	0.07	.
Probability of labor force participation	0.23	0.23	0.26	.	0.25	0.26	0.26	0.27	.
Probability of labor force participation (separate probit by South/non-South)	0.23	0.23	0.23	.	0.24	0.23	0.23	0.23	.
Outcomes									
In school	0.86	0.85	.	.	.	.	.	.	.
In labor force	0.22	0.24	.	.	.	.	.	.	.
Years of education	.	.	11.39	12.99	12.43	12.62	12.43	12.73	13.54
Observations									
Number of observations	1,136,536	1,290,044	8,805	23,624	3,219	4,663	6,087	17,234	6,102
Number of households	439,398	499,885	2,117	6,509	1,110	1,470	1,813	4,387	2,339

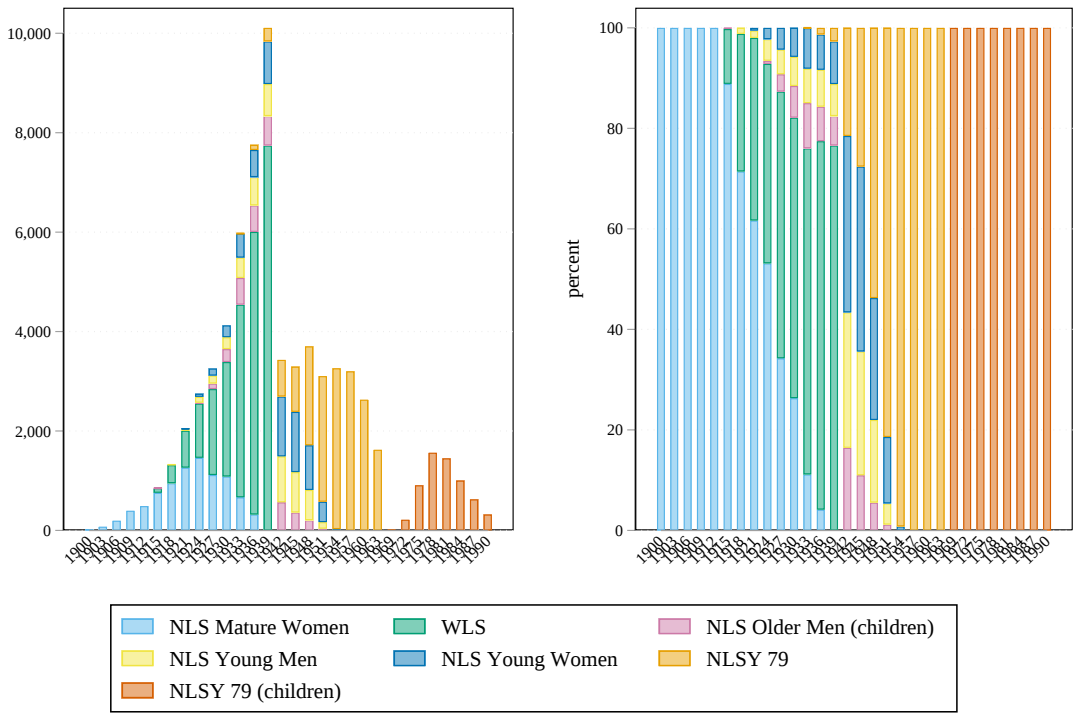
Notes: Census samples comprise all individuals in 1900 and 1910 Censuses for whom I observe age, sex, and school attendance for at least one sibling. Census samples are further restricted to individuals aged 10-17 who are children of the household head and who have mothers at least 40 years old at the time of the Census. NLS Original Cohorts, NLSY79, NLSY79 CYA, and WLS samples include respondents and siblings (or children, for NLSM Older Men) of respondents, for whom I observe age, sex, and years of education for at least one sibling. NLS and WLS samples are further restricted to individuals at least aged at least 23 when asked about education.

Table 1.2: Summary statistics by probability of child labor

	Below median	Above median
Background		
White	0.99	0.85
Black	0.00	0.15
Childhood		
Total # of surviving children born to mother (Census)	4.31	5.41
Father's highest grade completed (NLS)	10.84	8.63
Mother's highest grade completed (NLS)	10.96	9.35
Urban	0.35	0.40
Farm	0.38	0.46
South	0.07	0.52
Foreign-born parent	0.16	0.46
Probability of child labor	0.16	0.30
Outcomes		
In school (Census)	0.89	0.82
In labor force (Census)	0.16	0.30
Years of education (NLS)	12.79	11.86
Observations		
Number of observations	1,236,402	1,229,589
Number of households	500,903	449,118

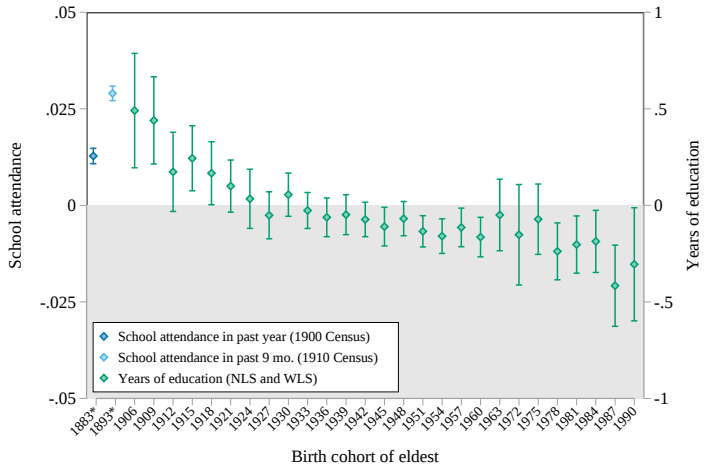
Notes: Sample comprises all individuals in 1900 and 1910 Censuses, NLS Original Cohorts, and NLSY79 samples for whom I observe age, sex, and years of education for at least one sibling. Sample is further restricted to individuals for whom I observe race, childhood urban/rural status, childhood farm/non-farm status, childhood region of residence, and birthplaces of parents. Respondents in WLS were not included in sample due to lack of data on race. Respondents in NLSY79 CYA were not included due to lack of data on childhood farm/non-farm status and birthplace of father. Table reports summary statistics of sample split by above- and below-median propensity for child labor using predicted propensities from equation (1.3), run separately for the South and the non-South. Regression results from equation (1.3) are presented in Appendix Table 1.A.4.

Figure 1.1: Composition of sample by birth cohort of eldest child (NLS and WLS)

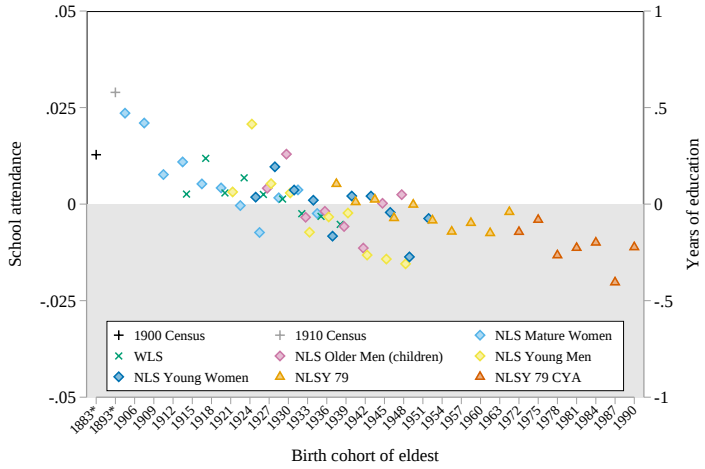


Notes: This figure reports the total number of observations in each cohort, and percentage of observations in each cohort, from each NLS and WLS data source. Cohorts are defined as three-year bins based on the birthyear of the eldest child in a family.

Figure 1.2: Effect of birth order on educational attainment over time



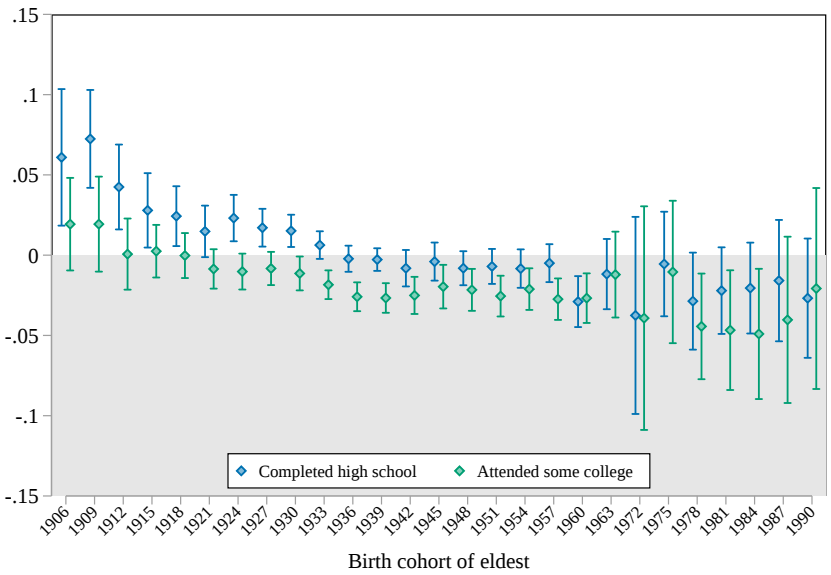
(a) Full sample



(b) Separate regressions by data source

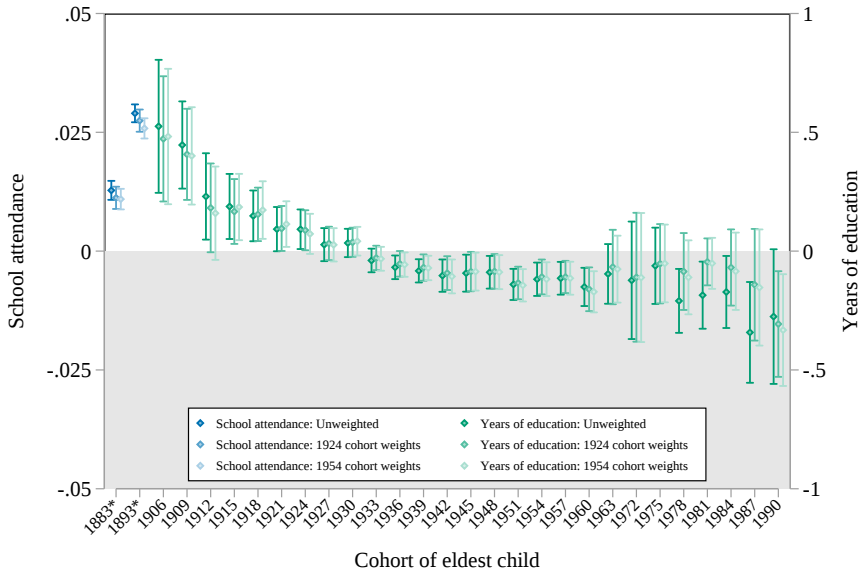
Notes: This figure presents coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. The 1883 and 1893 cohorts comprise all families from the 1900 and 1910 Censuses respectively; the outcome variable is school attendance. The outcome variable for post-1900 cohorts is years of education. Post-1900 cohorts in panel (a) include all families with eldest siblings born in each three-year bin from the pooled NLS and WLS sample; panel (b) presents coefficients from separate regressions by data source. Cohorts with fewer than 100 observations (1900, 1903, 1966, and 1969) are not shown. All regressions include sex, birthyear, and household fixed effects. NLS and WLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.3: Birth-order effects on high school completion and college attendance



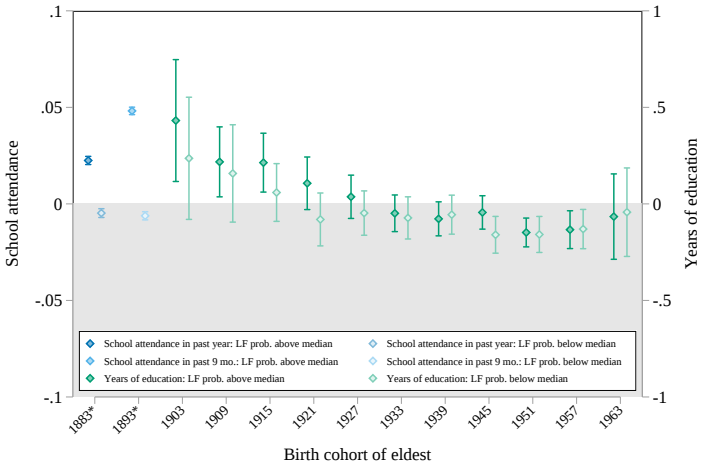
Notes: This figure presents coefficients on birth-order from equation (1.2) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals from the pooled NLS and WLS data for whom I observe age, sex, and years of education for at least one sibling. "Completed high school" is defined as having completed at least 12 years of education; and "Attended some college" is defined as having completed 13 or more years of education. Cohorts with fewer than 100 observations (1900, 1903, 1966, and 1969) are not shown. All regressions include sex, birthyear, and household fixed effects. Observations are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.4: Main results reweighted by family size distribution

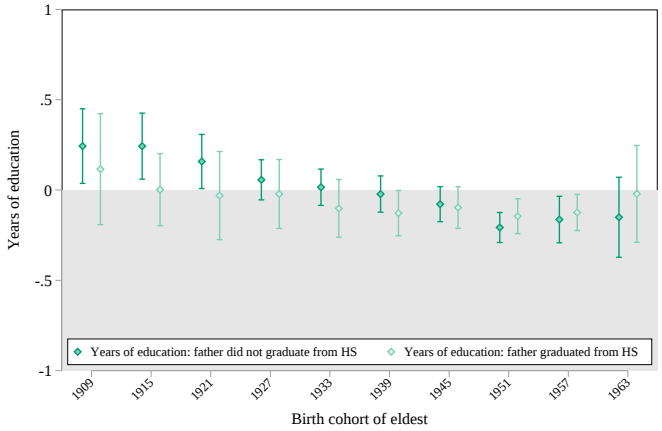


Notes: This figure presents reweighted coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family. Regressions are reweighted such that the family size distribution of each cohort matches those of the 1924 and 1954 cohorts. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. The 1883 and 1893 cohorts comprise all families from the 1900 and 1910 Censuses respectively; the outcome variable is school attendance. Post-1900 cohorts include all families with eldest siblings born in each three-year bin from the pooled NLS and WLS sample; the outcome variable is years of education. Cohorts with fewer than 100 observations (1900, 1903, 1966, and 1969) are not shown. All regressions include sex, birthyear, and household fixed effects. Standard errors are clustered at the household level.

Figure 1.5: Heterogeneity analysis



(a) Sample split by propensity for child labor



(b) Sample split by father's education

Notes: This figure presents coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals from the 1900 and 1910 Censuses and the NLS for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. Panel (a) presents results after splitting the sample into families with above- and below-median propensities for child labor; panel (b) also includes families in the WLS and presents results for families whose fathers did vs. did not graduate from high school. All regressions include sex, birthyear, and household fixed effects. The top panel also includes birthyear x South fixed effects. NLS and WLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Appendix

1.A Appendix tables and figures

Table 1.A.1: Coefficients in birth-order literature

Paper	Country	Sample	Outcome variable	Coefficent on birth order (Standard errors in parentheses)
Tenikue and Verheyden (2010)	Benin	Children aged 6-18 in 2006	Years of education	0.51***
	Burkina Faso	Children aged 6-18 in 2003	Years of education	0.84***
	Cameroon	Children aged 6-18 in 2004	Years of education	0.60***
	Ghana	Children aged 6-18 in 2003	Years of education	0.34*
	Kenya	Children aged 6-18 in 2003	Years of education	0.04
	Mali	Children aged 6-18 in 2006	Years of education	0.80***
	Niger	Children aged 6-18 in 2006	Years of education	0.84***
	Senegal	Children aged 6-18 in 2005	Years of education	0.87***
	Tanzania	Children aged 6-18 in 2004	Years of education	-0.33***
	Uganda	Children aged 6-18 in 2006	Years of education	-0.03**
	Zambia	Children aged 6-18 in 2007	Years of education	0.06
	Zimbabwe	Children aged 6-18 in 2005	Years of education	-0.11
Weng et al. (2020)	China	Individuals aged 25-65 in 2014	Years of education	0.063*** (0.022)
Kumar (2016)*	India	District Level Household Survey	Years of education	0.016** (0.007)
Botzet et al. (2021)	Indonesia	Indonesian Family Life Survey	Years of education	0.01578** (0.006262)
Kim (2020)*	U.S.	WLS original respondents	Years of education	-0.092 (0.098)
		Children of original WLS respondents	Years of education	-0.223*** (0.077)
Wang et al. (2020)	China	China Family Panel Studies 2010 (rural)	Years of education	-0.166*** (0.028)
		China Family Panel Studies 2010 (urban)	Years of education	-0.226*** (0.037)
Bagger et al. (2021)	Denmark	Individuals aged 15-70 between 1980-2006	Years of education	-0.197*** (0.006)
Mechoulan and Wolff (2015)*	France	Individuals born between 1920-1986	Years of education	-0.198*** (0.0160)
Booth and Kee (2009)	U.K.	Individuals aged 28-55 in 2003	Years of education	-0.26*** (0.036)
De Haan (2010)	U.S.	WLS	Years of education	-0.26*** (0.07)
Barclay (2018)*	Sweden	Women born between 1960-1982	Years of education	-0.29*** (0.01)
		Men born between 1960-1982	Years of education	-0.27*** (0.01)
Black et al. (2005)*	Norway	Individuals aged 16-74 between 1986-2000	Years of education	-0.318*** (0.007)
Härkönen (2014)*	Germany	Individuals born between 1945-1969	Years of education	-0.354*** (0.093)
Lehmann et al. (2018)*	U.S.	NLSY79	Years of education	-0.496** (0.200)
Esposito et al. (2020)	Mexico	Children aged 6-17 in 2010	On-track grade enrollment	-0.341*** (0.00697)
Handy and Shester (2021)*	U.S.	Health and Retirement Survey	High school attendance	-0.026*** (0.008)
			College attendance	-0.049*** (0.014)
Iacovou (2008)*	U.K.	National Child Development Study	Highest qualification at age 23	-0.1353*** (0.03292)
Emerson and Souza (2008)*	Brazil	Girls aged 7-16 in 1998	School attendance	0.112 (0.084)
			Child labor	-0.014 (0.069)
		Boys aged 7-16 in 1998	School attendance	0.062 (0.072)
			Child labor	-0.117** (0.056)
De Haan et al. (2014)*	Ecuador	Children aged 12-18 between 2001-2006	School enrollment	0.008*** (0.003)
			Child labor	-0.022*** (0.003)
Dammert (2010)	Guatemala	Children aged 7-14 in 2000	School attendance	-0.005 (0.02)
			Market work	-0.041*** (0.04)
			Domestic work	-0.054*** (0.02)
			Any work	-0.067*** (0.02)
	Nicaragua	Children aged 7-14 in 2001	School attendance	0.007 (0.01)
			Market work	-0.007 (0.01)
			Domestic work	0.007 (0.01)
			Any work	-0.001 (0.01)
Moeshoeshoe (2019)*	Lesotho	Children aged 6-18 in 2006	School enrollment	-0.0107*** (0.00365)
Seid and Gurmu (2015)	Ethiopia	Children aged 5-17 in 2006 or 2009	School attendance	-0.026*** (0.06)
			Child labor	-0.252*** (0.05)
Khanam and Rahman (2007)	Bangladesh	Children aged 5-17	Child labor	-0.334 (0.234 p-value)
Edmonds (2006)	Nepal	Children aged 6-15 in 1998	Total hours worked	-2.187*** (0.617)

Notes: This table presents estimates of birth-order effects from the current literature. When available, I report the coefficient on birth-order index. Papers marked with an asterisk do not report a single coefficient on birth order and instead estimate effects for each individual birth-order position; for these papers, I report the coefficient on having a birth-order position of 2 in a three-child family (i.e., the effect of being the second-born child, relative to the eldest).

Table 1.A.2: Summary statistics for 1870 to 1940 Census sample

	1870 Census	1880 Census	1900 Census	1910 Census	1920 Census	1930 Census	1940 Census
Background							
Year of birth	1856.99	1867.09	1886.90	1896.85	1906.89	1916.81	1926.66
Male	0.51	0.51	0.51	0.51	0.51	0.51	0.51
White	0.90	0.87	0.90	0.90	0.90	0.92	0.92
Black	0.10	0.13	0.09	0.09	0.09	0.07	0.07
Childhood							
Total # of children in household	5.02	5.01	4.81	4.72	4.55	4.28	3.91
Father's age in childhood Census	42.74	42.66	42.31	42.35	42.31	41.71	41.88
Mother's age in childhood Census	37.41	37.35	37.47	37.41	37.38	37.36	37.36
Urban	0.25	0.23	0.34	0.39	0.42	0.50	0.48
Farm	0.45	0.49	0.43	0.41	0.39	0.31	0.29
South	0.29	0.33	0.31	0.34	0.35	0.29	0.32
Foreign-born parent	0.31	0.28	0.26	0.27	0.25	0.27	0.17
Outcomes							
In school	0.65	0.64	0.87	0.87	0.84	0.89	0.90
In labor force	0.17	0.24	0.20	0.24	0.14	0.08	0.04
Observations							
Number of observations	879,208	958,680	1,538,915	1,870,896	2,269,316	2,828,059	2,516,412
Number of households	341,156	379,775	598,003	724,216	889,138	1,114,884	1,010,793

Notes: Samples comprise all native-born individuals aged 10-17 in the 1870 to 1940 Censuses for whom I observe age, sex, and school attendance for at least one sibling; who are children of the household head; and who have mothers between 35-40 years old at the time of the Census.

Table 1.A.3: Summary statistics for 1900 and 1910 Census subsample matched to 1940 Census

	1900 Census	1910 Census
Background		
Year of birth	1886.58	1896.48
Male	1.00	1.00
White	0.98	0.98
Black	0.02	0.02
Childhood		
Total # of children in household	4.61	4.42
Total # of surviving children born to mother	5.23	5.02
Father's age in childhood Census	46.55	46.88
Mother's age in childhood Census	43.62	43.60
Urban	0.31	0.39
Farm	0.49	0.43
South	0.19	0.19
Foreign-born parent	0.33	0.34
Outcomes		
In school	0.87	0.87
In labor force	0.28	0.30
Years of education	8.75	9.45
Observations		
Number of observations	32,885	55,729
Number of households	15,816	26,587

Notes: Samples comprise all native-born individuals aged 10-17 in the 1900 and 1910 Censuses who can be matched to the 1940 Census, for whom I observe age, sex, school attendance in 1900 or 1910, and years of education in 1940 for at least one sibling. Individuals must be children of the household head and have mothers at least 40 years old at the time of the 1900 or 1910 Census.

Table 1.A.4: Probit regression results for probability of labor force participation – Pooled Census CHSURV sample

Outcome: In labor force	Full sample	Non-South only	South only
Total # of surviving children born to mother	0.0296***	0.0298***	0.0237***
	(0.000507)	(0.000644)	(0.000837)
Black	0.522***	0.215***	0.539***
	(0.00332)	(0.0101)	(0.00363)
Other race	0.00805	0.0569*	-0.125***
	(0.0206)	(0.0241)	(0.0389)
Urban	0.121***	0.144***	-0.206***
	(0.00257)	(0.00295)	(0.00563)
Farm	0.218***	0.0644***	0.409***
	(0.00239)	(0.00311)	(0.00386)
Foreign-born parent	0.157***	0.0656***	-0.0296
	(0.00220)	(0.00234)	(0.00742)
South	0.462***	.	.
	(0.00225)	.	.
Constant	-1.259***	-1.214***	-0.850***
	(0.00306)	(0.00364)	(0.00511)
Number of observations	2,426,733	1,716,898	709,835
Pseudo R^2	0.0497	0.0074	0.0536

Notes: *** p<0.01, ** p<0.05, * p<0.1. This table reports probit regression results from equation (1.3). Each column represents a separate regression. The sample comprises all individuals aged 10-17 in the 1900 and 1910 Censuses for whom I observe age, sex, and labor force participation for at least one sibling; who are children of the household head; and who have mothers at least 40 years old at the time of the Census. The sample is further restricted to individuals for whom I observe race, childhood urban/rural status, childhood farm/non-farm status, childhood region of residence, and birthplaces of parents. Column (1) reports results from the entire sample; columns (2) and (3) report results separately by South/non-South residency.

Table 1.A.5: Results by family size and individual birth-order position – Pooled NLS and WLS sample

Panel (a)

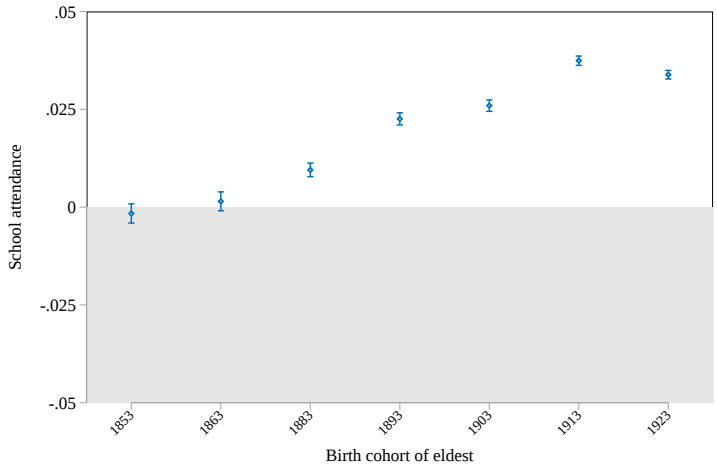
Outcome:		Number of children in family					
Years of education	Full sample	2	3	4	5	6	7+
Birth order	-0.0677***	-0.327***	-0.241***	-0.181***	-0.0407	-0.0547	-0.0890**
	(0.0205)	(0.0892)	(0.0531)	(0.0505)	(0.0622)	(0.0765)	(0.0388)
Constant	12.74***	14.05***	13.68***	13.18***	12.45***	12.15***	11.66***
	(0.0541)	(0.134)	(0.103)	(0.120)	(0.177)	(0.252)	(0.162)
Observations	68,986	10,278	14,503	13,090	10,079	7,108	13,910
R-squared	0.628	0.748	0.666	0.594	0.553	0.543	0.528

Panel (b)

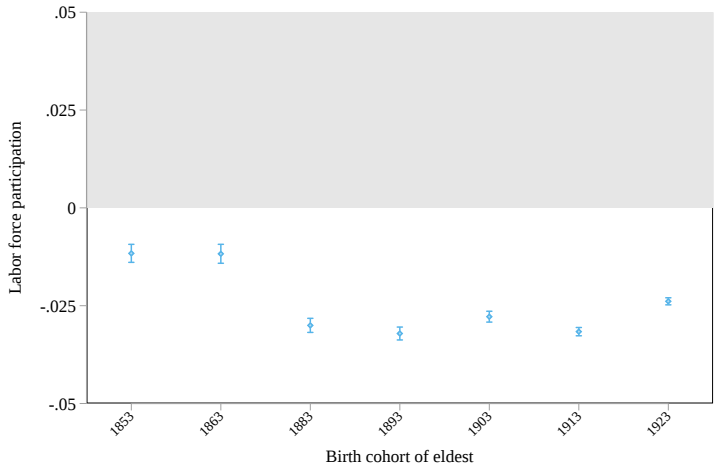
Outcome:		Number of children in family					
Years of education	Full sample	2	3	4	5	6	7+
2nd child	-0.220***	-0.327***	-0.250***	-0.316***	-0.00731	-0.248**	-0.212**
	(0.0336)	(0.0892)	(0.0637)	(0.0768)	(0.0999)	(0.123)	(0.0984)
3rd child	-0.327***		-0.478***	-0.389***	-0.122	-0.237	-0.317***
	(0.0517)		(0.108)	(0.106)	(0.137)	(0.180)	(0.123)
4th child	-0.337***			-0.503***	-0.0706	-0.241	-0.387***
	(0.0703)			(0.155)	(0.183)	(0.237)	(0.141)
5th child	-0.416***				-0.189	-0.314	-0.356**
	(0.0955)				(0.277)	(0.300)	(0.169)
6th child	-0.389***					0.0680	-0.534***
	(0.113)					(0.422)	(0.197)
7th or younger child	-0.263*						-0.379
	(0.135)						(0.247)
Constant	12.77***	13.72***	13.44***	13.03***	12.40***	12.14***	11.60***
	(0.0359)	(0.0446)	(0.0502)	(0.0721)	(0.117)	(0.178)	(0.126)
Observations	68,986	10,278	14,503	13,090	10,079	7,108	13,910
R-squared	0.629	0.748	0.666	0.595	0.553	0.545	0.529

Notes: *** p<0.01, ** p<0.05, * p<0.1. This table reports results from regressions of years of education on birth order for the pooled NLS and WLS sample for families of each size. Each column represents a separate regression. Panel (a) reports coefficients on the birth-order index; panel (b) reports coefficients on indicators for each individual birth-order position. The sample includes individuals from the pooled NLS and WLS data for whom I observe age, sex, and years of education for at least one sibling. All regressions include sex, birthyear, and household fixed effects. Observations are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.A.1: School attendance and labor force participation results using 1870 to 1940 Censuses



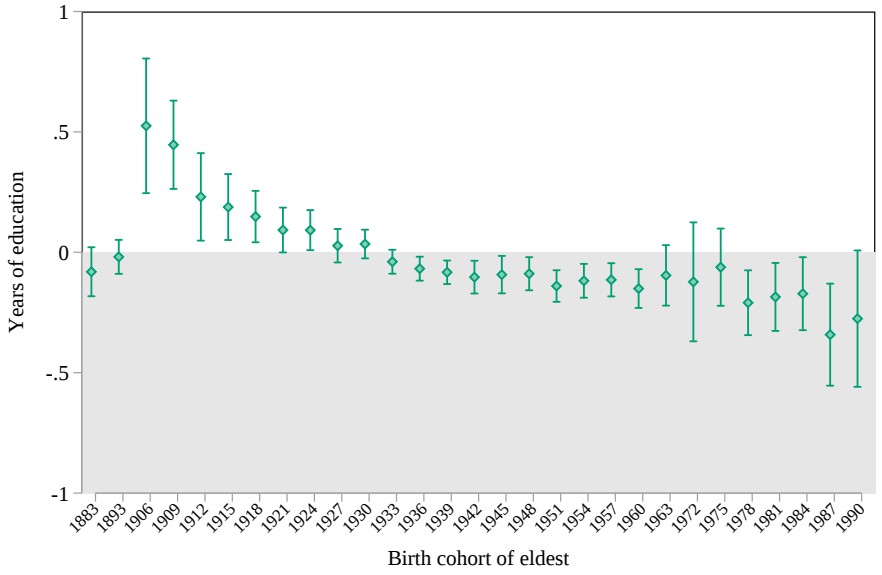
(a) School attendance



(b) Labor force participation

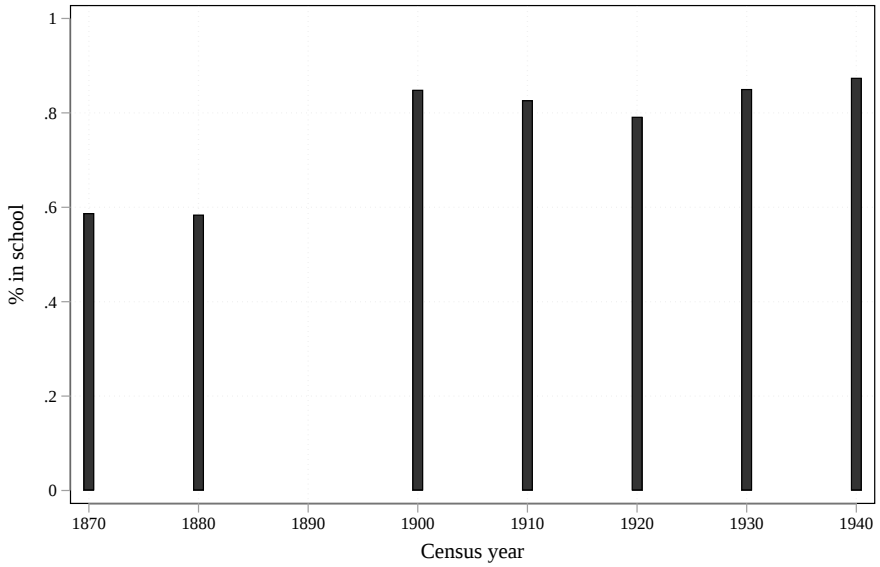
Notes: This figure presents coefficients on birth-order from equation (1.1) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals all native-born individuals aged 10-17 in the 1870 to 1940 Censuses for whom I observe age, sex, and school attendance or labor force participation for at least one sibling; who are children of the household head; and who have mothers between 35-40 years old at the time of the Census. Definitions of "school attendance" and "labor force participation" change slightly between Censuses; see Data section for further discussion. All regressions include sex, birthyear, and household fixed effects. Standard errors are clustered at the household level.

Figure 1.A.2: Years of education results for matched Census subsample



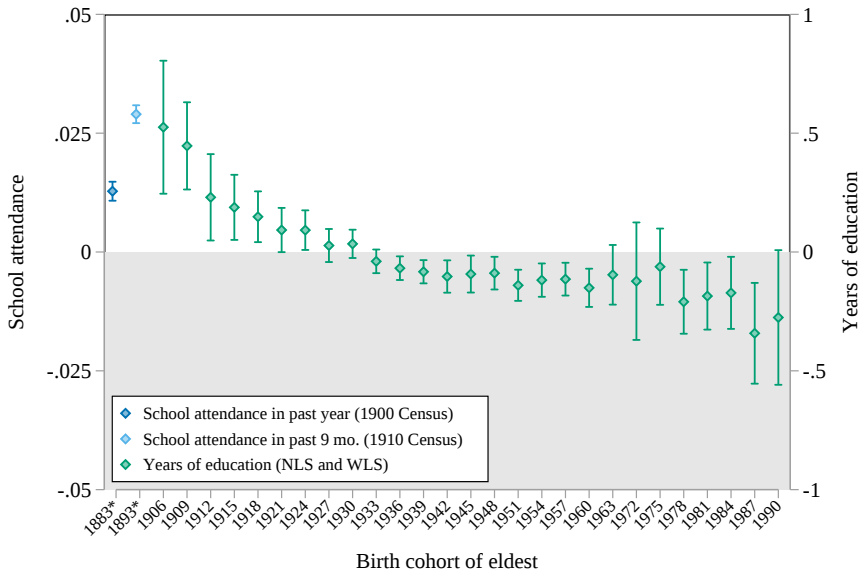
Notes: This figure presents coefficients on birth-order from equation (1.2) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals for whom I can determine birth order and for whom I observe age, sex, and years of education for at least one sibling. The 1883 and 1893 cohorts include families with children aged 10-17 in the 1900 and 1910 Censuses respectively who can be matched to the 1940 Census. Post-1900 cohorts include families with eldest siblings born in each three-year bin from the pooled NLS and WLS sample. Cohorts with fewer than 100 observations (1900, 1903, 1966, and 1969) are not shown. All regressions include sex, birthyear, and household fixed effects. NLS and WLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.A.3: Share of 10- to 17-year-old children in 1870 to 1940 Censuses in school



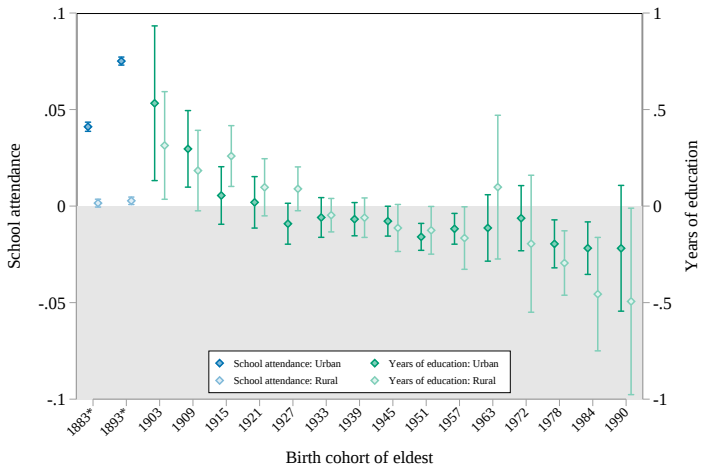
Notes: This figure reports the share of children aged 10 to 17 and living with their parents in the 1870 to 1940 Censuses who report being in school; see Appendix 1.B for definitions of schooling across Censuses. There is no data from the 1890 Census.

Figure 1.A.4: Unweighted results for Figure 1.5

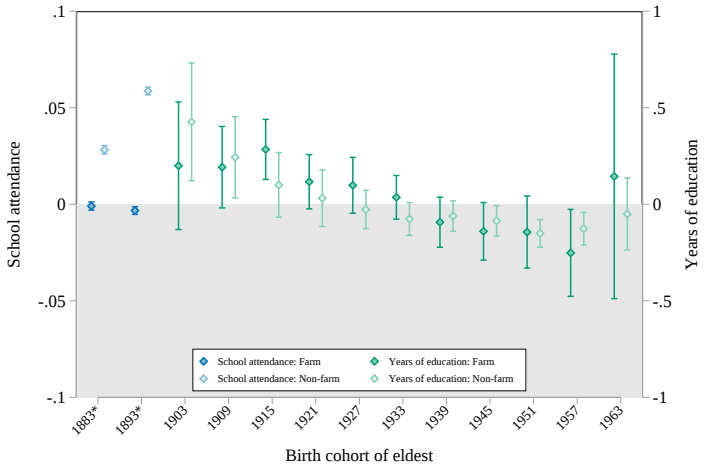


Notes: This figure presents unweighted coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. The 1883 and 1893 cohorts comprise all families from the 1900 and 1910 Censuses respectively; the outcome variable is school attendance. Post-1900 cohorts include all families with eldest siblings born in each three-year bin from the pooled NLS and WLS sample; the outcome variable is years of education. Cohorts with fewer than 100 observations (1900, 1903, 1966, and 1969) are not shown. All regressions include sex, birthyear, and household fixed effects. Standard errors are clustered at the household level.

Figure 1.A.5: Heterogeneity analysis



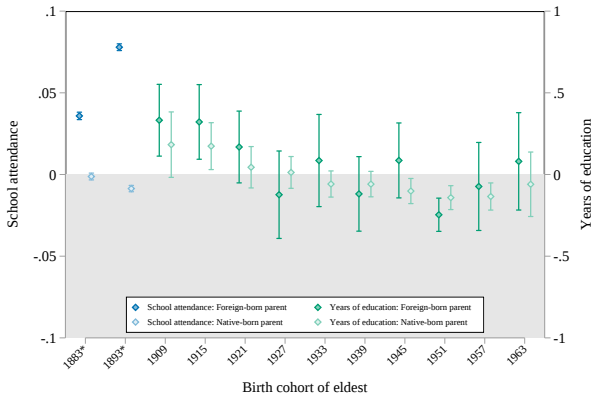
(a) Urban vs. rural



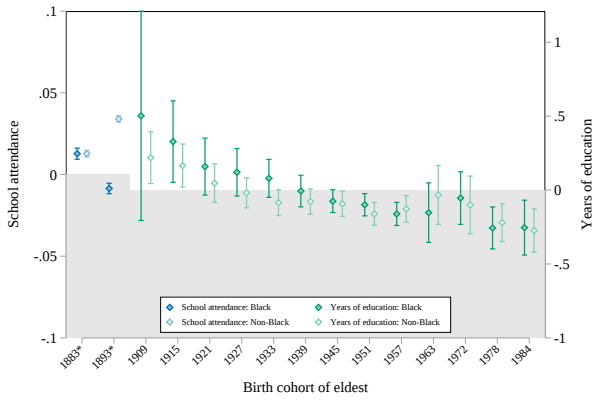
(b) Farm vs. non-farm

Notes: This figure presents coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family, where the sample is split into the subgroups listed. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals from the 1900 and 1910 Censuses, the NLS, and the WLS for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. All regressions include sex, birthyear, and household fixed effects. NLS and WLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

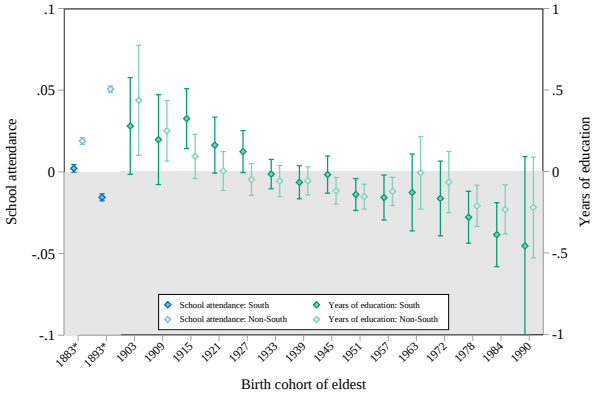
Figure 1.A.5: Heterogeneity analysis (cont.)



(c) Foreign- vs. native-born parent



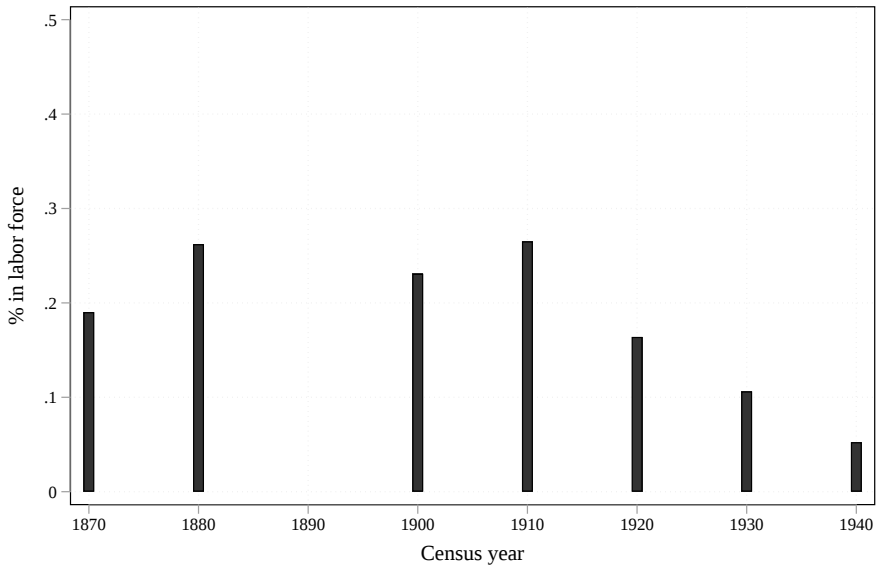
(d) Black vs. non-Black



(e) South vs. non-South

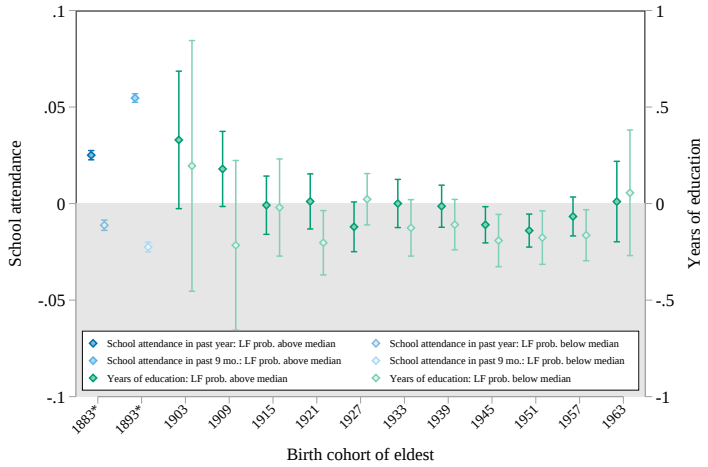
Notes: This figure presents coefficients on birth-order from equations (1.1) and (1.2) by the birth cohort of the eldest sibling in the family, where the sample is split into the subgroups listed. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals from the 1900 and 1910 Censuses, the NLS, and the WLS for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. All regressions include sex, birthyear, and household fixed effects. NLS and WLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.A.6: Share of 10- to 17-year-old children in 1870 to 1940 Censuses in labor force

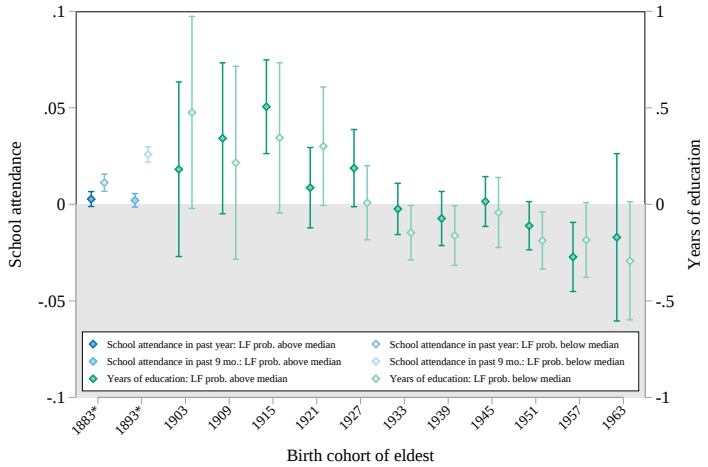


Notes: This figure reports the share of children aged 10 to 17 and living with their parents in the 1870 to 1940 Censuses who report having an occupation. There is no data from the 1890 Census.

Figure 1.A.7: Sample split by propensity for child labor, separately by region



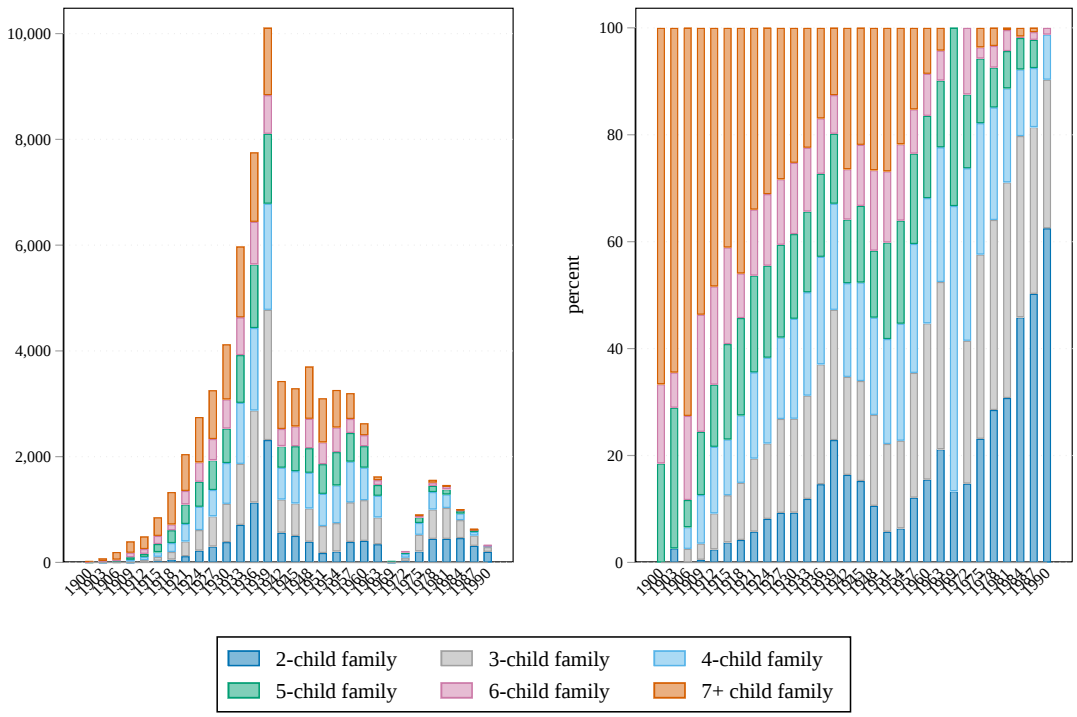
(a) Non-South only



(b) South only

Notes: This figure presents coefficients on birth-order from equations (1.1) and (1.2) after splitting the sample in families with above- and below-median propensities for child labor, separately for Southern and non-Southern families. Coefficients in the shaded area indicate cohorts in which older siblings have better educational outcomes than their younger siblings. The sample includes individuals from the 1900 and 1910 Censuses and the NLS for whom I can determine birth order and for whom I observe age, sex, and school attendance or years of education for at least one sibling. All regressions include sex, birthyear x South, and household fixed effects. NLS data are weighted to match the full U.S. native-born population for each birthyear on race, urban/rural status, farm status, residency in the South, and foreign-born parentage. Standard errors are clustered at the household level.

Figure 1.A.8: Composition of sample by family size (NLS and WLS)



Notes: This figure reports the total number of observations in each cohort, and percentage of observations in each cohort, from each family size in the pooled NLS and WLS data. Cohorts are defined as three-year bins based on the birthyear of the eldest child in a family.

1.B Data appendix

I provide further detail on the data sources I use below.

1.B.1 Data using school attendance as outcome variable

1900 and 1910 complete-count Censuses

I use the digitized complete-count U.S. Census data available from NBER. The Census is taken every 10 years and includes all people living in the United States at the time. Full-count data is available for Censuses from 1850 to 1940. However, women were only asked about their total number of children surviving in 1900 and 1910 (CHSURV); I use this question to determine completed family size and birth order. Thus, my preferred sample includes children from the 1900 and 1910 Censuses only.

I classify a child as "in school" if they report that they have attended school in a time period specified by the Census; this time period is defined as "within the past year" in the 1900 Census (when respondents were asked on June 1st) and "since September 1" of the previous year in the 1910 Census (when respondents were asked on April 15th).

I classify a child as "in the labor force" if they report any occupation in any Census. Like the schooling question, the universe and comparability of this question has varied over the years. All respondents were asked about their occupation in 1870 and 1910 to 1930, while this question was only asked to individuals aged 10 or older in 1880 and 1900 and to individuals aged 14 or older in 1940. Note that using these variable definitions, "in school" and "in the labor force" are not mutually exclusive.

I restrict the sample to children whose mothers are at least 40 years old; these mothers have likely finished having children. I then exclude families with multiple children born in the same year. I further restrict to only families where the number of

children currently living in the household is equal to or less than the total number of surviving children born to the mother as recorded in the CHSURV variable. CHSURV reports the number of children born to each ever-married woman who are still alive on Census day. I drop families where number of children in the household is larger than the mother's number of children surviving (likely adopted or stepchildren) for consistency, as I am only able to observe education for full siblings or biological children for most of my data. This restriction drops around 5% of children from the sample.

To determine each child's position in the birth order, I rank all individuals who are classified as a child of the household head, regardless of age, by birthyear. If the number of children in the household is less than the number of surviving children born to the mother, I assume that the missing children are older and have already left their parents' household and adjust the birth-order rank for observed children accordingly.

I keep all 10- to 17-year-olds still living in their childhood household at the time of each Census. Finally, I restrict my sample to only families who have two or more children in the age range, and to families where all children were born in the U.S. My final preferred sample consists of 1,136,536 observations from 439,398 families in the 1900 Census and 1,290,044 observations from 499,885 families in the 1910 Census.

I can also perform my analysis using a longer sample from the 1870 to 1940 Censuses. However, because CHSURV was not asked in other Census years, I use only families whose mothers were aged 35-40 in each Census (including 1900 and 1910 for consistency). This restriction yields total sample sizes of 879,208 (341,156 families) in 1870; 958,680 (379,775 families) in 1880; 1,538,915 (598,003 families) in 1900; 1,870,896 (724,216 families) in 1910; 2,269,316 (889,138 families) in 1920; 2,828,059 (1,114,884 families) in 1930; and 2,516,412 (1,010,793 families) in 1940. I do not use the 1850 and 1860 Censuses, as labor force data in those years were

collected only for free males and free persons aged 15 or older respectively. I am unable to include data from 1890, as Census records from that year were damaged in a fire.

Finally, I also constructed a matched Census sample to look at effects on highest grade completed. The Census does not ask respondents for their highest grade completed until 1940. I thus match individuals in my preferred Census sample from their childhood Census to the 1940 Census using the Census Linking Project (Abramitzky, Boustan, and Rashid 2020) and drop all singletons. The CLP is only able to link men between Censuses; thus, matching restricts my sample to men (and thus families with at least two sons) only. The matched sample consists of 32,885 observations from 15,816 families in the 1900 Census and 55,729 observations from 26,587 families in the 1910 Census.

1.B.2 Data using years of education as outcome variable

NLS Original Cohorts

The NLS Original Cohorts comprise two cohorts of men from the National Longitudinal Survey of Older and Young Men (NLSM) and two cohorts of women from the National Longitudinal Survey of Mature and Young Women (NLSW). I observe age, sex, and years of schooling for respondents and their siblings from three of the four Original Cohorts: the NLSM Young Men, the NLSW Mature Women, and the NLSW Young Women. Respondents in the NLSM Older Men cohort were not asked about their siblings; they were asked about the educational attainment of each of their children (including nonresident children). I include the children of the NLSM Older Men in my analysis rather than the respondents themselves.

The Older Men cohort of the NLSM consists of 5,020 men born between 1901 and 1926 (aged 45 to 59 when first interviewed in 1966). In 1976, they were asked about the age, sex, and completed years of schooling for each child (including nonresident

children). Of the original respondents from 1966, I drop 1,533 men due to nonresponse in 1976 and 158 did not have any children.

The Younger Men cohort of the NLSM consists of 5,225 men born between 1942 and 1952 (aged 14 to 24 when first interviewed in 1966). In 1976, men were asked about the highest grade of school they completed, the number of siblings they had (including surviving nonresident siblings), and the age, sex, and completed years of schooling for each sibling. Of the original respondents from 1966, I drop 1,551 men due to nonresponse in 1976 and 232 due to their being only children.

The Mature Women cohort of the NLSW consists of 5,083 women born between 1923 and 1937 (aged 30 to 44 when first interviewed in 1967). In 1977, women were asked about the highest grade of school they completed, the number of siblings they had (including nonresident siblings), and the birthyear, sex, and completed years of schooling for each sibling. Of the original respondents from 1967, I drop 1,167 women due to nonresponse in 1977 and 308 due to their being only children.

The Younger Women cohort of the NLSW consists of 5,159 women born between 1944 and 1954 (aged 14 to 24 when first interviewed in 1968). In 1978, women were asked about the highest grade of school they completed, the number of siblings they had (including nonresident siblings), and the birthyear, sex, and completed years of schooling for each sibling in 1977. Of the original respondents from 1968, I drop 1,273 women due to nonresponse in 1978 and 196 due to their being only children.

NLSY79 and NLSY79 Child and Young Adult

The National Longitudinal Survey of Youth 1979 (NLSY79) consists of 12,686 respondents born between 1957 and 1965 (aged 14 to 22 when first interviewed in 1979). In 1993, respondents were asked about their highest grade completed and about their siblings (including sex of each sibling, whether each sibling was older or younger, how many years older or younger, and each sibling's level of education). Of

the original respondents from 1979, I drop 3,690 due to nonresponse in 1993 and 331 due to their being only children.

The National Longitudinal Survey of Youth 1979 Child and Young Adult (NLSY79 CYA) surveys record information for every child born to a woman in the original NLSY79 respondent group. The NLSY79 CYA began data collection in 1986 (and the Young Adult survey for children over age 15 was introduced in 1994) and is ongoing as of 2020. The latest survey wave I use is 2018. The NLSY79 CYA is thus able to accurately determine birth order for each child, including children who have moved out of the mother's household and children who have died. The full sample includes 11,545 children from 4,941 mothers from the NLSY79; I drop 1,178 individuals who were only children at the time of the 2018 survey. While questions about education in my other sources of modern data are retrospective, the NLSY79 CYA collects information about each child's education every two years. For years up to 2012, young adult respondents are asked about the highest grade they have completed. However, the wording and coding of this question changed in 2014; responses were grouped into categories rather than exact grades as they had been in prior years. I recode the post-2012 categories into grade levels as described below. If a respondent's imputed highest grade completed in any post-2012 survey is the higher than their highest grade completed in 2012, I denote that respondent's completed years of schooling as the imputed post-2012 grade. Otherwise, I denote a respondent's completed years of schooling as the reported grade from the 2012 survey or from their most recent pre-2012 response.

Wisconsin Longitudinal Study

The Wisconsin Longitudinal Study is a study of of 10,317 respondents who graduated from Wisconsin high schools in 1957. Survey data were collected in various years between 1957 and 2011. In 1975, respondents were asked to provide a full sib-

Recoding of post-2012 education variable in NLSY79 CYA

NLSY79 CYA response	Recoded years of education
"8th grade or less"	8
"Some high school"	10
"High school graduate"	12
"Some vocational/technical training (after high school)"	13
"Completed vocational/technical training (after high school)"	13
"Some college"	13
"Completed college (associate's degree)"	14
"Completed college (bachelor's degree)"	16
"Some graduate school"	17
"Completed master's degree"	17
"Some grad school beyond master's degree"	17
"Completed doctoral degree (PhD)"	17
"Some post-baccalaureate professional education (e.g., law school, med school, nursing)"	17
"Completed post-baccalaureate professional education (e.g., law school, med school, nursing)"	17

ling roster (including the respondents themselves) with information about the age, sex, and the highest grade of schooling each sibling had completed. Of the original respondents from 1957, I drop 754 due to nonresponse in 1975 and 646 due to their being only children.

Sample construction and restrictions

To construct my modern sample, I first pool data from the NLS and WLS. I keep families with at least two children for whom I have information on birthyear or age for all siblings; I need to observe age for all siblings in order to accurately determine birth order within a family. For multiple-respondent families, I count each family only once. I further restrict the sample to families where all children were born in the U.S. This preliminary sample yields 98,398 individuals from 24,493 families.

I then construct the birth order variable for each sibling within a family by ranking children in each family by age or birthyear and topcode birth order at 7 for the 1,174 families with more than seven children. I drop families with multiple births (e.g. twins, triplets, etc.) or children born in the same year and any families with an age gap of more than 25 years between the oldest and youngest child. I also drop individuals who are younger than 23 at the time of the survey. Finally, I keep only

families for whom I am able to observe the completed education for at least two siblings and topcode years of schooling at 17.

To adjust for any sampling biases, I weight all regressions using the NLS or WLS to match the US native-born population by race (White, Black, or other), urban status, farm status, geographic region (South or non-South), and whether any parent was foreign-born. Weights are calculated separately for each birthyear, based on population counts from the Census in which individuals born that year were aged ten or younger (e.g., weights for individuals born in 1929 were calculated using all native-born 9-year-old children in the 1920 Census). For the WLS, because I cannot observe race or immigration status, I weight by urban status and farm status to match Wisconsin only. For the NLSY79 CYA, because I do not observe farm or immigration status, I weight by race, urban status, and geographic region only.

These restrictions yield a preferred sample of 69,734 individuals from 19,745 families, with one family per survey respondent for non-NLSY79 CYA surveys.

Chapter 2

Green Card Eligibility of High-Skilled Workers and U.S. Innovation[1]

2.1 Introduction

Policymakers and economists have long discussed the effect of high-skilled immigration on the U.S. economy. Proponents of increasing high-skilled immigration argue that immigrants contribute positively to U.S. invention, research and development, and job creation. For example, while immigrants make up around 14 percent of the U.S. population, 20.4 percent of Fortune 500 companies in 2017 had at least one immigrant co-founder (Hathaway, 2017). Immigrants also accounted for 26 percent of U.S.-based Nobel Prize winners between 1990 and 2000 and produced around 23 percent of total innovation between between 1990 and 2016 (Bernstein et al., 2022

[1]I am incredibly grateful to Leah Boustan and Ilyana Kuziemko for their guidance and support in this project. I also thank Dorisz Albrecht, Jiwon Choi, Ellora Derenoncourt, Hank Farber, David Lee, Alex Mas, Christiane Szerman, Carolyn Tsao, and participants in the Princeton Labor Economics Seminar and Industrial Relations Section Labor Lunch for helpful comments. All errors are my own.

WP). However, others argue that immigrants may crowd out native-born workers by competing for employment opportunities and may even discourage natives from pursuing higher education.

In this paper, I examine how granting Lawful Permanent Residence (green card) status to high-skilled foreign nationals on temporary visas affects U.S. innovation. In particular, I exploit an exogenous shock in LPR eligibility: the passage of the 1992 Chinese Student Protection Act after the 1989 Tiananmen Square Incident, which granted green card eligibility to People's Republic of China (PRC) nationals living in the U.S. at the time. To estimate the causal effect of increasing green card eligibility of foreign students on U.S. innovation, I compare patenting in research areas that were more or less exposed to the CSPA. I use variation in counties and research fields of pre-1990 PRC inventors to determine the level of CSPA exposure for each county-field cell.

I find that patenting increased in cells that were more exposed to the CSPA and thus more likely to have experienced an increase in immigrant inventors after 1992. Specifically, county-field cells with a 1-percentage-point higher share of pre-period PRC inventors experienced a 3 to 4 percent increase in patent applications. I then explore mechanisms through which green card eligibility may have affected patenting. Receiving a green card may allow foreign graduate students access to a wider range of employment options in the U.S. post-graduation – particularly in non-academic private sector positions, for which employment-based temporary visas and immigration are capped. Consistent with evidence (e.g., Lan, 2013) that green cards increase foreign graduate students' employment in industry, I find that the increase in patenting seemed driven by growth in non-academic private-sector patenting rather than university patenting. Furthermore, the increase in patenting occurred primarily through the extensive margin: cells with a 1-percentage-point higher share of pre-period PRC inventors experienced a 2 to 3 percent increase in the number of new

inventors. Thus, providing green card eligibility to high-skilled foreign-born nationals appeared to have a positive effect on innovation through encouraging their entry into U.S. patenting.

This paper contributes to the large literature on the effects of immigration on receiving countries' economies. Hunt and Gauthier-Loiselle (2010), Akcigit et al. (2017), Bernstein et al. (2022 WP) find that immigrants contribute positively to U.S. innovation. Many studies – e.g., Kerr and Lincoln (2010), Peri et al. (2013), Dimmock et al. (2021), and Doran et al. (2022), among others – leverage variation in H-1B visa policy to determine the causal effect of high-skilled foreign-born temporary workers on various measures of productivity, such as patenting, research and development, employment, and wages.

Increases in long-term or permanent immigration may affect innovation more so than temporary work visas such as the H-1B, but its effects have been relatively less explored. This paper contributes to the literature by exploiting a unique shock in the green card eligibility of foreign students. Several papers use historical events as "natural experiments" to determine the effect of permanent immigration on productivity: Borjas and Doran (2012) argue that the influx of Soviet mathematicians into the U.S. after 1992 had negative effects on the publishing activity of their native-born counterparts, although overall productivity remained the same because the Soviet immigrants compensated for the decline. Moser et al. (2014) find that German-Jewish scientists who left Nazi Germany for the U.S. in the 1930s increased American patenting in their research fields. Similarly, Moser and San (2020 WP) find that the restrictions on immigration placed by the Quota Acts in 1920 resulted in decreases in patenting, again suggesting a positive contribution of immigrants to U.S. innovation. However, these papers primarily examine the effect of relatively established scholars on innovation and publication. On the other hand, I focus on a policy change that

affected mostly students, a group whose immigration pathways have been the focus of much of the recent policy discussion.

Finally, this paper also adds to the small body of literature on the Chinese Student Protection Act itself. Previous studies exploring its impact have primarily focused on the effects of obtaining permanent residency on the beneficiaries of the Act themselves: Lan finds that Chinese PhD students who had been eligible for green cards through the CSPA were less likely to take lower-paying post-doctorate positions after graduation (2013), were more likely to work in non-academic private-sector jobs, and eventually produced more patents (2012) than other international students on temporary visas. Orrenius et al. (2012) find that Chinese-born individuals in the Census earned higher wages and were more likely to be employed than similar foreign-born individuals; the positive effect of green cards on employment was higher for women, who, as the wives or dependents of international students, may have faced more restrictions in finding employment as a condition of their pre-CSPA visas. Finally, Su et al. (2019) find that the CSPA had positive effects on the human capital development of second-generation immigrants whose mothers received a green card through the Act. This paper extends the current literature by focusing on the effect of the CSPA not only its direct beneficiaries but instead *on the U.S. as a whole*: that is, how increasing high-skilled immigration may affect overall U.S. innovation.

This paper proceeds as follows. In Section 2.2, I provide historical background on the Chinese Student Protection Act. I describe the data in Section 2.3 and outline my empirical strategy in Section 2.4. Section 2.5 presents results. Lastly, I explore potential mechanisms in Section 2.6 and conclude in Section 2.7.

2.2 Background on the Chinese Student Protection Act

After the Tiananmen Square protests in June 1989, in which university students in Beijing demonstrated against the Chinese Communist Party, the United States sought to prevent the political persecution of Chinese students living in the U.S. at the time. Thus, on April 11, 1990, President George H. W. Bush issued Executive Order 12711, which allowed PRC nationals who had been living in the U.S. between June 5, 1989, and April 11, 1990, to remain in America regardless of their visa status. The Executive Order also granted them employment authorization through January 1, 1994.

This change in immigration status was made permanent when the Chinese Student Protection Act was passed by Congress and signed into law on October 9, 1992. The Act took effect on July 1, 1993, and allowed all PRC nationals to apply for permanent residency through July 1, 1994, if they had been living continuously in the U.S. since April 11, 1990. 52,531 PRC nationals (and 478 dependents) received green cards through the CSPA, many of whom were students.

Executive Order 12711, and the CSPA that followed, was a large and unexpected shock to the immigration status of Chinese nationals in the United States. Figure 2.1 shows the total number of green cards issued to PRC nationals over time, using microdata from the Immigration and Naturalization Service's Immigration Yearbooks (collected annually between 1972 and 2000). Little immigration to the US from the PRC occurred prior to 1978, due to the prohibition of emigration out of mainland China. Immigration jumped after 1979, when the PRC relaxed its emigration restrictions, and increased steadily until 1991. In 1992, however, there was a large spike in the number of green cards issued. Nearly all of this spike can be attributed to the CSPA: immigration from PRC nationals more than doubled in 1992 relative to

the previous year, and the number of green cards issued due to the CSPA comprised around 10 percent of the total stock of green cards issued to PRC nationals between 1972 and 2000.[2] Post-1992, green card issuances to PRC nationals returned to their pre-1990 trend.[3]

Most CSPA beneficiaries had indeed been students or other high-skilled immigrants. Figure 2.2 presents the percent of CSPA green card recipients by the category of their most recent temporary visa, and the underlying percentages are presented in Appendix Table 2.A.1. Prior to their receiving green cards, around 40 percent of CSPA beneficiaries been in the U.S. on student (F1) visas; most of these were graduate students. Furthermore, 9 percent had been on temporary worker (H1) visas and 8 percent on exchange (J1) visas issued to foreign scholars amd researchers. Overall, more than 57 percent of PRC nationals who received green cards through the CSPA had been on these "high-skill" visas; an additional 18 percent had been their dependents (e.g. spouses and children). Thus, the CSPA increased the number of high-skilled immigrants to the U.S. by around 30,000.[4]

The CSPA did in fact increase the post-graduation stay rate of PRC students. Appendix Figure 2.A.1 presents the share of PRC science and engineering PhD students with plans to stay in the U.S. upon completion of their degree, using data from the National Science Foundation's annual Survey of Earned Doctorates. The share of students from the PRC with plans to remain in the U.S. increased from around 60 percent to more than 90 percent between 1990, when the Executive Order was

[2]The INS data records whether an individual received a green card as a result of the CSPA, either as a direct beneficiary or as a dependent. I show the classes of admission for all PRC green card recipients between 1972 and 2000 in Appendix Table 2.A.2.

[3]Immigration from Eastern European countries also increased over this time period after the collapse of the Soviet Union. Most of these immigrants received green cards due to family reunification or refugee status. I consider possible effects of other immigrant groups on U.S. patenting in Section 2.5.1.

[4]Because eligibility for permanent residency under the CSPA did not require a valid visa at the time of application, the Act also granted green cards to a number of undocumented PRC migrants who had been living in the U.S. at the time. While this paper examines the effects of providing green card eligibility to high-skilled immigrants, the CSPA may provide an interesting setting to study the effects of low-skilled or low-wage immigration as well.

issued, and 1992, when the CSPA was passed. No similar sharp increase is observed for international graduate students from other countries.[5]

Furthermore, CSPA beneficiaries were also more likely to become naturalized citizens of the United States: Appendix Figure 2.A.2 shows this using data from the 2000, 2005, 2010, and 2015 waves of the American Community Survey. Of PRC-born individuals in the ACS with more than 16 years of schooling, those who arrived in the U.S. between 1985 and 1990 have consistently higher naturalization rates than those who arrived in other years.

Taken together, this descriptive evidence is consistent with the story proposed by policymakers today – and by Lan (2013) and Orrenius et al. (2012) in the context of the CSPA – that permanent residency opens up a wider set of post-graduation employment opportunities to international students in the U.S. Without green cards, foreign students' options are to either seek employment with companies willing sponsor them for one of the limited number of H-1B visas granted each year, or to remain in academia, in which there are no restrictions placed on the number of H-1B visas available but far fewer job openings.

2.3 Data

I now turn to estimating the *causal* effect of high-skilled immigration on U.S. innovation. Specifically, I estimate the effect of the Chinese Student Protection Act on U.S. innovation through changes in patenting: my empirical strategy compares patenting in different areas, defined as county-research field cells, that were differentially exposed to the CSPA. This section describes the sources and construction of the data used in my main analysis.

[5]Finn (2010) matches graduate student names in the Survey of Earned Doctorates to IRS tax returns to determine how many foreign-born students *actually* remained the the U.S. after five years. A similar pattern emerges: after the issuance of Executive Order 12711, students from the PRC were more likely to still be in the U.S. five years after graduation than students from other countries.

2.3.1 U.S. patent data

The main data source used in this paper is patent data from the United States Patent and Trademark Office (USPTO), accessed through the PatentsView website. The USPTO provides full-text data on all patents granted from 1976 to the present, including information on each patent's application and issue dates, patents cited, number of citations, detailed IPC patent classification categories, and NBER technology categories. The data also includes information on inventors and assignees and their names and locations (cities, counties, and states). Furthermore, inventor and assignee data are disambiguated, allowing the tracking of inventors over time.[6]

2.3.2 Patent applications

I use patent applications as the outcome variable in my main analysis. I restrict my analysis to successful patent applications filed between 1976 and 2007 in my analysis; this shorter time frame removes possible concerns about changes in patenting after the Great Recession. Because patents may face delays of several years between the submission of the application and the grant date; I use the year of the patent application, rather than the year in which the patent was granted. Patents are classified as either utility, design, or plant patents; I do not include the latter two categories in my analysis.[7] Most patent applications filed are for utility patents; utility patent applications consisted of around 90 percent of all patent applications filed during this time period. Finally, I also consider only patent applications with at least one U.S.-based inventor or assignee – around 45 percent of all utility patent applications filed between 1976 and 2007 – as this paper focuses on the effects on U.S. innovation.[8]

[6]For more detail on the NBER classification scheme and disambiguation algorithm, see Hall, Jaffe, and Trajtenberg (2001) and Marco et al. (2015).

[7]Design patents are granted to protect the visual appearance or ornamentality of a product, rather than its use. Plant patents are granted for the discovery or creation of new plant species.

[8]This excludes patents with only foreign-based inventors or assignees, e.g., patents from Japanese companies by Japanese inventors.

I then aggregate the patent-level data into counts of patent applications per year for county-by-research field cells. I define each patent's county as the county of residence of its first named inventor.[9] I define each patent's research field as its NBER technology subcategory as assigned in the PatentsView database. The NBER classified all patents between 1976 and 2015 into one of 6 technology categories and one of 36 technology subcategories listed in Table 2.1. I show the distribution of patent applications across NBER technology categories in Appendix Table 2.A.4 and trends in patenting across categories in Appendix Figure 2.A.3.

2.3.3 Determining CSPA exposure

Ideally, to determine the effect of the CSPA on U.S. patenting, I would compare patenting between research fields with different shares of *pre-period graduate students* from the PRC. That is, I would leverage variation in CSPA exposure across different research fields chosen by the students while still in graduate school, before they knew they would receive green cards through the CSPA. However, the largest issue with this approach is that research fields in graduate school do not map cleanly to the either the NBER or IPC patent classification schemes defined in the USPTO data. Thus, the only data on research field I have are the NBER technology categories in which inventors would eventually patent. That is, even if I were able to identify the exact inventors who were beneficiaries of the CSPA in the patent data, I would only be able to observe their fields from categories in which they endogenously chose.

Given the data available, I use instead the share of *pre-CSPA* inventors from the PRC in each county-NBER technology subcategory cell as a proxy for CSPA exposure. Specifically, I calculate the level of CSPA exposure for each cell as the share of inventors between 1984 and 1988 with PRC nationality:

[9] Trends in patenting by county, for counties with more than 1,200 total patents, are presented in Appendix Figure 2.A.4.

$$PercentPRC_{ij} = \frac{\#\text{ of U.S.-based PRC inventors between 1984-1988}}{\#\text{ of U.S.-based inventors between 1984-1988}} \tag{2.1}$$

Thus, I exploit pre-period variation in geography and NBER technology subcategory PRC inventors to determine the level of CSPA exposure for each county-field cell.

Predicting PRC nationality

To determine the nationality of inventors in my data, I use the inventor names available in the PatentsView data. However, most current ethnicity predictors only determine ethnicity or race, rather than nationality, and may not distinguish between PRC and other East Asian inventors (e.g., Taiwanese, Korean). Thus, I use a three-step process to distinguish PRC names in my dataset.

First, I use the publicly available ethnicolr package developed by Laohaprapanon, Sood, and Naji to determine the *ethnicity* of each inventor. The ethnicolr package includes two pre-trained machine learning models. The first model is trained on data from the 2000 Census and predicts race (White, Black, API, and Hispanic) from last names. I use this model in my main classification of PRC names.[10] I denote all names classified by ethnicolr as White, Black, and Hispanic as non-PRC.

Secondly, for the inventor names classified as API by ethnicolr, I further distinguish between PRC and non-PRC API names using unique attributes of Chinese names (e.g., name length, character substrings). For example, Taiwanese immigrants and immigrants who arrived during earlier waves of Chinese emigration (predominantly from Hong Kong or Macau rather than mainland China) would not have benefited from the CSPA. They were also culturally different from immigrants from the

[10]The second model is trained on data from Wikipedia collected by Skiena et al. and classifies names into more fine-grained ethnicity categories; I use these categories for robustness checks.

PRC and may have specialized in different areas of research. To distinguish between these groups and the newer (post-1978) PRC immigrants, to whom CSPA beneficiaries would have been more similar, I exploit differences between the romanization systems used by each group: Hanyu Pinyin for Mandarin names, Wade-Giles for Taiwanese, and Yale for Cantonese. I provide further detail on the specific characteristics of Hanyu Pinyin I use in Appendix 2.B.

Finally, I predict *nationality* for the remaining unclassified API names using proprietary NamSor software to identify PRC inventors. The NamSor software uses artificial intelligence trained on different data sources and is able to distinguish between mainland Chinese-origin names and those originating from Taiwan, Hong Kong (Cantonese), and other East Asian countries.[11] Thus, I am able to predict which inventors actually originated from the PRC at a more finer and accurate level than previous studies using inventor names from patent data. The most common surnames for each nationality are listed in Appendix Table 2.A.3.

Support for measure of CSPA exposure

The measure of exposure I use, as defined in equation (2.1), relies on the assumption that PRC inventors who received green cards due to the CSPA would have patented in similar fields as the PRC inventors immediately before them.

I provide support for my measure of CSPA exposure in two ways. First, I show that mainland Chinese *PhD students* who were eligible for green cards under the CSPA concentrated in similar fields as those who graduated slightly earlier and thus were not eligible for green cards upon graduation. That is, I show that the *pre-period research field choices* of students who did and did not benefit from the CSPA were similar.

[11] Details on the NamSor algorithm can be found at https://namsor.app.

I use data from the ProQuest Dissertation database, which provides the names, fields, and dissertation subjects of nearly all PhD students to the present day. Following the procedure described in Section 2.3.3, I predict PRC nationality for each PhD student who received a degree between 1984 and 1994.

In Appendix Figure 2.A.5, I plot the correlation between the share of all PRC PhD students receiving a degree in each field, for students graduating between 1984-1988 and students graduating between 1989-1994. The PRC students graduating between 1984 and 1988 would not have received green cards upon graduation, while those graduating between 1989 and 1994 would have. The correlation between the two groups is 0.916, and most points lie close to the dashed 45-degree line. For example, around 7.5 percent of PRC PhD students graduating between 1984 and 1988 received a degree in electrical engineering, which is almost exactly the same as share of PRC students receiving EE degrees between 1989 and 1994.

Secondly, in Appendix Figure 2.A.6, I show that PRC inventors who applied for patents immediately before and after the CSPA did so in similar fields as well. I plot the correlation between the share of all PRC inventors with patents in each field between 1984-1988 and the share of all PRC inventors with patents in each field between between 1990-1998. Again, the field choices of pre- and post-CSPA PRC inventors are highly correlated, at 0.732[12] Taken together, these suggest that the initial "shares" of PRC inventors in each cell provide a good measure of exposure to the CSPA, in that cells with high pre-1990 shares of PRC inventors were likely to receive more PRC inventors after the CSPA as well.

2.3.4 Characteristics of final dataset

The final dataset is a balanced panel of 2,216 county-NBER technology subcategory cells with at least one patent application or grant in all 32 years between 1976

[12]The Drugs category is an outlier; there was a policy change for drug patents in 1995 that resulted in changes in classification for this category (Marco et al., 2015).

and 2007, for a total of 70,912 observations at the county-field-year level. This sample yields 319 distinct counties and 36 distinct NBER technology subcategories; however, I do not require that every county in the sample must have patents in *all* NBER subcategories every year, or vice versa.[13] The variation in this measure of exposure, for a selected number of counties and at a coarser level of research field (NBER one-digit technology categories instead of two-digit subcategories), is shown in the heat map in Figure 2.3.[14]

Summary statistics at the cell level are presented in Table 2.2. On average, each cell had 14 patent applications per year, most of which were private-sector patents, and 35 inventors. 0.192 percent of inventors who applied for a patent between 1984 and 1988 were from the PRC; although this share may seem small, it is not unreasonable given that emigration out of mainland China was restricted before 1978. 8.5 percent of cells (190 out of 2,216) had at least one PRC inventor between 1984 and 1988; in these cells, 2.2 percent of inventors were from the PRC. Cells with at least one PRC inventor are on average larger, both in terms of patent applications and number of inventors, and contained higher shares of inventors from other immigrant groups as well.

2.4 Empirical strategy

To determine the effect of increasing green card eligibility on patenting, I compare changes in patenting pre- and post-1990 in county-NBER subcategory cells that were more or less exposed to the CSPA. Specifically, I estimate the difference-in-differences

[13] The share of PRC inventors and total patent application counts between 1984 and 1988 are presented for each NBER technology subcategory in Appendix Table 2.A.5 and for the 36 largest counties in Appendix Table 2.A.6.

[14] I show geographic variation in CSPA exposure (i.e., the share of pre-period PRC inventors) for Chemical, Drug, Electrical, and Computer patents in Appendix Figure 2.A.7.

regression:

$$LogPatents_{ijt} = Post_t + PercentPRC_{ij} + \beta * Post_t * PercentPRC_{ij} + \gamma_i\delta_j + \gamma_i\eta_t + \delta_j\eta_t + \gamma_i + \delta_j + \eta_t + \epsilon_{ijt} \quad (2.2)$$

where $LogPatents_{ijt}$ is the log number of patent applications filed (that were later granted) by U.S. inventors in county i, NBER technology subcategory j, and year t. $Post_t$ is an indicator for the year 1990 and after, and $PercentPRC_{ij}$ is the percentage of inventors with PRC names in county i and NBER subcategory j between 1984 and 1988 as defined in (2.1).

Furthermore, to address trends in patenting over time that may differ by field or county, I include flexible controls γ_i, δ_j, and η_t for county, NBER technology subcategory, and year respectively. The $\gamma_i\eta_t$ interactions control for differences in patenting in counties over time; over this time period, for example, Santa Clara experienced much larger increases in patenting than any other county in the U.S. The $\delta_j\eta_t$ interactions control for changes in patenting by field over time, such as technological advances that may have led to increased patenting in certain research areas (e.g., computers). Thus, the effect is identified off of differences in the share of PRC inventors across fields within a county. Additionally, so that cells with a large number of patents (that thus account for a larger portion of U.S. patenting as a whole) are not treated the same as cells with very few patents, I weight each cell by its total number of patent applications between 1984 and 1988. Finally, I cluster standard errors by county.

The coefficient of interest is β, which is an estimate of the percent change in patenting after 1990, for cells with a 1-percentage-point higher share of pre-period PRC inventors. The identification assumption is that, after flexibly controlling for

county, field, and year, patenting in cells with higher and lower pre-1990 shares of PRC inventors would have evolved similarly in the absence of the CSPA.

Finally, I examine dynamics over time. To increase power, I collapse time into three-year bins and use the following event-study specification:[15]

$$LogPatents_{ijt} = \sum_{t=1976}^{2007} \beta_t * PercentPRC_{ij} * Year_t + \gamma_i\delta_j + \gamma_i\eta_t + \delta_j\eta_t + \gamma_i + \delta_j + \eta_t + \epsilon_{ijt} \tag{2.3}$$

2.5 Results

I find that patenting did increase in cells with a higher share of pre-period PRC inventors. The difference-in-difference estimates from equation (2.2) are presented in Table 2.3. I present the baseline results in column 1. County-field cells with a 1-percentage-point higher exposure to the CSPA (i.e., with a 1-percentage-point higher share of PRC inventors) experienced a 3.6 percent increase in patent applications. Thus, patent applications in county-field cells in the top quartile of CSPA exposure increased around 20 percent relative to cells in the lowest quartile of CSPA exposure.[16]

It is possible that changes in the flow of high-skilled immigrants from other ethnic groups around the same time period may have contributed to increases in patenting as well and, if immigrants from all groups tend to patent in similar fields and settle in similar counties, upwardly bias the estimate in column 1. The Immigration Act of 1990 introduced the H-1B visa and thus more than doubled the number of employment-based visas available; the number of Indians on temporary work visas in the U.S. also increased during this time period. Furthermore, the Eastern European immigration to the U.S. rose around collapse of the Soviet Union in 1991 as well. To

[15] The bins start in 1976 and end in 2007; the last bin consists of only two years: 2006 and 2007.

[16] Cells in the top quartile of CSPA exposure had on average 5.9 percent PRC inventors, relative to 0.38 percent in the lowest quartile. Thus, the CSPA increased patent applications by (5.9 - 0.38) * 0.036 = 20% between the top and bottom quartiles.

alleviate these concerns, in column 2, I present results after controlling for the share of pre-period Indian, Jewish, and Eastern European inventors in each cell interacted with year. The coefficient on $Post * PercentPRC$ remains similar to column 1, suggesting that the effect is not driven by an increase in patenting by other ethnic groups who may have entered the U.S. at the same time.

Columns 3 to 5 of Table 2.3 address potential concerns that this result may be disproportionately driven by a few select county-field cells. For example, as seen in Appendix Figure 2.A.4, patent applications increased in Santa Clara County, California, much faster than any other county in the U.S.; it is also home to a large mainland Chinese population. I present results after dropping Santa Clara from the regressions in column 3 of Table 2.3. Columns 4 drops cells with the largest pre-period patent counts to ensure that the effect is not driven by the largest county-field cells, while column 5 drops cells with the highest pre-period PRC inventor share. Reassuringly, the coefficient on $Post * PercentPRC$ remains similar in magnitude and statistically significant.

The event-study specification in equation (2.3) provides further insight into the timing of this increase in patent applications. The β coefficients from the baseline event study are presented in Figure 2.5, plotted relative to the 1988-1990 bin.[17] There is a relatively flat pretrend in the years prior to 1988. Patent applications then increased after 1990 in cells with a higher percentage of pre-period PRC inventors. The effect remains high in the first decade after the CSPA, tapering off towards the later years in my sample. Panel (a) of Figure 2.6 controls flexibly for shares of other ethnic groups over time in the baseline specification and yields similar estimates. Panel (b) of Figure 2.6 presents event-study coefficients from the same cells as in columns (3) to (5) in Table 2.3; the results are similar to those from the main specification in Figure 2.5.

[17] Results with coefficients for each year are presented in Appendix Figure 2.A.8.

2.5.1 Robustness

I further explore robustness to alternative explanations and specifications.

First, it is possible that other events that occurred around the same time may have affected patenting in cells with higher CSPA exposure. As mentioned before, the establishment of the H-1B visa from the Immigration Act of 1990 resulted in a large increase in Indian immigrants. Additionally, the collapse of the Soviet Union in 1991 increased immigration from Eastern Europe. If these immigrants patented in similar fields and settled in similar counties as mainland Chinese immigrants, then it is possible that the effect found in Figure 2.5 is overstated. While panel (a) of Figure 2.6 controls flexibly for these other groups, I show the effect of these groups on patent applications directly. That is, I use the share of pre-period Indian inventors, the share of pre-period Jewish inventors, and the share of pre-period Eastern European inventors as my "exposure" variable. I present the difference-in-difference results in Appendix Table 2.A.7 and the event study in Figure 2.7. For each alternative ethnic group, I find little effect on patent applications in more-exposed cells post-1990: the coefficients are much smaller in magnitude and not statistically significant.

The main results are also robust to alternative treatments of the outcome variable: in Appendix Table 2.A.8 and Appendix Figure 2.A.9, I show that the coefficient on $Post * PercentPRC$ remains statistically significant after adding 0.1 or 0.001 to zeros for log patent applications or using the inverse hyperbolic sine of the number of patent applications.

Finally, it is possible that the results I find are picking up an overall upward trend in patenting in cells with a higher share of PRC inventors, rather than the effect of the CSPA. To assess whether this is the case, I perform a placebo check: specifically, I restrict the sample to years before 1990 only and simulate three "placebo" CSPAs in 1979, 1982, and 1985. The results of this exercise are shown in Appendix Table 2.A.9 and Appendix Figure 2.A.10. I find no significant difference between treated

cells (those with a higher exposure to the actual CSPA) and control cells for any of these three placebo CSPAs.

2.6 Mechanisms

What are the channels through which innovation could have grown as a result of granting permanent residency to high-skilled foreign nationals? I explore two possible mechanisms: through expanding career opportunities for immigrants in the non-academic private sector, and through the extensive (increasing the number of inventors in the U.S.) and intensive margins (increasing the productivity of inventors).

2.6.1 Academic vs. non-academic patenting

One way receiving a green card may affect the patenting of immigrants is through opening up job prospects in industry. Without a green card, graduate students need a temporary worker visa (usually an H-1B) to remain in the U.S. While universities and other academic or research instutions have no limit to the number of H-1B visas they may apply for and receive, the same is not true for private-sector companies. The quota and application process for an H-1B thus may prevent foreign students from seeking and receiving employment in the private sector. With a green card, companies are more willing to hire foreign-born workers, and the workers themselves may be more willing to move to the private sector as well.

To explore this channel, I run my analysis separately for academic and non-academic patents. I classify a patent as "academic" if its assignee has "University" or "College" in its name. I classify the remaining patents with private-sector assignees as "non-academic". Due to the large number of zeros in the data, especially

for academic patents, I keep only county-field cells with at least one patent in each three-year bin between 1976 and 2007.[18]

I show difference-in-difference results from this analysis in Table 2.4: baseline results are presented in column 1, and results controlling flexibly for the pre-period share of other ethnic (Indian, Eastern European, and Jewish) inventors in each cell are shown in column 2. The event study is shown in Figure 2.8.

I find the increase in patenting after the CSPA is primarily driven by non-academic, private-sector patents: cells with a 1-percentage-point higher share of pre-period PRC inventors experienced a 2 percent increase in industry patenting post-1990. The coefficient on $Post * PercentPRC$ for academic patents, while negative, is not statistically significant. Thus, the CSPA increased innovation by private-sector firms, suggesting that green cards may open up additional career opportunities to foreign students in industry, without having a detrimental effect on innovation by universities.

2.6.2 Extensive vs. intensive margin

Finally, I examine the margins along which the CSPA may have contributed to U.S. patenting. By enabling green card holders from the PRC to stay in the U.S., patenting could have increased on the extensive margin – that is, by directly increasing the number of inventors in the U.S. Furthermore, knowledge spillovers between foreign- and native-born collaborators could result in increased patenting along the intensive margin as well, by making individual inventors more productive.

I first focus on the *extensive* margin: that is, did the CSPA increase the number of active inventors in the U.S.? To answer this question, I re-run the regressions in equations (2.2) and (2.3) using as the outcome variable the log number of unique

[18]Results are robust to using the inverse hyperbolic sine, rather than the natural log, of patents as the outcome variable.

inventors in each county-field cell. I present the difference-in-difference results in Table 2.5 and the event study in panel (a) of Figure 2.9.

I find that there is a positive and significant increase in the number of inventors after the CSPA. County-field cells with 1-percentage-point higher CSPA exposure experienced a 2 percent increase in the number of unique inventors. That is, the number of inventors in cells in the top quartile of CSPA exposure was 13 percent higher after 1990 relative to cells in the bottom quartile. The results are robust to the inclusion of flexible controls for the pre-period share of other ethnic inventors in each cell.

Furthermore, I investigate further whether the increase in patenting was driven by *new* inventors entering U.S. innovation. That is, I use as the outcome variable the number of unique inventors with no prior patent history. I find that there was indeed a 2 percent increase in patent applications county-field cells with a 1-percentage-point higher exposure to the CSPA. The timing of this increase can be seen in the event study in panel (b) of Figure 2.9 The effect is largest for the 1994-1996 and 1997-1999 year bins, consistent with the story that these years are when students who received green cards through the CSPA may have first entered U.S. innovation.

Finally, I consider whether the CSPA contributed to the *intensive* margin of patenting, i.e. whether each inventor became more productive after working with immigrant collaborators. Bernstein et al. (2022 WP) find that immigrant inventors contribute more to the productivity of their collaborators relative to native-born inventors, due in part to their sources of foreign knowledge (e.g., foreign markets, technologies, and collaborators). I explore this channel in the context of this paper by examining the effect of the CSPA on *patents per inventor*. In particular, I re-run my analysis using the natural logarithm of the average number of patents per inventor in each county-field cell as the outcome variable in equations (2.2) and (2.3).

Results from this exercise are presented in Table 2.5 and Appendix Figure 2.A.11. I do not find any significant effect of CSPA exposure on patents per inventor, suggesting the CSPA increased patenting primarily through the extensive margin of the entrance of new inventors into U.S. innovation. Taken together, these results suggest that extending green card access to highly-skilled foreign students enables them to find employment (especially in private sector jobs that may have been previously unavailable to them) and remain in the U.S., contributing positively to U.S. innovation.

2.7 Conclusion

The effects of extending green cards to high-skilled foreign nationals, especially graduate students, on the U.S. economy remains an important policy question today. Permanent residency may allow international graduate students to remain in the U.S. with few or no restrictions, relative to temporary work visas such as the H-1B, on their career options after completing their degree. As a result of these additional career opportunities, foreign-born students may be more likely to contribute to U.S. innovation, especially in non-academic, private-sector companies.

This paper uses the passage of the 1992 Chinese Student Protection Act to examine the causal effect of granting green card eligibility to foreign students on U.S. patenting. I find that patenting increased in county-field cells that were more exposed to the CSPA. Specifically, cells with a 1-percentage-point higher share of pre-period PRC inventors experience a 3 to 4 percent increase in patent applications. The increase in patenting seems driven by growth in private-sector patenting and by new inventors patenting for the first time in the U.S.

One caveat to the analysis conducted in this paper is that I examine the effect of a one-time, unanticipated shock in green card eligibility. The effect of a more permanent program to extend this eligibility to high-skilled immigrants would also

need to address factors that may change in the long run, such as the effect on native-born Americans.

Nevertheless, the findings in this paper may have important implications for immigration policy today. Future avenues of research may explore the effects of green cards and the CSPA on other measures of innovation, such as the publication of academic papers or entrepreneurship. Finally, as immigrants from mainland China are the fastest-growing immigrant group – and make up an increasingly large share of foreign-born population – in the U.S. today, it may be interesting to examine network effects (e.g., through increases in information, connections to U.S. universities or companies, or family reunification) that may stem from the initial wave of green cards granted by the CSPA.

2.8 Tables and figures

Table 2.1: NBER technology categories and subcategories

1 Chemical
11 Agriculture, food, textiles
12 Coating
13 Gas
14 Organic compounds
15 Resins
19 Misc. (chemical)
2 Computers & Communications
21 Communications
22 Computer hardware and software
23 Computer peripherals
24 Information storage
25 Electronic business methods and software
3 Drugs & Medical
31 Drugs
32 Surgery and medical instruments
33 Biotechnology
39 Misc. (drugs & medical)
4 Electrical & Electronics
41 Electrical devices
42 Electrical lighting
43 Measuring, testing
44 Nuclear, X-rays
45 Power systems
46 Semiconductor devices
49 Misc. (electrical)
5 Mechanical
51 Materials processing and handling
52 Metal working
53 Motors, engines, and parts
54 Optics
55 Transportation
59 Misc. (mechanical)
6 Others
61 Agriculture, husbandry, food
62 Amusement devices
63 Apparel and textile
64 Earth working and wells
65 Furniture, house fixtures
67 Pipes and joints
68 Receptacles
69 Misc. (others)

Notes: This table lists the one-digit technology categories and two-digit technology subcategories as defined for patent classification by the NBER in Hall, Jaffe, and Trajtenberg (2001). I define field of research by two-digit technology subcategory.

Table 2.2: Summary statistics by county-NBER subcategory cell

				Cells with at least 1 PRC inventor between 1984-88			
	All	No PRC inventors between 1984-88	At least 1 PRC inventor between 1984-88	Bottom quartile % PRC	2nd quartile % PRC	3rd quartile % PRC	Top quartile % PRC
Average patent applications per year	14.00	11.42	41.51	83.83	56.90	16.88	8.071
Private-sector patent applications per year	13.08	10.72	38.23	77.94	51.48	15.63	7.501
University patent applications per year	0.374	0.306	1.098	2.305	1.105	0.584	0.385
Average number of inventors per year	35.55	28.31	113.2	231.1	159.2	42.54	19.88
Total patent applications, 1984-1988	41.02	35.52	99.59	221.4	102.4	50.92	22.19
Total patent applications, 1990-2007	335.1	266.8	1063.3	2094.1	1556.8	401.2	193.3
Total inventors, 1984-1988	76.71	65.65	194.7	437.9	203.0	94.65	40.04
Total inventors, 1990-2007	855.9	672.5	2812.2	5651.5	4080.3	1012.0	482.7
% PRC inventors, 1984-1988	0.192	0	2.243	0.376	0.911	1.816	5.917
% East Asian inventors, 1984-1988	3.214	2.819	7.429	6.641	6.632	7.130	9.336
% Indian inventors, 1984-1988	2.395	2.328	3.118	3.544	3.071	3.427	2.415
% Jewish inventors, 1984-1988	9.492	9.570	8.661	8.916	8.794	9.165	7.754
% Eastern European inventors, 1984-1988	1.878	1.846	2.220	1.905	2.362	2.341	2.278
N	70912	64832	6080	1536	1504	1536	1504

Notes: This table presents summary statistics at the county-NBER technology subcategory cell level for the sample used in my main analysis. The sample includes cells in which there has been at least one patent application in all year bins.

Table 2.3: Effect of the CSPA on patent applications

Outcome: Log patent applications	(1)	(2)	(3)	(4)	(5)
Post-1990 x % PRC	0.036***	0.037***	0.037***	0.043***	0.044**
	(0.01)	(0.01)	(0.01)	(0.01)	(0.02)
Constant	2.684***	2.684***	2.581***	2.616***	2.685***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
N	70912	70912	69920	70752	70752
N (weighted)	2908480	2908480	2754080	2825120	2905216
Controls for % pre-period ethnic inv. x year		✓			
Drop Santa Clara			✓		
Drop cells with largest 5 pre-period patent counts				✓	
Drop cells with highest 5 pre-period PRC inventor shares					✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % PRC inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2). Each column represents a separate regression. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Table 2.4: Patent applications for non-academic vs. academic assignees

Outcome:	Log non-academic applications		Log academic applications	
Post x % PRC	0.022**	0.023**	-0.031	-0.031
	(0.01)	(0.01)	(0.04)	(0.03)
Constant	2.689***	2.689***	0.735***	0.735***
	(0.00)	(0.00)	(0.01)	(0.01)
N	67088	67088	9941	9941
N	2854922	2854922	723850	723850
Controls for % pre-period ethnic inv. x year		✓		✓

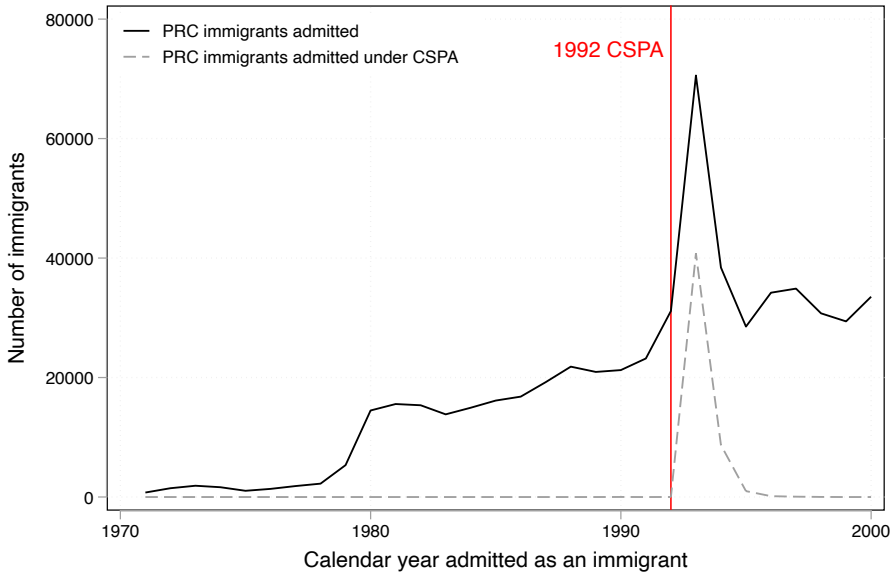
Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % PRC inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2). Each column represents a separate regression. The outcome is log patent applications with non-academic private-sector assignees in the first two columns and log patent applications with academic assignees in the second two columns. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Table 2.5: Results for extensive vs. intensive margin

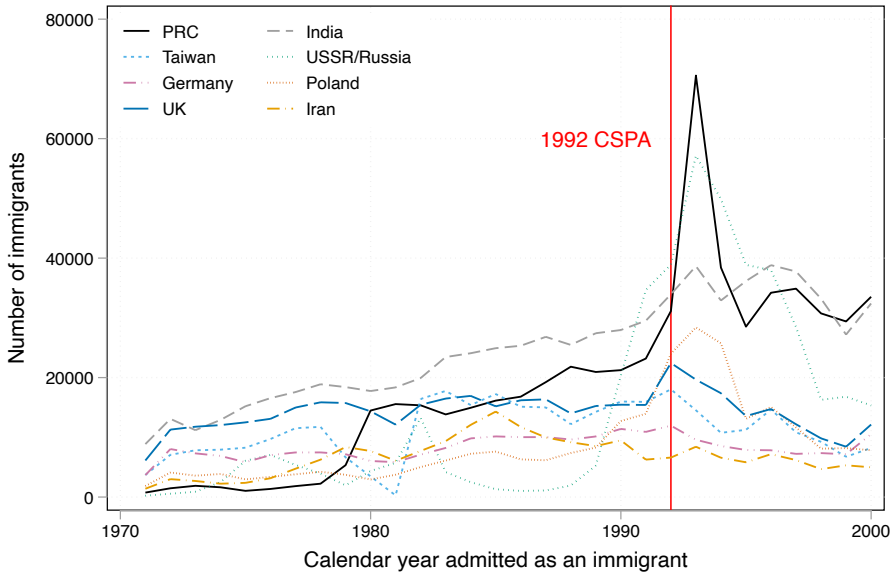
Outcome:	Log inventors		Log new inventors		Patents per inventor	
Post-1990 x % PRC	0.023**	0.024***	0.023**	0.023**	0.002	0.002
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Constant	3.283***	3.283***	2.212***	2.212***	0.238***	0.238***
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
N	68368	68368	62372	62372	68368	68368
N (weighted)	2874053	2874053	2767221	2767221	2874053	2874053
Controls for % pre-period ethnic inv. x year		✓		✓		✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % PRC inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2). Each column represents a separate regression. The outcome is log unique inventors in the first two columns, log unique new inventors in the middle two columns, and patents per inventor in the last two columns. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.1: Number of immigrants admitted to the U.S.



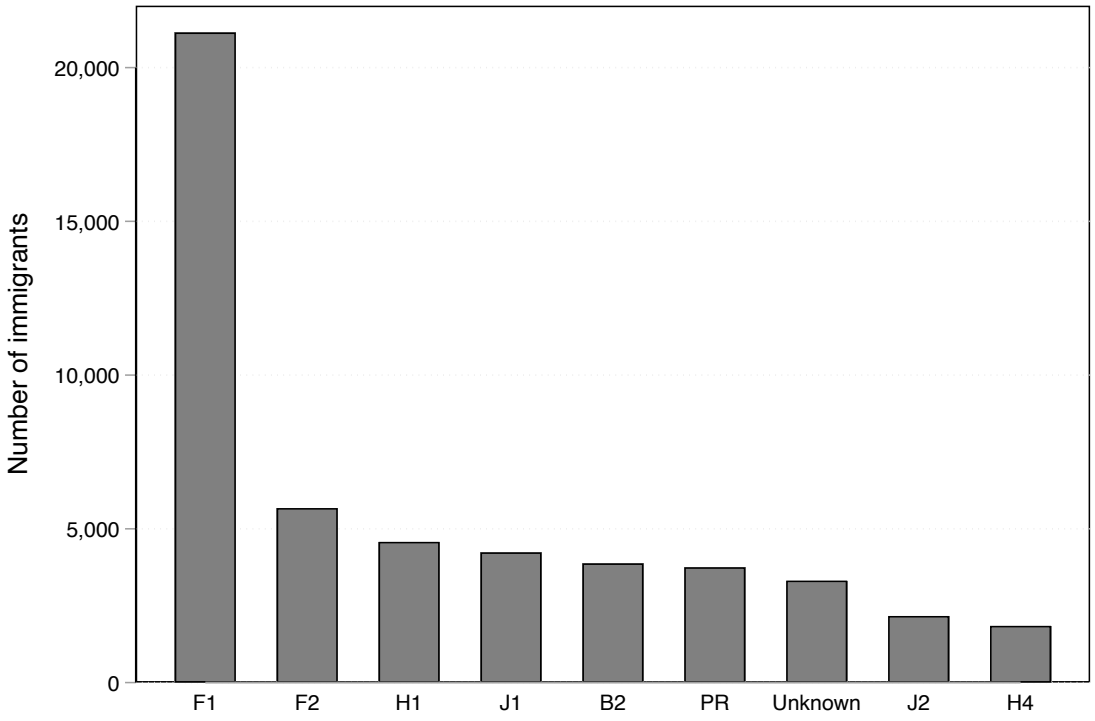
(a) Number of PRC nationals admitted



(b) Number of immigrants admitted

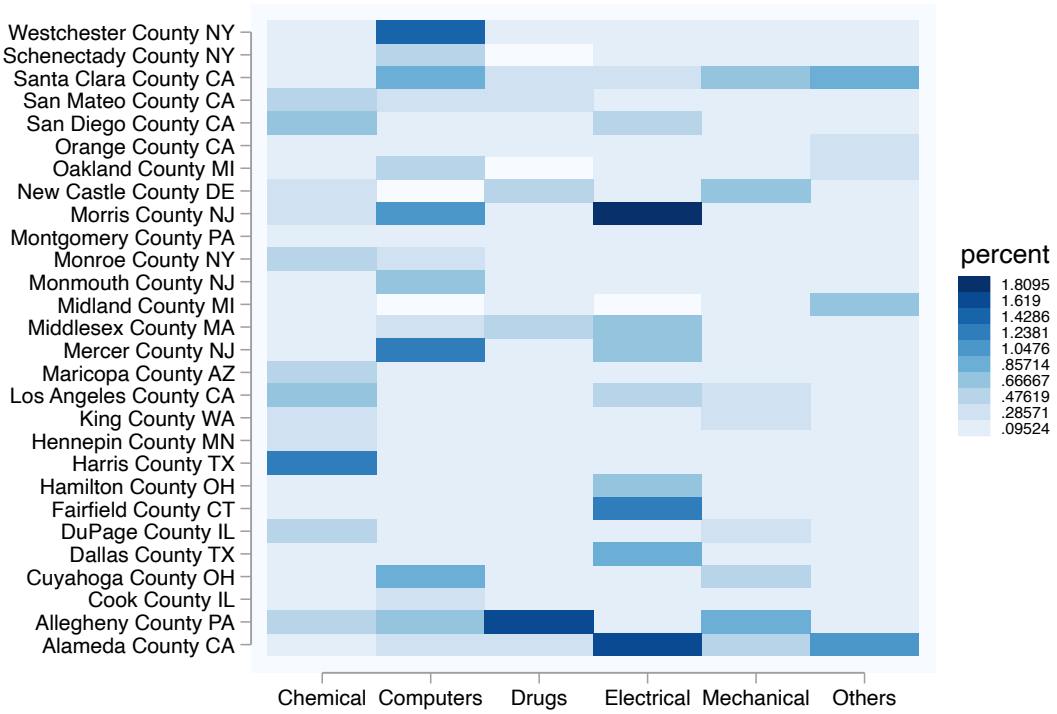
Source: Immigration and Naturalization Service, Immigrants Admitted to the United States 1972-2000. *Notes:* This figure presents the number of immigrants admitted to the U.S. each year from the PRC and other sending countries. The red line denotes the passage of the CSPA in 1992.

Figure 2.2: CSPA recipients by class of non-immigrant entry



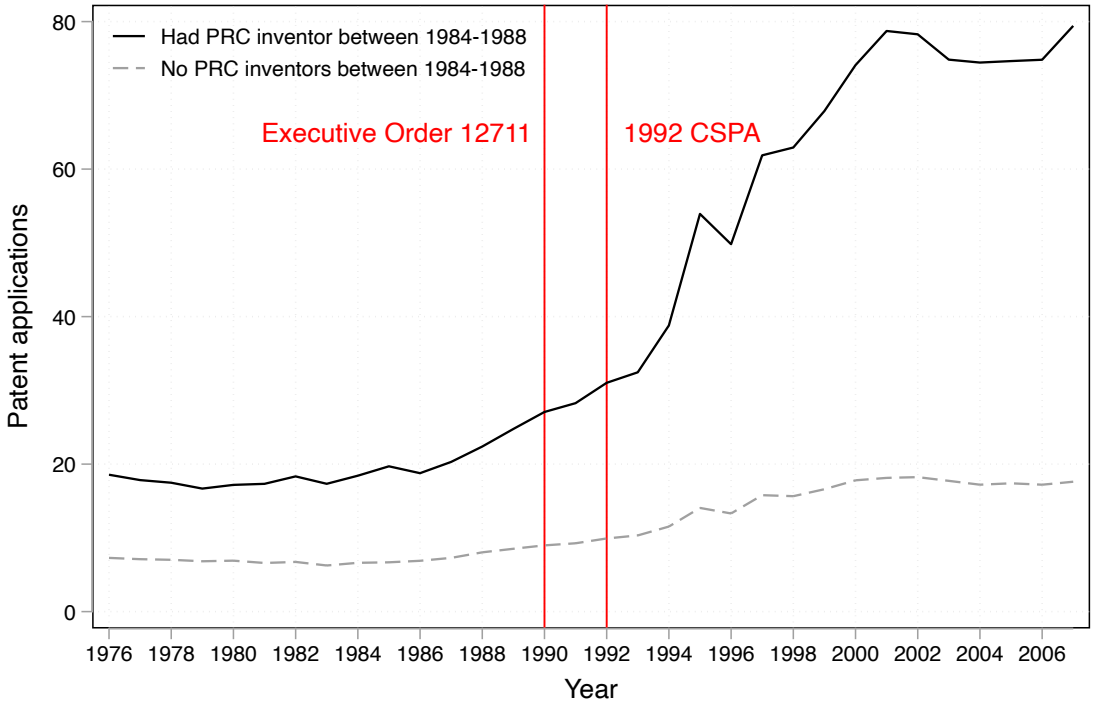
Source: Immigration and Naturalization Service, Immigrants Admitted to the United States 1972-2000. *Notes:* This figure presents the number of immigrants admitted to the U.S. under the CSPA who had last entered the U.S. in each class of non-immigrant entry.

Figure 2.3: Variation in pre-period share of PRC inventors by county and NBER category



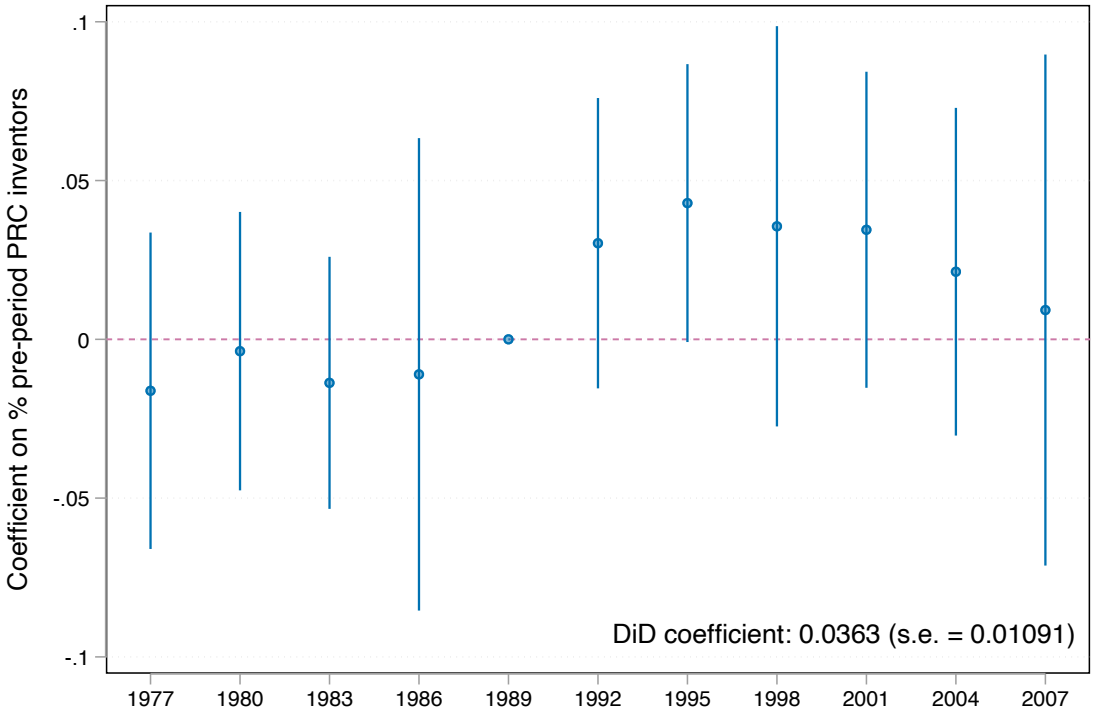
Notes: This figure shows variation in pre-period shares of PRC inventors across county-NBER technology category cells. Higher shares of PRC inventors are represented by darker shades in each cell.

Figure 2.4: Raw means of patent applications per cell



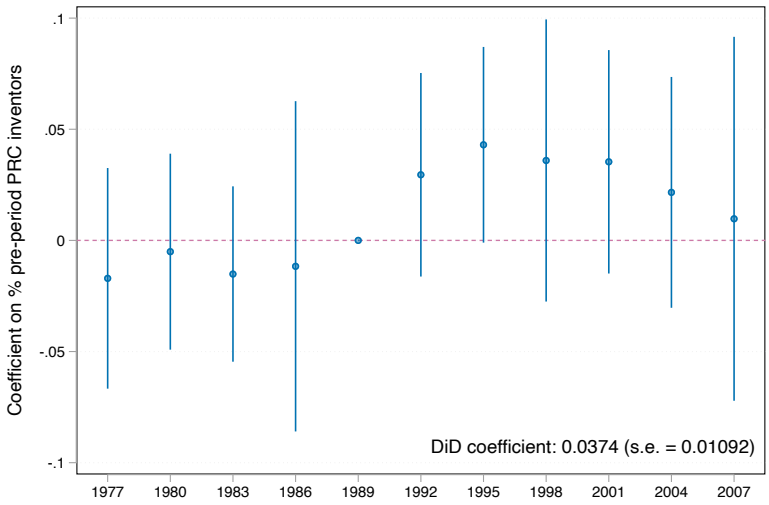
Notes: This figure plots the average number of patent applications per county-NBER technology subcategory cell, for cells with and without at least one PRC inventor between 1984 and 1988. The first red line marks 1990, the year Executive Order 12711 was signed. The second red line marks 1992, the year the CSPA was passed.

Figure 2.5: Effect of CSPA on log patent applications

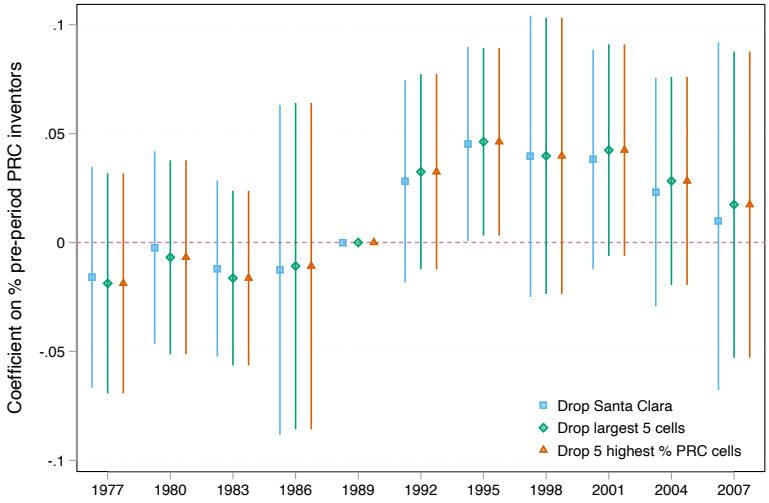


Notes: This figure presents coefficients on Year x % PRC from equation (2.3), where % PRC is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality and Year represents one of 11 three-year bins between 1976 and 2007. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome variable is the log number of patent applications in each cell. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.6: Robustness checks



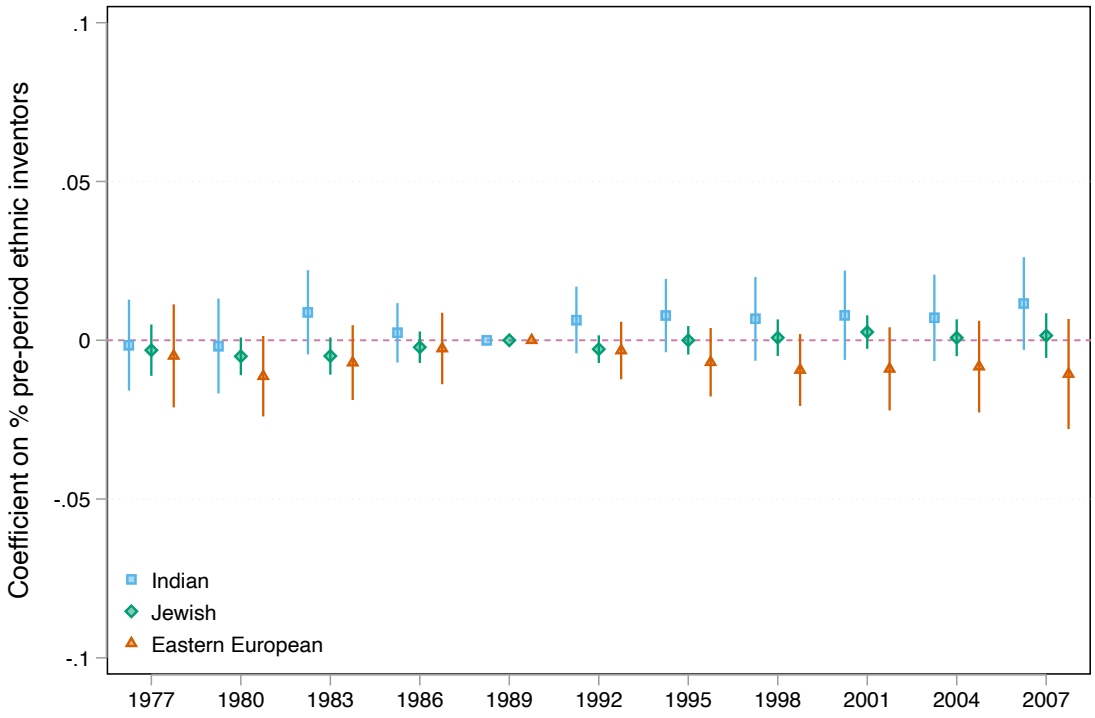
(a) Main results with controls for ethnicity x year



(b) Main results dropping different cells

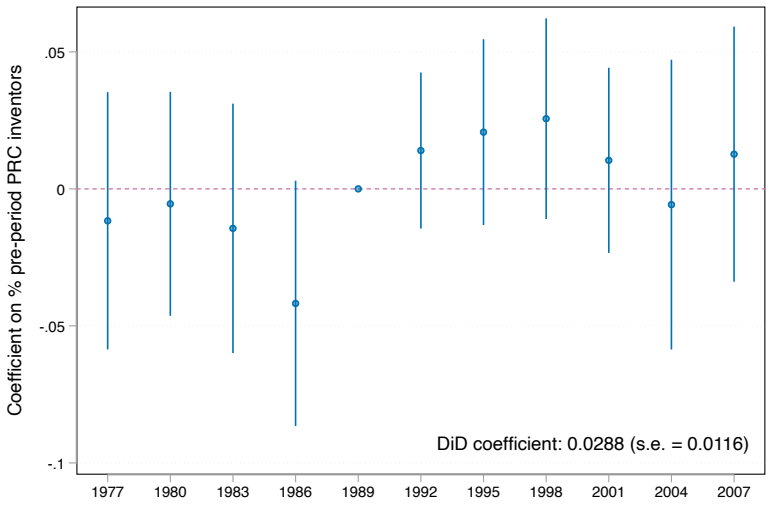
Notes: This figure presents coefficients on $PercentPRC * Year$ from equation (2.3), where $PercentPRC$ is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality and $Year$ represents one of 11 three-year bins between 1976 and 2007. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome variable is the log number of patent applications in each cell. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. The top panel further controls for $PercentIndian * Year$, $PercentJewish * Year$, and $PercentEasternEuropean * Year$. The bottom panel drops the listed cells from the sample. Standard errors are clustered by county.

Figure 2.7: Effect of other immigrant groups on patent applications

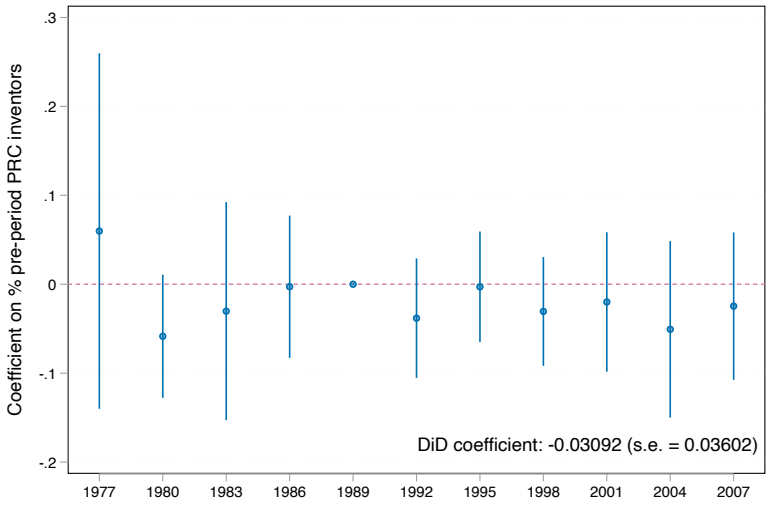


Notes: This figure presents coefficients on Year x % Indian, Jewish, and Eastern European from equation (2.3), where Year represents one of 11 three-year bins between 1976 and 2007 and % Indian, Jewish, and Eastern European is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with Indian, Jewish, or Eastern European nationality. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome variable is the log number of patent applications in each cell. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988.

Figure 2.8: Effects of CSPA on non-academic vs. academic patenting



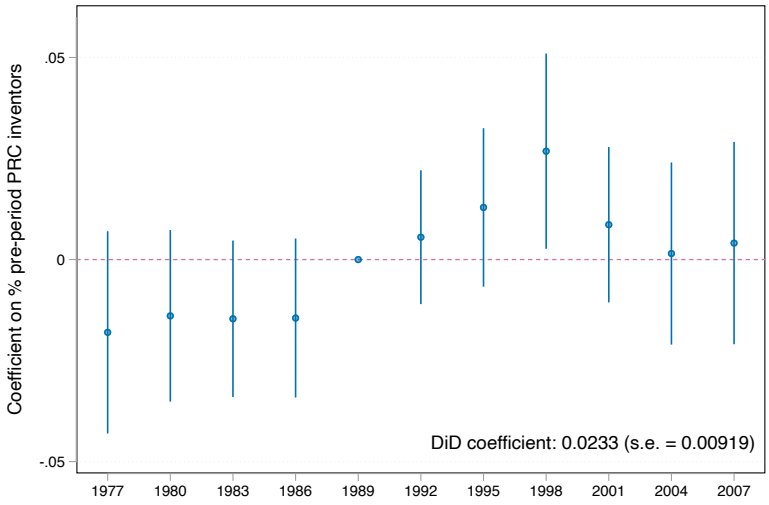
(a) Patent applications with non-academic private-sector assignees



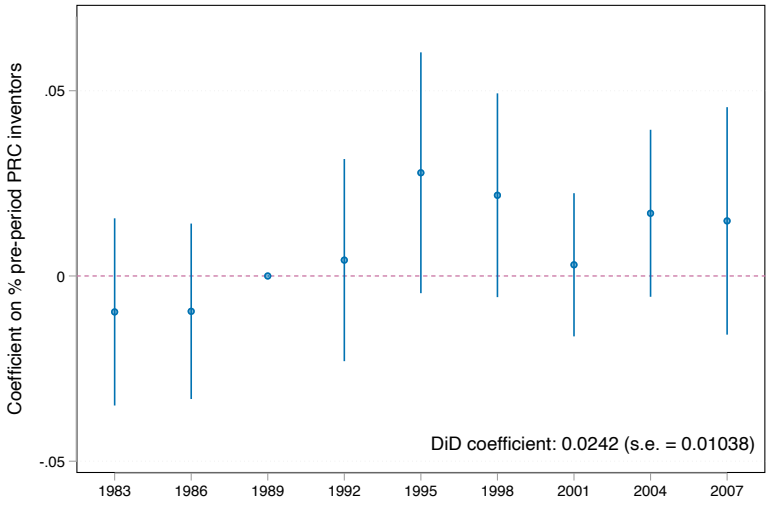
(b) Patent applications with academic assignees

Notes: This figure presents coefficients on $PercentPRC * Year$ from equation (2.3), where $PercentPRC$ is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality and $Year$ represents one of 11 three-year bins between 1976 and 2007. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome is log patent applications with non-academic private-sector assignees in the top panel and log patent applications with academic assignees in the bottom panel. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.9: Effect of CSPA on inventors



(a) Effect on log number of inventors



(b) Effect on log number of *new* inventors

Notes: This figure presents coefficients on $PercentPRC * Year$ from equation (2.3), where $PercentPRC$ is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality and $Year$ represents one of 11 three-year bins between 1976 and 2007. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome is log unique inventors in the top panel and log unique *new* inventors in the bottom panel. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Appendix

2.A Appendix tables and figures

Table 2.A.1: Top 15 most frequent non-immigrant visa classes of CSPA LPR recipients

Visa class	Freq.	Percent	Cum.
F1	21,141	40.24	40.24
F2	5,675	10.8	51.05
H1	4,576	8.71	59.76
J1	4,233	8.06	67.82
B2	3,875	7.38	75.19
Parolee	3,748	7.13	82.33
Unknown	3,316	6.31	88.64
J2	2,166	4.12	92.76
H4	1,845	3.51	96.28
B1	904	1.72	98.00
Without inspection	296	0.56	98.56
C1	199	0.38	98.94
L1	153	0.29	99.23
G4	151	0.29	99.52
L2	67	0.13	99.65

Source: Immigration and Naturalization Service, Immigrants Admitted to the United States 1972-2000. *Notes:* This table presents summary statistics for the number of immigrants admitted to the U.S. under the CSPA who had last entered the U.S. in each class of non-immigrant entry.

Table 2.A.2: Top 15 most frequent classes of admission of PRC green card recipients between 1972-2000

Class of immigrant admission	Freq.	Percent	Cum.
Alien covered by the Chinese Student Protection Act	50,699	10.32	10.32
Parent of U.S. citizen	42,508	8.66	18.98
Unmarried son or daughter of lawful permanent resident alien	28,647	5.83	24.81
Child of alien classified P51 or P56	23,219	4.73	29.54
Spouse of U.S. citizen	17,744	3.61	33.15
Brother or sister of U.S. citizen (age 21 or over)	16,612	3.38	36.54
Parent of U.S. citizen	14,474	2.95	39.48
Spouse of alien classified P51 or P56	14,126	2.88	42.36
Spouse of lawful permanent resident	13,661	2.78	45.14
Member of profession holding advanced degree or exceptional ability	12,824	2.61	47.75
Orphan adopted abroad by U.S. citizen	12,567	2.56	50.31
Child of alien classified P41 or P46	12,484	2.54	52.85
Spouse of alien classified E21 or E26	10,185	2.07	54.93
Spouse of U.S. citizen	9,719	1.98	56.91
Child of skilled worker or professional classified as E31, E36, E32, or E37	9,409	1.92	58.82

Source: Immigration and Naturalization Service, Immigrants Admitted to the United States 1972-2000. *Notes:* This table presents summary statistics for the class of immigrant admission for all immigrants admitted to the U.S. from the PRC (i.e., with former PRC nationality) between 1972 and 2000.

Table 2.A.3: Most common surnames on patent applications between 1984-1988 by ethnicity

PRC	N	Indian	N	Other East Asian	N	Eastern European	N	Jewish	N
Wang	1663	Patel	1675	Chen	4135	Novak	109	Cohen	1262
Zhang	984	Shah	1042	Lee	3576	Horodysky	103	Klein	971
Li	904	Singh	972	Chang	3002	Prevorsek	95	Hoffman	876
Liu	697	Gupta	943	Wang	2226	Markovich	91	Schwartz	750
Yang	643	Kumar	901	Wong	2216	Khandros	86	Meyer	737
Xu	565	Rao	808	Lin	2063	Metlitsky	84	Rosen	705
Wu	534	Desai	469	Kim	1621	Korsunsky	82	Kaufman	609
Chen	516	Sharma	457	Wu	1572	Haluska	72	Katz	601
Sun	451	Reddy	452	Liu	1444	Dvorkis	71	Goldstein	599
Zhu	403	Jain	438	Huang	1426	Uskokovic	69	Fischer	599
Zhou	395	Mehta	433	Chan	1395	Rubin	67	Stern	538
Yu	381	Agrawal	362	Hsu	1220	Sachdev	65	Goldberg	533
Jiang	381	Das	355	Chu	1140	Hynecek	61	Weiss	531
Zhao	365	Chatterjee	349	Li	1065	Klemarczyk	60	Wolf	504
Lu	316	Srinivasan	325	Cheng	932	Dvorak	58	Kaplan	492

Notes: This table presents the most common surnames from inventors listed on patent applications between 1984 and 1988 in each ethnicity. The methodology for ethnicity classification is described briefly in Section 2.3.3 and in more detail in Appendix 2.B.

Table 2.A.4: Distribution of inventors across NBER categories by ethnicity

	All inventors			PRC			Indian			Other East Asian			Eastern European			Jewish		
	All	1984-1988	Post-1990	All	1984-1988	Post-1990	All	1984-1988	Post-1990	All	1984-1988	Post-1990	All	1984-1988	Post-1990	All	1984-1988	Post-1990
Chemical	20.90	24.74	18.75	24.55	30.17	24.13	24.18	32.04	21.61	26.61	37.08	22.52	22.56	27.41	20.10	20.41	23.69	18.30
Computers	18.75	11.92	21.97	22.08	17.79	22.34	26.42	12.45	30.30	23.15	13.20	26.67	17.66	10.43	20.56	18.21	11.82	21.64
Drugs/Medical	13.61	10.41	15.63	19.49	7.35	20.14	14.31	14.12	14.92	13.69	12.27	14.65	14.29	11.53	15.97	16.08	12.35	18.53
Electrical	17.63	18.52	17.20	22.03	22.44	22.01	18.21	17.50	18.37	21.75	18.55	22.78	19.11	20.44	18.73	16.77	18.47	15.81
Mechanical	14.22	16.64	12.97	6.20	11.99	5.93	8.76	12.11	7.69	7.01	8.73	6.39	14.25	15.36	13.58	13.83	16.23	12.55
Others	14.89	17.78	13.48	5.65	10.25	5.45	8.12	11.77	7.10	7.80	10.16	6.99	12.13	14.82	11.06	14.71	17.44	13.17
Total	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00	100.00

Notes: This table presents the distribution of ethnicities for inventors listed on patent applications across NBER technology categories in the pre-period (1984 to 1988) and post-period (1990 to 2007).

Table 2.A.5: PRC inventors and patent applications by NBER subcategory, 1984-1988

NBER subcategory	% PRC	Patent applications	NBER subcategory	% PRC	Patent applications
Apparel and textile	0.000	16.4	Power systems	0.163	33.4
Gas	0.000	19.1	Misc. (drugs & med)	0.167	20.5
Agriculture, husbandry, food	0.000	29.5	Misc. (chem)	0.169	63.6
Agriculture, food, textiles	0.000	55.6	Metal working	0.172	32.5
Amusement devices	0.000	29.6	Resins	0.175	60.7
Electronic business methods and software	0.000	37.0	Measuring, testing	0.177	33.9
Pipes and joints	0.000	27.1	Electrical devices	0.209	36.7
Transportation	0.010	28.8	Information storage	0.218	45.2
Motors, engines, and parts	0.026	32.6	Communications	0.256	47.4
Optics	0.028	28.0	Misc. (elec)	0.296	36.6
Receptacles	0.063	24.3	Organic compounds	0.305	51.3
Misc. (mechanical)	0.074	30.6	Heating	0.347	23.4
Nuclear, X-rays	0.090	36.8	Computer hardware and software	0.371	43.1
Drugs	0.099	72.6	Coating	0.379	22.4
Furniture, house fixtures	0.113	20.7	Semiconductor devices	0.391	71.9
Misc. (others)	0.138	41.6	Computer peripherals	0.504	30.6
Surgery, medical instruments	0.138	41.8	Earth working and wells	0.579	70.0
Materials processing and handling	0.154	30.3	Electrical lighting	1.060	29.6

Notes: This table presents the share of PRC inventors and the total number of patent applications in each NBER subcategory in my sample between 1984 and 1988.

Table 2.A.6: PRC inventors and patent applications by county, 1984-1988

County	% PRC	Patent applications	County	% PRC	Patent applications
Gloucester County NJ	0.000	121.0	San Mateo County CA	0.146	49.5
Cumberland County PA	0.000	96.0	Cuyahoga County OH	0.167	54.7
Rockland County NY	0.000	53.4	Somerset County NJ	0.169	48.8
Stephens County OK	0.000	120.0	Los Angeles County CA	0.177	101.0
Sullivan County TN	0.000	53.8	San Diego County CA	0.186	51.8
Kanawha County WV	0.000	49.7	Dallas County TX	0.230	61.8
New London County CT	0.000	51.7	Middlesex County MA	0.241	70.0
Winnebago County IL	0.000	48.3	New Castle County DE	0.271	105.8
Dauphin County PA	0.000	67.0	East Baton Rouge Parish LA	0.311	116.5
Bradford County PA	0.000	87.0	DuPage County IL	0.398	65.5
Maricopa County AZ	0.022	57.0	Allegheny County PA	0.494	68.7
Cook County IL	0.066	118.8	Brazoria County TX	0.522	114.8
Orange County CA	0.087	78.9	Santa Clara County CA	0.531	155.6
Monroe County NY	0.103	96.9	Morris County NJ	0.640	48.7
Oakland County MI	0.111	80.2	Harris County TX	0.675	103.8
Midland County MI	0.111	114.9	Mercer County NJ	0.688	56.0
King County WA	0.131	59.0	Camden County NJ	1.190	65.7
Hennepin County MN	0.145	48.0	Washington County OK	4.395	167.0

Notes: This table presents the share of PRC inventors and the total number of patent applications in the 36 largest counties in my sample between 1984 and 1988.

Table 2.A.7: Results using share of other pre-period ethnic inventors as exposure measure

Outcome: Log patent applications	(1)	(2)	(3)
Post x % Indian	0.005		
	(0.01)		
Post x % Jewish		0.004	
		(0.00)	
Post x % Eastern European			-0.001
			(0.00)
Constant	2.680***	2.670***	2.690***
	(0.01)	(0.01)	(0.01)
N	70912	70912	70912
N	2908480	2908480	2908480

Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % Indian, Jewish, and Eastern European inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2). Each column represents a separate regression. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Table 2.A.8: Robustness to alternative definitions of outcome variable

Outcome:	ln(patent applications + 0.1)	ln(patent applications + 0.01)	arcsinh(patent applications)
Post x % PRC	0.031***	0.041***	0.027***
	(0.01)	(0.01)	(0.01)
Constant	2.723***	2.655***	3.435***
	(0.00)	(0.00)	(0.00)
N	70912	70912	70912
N (weighted)	2908480	2908480	2908480

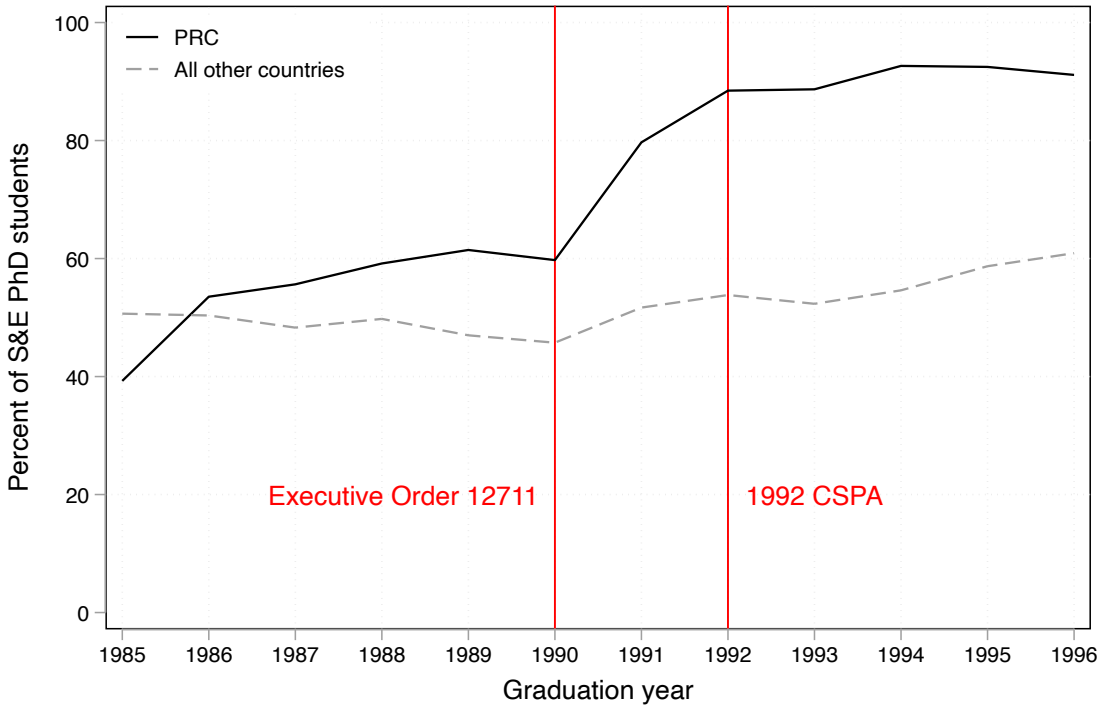
Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % PRC inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2) for alternative definitions of the outcome variable. Each column represents a separate regression. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Table 2.A.9: Results for placebo CSPAs in pre-period

Outcome: Log patent applications	(1)	(2)	(3)
Post:	1979	1982	1985
Post x % PRC	0.010	0.002	0.007
	(0.02)	(0.01)	(0.01)
Constant	2.377***	2.378***	2.378***
	(0.00)	(0.00)	(0.00)
N	31024	31024	31024
N (weighted)	1272460	1272460	1272460

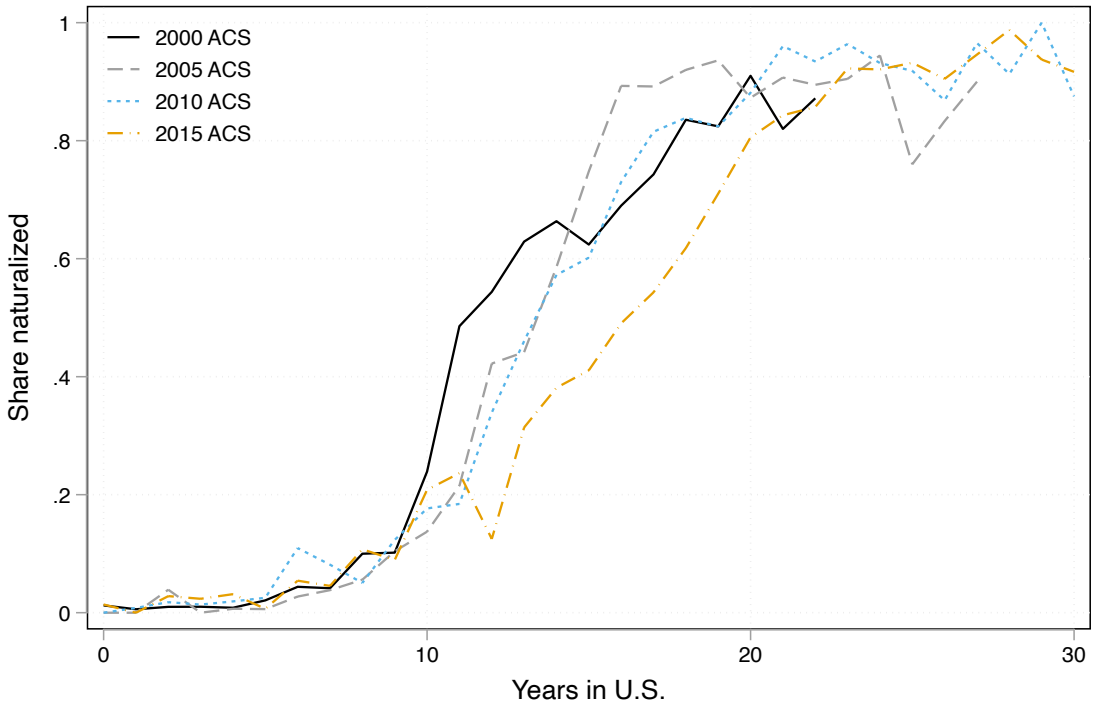
Notes: *** p<0.01, ** p<0.05, * p<0.1. Table reports the coefficient on Post x % PRC inventors in each county-NBER subcategory cell between 1984-1988 from equation (2.2) for different placebo CSPAs using only data from before 1990. Each column represents a separate regression. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.A.1: Share of foreign S&E PhD students with plans to stay in U.S.



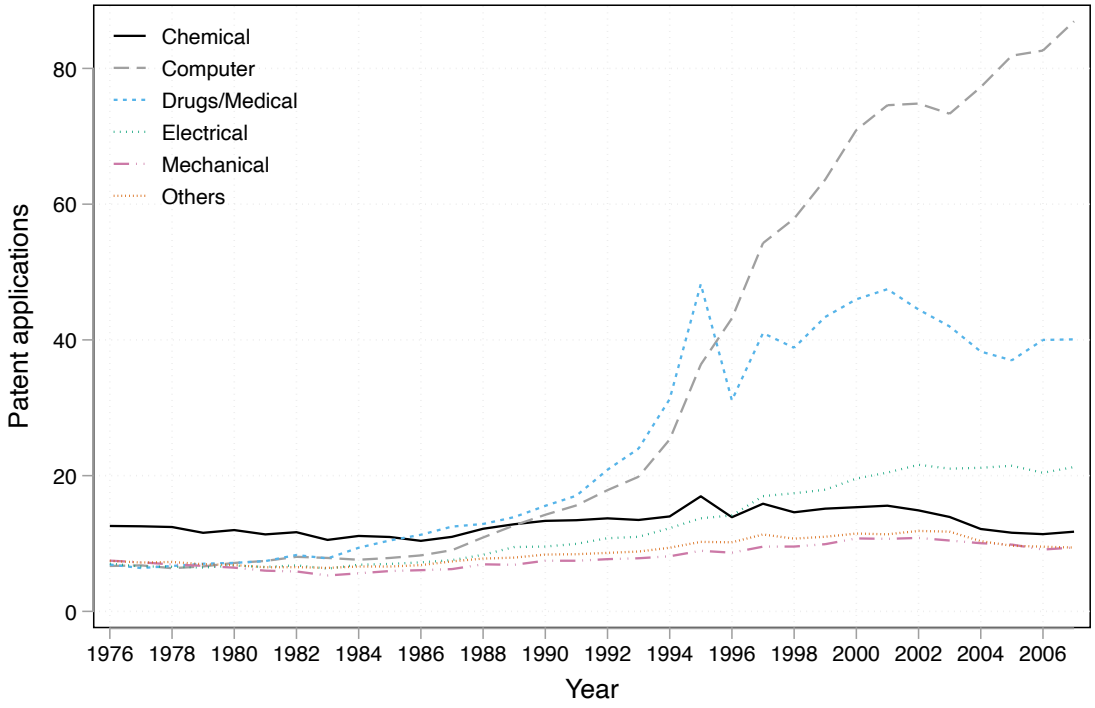
Source: NSF Survey of Earned Doctorates, Johnson (1998). *Notes:* This figure plots the share of foreign-born PhD students attending U.S. universities in science and engineering fields with plans to stay in the U.S., separately for PRC and non-PRC students. Graduation year is represented on the x axis. The first red line marks 1990, the year Executive Order 12711 was signed. The second red line marks 1992, the year the CSPA was passed.

Figure 2.A.2: Share of PRC-born individuals with > 16 years of education in ACS who are naturalized citizens



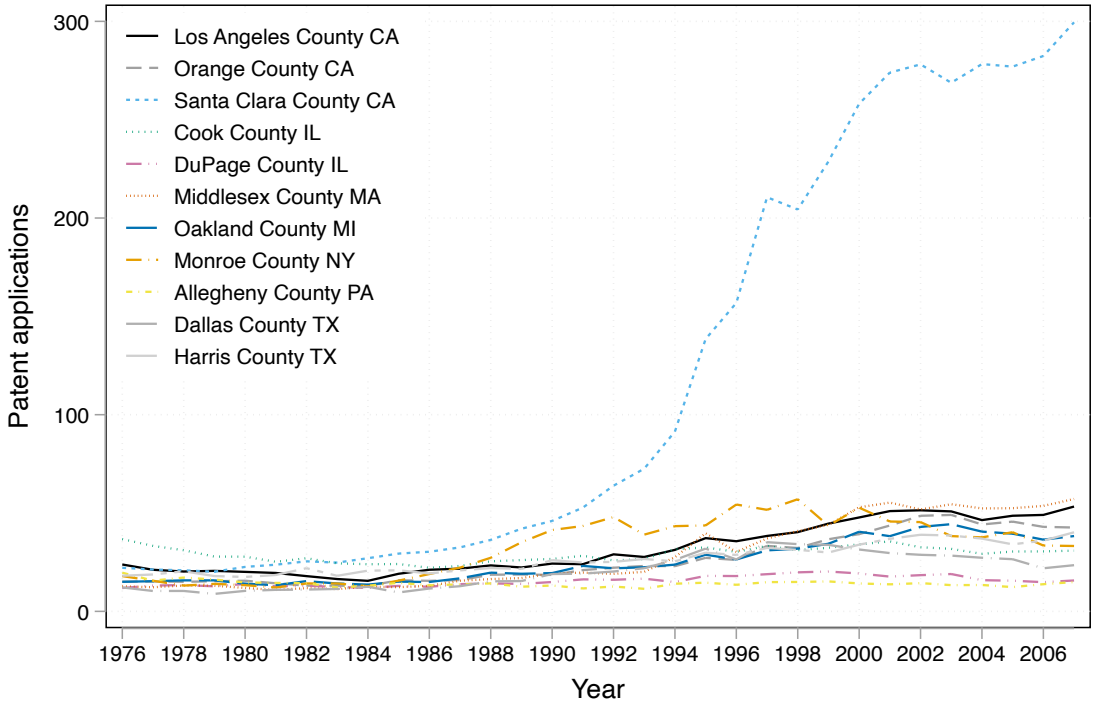
Source: ACS (2000, 2005, 2010, 2015). *Notes:* This figure shows the share of PRC-born individuals with more than 16 years of education in the ACS who are naturalized citizens, by years in the U.S. Each line denotes a separate ACS survey.

Figure 2.A.3: Raw means of patent applications per cell over time, by NBER category



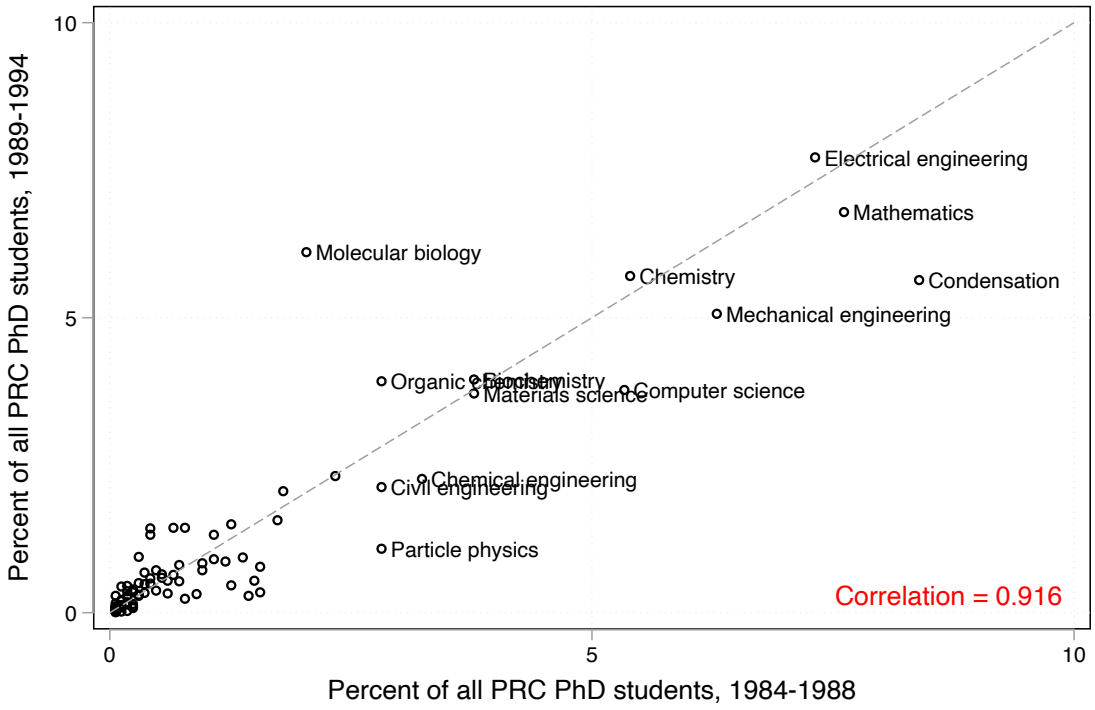
Notes: This figure plots the average number of patent applications per cell in each NBER technology category between 1976 and 2007.

Figure 2.A.4: Raw means of patent applications per cell over time, by county



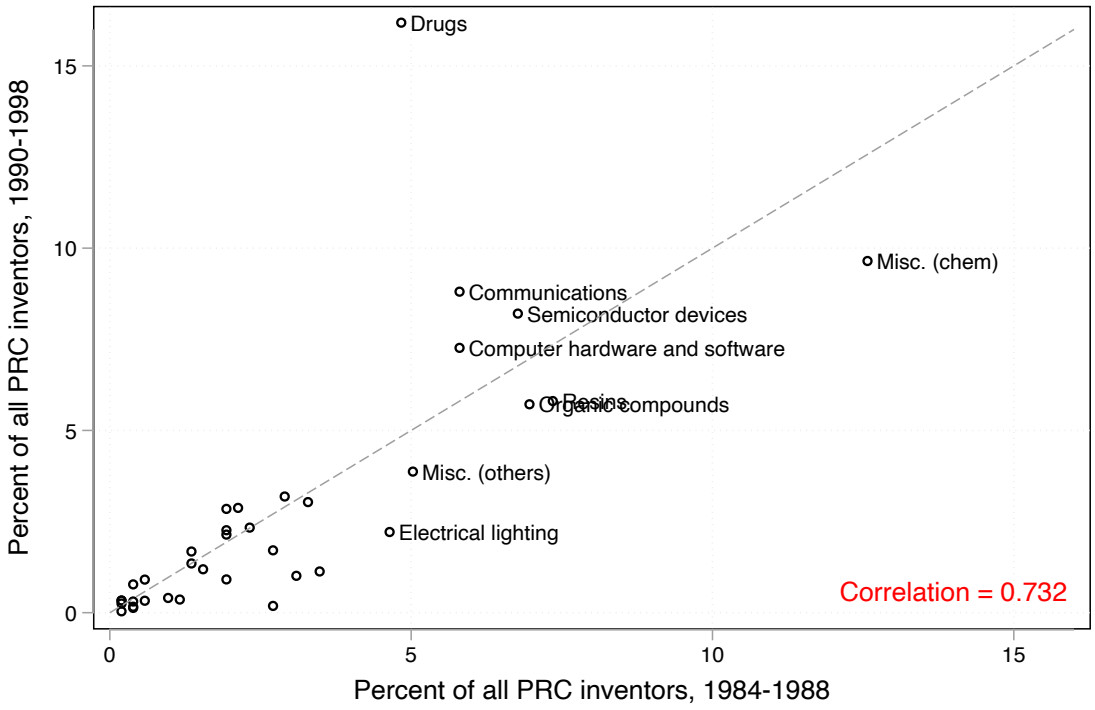
Notes: This figure plots the average number of patent applications per cell in each listed county between 1976 and 2007.

Figure 2.A.5: Correlation between research fields of pre-period and CSPA PRC PhD students



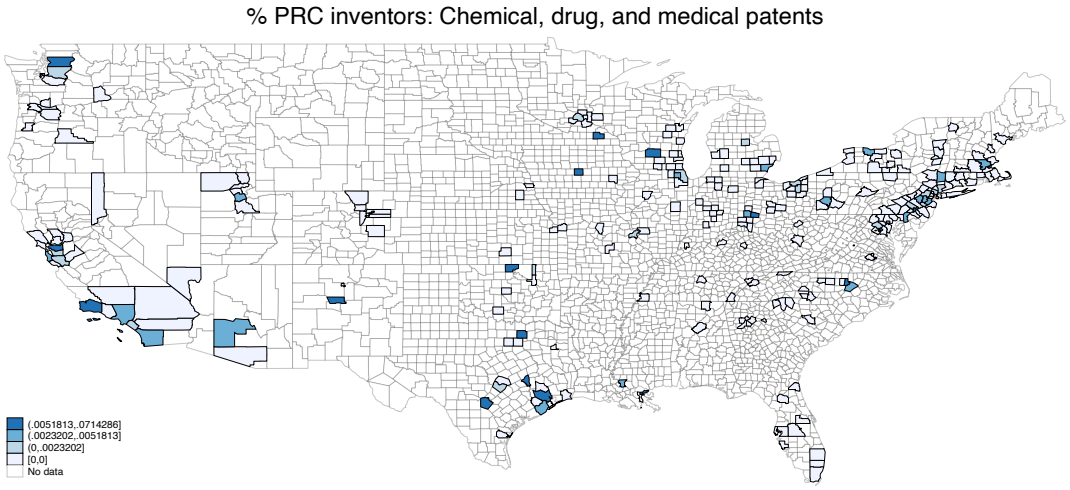
Notes: This figure plots the correlation between the share of all PRC PhD students receiving a degree in each field, for students graduating between 1984-1988 (on the x-axis) and students graduating between 1989-1994 (on the y-axis). The 45-degree line is denoted by dashes.

Figure 2.A.6: Correlation between NBER technology subcategories of pre- and post-CSPA PRC inventors

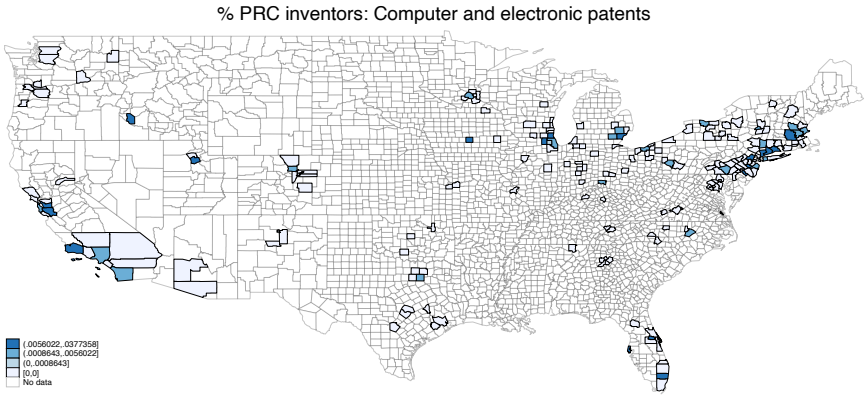


Notes: This figure plots the correlation between the share of all PRC inventors with patents in each field between 1984-1988 (on the x-axis) and the share of all PRC inventors with patents in each field between between 1990-1998 (on the y-axis). The 45-degree line is denoted by dashes.

Figure 2.A.7: Geographic variation in pre-period PRC inventor share



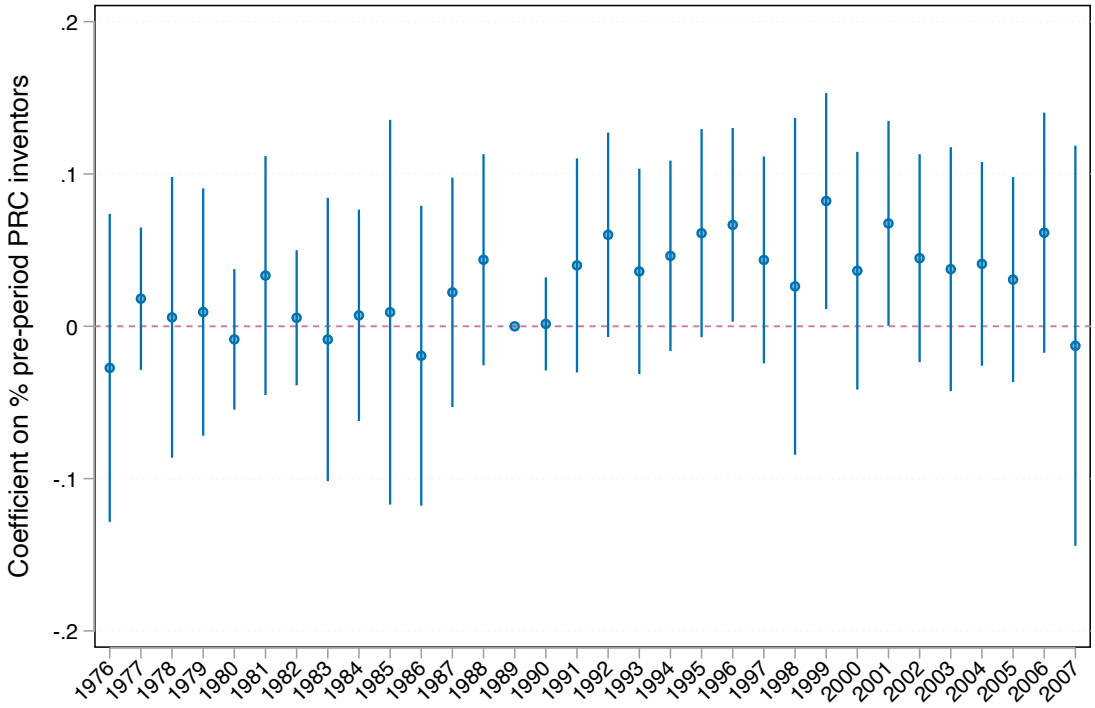
(a) Share of pre-period PRC inventors by county, Chemical/Drug patents



(b) Share of pre-period PRC inventors by county, Electrical/Computer patents

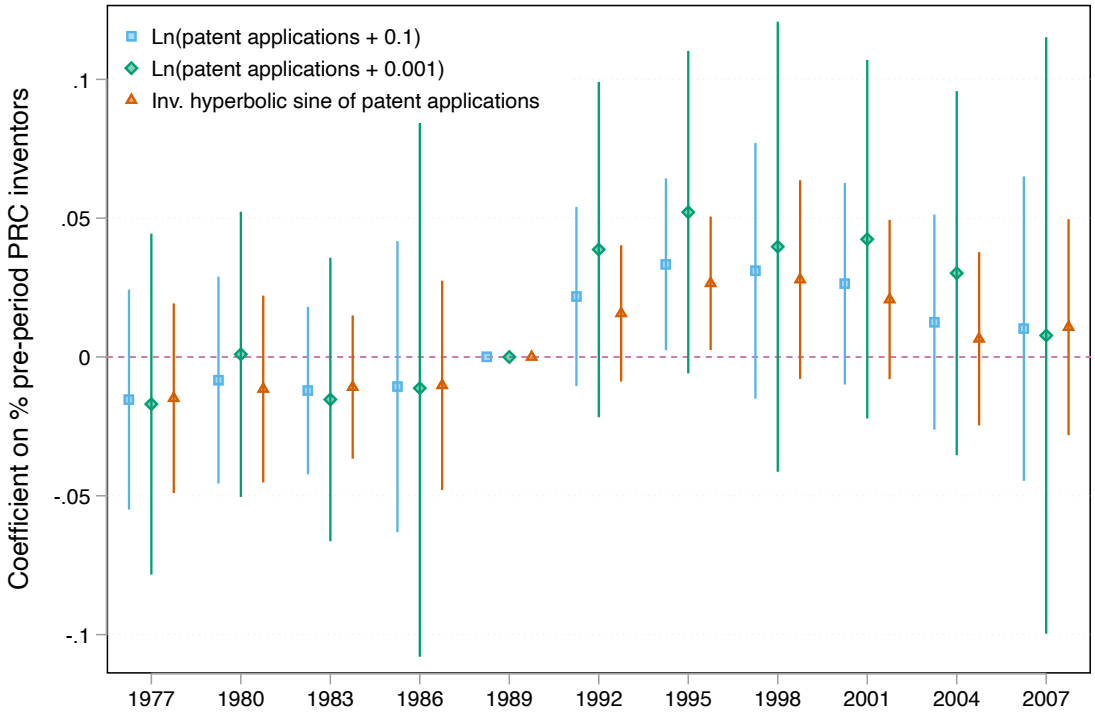
Notes: This figure shows geographic variation in pre-period shares of PRC inventors across county-NBER technology category cells. Higher shares of PRC inventors are represented by darker shades in each cell. The first panel shows the share of inventors listed on Chemical and Drug patent applications with PRC nationality. The second panel shows the share of inventors listed on Electrical and Computer patent applications with PRC nationality. Counties outlined in gray are not included in my sample.

Figure 2.A.8: Main results by year



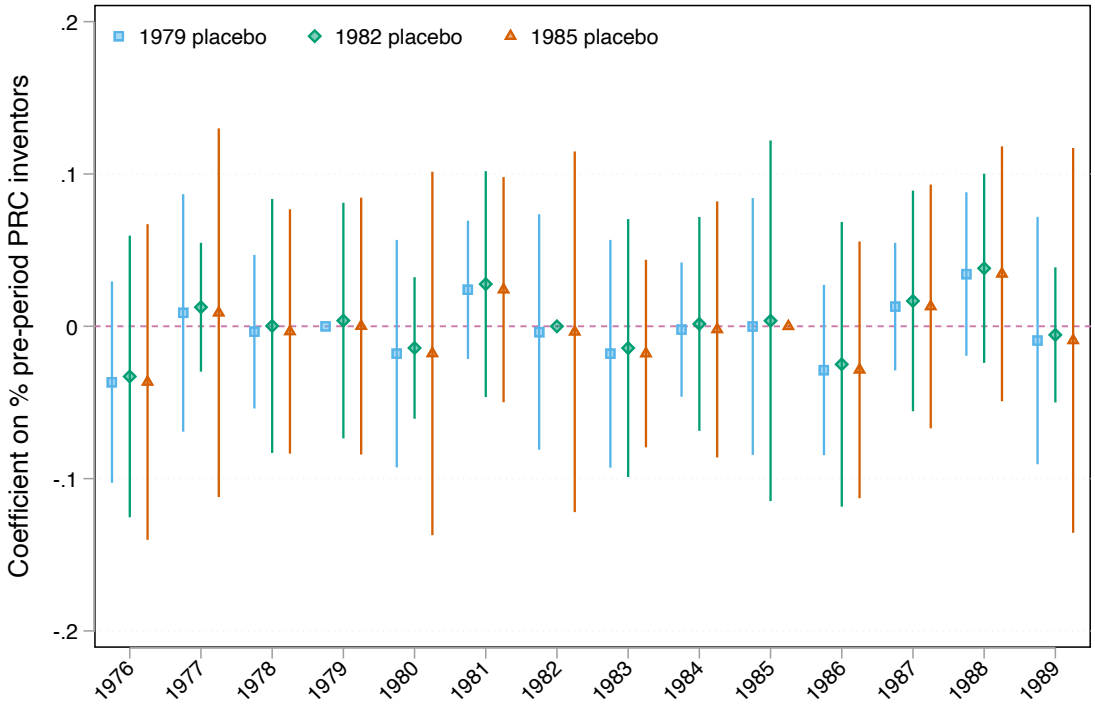
Notes: This figure presents coefficients on Year x % PRC from equation (2.3), where % PRC is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all years. The outcome variable is the log number of patent applications in each cell. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.A.9: Robustness to alternative treatments of outcome variable



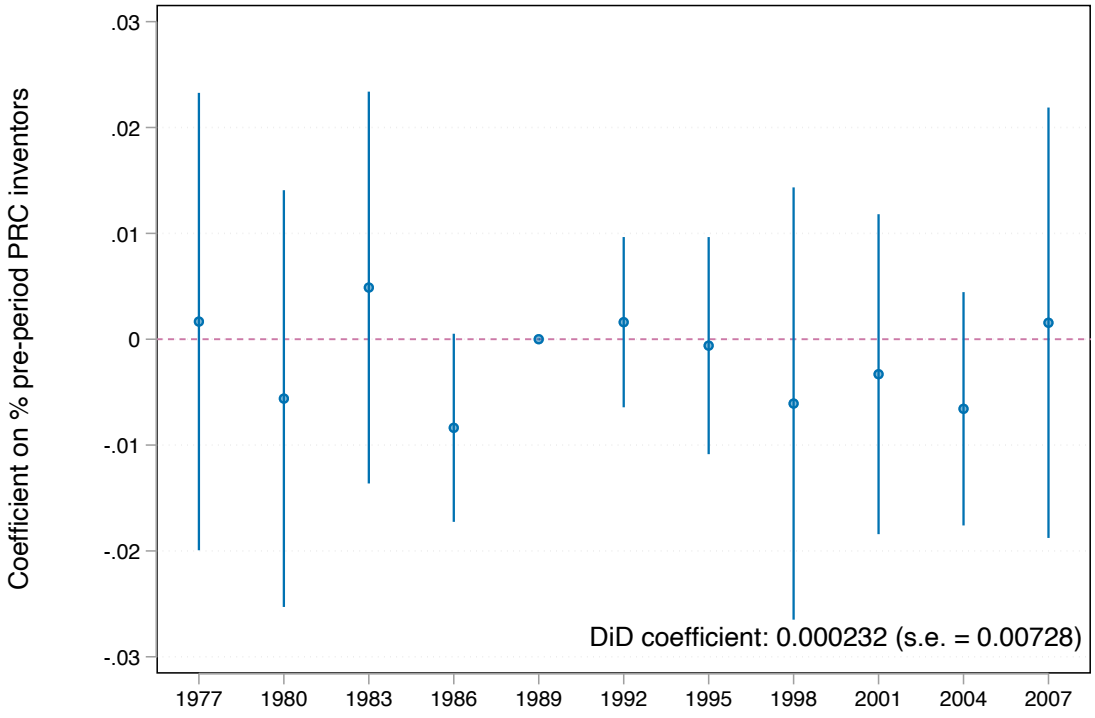
Notes: This figure presents coefficients on Year x % PRC from equation (2.3) for different definitions of the outcome variable. Year represents one of 11 three-year bins between 1976 and 2007 and % is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.A.10: Results for placebo CSPAs in pre-period



Notes: This figure presents coefficients on Year x % PRC from equation (2.3) for different placebo CSPAs using only data from before 1990. Year represents one of 11 three-year bins between 1976 and 2007 and % is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

Figure 2.A.11: Patents per inventor



Notes: This figure presents coefficients on Year x % PRC from equation (2.3), where Year represents one of 11 three-year bins between 1976 and 2007 and % PRC is the share of inventors between 1984-1988 in each county-NBER technology subcategory cell with PRC nationality. The sample includes county-NBER technology subcategory cells in which there has been at least one patent application in all year bins. The outcome variable is patents per inventor. All regressions control flexibly for county, NBER subcategory, and year and are weighted by total patents in each cell between 1984-1988. Standard errors are clustered by county.

2.B Data appendix

My main measure of CSPA exposure is the share of inventors from the PRC between 1984 and 1988 in each county-NBER technology subcategory cell. To construct this measure, I first need to determine the nationality (PRC vs. non-PRC) of each inventor. To predict inventors' nationalities, I use their full names as recorded in the PatentsView USPTO data.

I first use the ethnicolr Python Package (developed by Suriyan Laohaprapanon, Gaurav Sood and Bashar Naji; more detail is available at https://pypi.org/project/ethnicolr/) to predict race and ethnicity for each inventor surname.

The race prediction algorithm is trained on data from the 2000 U.S. Census to predict race from last names. It predicts the probability that a given last name is white, Black, Hispanic, or API, and provides means, standard errors, and confidence intervals for each prediction. It then classifies last names into race based on the category with the highest probability. I use these categories in my main analysis to construct my CSPA exposure measure.

For inventors classified as API using ethnicolr, I categorize PRC and non-PRC names as follows:

1. Last names ending in any consonant except *n* or *ng* as non-PRC
2. Last names with more than 6 letters as non-PRC
3. Last names with [vowel, consonant, vowel] pattern as non-PRC
4. Last names containing more than one space, apostrophe, or hyphen as non-PRC
5. Last names containing the following character n-grams as non-PRC: *v; aa, ae, ah, ak, ar, au, ea, eh, ek, eo, ih, ii, ik, oa, oh, oi, ok, oo, uh, uk, uu; by, hs, hw, hy, ja, je, jo, jy, kw, my, py, si, sz, ts, tz; ien, uen, ung*

6. First names starting or ending with the above n-grams and containing at least one space, hyphen, or apostrophe as non-PRC

7. First or last names containing the following character n-grams as PRC: *qi, xi, xu, zh*

Finally, I use a proprietary algorithm from NamSor (details available at https://namsor.app) to predict the most likely country of origin for any remaining unclassified API names. The NamSor algorithm is able to distinguish between East Asian countries, and separates names into origins including, e.g., mainland China, Taiwan, Hong Kong, Vietnam, and South Korea.

Finally, in robustness checks, I use the share of inventors from other ethnic immigrant groups as placebo checks. To determine the share of inventors from different ethnicities, I use another *ethnicity* prediction algorithm available in the ethnicolr package. The ethnicity prediction algorithm is trained on Wikipedia data to predict race from both first *and* last names. It predicts the probability that a given full name is Japanese, other East Asian, Indian, African, Muslim, British, Eastern European, Jewish, French, German, Hispanic, Italian, or Nordic; the algorithm again provides means, standard errors, and confidence intervals for each prediction. It then classifies last names into ethnicities based on the category with the highest probability. I use the Indian, Eastern European, and Jewish categories in my supplementary analysis.

Chapter 3

Discrimination and the Returns to Cultural Assimilation in the Age of Mass Migration[1]

3.1 Introduction

Immigrants who are identifiably foreign due to their accent, dress, name or other ethnic markers may be subject to discrimination at school or in the labor and housing markets. Because of this cost to retaining their cultural identity, immigrants often take steps to culturally assimilate into their adoptive countries. Indeed, in the early twentieth century, many immigrants to the U.S. changed their own names and/or chose more American sounding names for their children as they adapted to the country (Biavaschi, Giulietti, and Siddique, 2017; Abramitzky, Boustan, and Eriksson, 2020). In this paper, we study whether there is a negative association between for-

[1]This chapter was co-authored with Ran Abramitzky, Leah Boustan, and Katherine Eriksson and was published in *AEA Papers and Proceedings* in May 2020. We acknowledge generous research support from the Russell Sage Foundation. We thank Michael Clemens for his insightful discussion at the 2020 ASSA meetings, as well as numerous colleagues and seminar participants who provided comment on an earlier version of related work, which circulated under the title "Cultural Assimilation During the Age of Mass Migration."

eign sounding first names and outcomes, and whether this negative association is indicative of a causal effect of having an ethnic-sounding first name, or may instead reflect an economic return to cultural assimilation.

We start by constructing a large sample of the children of immigrants observed in their childhood households in the 1920 Census linked to adult outcomes in the 1940 Census. Parents of these children immigrated to the United States during the Age of Mass Migration from Europe (1850-1913), one of the largest migration episodes in American history. During this era, about 30 million immigrants from countries like Ireland, Germany, Italy and Russia moved across the Atlantic (Abramitzky, Boustan, and Eriksson, 2012, 2014, 2016; Abramitzky and Boustan, 2017). In a raw comparison, we find that men who were given more foreign-sounding first names attained fewer years of schooling, earned less, and experienced lower levels of employment than those with more American-sounding names. Furthermore, these men were more likely to marry women who were themselves less culturally assimilated (a foreign spouse or a spouse with a more foreign name). These baseline comparisons could reflect two channels: differences in the way that teachers, employers and others treat a Luigi versus a John (discrimination, both taste-based and statistical), and/or the family attributes that are correlated with the choice of an Italian name (particularly, the degree of cultural assimilation). Our paper tries to disentangle the role of family attributes from the role of discrimination against students or workers with foreign sounding names.

We find that the associations between name foreignness and adult outcomes remain economically meaningful and statistically significant when adding an increasingly detailed set of fixed effects – starting with country-of-origin, and then moving to surname, and childhood county of residence (in 1920) – implying that second generation immigrants with foreign first names fared worse than others who shared similar backgrounds.

Yet, the relationship between name foreignness and adult outcomes diminishes substantially when adding household fixed effects. This pattern suggests that observed disparities in economic outcomes by name foreignness were largely driven by differences between households that took steps to assimilate (e.g., by giving their children more American names) and those that did not. In particular, comparisons between brothers rules out discrimination in the labor market on the basis of first names but this null result still could be consistent with ethnic discrimination on the basis of other ethnic signals (including last name, accent, neighborhood of residence, etc.).

Our paper contributes to the literature on discrimination against immigrants and other demographic groups, which is often measured using differential treatment by name. The challenge in this literature (recognized by authors) is that names or name changes could be correlated with other aspects of family background or individual motivation (Arai and Thoursie, 2009; Moser, 2012; Goldstein and Stecklov, 2015; Carneiro, Lee, and Reis, 2015; Biavaschi, Giulietti, and Siddique, 2017; in another context, see Rubinstein and Brenner, 2014). Our findings suggest that the negative association between ethnic names and outcomes at least partially reflects the return to cultural assimilation or household selection into assimilation, rather than external discrimination. Consistent with our results, Oreopoulos (2011) does not find lower callback rates for workers with foreign-sounding first names in a resume audit study, but does find lower rates for workers with foreign-sounding last names.[2] We complement Oreopoulos by looking at real labor market success as measured by wages and employment, by considering a variety of other outcomes like education and matching

[2]In a related literature, Bertrand and Mullainathan (2004) document lower call-back rates for resumes assigned a distinctively black name, whereas Fryer and Levitt (2004) instead find that having a black name is not associated with lower levels of education or early childbearing after controlling for measures of family background. Cook, Logan, and Parman (2016) find a mortality advantage for men who received distinctively black names in the nineteenth and early twentieth centuries and point to suggestive evidence that receiving a black name is correlated with cultural factors at the household level.

in the marriage market, and by studying an important period in U.S. history, during which ethnicity is said to have played a large role in employment (the classic "No Irish Need Apply" signs) and before anti-discrimination laws were passed.

3.2 Data and definitions

3.2.1 Measuring the foreignness of given names

Historical Census data contain individual records with information on first and last name and country of birth for the full population. The Census Bureau releases these complete manuscripts after 72 years. This data has been made available to researchers by the Minnesota Population Center and Ancestry.com. To develop a systematic measure of name foreignness, we use the newly digitized complete-count 1920 and 1940 U.S. census to calculate an index of the relative probability that a given name was held by a foreigner versus an American (used also in Abramitzky, Boustan, and Eriksson (2020) and based on Fryer and Levitt's (2004) index of black names). The Foreignness Index is defined as:

$$ForeignnessIndex_{name} = 100 * \frac{\frac{\text{\# foreigners}_{name}}{\text{total \# foreigners}}}{\frac{\text{\# foreigners}_{name}}{\text{total \# foreigners}} + \frac{\text{\# natives}_{name}}{\text{total \# natives}}} \tag{3.1}$$

and ranges from zero to 100, with a value of zero reflecting the fact that no men in the U.S. with a given first name were foreign born (i.e., a distinctively American name) and a value of 100 assigned to a child whose first name is distinctively foreign.

In our main analysis, the F-index is calculated by contrasting the names of all foreign-born residents to all U.S.-born residents for the twenty years prior to an individual's birth year (to account for changes in naming trends). We also consider alternative measures that instead create country-specific name indices (e.g., all German-born versus all non-German-born). To address the entry of second-generation

immigrants into the pool of the U.S. born over time, we present results using an F-index that is fixed at a point in time (calculated for the 1895 to 1905 birth cohorts), rather than varying by birth year. In all cases, we focus on the sample of men with first names held by at least 100 others in the relevant years to calculate a reliable F-index.

3.2.2 Creating a linked Census sample

We create a matched dataset of the U.S.-born sons of immigrants observed in childhood in the 1920 Census and in adulthood in the 1940 Census. Siblings are observed in the same household in 1920. We link men across censuses by first and last name, age, and state of birth using the Abramitzky, Boustan, and Eriksson algorithm (2012, 2014), and we explore alternative matching algorithms in the online appendix (see Abramitzky et al., 2019).[3] We restrict our attention to men between the ages of 3 and 15 in 1920, who were young enough to be living at home with their parents in 1920 and would have been of prime labor market age in 1940. Our primary linked sample contains more than 800,000 men, around 300,000 of whom are in sets of matched brothers. We achieve a match rate of 25.5 percent, which is standard for historical matched samples (see Abramitzky et al., 2019).

Appendix Table 3.A.1 compares the men in our matched sample to all white second-generation immigrant men in 1940 on a number of baseline characteristics. As is typical for matched samples, we find that men in our linked dataset have higher socio-economic status than in the full population, as measured by employment and income. We present results that reweight the matched sample to match the population on family background. Furthermore, we find that the OLS association between name foreignness and economic outcomes is similar in the matched sample and the full

[3]We note that the presence of false positive links would bias the coefficients from regressions with household fixed effects toward the OLS coefficients because picking two random people due to bad matches and calling them brothers would tend to recover the population-wide estimates.

population (compare Appendix Table 3.A.2 to Table 3.1, which will be described in more detail below).

3.3 Foreign names and outcomes in the 1940 labor and marriage markets

Having a foreign name may have a negative effect on one's life trajectory for a number of reasons. First, ethnic identity might be more salient to children with foreign names, and if they expect to be treated less favorably because of their foreign identity, they may perceive lower returns to education and exert less effort in school (Jensen, 2010). Second, teachers or employers may use names as a signal of ethnicity and discriminate against children or workers with foreign-sounding names (Figlio, 2005). A similar dynamic may play out in the marriage market: men with foreign names might identify more strongly with their ethnic group and prefer to find a spouse within their own ethnic community, or these men may have been overlooked or rejected by spouses from other backgrounds. We start by regressing adult outcomes of son i from family j (y_{ij}) – namely education, earnings, employment, and spouse characteristics – on the foreignness of the son's name at birth, controlling for a vector of dummies for the son's birth year (γ_{ij}):

$$y_{ij} = \alpha_j + \beta_1 FIndex_{ij} + \beta_2 BirthOrder_{ij} + \beta_3 YearsUS_{ij} + \beta_4 \gamma_{ij} + \beta_5 X_{ij} + \epsilon_{ij} \quad (3.2)$$

We control for parental years in the U.S. and child's place in the birth order to absorb the association between name foreignness and other aspects of family structure or parental cultural assimilation (see Abramitzky, Boustan, and Eriksson, 2020). We then augment this core specification with sets of fixed effects included in the vector X, starting with parental country of origin, and then surname, and childhood county

of residence. Finally, we add household fixed effects (α_j) in order to compare brothers who were raised in the same family but received names with a different foreignness index.

Table 3.1 documents that name foreignness is negatively related to educational attainment, employment and earnings. In OLS, a 20-point shift in the F-index, the typical gap between the children of immigrants and U.S. born, is associated with 2 fewer months of schooling (relative to mean years of schooling of 10 years); a 2.2-percent decline in annual earnings; and a 1-percentage point increase in the probability of unemployment (column 1). The estimated effects decline (by 15 to 70 percent) but remain economically meaningful and statistically significant when adding country of origin, surname, and county fixed effects (columns 2-4). These controls address the fact that immigrants from some regions, particularly those more culturally distant from the U.S., gave their children more foreign-sounding names and also may have had worse economic outcomes. Furthermore, more recently arrived and less assimilated immigrants clustered in particular locations, which may itself have disadvantaged their children (Eriksson and Ward, 2019).

We then compare brothers who were given names with different ethnic content. Table 3.1 (column 5) demonstrates that including household fixed effects substantially reduces the estimated effects of name foreignness on adult outcomes (estimates become statistically insignificant in most cases).[4] Foreign names had a small and significant effect on educational attainment within brother pairs (20 points of F-index associated with two fewer weeks of schooling) and no effect on earnings and unemployment. The population-based estimates pick up other differences between families that

[4]In an earlier version of this paper, we found an association between name foreignness and adult outcomes, even within pairs of brothers (Abramitzky, Boustan, and Eriksson, 2016). These results were driven by a few households with co-resident children who were likely not related and could not be easily identified with Census data provided by Family Search (e.g., two families living within a single household, or borders living with a family). The updated results in this draft are instead based on Census data from the Minnesota Population Center, which provides detailed information on families, even within households, as coded by the variable "serial number." With more accurate measures, we find little association between name and adult outcomes within brother pairs.

chose foreign or American sounding names, rather than the causal effect of foreign names themselves.[5]

The ethnic signal of names that parents select for their children at birth can be attenuated or augmented as the name becomes more/less popular among certain groups. We find similar results when using an F-index calculated for the 1895 to 1905 birth cohorts, which measures name foreignness for a fixed set of birth cohorts born (Appendix Figure 3.A.1). More relevant to an employer's perception of a worker's ethnic identity might be the Foreignness Index of his name at the time of labor market entry. In Appendix Table 3.A.3, we try a specification that includes two F-indices on the right-hand side – one calculated at birth and the other at labor market entry – however the two measures are highly correlated (corr. $= 0.98$) and so no clear pattern emerges. Note that our matched sample contains disproportionate numbers of men with uncommon names. One possibility is that employers cannot discern the ethnic content of rare names. However, we find similar results if we split our sample by name frequency (see Appendix Tables 3.A.4 and 3.A.5).

Beyond the labor market, having a foreign name may have influenced men's marriage market outcomes. Table 3.1 considers two measures of the foreignness of a man's spouse: whether the spouse, herself, was born abroad and the (gender-specific) Foreignness Index of spouse's first name, an indicator of either being born abroad or being raised in a less culturally assimilated family in the U.S.[6] In both cases, we find that men with foreign names are more likely to marry women with a stronger ethnic identity. A 20-point difference in a man's F-index is associated with a 0.8-percentage-point increase in the probability of having a foreign-born spouse (on a base of 5.2 percent) and a 2.4-point increase in the F-index of his spouse's first name

[5]Estimates are little changed by weighting the distribution of father's country-of-origin in the linked sample to match the full population in 1920 or 1940; see Appendix Figure 3.A.4.

[6]Until 1930, the Census asked all respondents about parents' birthplace, which would have allowed us to classify whether spouses were second-generation immigrants. However, in the 1940 Census, the question about parental birthplace became a 'sample-line characteristic' asked of only five percent of the population and so we instead focus on whether a spouse is a first-generation immigrant herself.

(on a base of 45 points).[7] Yet, as with the labor market outcomes, neither relationship remains statistically significant after adding household fixed effects (although confidence intervals are large).

Appendix Figure 3.A.2 considers a series of additional labor market outcomes, including hours and weeks worked per year and the log weekly wage, as well as whether or not an individual was self-employed or employed by the government in a public works job.[8] Consistent with the negative association between name foreignness and employment at the time of the Census, men with more foreign names work less time during the year in both hours and weeks and receive a lower weekly wage in the OLS specifications. Men with foreign names were also more likely to hold a public works job through the New Deal, which could be an indication of weak attachment to the labor force. In all cases, the relationship between name foreignness and economic outcomes is reduced when comparing brothers within households and nearly disappears in all cases except weekly wages. These results are robust to alternative linking algorithms that increasingly introduce more conservative requirements on what is considered a successful match.[9]

We present separate results in Figure 3.1 by country of origin for all sending countries with at least 1000 brother pairs in order to differentiate between groups like the Brits that may have easily assimilated with new immigrant groups like Italians and Russians that may have been targets of discrimination in the labor market. We do find one outcome – education – that seems to be associated with name foreignness even within pairs of brothers. The negative association between name foreignness and

[7]Results on spouse characteristics are restricted to the subsample of men who were 25 years or older in 1940 and who were married in that year. Men with a more foreign name are less likely to be married by 1940, but this effect is economically small. 20 points on the F-index is associated with a 1-percentage-point decline in the probability of being married (on a base of 68 percent).

[8]The 1940 Census only contains information on wage and salary income. As a result, results on annual earnings exclude the self-employed. Appendix Figure 3.A.2 and Appendix Table 3.A.6 shows that shifts in the F-index have only a very small effect on the probability of self-employment.

[9]See Appendix Figures 3.A.5-3.A.9 and Appendix Tables 3.A.7-3.A.10. See Abramitzky et al. (2019) for a detailed description of alternative matching approaches.

education (panel (a)) is statistically significant or nearly so in regressions with household fixed effects for immigrants from Italy, Russia, Ireland, and Austria, countries that by historical account may have suffered more discrimination and worse outcomes.

The negative association between name foreignness and other economic or marriage market outcomes is not statistically significant in specifications with household fixed effects, even for immigrants from newer sending countries (panels (b)-(c)). However, because the confidence intervals are large when we subdivide the sample by country and include household fixed effects, we note that, in almost all cases, we cannot reject the hypothesis that the coefficients from OLS regressions and regressions with household fixed effects are the same. These patterns generally hold when we use country-specific foreignness index measure instead (see panels (a)-(e) of Appendix Figure 3.A.3).

3.4 Conclusion

In the early twentieth century, receiving an American-sounding name was associated with a series of positive outcomes for the children of immigrants, including more education, increased income, and a higher probability of employment. However, within households, brothers with more and less foreign names did not seem to attain very different levels of schooling, earnings, and employment. Our findings suggest that the negative association between ethnic names and outcomes may not be indicative of a causal effect of having an ethnic-sounding first name, but may instead reflect the return to cultural assimilation, as the children of families that chose to assimilate had better outcomes than those that retained their cultural markers.

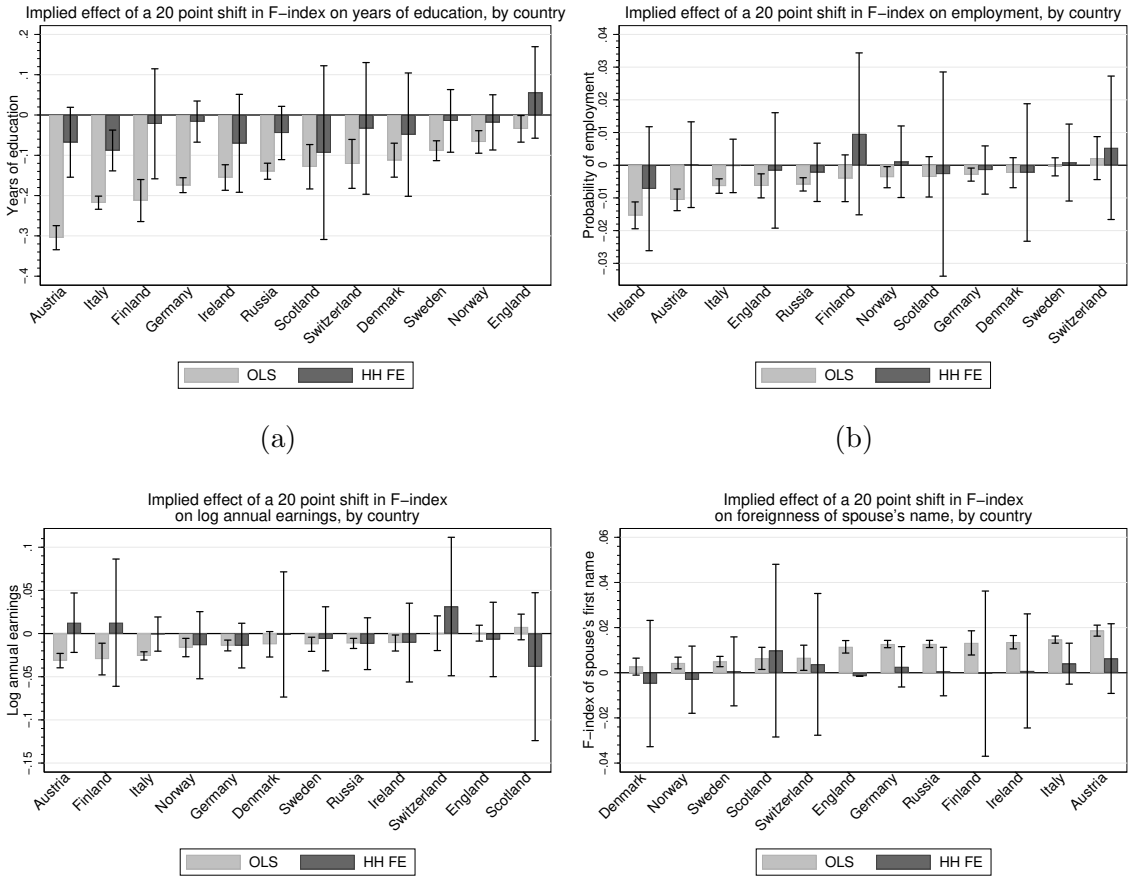
3.5 Tables and figures

Table 3.1: Effect of given name foreignness on labor and marriage market outcomes

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.16	-1.169***	-0.905***	-0.814***	-0.797***	-0.215***	-0.180***
		(0.0165)	(0.0175)	(0.0226)	(0.0222)	(0.0527)	(0.0488)
N	802519	709727	709727	709723	709723	709727	113239
Employed	0.865	-0.0534***	-0.0260***	-0.0203***	-0.0159***	-0.00409	-0.0172*
		(0.00190)	(0.00203)	(0.00268)	(0.00269)	(0.00798)	(0.00757)
N	820294	725300	725300	725295	725295	725300	115601
Log annual income	6.846	-0.111***	-0.0929***	-0.0769***	-0.0944***	-0.0234	-0.0302
		(0.00495)	(0.00534)	(0.00742)	(0.00725)	(0.0230)	(0.0213)
N	629969	555406	555406	555403	555403	555406	87238
F-index of spouse's name	0.450	0.122***	0.0571***	0.0479***	0.0399***	0.0118	0.0122
		(0.00145)	(0.00154)	(0.00226)	(0.00227)	(0.00954)	(0.00892)
N	435802	381440	381440	381438	381438	381440	60076
Foreign-born spouse	0.0522	0.0410***	0.0269***	0.0251***	0.0228***	0.00729	0.00228
		(0.00167)	(0.00180)	(0.00262)	(0.00265)	(0.0102)	(0.00978)
N	459209	401810	401810	401808	401808	401810	63088
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Figure 3.1: Effect of a 20-point shift in F-index



Notes: *** p<0.01, ** p<0.05, * p<0.1. In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was born in one of the countries listed.

Appendix

3.A Appendix tables and figures

Table 3.A.1: Comparing mean characteristics of matched and unmatched second-generation men

	1940 full population (5% sample line)	Matched Sample	Difference
Age	28.52	28.73	0.214***
N	121181	820294	
Years of education	9.974	10.16	0.190***
N	118905	802519	
Employed	0.846	0.865	0.0184***
N	121181	820294	
Public emergency worker	0.0318	0.0299	-0.00197***
N	121181	820294	
Self-employed	0.118	0.138	0.0201***
N	121181	820294	
Hours worked per week	33.92	35.43	1.509***
N	121181	820294	
Weeks worked per year	38.69	39.95	1.258***
N	121181	820294	
Log annual income	6.810	6.846	0.0358***
N	94669	629969	
Log weekly income	3.178	3.188	0.0103***
N	78283	527481	
F-index of spouse's first name	0.468	0.450	-0.0187***
N	58902	435802	
Foreign-born spouse	0.0673	0.0522	-0.0150***
N	62242	459209	

Notes: *** p<0.01, ** p<0.05, * p<0.1. 1940 full population sample includes native born white men with foreign-born fathers from the 1940 complete-count Census, aged 23-35 in 1940, living outside the South. Matched sample includes native-born men matched between the 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Table 3.A.2: Effect of given name foreignness on labor and marriage market outcomes: 5% of full 1940 population

Outcome:	Mean	OLS
Years of education	9.974	-1.409***
		(0.0422)
N	118905	112350
Employed	0.846	-0.0603***
		(0.00511)
N	121181	114480
Public emergency worker	0.0318	0.0102***
		(0.00252)
N	121181	114480
Self-employed	0.118	-0.0391***
		(0.00495)
N	121181	114480
Hours worked per week	33.92	-5.983***
		(0.314)
N	121181	114480
Weeks worked per year	38.69	-4.447***
		(0.258)
N	121181	114480
Log annual income	6.810	-0.0876***
		(0.0126)
N	94669	89584
Log weekly income	3.178	-0.00987
		(0.0104)
N	78283	74082
F-index of spouse's name	0.468	0.144***
		(0.00386)
N	58902	55874
Foreign-born spouse	0.0673	0.0664***
		(0.00490)
N	62242	58914

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born white men with foreign-born fathers in the 1940 Census, aged 23-35 in 1940, living outside the South. Father's birthplace is a 5% sample line in the 1940 Census. All regressions control for age.

Table 3.A.3: Effect of F-indices calculated at birth and at age 20 on labor and marriage market outcomes

	(1)		(2)	
F-index	At birth	At age 20	At birth	At age 20
Years of education	2.019***	-3.269***	-0.113	-0.105
	(0.0777)	(0.0777)	(0.224)	(0.223)
N	709727		709727	
Employed	-0.100***	0.0481***	-0.00255	-0.00158
	(0.00920)	(0.00926)	(0.0344)	(0.0344)
N	725300		725300	
Log annual income	0.290***	-0.412***	0.0253	-0.0503
	(0.0231)	(0.0233)	(0.0956)	(0.0961)
N	555406		555406	
F-index of spouse's first name	0.220***	-0.101***	0.0579	-0.0479
	(0.00707)	(0.00716)	(0.0410)	(0.0414)
N	381440		381440	
Foreign-born spouse	0.0848***	-0.0455***	0.0113	-0.00421
	(0.00802)	(0.00798)	(0.0447)	(0.0444)
N	401810		401810	
Household FE			✓	

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.4: Effect of given name foreignness on labor and marriage market outcomes: men with names below median frequency

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.27	-1.122***	-0.960***	-0.845***	-0.851***	-0.136	-0.144
		(0.0203)	(0.0221)	(0.0314)	0.0311)	(0.119)	(0.114)
N	384364	347644	347644	347643	347643	347644	53750
Employed	0.867	-0.0506***	-0.0243***	-0.0203***	-0.0162***	0.00560	-0.0107
		(0.00231)	(0.00251)	(0.00364)	(0.00367)	(0.0179)	(0.0175)
N	393062	355473	355473	355472	355472	355473	54888
Log annual income	6.845	-0.0846***	-0.0824***	-0.0738***	-0.101***	-0.0486	-0.0389
		(0.00615)	(0.00683)	(0.0105)	(0.0103)	(0.0556)	(0.0525)
N	296349	267067	267067	267066	267066	267067	40488
F-index of spouse's name	0.448	0.123***	0.0583***	0.0490***	0.0398***	0.0163	0.0149
		(0.00178)	(0.00194)	(0.00316)	(0.00320)	(0.0226)	(0.0215)
N	208355	186403	186403	186403	186403	186403	28656
Foreign-born spouse	0.0535	0.0440***	0.0297***	0.0283***	0.0239***	0.00756	0.00340
		(0.00202)	(0.00222)	(0.00360)	(0.00369)	(0.0232)	(0.0214)
N	219442	196328	196328	196328	196328	196328	30093
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. Men must have names below the sample median name frequency, with name frequency calculated over the entire US population at the time of the 1940 Census. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.5: Effect of given name foreignness on labor and marriage market outcomes: men with names at or above median frequency

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.06	-1.716***	-1.242***	-1.038***	-0.959***	-0.223	-0.168
		(0.0301)	(0.0317)	(0.0484)	(0.0479)	(0.153)	(0.146)
N	397108	355642	355642	355639	355639	355642	58435
Employed	0.863	-0.0710***	-0.0372***	-0.0216***	-0.0181**	-0.0132	-0.0282
		(0.00358)	(0.00380)	(0.00588)	(0.00593)	(0.0236)	(0.0229)
N	405685	363258	363258	363254	363254	363258	59641
Log annual income	6.850	-0.195***	-0.154***	-0.110***	-0.109***	-0.0575	-0.0385
		(0.00898)	(0.00956)	(0.0158)	(0.0155)	(0.0650)	(0.0615)
N	317658	283510	283510	283508	283508	283510	45984
F-index of spouse's name	0.452	0.137***	0.0605***	0.0472***	0.0414***	0.00506	0.0248
		(0.00268)	(0.00282)	(0.00503)	(0.00508)	(0.0281)	(0.0270)
N	216266	191494	191494	191492	191492	191494	30856
Foreign-born spouse	0.0511	0.0317***	0.0174***	0.0178**	0.0196***	0.0128	-0.000939
		(0.00309)	(0.00327)	(0.00575)	(0.00586)	(0.0308)	(0.0312)
N	227885	201742	201742	201740	201740	201742	32393
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. Men must have names at or above the median name frequency, with name frequency calculated over the entire US population at the time of the 1940 Census. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.6: Effect of given name foreignness on probability of self-employment

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Self-employed	0.138	-0.0185***	-0.00852***	-0.00528	0.00396	-0.00100	-0.00476
		(0.00202)	(0.00217)	(0.00290)	(0.00286)	(0.00797)	(0.00741)
N	820294	725300	725300	725295	725295	725300	115601
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.7: Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 3-year band, age difference within 5-year band

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.18	-1.159***	-0.910***	-0.816***	-0.802***	-0.150*	-0.100
		(0.0185)	(0.0195)	(0.0265)	(0.0261)	(0.0617)	(0.0585)
N	606268	535424	535424	535424	535424	535424	91370
Employed	0.869	-0.0480***	-0.0218***	-0.0162***	-0.0124***	-0.000104	-0.0139
		(0.00211)	(0.00224)	(0.00312)	(0.00313)	(0.00962)	(0.00940)
N	619475	547020	547020	547020	547020	547020	93211
Log annual income	6.849	-0.112***	-0.0937***	-0.0728***	-0.0921***	-0.0176	-0.0163
		(0.00557)	(0.00598)	(0.00883)	(0.00861)	(0.0280)	(0.0268)
N	473222	416581	416581	416581	416581	416581	69925
F-index of spouse's name	0.448	0.115***	0.0516***	0.0420***	0.0343***	0.00694	0.00858
		(0.00163)	(0.00172)	(0.00271)	(0.00272)	(0.0117)	(0.0112)
N	331489	289949	289949	289949	289949	289949	48461
Foreign-born spouse	0.0519	0.0394***	0.0262***	0.0252***	0.0234***	0.00511	0.00450
		(0.00187)	(0.00200)	(0.00311)	(0.00316)	(0.0123)	(0.0121)
N	349214	305334	305334	305334	305334	305334	50861
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.8: Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 3-year band, age difference within 3-year band

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.21	-1.141***	-0.902***	-0.803***	-0.784***	-0.109	-0.0536
		(0.0190)	(0.0201)	(0.0274)	(0.0270)	(0.0648)	(0.0616)
N	566433	500877	500877	500877	500877	500877	86559
Employed	0.871	-0.0459***	-0.0204***	-0.0146***	-0.0109***	-0.0000833	-0.0141
		(0.00216)	(0.00230)	(0.00321)	(0.00323)	(0.0101)	(0.00997)
N	578637	511630	511630	511630	511630	511630	88287
Log annual income	6.855	-0.111***	-0.0932***	-0.0739***	-0.0934***	-0.0231	-0.0129
		(0.00574)	(0.00616)	(0.00914)	(0.00890)	(0.0297)	(0.0285)
N	442063	389645	389645	389645	389645	389645	66209
F-index of spouse's name	0.447	0.114***	0.0509***	0.0415***	0.0338***	0.00607	0.00466
		(0.00168)	(0.00177)	(0.00280)	(0.00282)	(0.0123)	(0.0119)
N	311505	272801	272801	272801	272801	272801	46145
Foreign-born spouse	0.0514	0.0382***	0.0249***	0.0248***	0.0230***	0.00567	0.00208
		(0.00191)	(0.00205)	(0.00321)	(0.00326)	(0.0130)	(0.0129)
N	328068	287189	287189	287189	287189	287189	48427
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.9: Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 5-year band, age difference within 5-year band

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.20	-1.162***	-0.922***	-0.815***	-0.794***	-0.160*	-0.152*
		(0.0197)	(0.0208)	(0.0292)	(0.0287)	(0.0681)	(0.0652)
N	520140	458745	458745	458745	458745	458745	80310
Employed	0.871	-0.0452***	-0.0203***	-0.0164***	-0.0129***	-0.00173	-0.0157
		(0.00224)	(0.00238)	(0.00343)	(0.00344)	(0.0107)	(0.0105)
N	531375	468603	468603	468603	468603	468603	81923
Log annual income	6.851	-0.114***	-0.0963***	-0.0744***	-0.0919***	-0.0264	-0.0203
		(0.00596)	(0.00639)	(0.00981)	(0.00957)	(0.0314)	(0.0303)
N	405149	356170	356170	356170	356170	356170	61314
F-index of spouse's name	0.446	0.111***	0.0492***	0.0391***	0.0321***	0.00131	0.0116
		(0.00174)	(0.00183)	(0.00302)	(0.00304)	(0.0131)	(0.0127)
N	284973	249041	249041	249041	249041	249041	42555
Foreign-born spouse	0.0513	0.0377***	0.0252***	0.0236***	0.0219***	0.000467	-0.00557
		(0.00198)	(0.00211)	(0.00343)	(0.00349)	(0.0136)	(0.0134)
N	300273	262318	262318	262318	262318	262318	44675
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.10: Effect of given name foreignness on labor and marriage market outcomes: Sample restricted to unique name matches within 5-year band, age difference within 3-year band

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	10.23	-1.145***	-0.916***	-0.803***	-0.779***	-0.109	-0.0930
		(0.0203)	(0.0214)	(0.0302)	(0.0297)	(0.0715)	(0.0685)
N	485267	428555	428555	428555	428555	428555	76002
Employed	0.873	-0.0432***	-0.0192***	-0.0153***	-0.0120***	-0.00170	-0.0160
		(0.00230)	(0.00244)	(0.00353)	(0.00354)	(0.0113)	(0.0111)
N	495612	437659	437659	437659	437659	437659	77512
Log annual income	6.856	-0.113***	-0.0962***	-0.0778***	-0.0956***	-0.0346	-0.0167
		(0.00614)	(0.00658)	(0.0102)	(0.00989)	(0.0333)	(0.0322)
N	377911	332662	332662	332662	332662	332662	57997
F-index of spouse's name	0.445	0.110***	0.0487***	0.0383***	0.0313***	0.000929	0.00873
		(0.00179)	(0.00189)	(0.00313)	(0.00315)	(0.0138)	(0.0135)
N	267464	234028	234028	234028	234028	234028	40484
Foreign-born spouse	0.0508	0.0363***	0.0239***	0.0231***	0.0216***	0.00137	-0.00680
		(0.00203)	(0.00216)	(0.00354)	(0.00361)	(0.0144)	(0.0143)
N	281730	246414	246414	246414	246414	246414	42495
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order.

Table 3.A.11: Effect of given name foreignness on labor and marriage market outcomes: EM matches

Outcome:	(1)	(2)	(3)	(4)	(5)	(6)
Years of education	-1.157***	-0.910***	-0.811***	-0.757***	-0.0768	-0.123
	(0.0242)	(0.0255)	(0.0380)	(0.0372)	(0.0858)	(0.0755)
	318840	318840	318840	318840	318840	41850
Employed	-0.0284***	-0.0138***	-0.00919*	-0.00833*	-0.00724	-0.0116
	(0.00254)	(0.00268)	(0.00414)	(0.00419)	(0.0126)	(0.0114)
	322769	322769	322769	322769	322769	42327
Log annual income	-0.0956***	-0.0967***	-0.0713***	-0.0804***	-0.000163	0.0154
	(0.00726)	(0.00774)	(0.0127)	(0.0124)	(0.0391)	(0.0340)
	243445	243445	243445	243445	243445	31032
F-index of spouse's name	0.102***	0.0472***	0.0421***	0.0345***	0.00848	0.00505
	(0.00208)	(0.00218)	(0.00387)	(0.00391)	(0.0161)	(0.0144)
	173856	173856	173856	173856	173856	22062
Foreign-born spouse	0.0311***	0.0211***	0.0198***	0.0189***	0.00191	0.00382
	(0.00225)	(0.00238)	(0.00416)	(0.00426)	(0.0160)	(0.0152)
	182905	182905	182905	182905	182905	23184
Country of origin FE		✓	✓	✓		
Surname FE			✓	✓		
1920 county of residence FE				✓		
Household FE					✓	✓
Brothers 2 or fewer years apart						✓

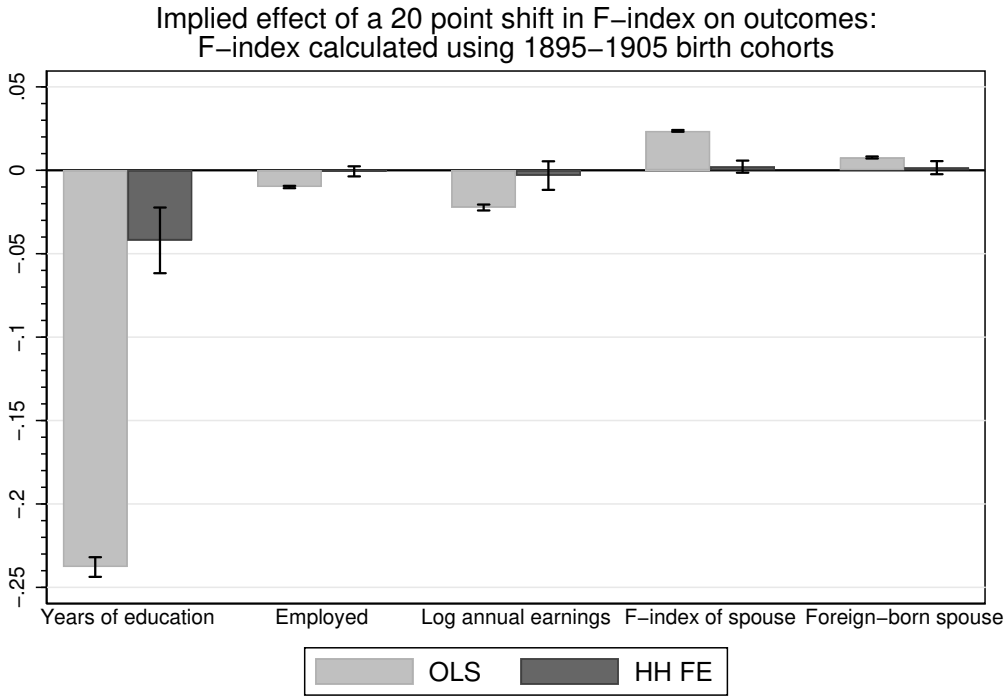
Notes: *** p<0.01, ** p<0.05, * p<0.1.

Table 3.A.12: Effect of given name foreignness on labor and marriage market outcomes: ABE-conservative matches restricted to states used in EM sample

Outcome:	Mean	(1)	(2)	(3)	(4)	(5)	(6)
Years of education		-1.155***	-0.856***	-0.765***	-0.729***	-0.177**	-0.177***
		(0.0189)	(0.0201)	(0.0261)	(0.0256)	(0.0564)	(0.0536)
		509764	509764	509760	509760	509764	86741
Employed		-0.0368***	-0.0197***	-0.0146***	-0.0132***	-0.00589	-0.0168*
		(0.00214)	(0.00228)	(0.00304)	(0.00305)	(0.00845)	(0.00825)
		520509	520509	520504	520504	520509	88478
Log annual income		-0.0996***	-0.0951***	-0.0787***	-0.0896***	-0.0316	-0.0461
		(0.00589)	(0.00634)	(0.00892)	(0.00868)	(0.0256)	(0.0242)
		392903	392903	392900	392900	392903	65552
F-index of spouse's name		0.110***	0.0521***	0.0432***	0.0368***	0.0104	0.0131
		(0.00170)	(0.00179)	(0.00265)	(0.00267)	(0.0103)	(0.00995)
		279801	279801	279799	279799	279801	46641
Foreign-born spouse		0.0324***	0.0198***	0.0171***	0.0157***	0.00112	-0.00156
		(0.00186)	(0.00198)	(0.00287)	(0.00291)	(0.0105)	(0.0105)
		294361	294361	294359	294359	294361	48961
Country of origin FE			✓	✓	✓		
Surname FE				✓	✓		
1920 county of residence FE					✓		
Household FE						✓	✓
Brothers 2 or fewer years apart							✓

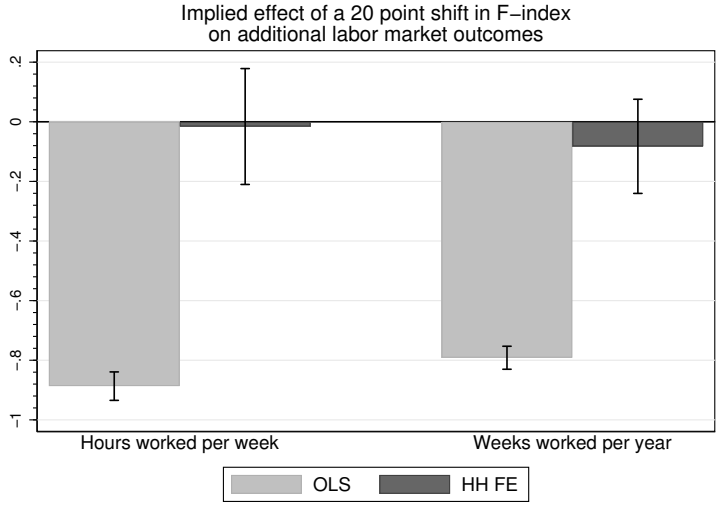
Notes: *** p<0.01, ** p<0.05, * p<0.1. Sample restricted to men whose states of residence in 1920 were used in the EM sample: excludes Alaska, New York, and Pennsylvania.

Figure 3.A.1: Results using F-index calculated from 1895-1905 birth cohorts

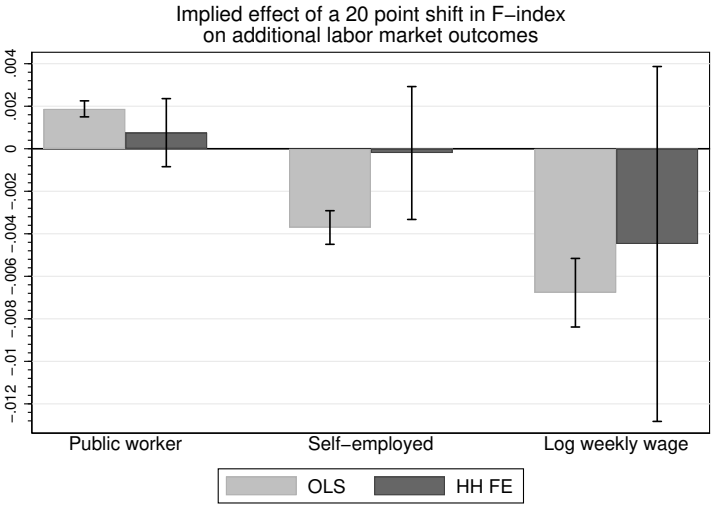


Notes: In this figure we estimate a regression of each outcome listed on the foreignness index of an individual's name. We report 0.2x the coefficient value of an individual's F-index calculated using the 1895-1905 birth cohorts. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.2: Results for F-index on additional labor market outcomes



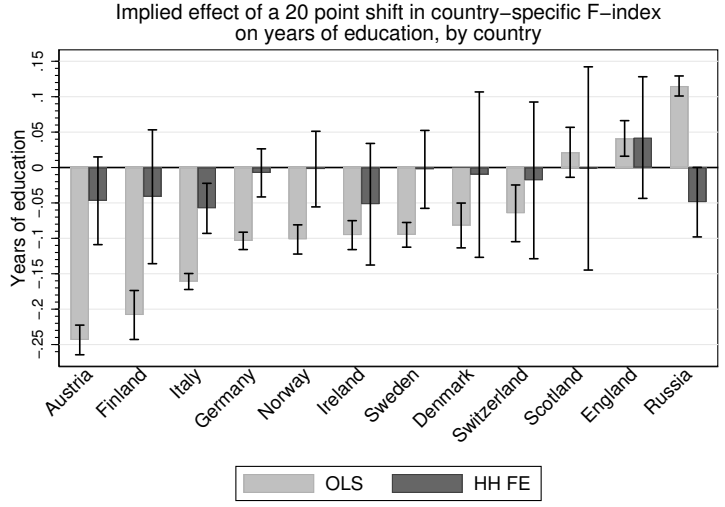
(a)



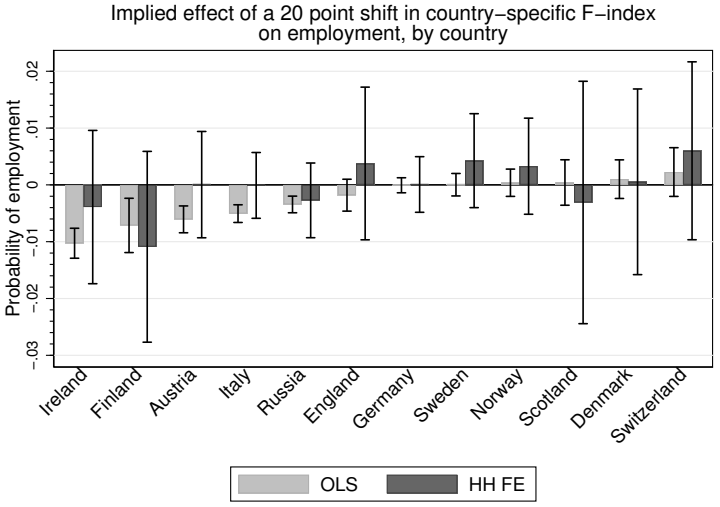
(b)

Notes: In this figure we estimate a regression of each outcome listed on the foreignness index of an individual's name. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.3: Results using country-specific F-indices



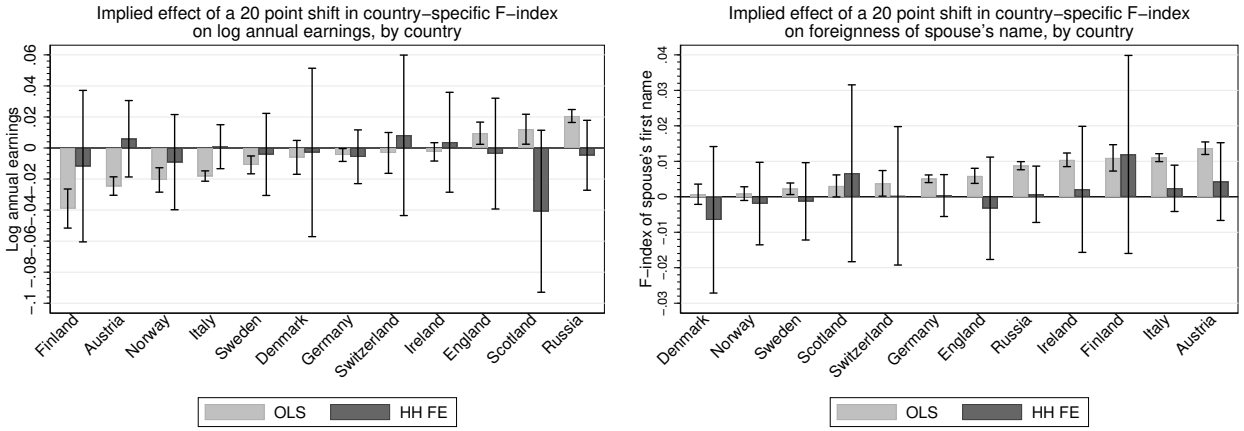
(a)



(b)

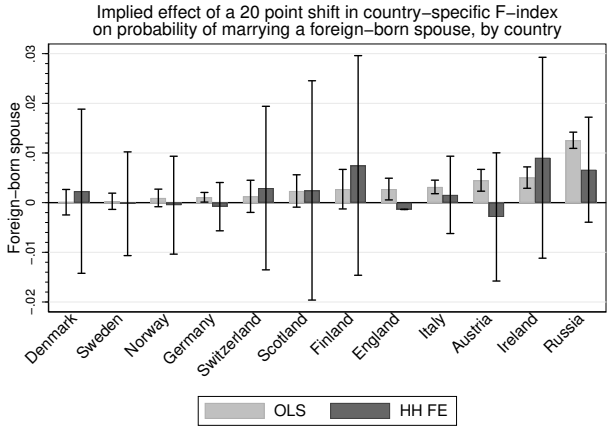
Notes: In these figures we estimate regressions of each outcome listed on the country-specific foreignness index of an individual's name for men from countries with more than 1000 observations at the household level. We report 0.2x the coefficient value of an individual's country-specific F-index calculated for his year of birth and the country of origin of his household head in 1920. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was born in one of the countries listed.

Figure 3.A.3: Results using country-specific F-indices (cont.)



(c)

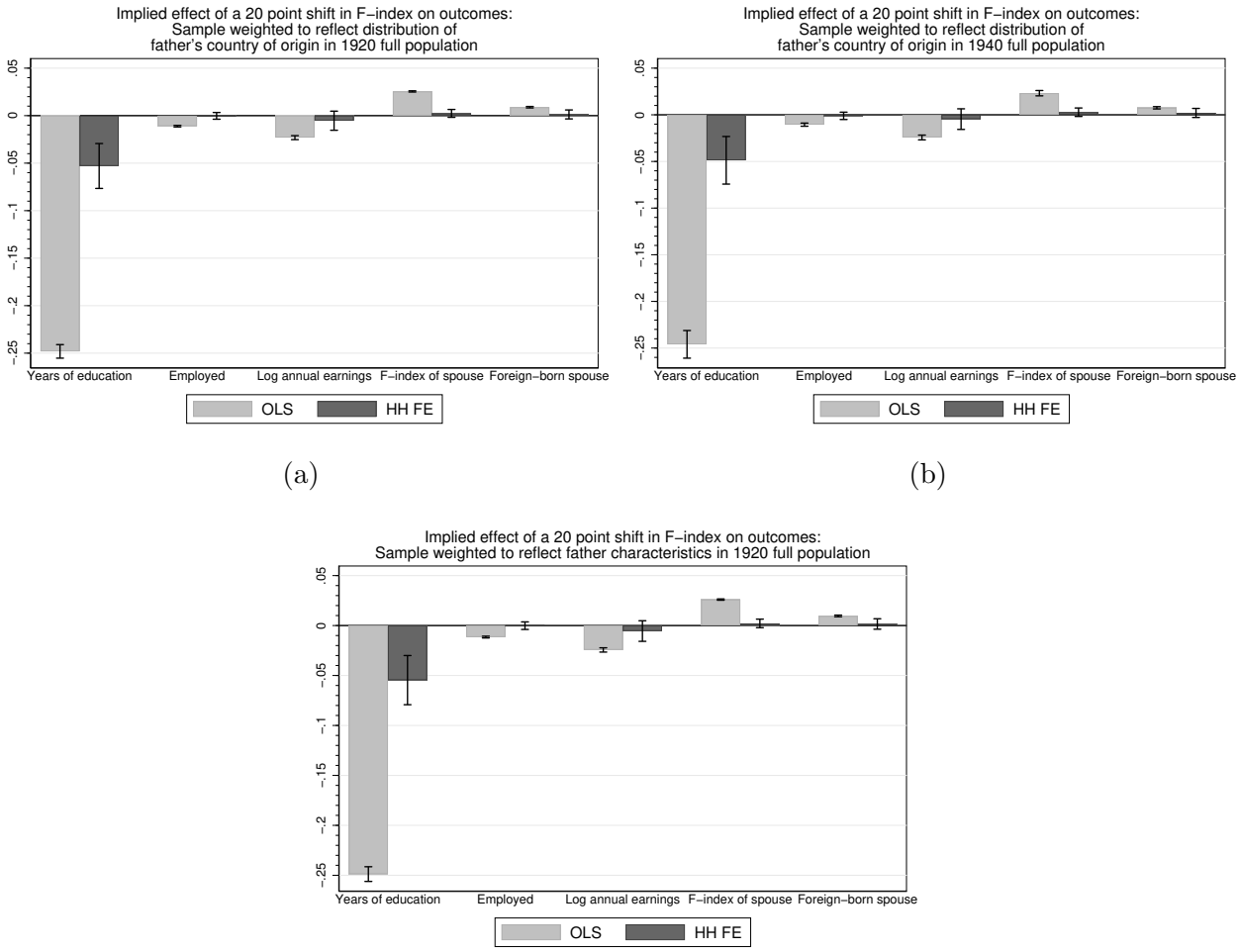
(d)



(e)

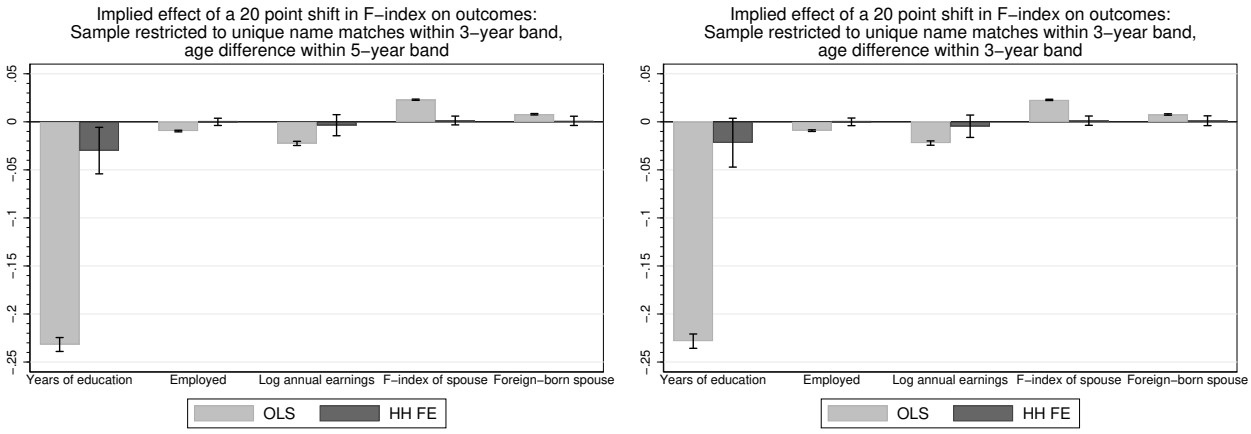
Notes: In these figures we estimate regressions of each outcome listed on the country-specific foreignness index of an individual's name for men from countries with more than 1000 observations at the household level. We report 0.2x the coefficient value of an individual's country-specific F-index calculated for his year of birth and the country of origin of his household head in 1920. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was born in one of the countries listed.

Figure 3.A.4: Results for weighted samples



Notes: In these figures we estimate weighted regressions of each outcome listed on the foreignness index of an individual's name. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. Panel (a) shows results from weighting the sample to reflect the distribution of father's birthplace in the 1940 population (father's birthplace is a 5% sample line in the 1940 Census). Panel (b) shows results from weighting the sample to reflect the distribution of father's birthplace in the full 1920 population. Panel (c) shows results from weighting the sample to match additional father characteristics to the full 1920 population. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.5: Results using more conservative linking algorithms

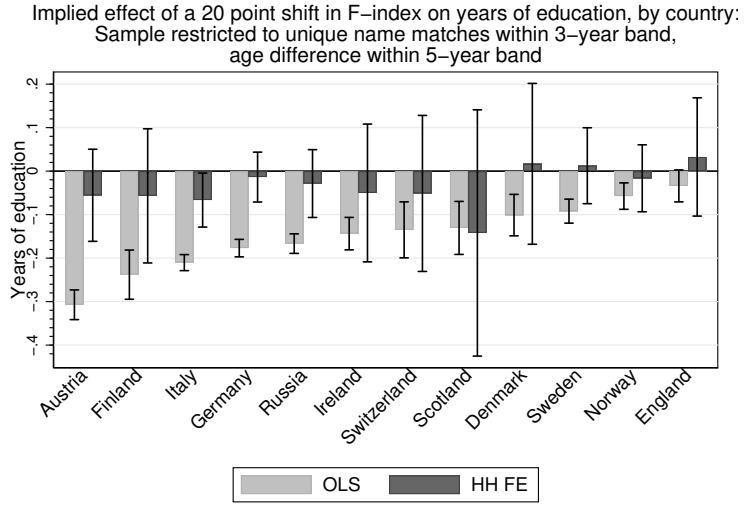


(a) (b)

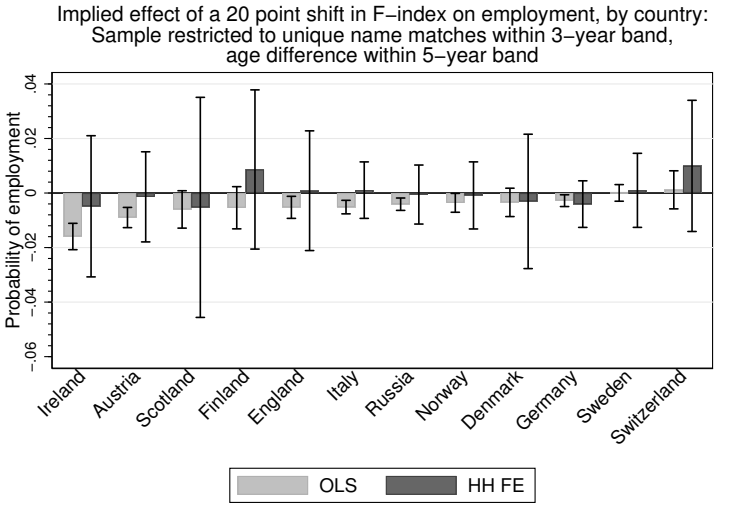
(c) (d)

Notes: In this figure we estimate a regression of each outcome listed on the foreignness index of an individual's name using more conservative linking algorithms. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born. Panel (a) shows results from using the ABE-conservative linking method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Panel (b) requires name uniqueness within a 3-year band and allows matched pairs to differ in reported age by up to 1 year on either side. Panel (c) requires name uniqueness within a 5-year band and allows matched pairs to differ in reported age by up to 2 years on either side. Panel (c) requires name uniqueness within a 5-year band and allows matched pairs to differ in reported age by up to 1 year on either side.

Figure 3.A.6: Results by country: sample matched using 3-year uniqueness bands, 5-year age bands



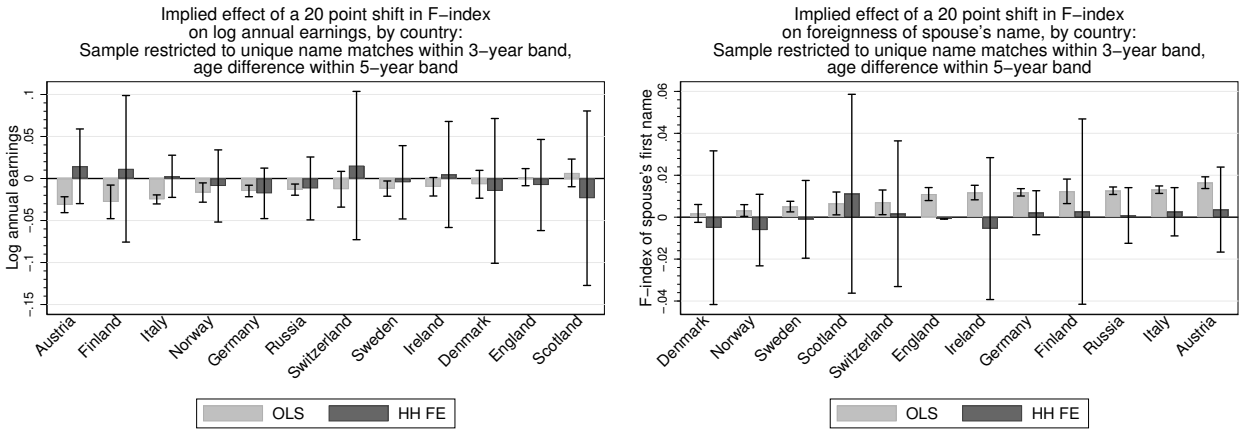
(a)



(b)

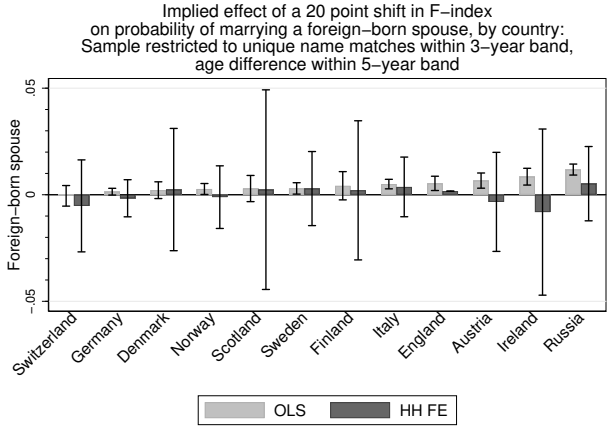
Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.6: Results by country: sample matched using 3-year uniqueness bands, 5-year age bands (cont.)



(c)

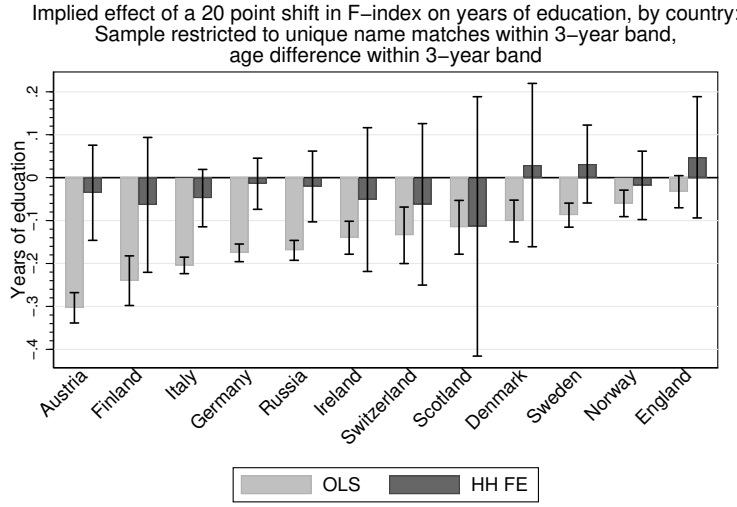
(d)



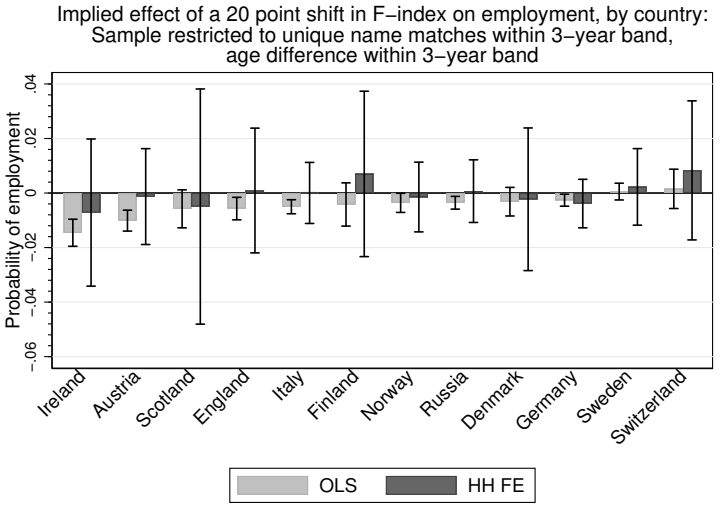
(e)

Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.7: Results by country: sample matched using 3-year uniqueness bands, 3-year age bands



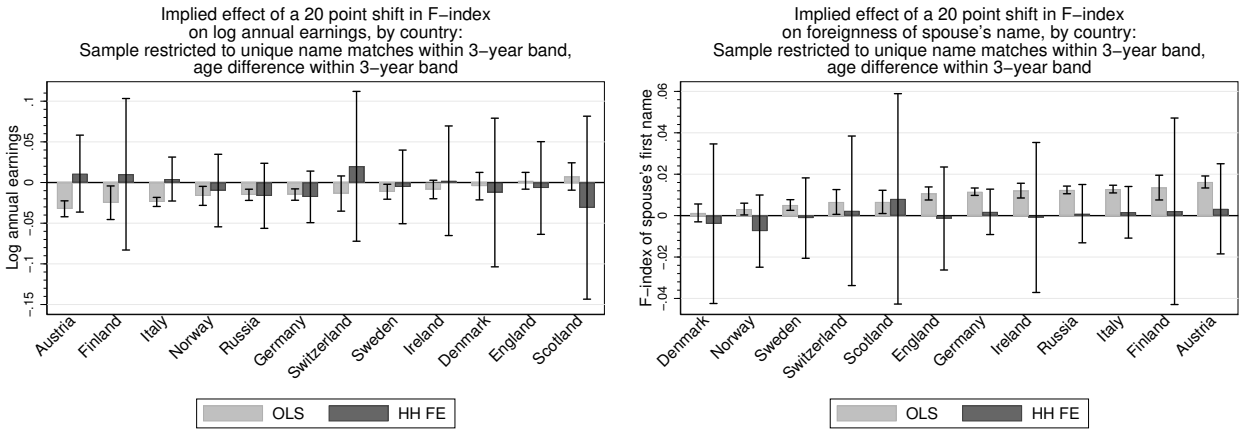
(a)



(b)

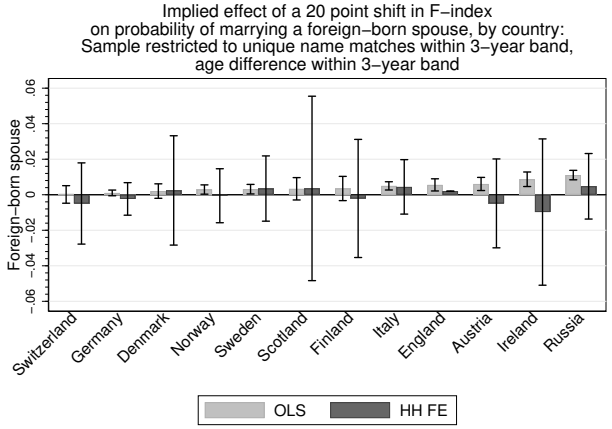
Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.7: Results by country: sample matched using 3-year uniqueness bands, 3-year age bands (cont.)



(c)

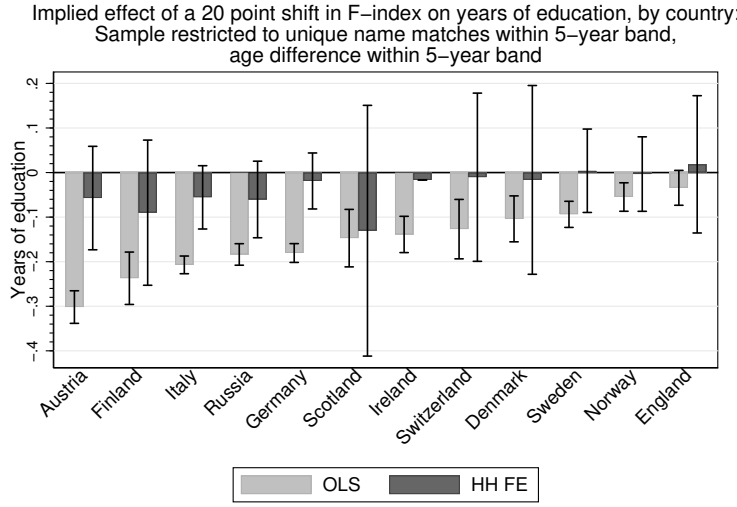
(d)



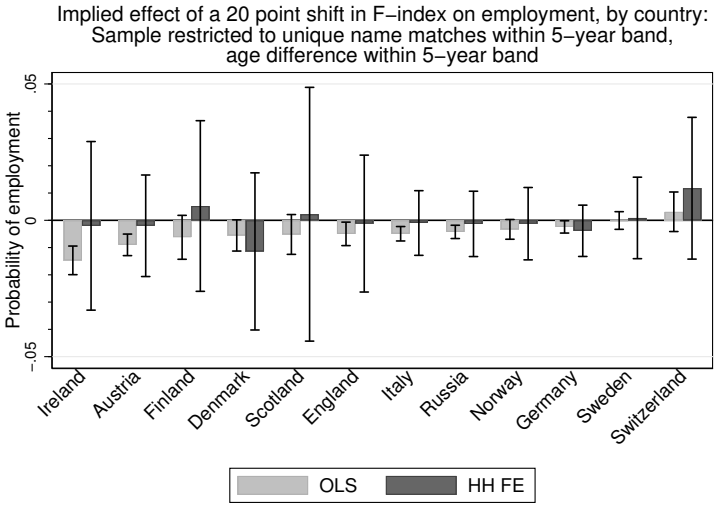
(e)

Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 3-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.8: Results by country: sample matched using 5-year uniqueness bands, 5-year age bands



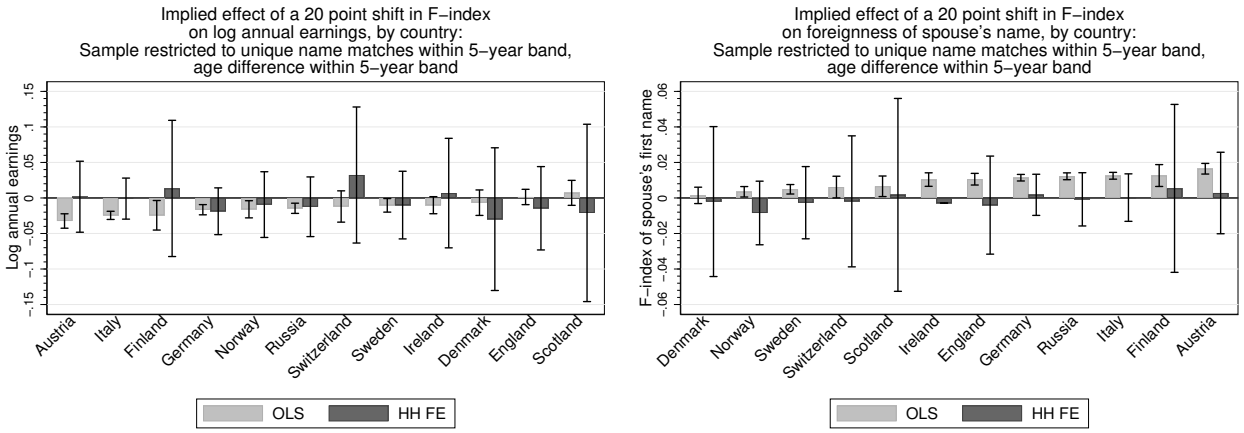
(a)



(b)

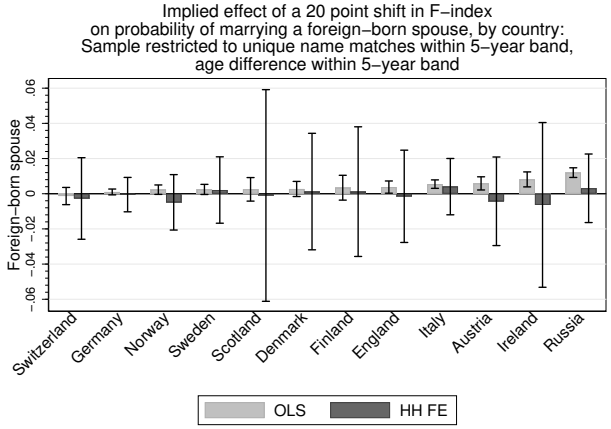
Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.8: Results by country: sample matched using 5-year uniqueness bands, 5-year age bands (cont.)



(c)

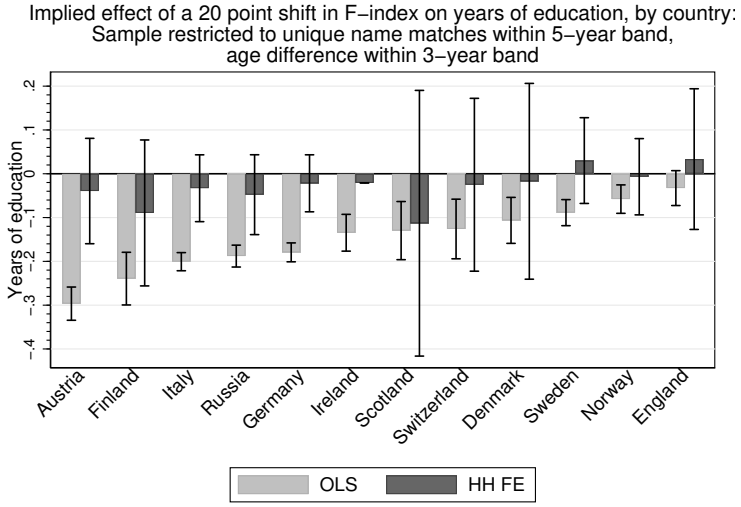
(d)



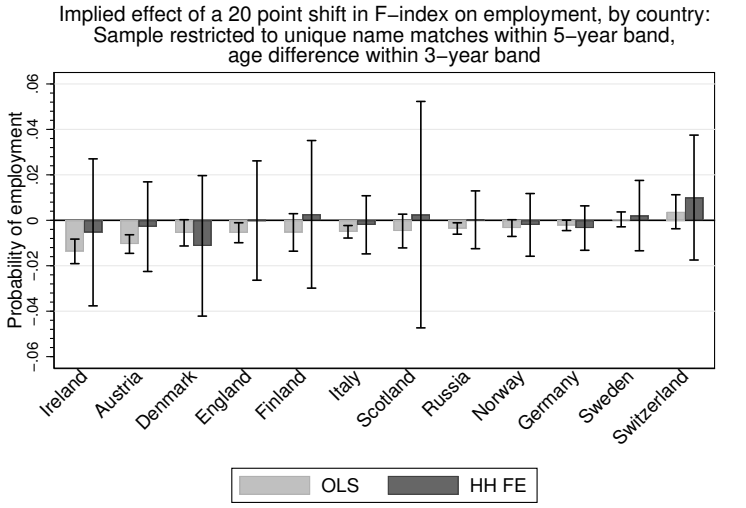
(e)

Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 2 years on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.9: Results by country: sample matched using 5-year uniqueness bands, 3-year age bands



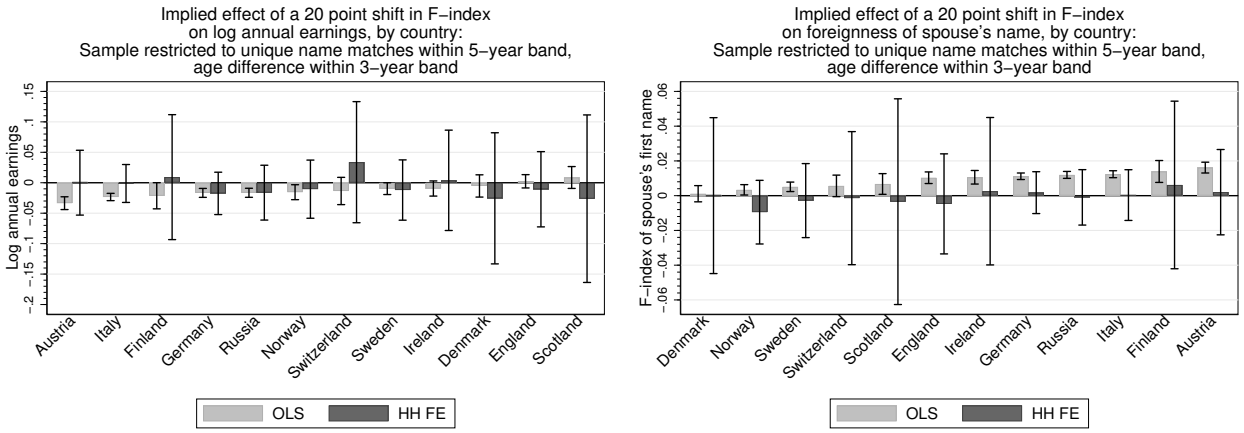
(a)



(b)

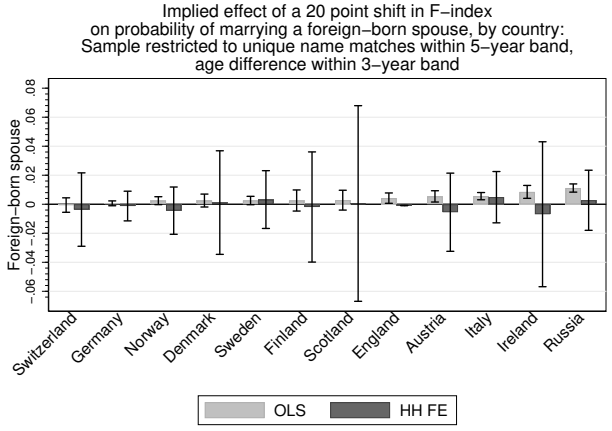
Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Figure 3.A.9: Results by country: sample matched using 5-year uniqueness bands, 3-year age bands (cont.)



(c)

(d)



(e)

Notes: In these figures we estimate regressions of each outcome listed on the foreignness index of an individual's name for men from countries with more than 1000 observations at the household level, using a more conservative linking algorithm. We report 0.2x the coefficient value of an individual's F-index calculated for his year of birth. All regressions control for a vector of dummies for child's age in 1920, parental years in the US, and child's rank in the birth order. Sample includes native-born men matched between 1920 and 1940 complete-count Censuses using the ABE-conservative method, requiring names to be unique within a 5-year band, with matched pairs allowed to differ in reported age by up to 1 year on either side. Men must be white, aged 3-15 in 1920, born outside the South, and living at home with parents in 1920 in a household whose head was foreign-born.

Bibliography

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2012. "Europe's Tired, Poor, Huddled Masses: Self-selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review* 102 (5): 1832-56.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2014. "A Nation of Immigrants: Assimilation and Economic Outcomes in the Age of Mass Migration." Journal of Political Economy 122 (3): 467-506.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2016. "Cultural Assimilation during the Age of Mass Migration." NBER Working Paper 22381.

Abramitzky, Ran, Leah Platt Boustan, and Katherine Eriksson. 2020. "Do Immigrants Assimilate More Slowly Today Than in the Past?" *American Economic Review: Insights* 2 (1): 125-141.

Abramitzky, Ran, Leah Platt Boustan, Katherine Eriksson, James Feigenbaum, Santiago Perez. 2019. "Automated Linking of Historical Data." NBER Working Paper 25825.

Akcigit, Ufuk, John Grigsby, and Tom Nicholas. 2017. "Immigration and the Rise of American Ingenuity." *American Economic Review* 107 (5): 327-31.

Arai, Mahmood, and Peter Skogman Thoursie. 2009. "Renouncing Personal Names: An Empirical Examination of Surname Change and Earnings." *Journal of Labor Economics* 27 (1): 127-147.

Bagger, Jesper, Javier A. Birchenall, Hani Mansour, and Sergio Urzúa. 2021. "Education, Birth Order, and Family Size." *The Economic Journal* 131 (633): 33-69.

Barclay, Kieron J. 2018. "The Birth Order Paradox: Sibling Differences in Educational Attainment." *Research in Social Stratification and Mobility* 54: 56-65.

Becker, Gary. 1981. *A Treatise on the Family.* Cambridge, Massachusetts: Harvard University Press.

Behrman, Jere, and Paul Taubman. 1986. "Birth Order, Schooling, and Earnings." *Journal of Labor Economics* 4 (3): 121-145.

Bernstein, Shai, Rebecca Diamond, Abhisit Jiranaphawiboon, Tim McQuade, and Beatriz Pousada. 2022. "The Contribution of High-Skilled Immigrants to Innovation in the United States." Working paper.

Bertrand, Marianne, and Sendhil Mullainathan. 2004. "Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination." *American Economic Review* 94 (4): 991-1013.

Biavaschi, Costanza, Corrado Giulietti, and Zahra Siddique. 2017. "The Economic Payoff of Name Americanization." *Journal of Labor Economics* 35 (4): 1089-1116.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2005. "The More the Merrier? The Effect of Family Size and Birth Order on Children's Education." *The Quarterly Journal of Economics* 120 (2): 669-700.

Black, Sandra E., Paul J. Devereux, and Kjell G. Salvanes. 2011. "Older and Wiser? Birth Order and IQ of Young Men." *CESifo Economic Studies* 57 (1): 103-120.

Blake, Judith. 1981. "Family Size and the Quality of Children." *Demography* 18 (4): 421-442.

Blake, Judith. 1989. *Family Size and the Achievement.* Berkeley, California: University of California Press.

Booth, Alison L., and Hiau Joo Kee. 2009. "Birth Order Matters: The Effect of Family Size and Birth Order on Educational Attainment." *Journal of Population Economics* 22 (2): 367-397.

Borjas, George J., and Kirk B. Doran. 2012. "The Collapse of the Soviet Union and the Productivity of American Mathematicians." *Quarterly Journal of Economics* 127 (3): 1143–1203.

Botzet, Laura J., Julia M. Rohrer, and Ruben C. Arslan. 2021. "Analysing Effects of Birth Order on Intelligence, Educational Attainment, Big Five and Risk Aversion in an Indonesian Sample." *European Journal of Personality* 35 (2): 234–248.

Bureau of Labor Statistics. 2019. "National Longitudinal Survey of Youth 1979 cohort, 1979-2016 (rounds 1-27)." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Bureau of Labor Statistics. 2004. "National Longitudinal Survey of Mature Women, 1967-2003 (rounds 1-22)." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Bureau of Labor Statistics. 2004. "National Longitudinal Survey of Young Women, 1968-2003 (rounds 1-23)." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Bureau of Labor Statistics. 1999. "National Longitudinal Survey of Older Men, 1967-1990 (rounds 1-13)." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Bureau of Labor Statistics. 1992. "National Longitudinal Survey of Young Men, 1966-1981 (rounds 1-12)." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Bureau of Labor Statistics. 2019. "Children of the NLSY79, 1979-2018." United States Department of Labor, Center for Human Resource Research (CHRR), Ohio State University [distributor]. Columbus, OH (accessed July 13, 2021).

Carneiro, Pedro, Sokbae Lee and Hugo Reis. 2015. "Please Call me John: Name Choice and the Assimilation of Immigrants in the United States, 1900-30." Institute for Fiscal Studies Working Paper.

Conley, Dalton, and Rebecca Glauber. 2006. "Parental Educational Investment and Children's Academic Risk: Estimates of the Impact of Sibship Size and Birth Order from Exogenous Variation in Fertility." *Journal of Human Resources* 41 (4): 722-737.

Cook, Lisa D., Trevon D. Logan and John M. Parman. 2016. "The Mortality Consequences of Distinctively Black Names." *Explorations in Economic History* 59: 114-125.

Dammert, Ana. 2010. "Siblings, Child Labor, and Schooling in Nicaragua and Guatemala." *Journal of Population Economics* 23 (1): 199-224.

De Haan, Monique. 2010. "Birth Order, Family Size and Educational Attainment." *Economics of Education Review* 29 (4): 576-588.

De Haan, Monique, Erik Plug, and José Rosero. 2014. "Birth Order and Human Capital Development: Evidence from Ecuador." *Journal of Human Resources* 49 (2): 359-392.

Dimmock, Stephen G., Jiekun Huang, and Scott J. Weisbenner. 2021. "Give Me Your Tired, Your Poor, Your High-Skilled Labor: H-1B Lottery Outcomes and Entrepreneurial Success." *Management Science* 68 (9): 6950-6970.

Doran, Kirk, Alexander Gelber, and Adam Isen. 2022. "The Effects of High-Skill Immigration Policy on Firms: Evidence from Visa Lotteries." *Journal of Political Economy* 130 (10): 2501-2770.

Edmonds, Eric V. 2006. "Understanding Sibling Differences in Child Labor." *Journal of Population Economics* 19 (4): 795–821.

Emerson, Patrick M., and André Portela Souza. 2002. "Birth Order, Child Labor and School Attendance in Brazil." Vanderbilt University Department of Economics Working Paper 0212.

Emerson, Patrick M., and André Portela Souza. 2008. "Birth Order, Child Labor, and School Attendance in Brazil." *World Development* 36 (9): 1647-1664.

Eriksson, Katherine and Zachary Ward. 2019. "The Residential Segregation of Immigrants in the United States from 1850 to 1940." *Journal of Economic History* 79 (4): 989-1026.

Esposito, Lucio, Sunil M. Kumar, and Adrián Villaseñor. 2020. "The Importance of Being Earliest: Birth Order and Educational Outcomes Along the Socioeconomic Ladder in Mexico." *Journal of Population Economics* 33 (1): 1069–1099.

Figlio, David N. 2005. "Names, Expectations and the Black-White Test Score Gap." NBER Working Paper 11195.

Finn, Michael G. 2010. "Stay Rates of Foreign Doctorate Recipients from U.S. Universities, 2007."

Fryer, Roland G., and Steven Levitt. 2004. "The Causes and Consequences of Distinctively Black Names." *Quarterly Journal of Economics* 119 (3): 767-805.

Grätz, Michael, Kieron J. Barclay, Øyvind N. Wiborg, Torkild H. Lyngstad, Aleksi Karhula, Jani Erola, Patrick Präg, Thomas Laidley, and Dalton Conley. 2021. "Sibling Similarity in Education Across and Within Societies." *Demography* 58 (3): 1011–1037.

Goldin, Claudia. 1998. "America's Graduation from High School: The Evolution and Spread of Secondary Schooling in the Twentieth Century." *The Journal of Economic History* 58 (2): 345-374.

Goldin, Claudia, and Lawrence F. Katz. 2008. "Mass Secondary Schooling and the State: The Role of State Compulsion in the High School Movement." In *Understanding Long-Run Economic Growth: Geography, Institutions, and the Knowledge Economy*, edited by Dora L. Costa and Naomi R. Lamoreaux, 275-310. Chicago, Illinois: University of Chicago Press.

Goldstein, Joshua R. and Guy Stecklov. 2015. "From Patrick to John F.: Ethnic Names and Occupational Success in the Last Era of Mass Migration." *American Sociological Review* 81 (1): 85-106.

Hall, Bronwyn H., Adam Jaffe, and Manuel Trajtenberg. 2001. "The NBER Patent Citations Data File: Lessons, Insights and Methodological Tools." Working paper.

Handy, Christopher, and Katharine L. Shester. 2022. "Birth Order and the Decline in College Completion Among the Baby Boom Generation." *Economic Inquiry* 60 (4): 1626-1643.

Hanushek, Eric A. 1992. "The Trade-off Between Child Quantity and Quality." *Journal of Political Economy* 100 (1): 84-117.

Härkönen, Juho. 2014. "Birth Order Effects on Educational Attainment and Educational Transitions in West Germany." *European Sociological Review* 30 (2): 166-179.

Hathaway, Ian. "Almost Half of Fortune 500 Companies Were Founded by American Immigrants or Their Children." Brookings. December 4, 2017. https://www.brookings.edu/blog/the-avenue/2017/12/04/almost-half-of-fortune-500-companies-were-founded-by-american-immigrants-or-their-children/

Hauser, Robert M. and William H. Sewell. 1985. "Birth Order and Educational Attainment in Full Sibships." *American Educational Research Journal* 22 (1): 1-23.

Hauser, Robert M., William H. Sewell, and Pamela Herd. 2020. "Wisconsin Longitudinal Study: 1957-2020, Version 13.08." University of Wisconsin-Madison. Madison, WI. http://www.ssc.wisc.edu/wlsresearch/documentation/ (accessed February 22, 2021).

Havari, Enkelejda and Marco Savegnago. 2022. "The Intergenerational Effects of Birth Order on Education." *Journal of Population Economics* 35 (1): 349–377.

Hotz, V. Joseph, and Juan Pantano. 2015. "Strategic Parenting, Birth Order, and School Performance." *Journal of Population Economics* 28 (4): 911–936.

Hunt, Jennifer. 2011. "Which Immigrants Are Most Innovative and Entrepreneurial? Distinctions by Entry Visa." *Journal of Labor Economics* 29 (3): 417–457.

Hunt, Jennifer, and Marjolaine Gauthier-Loiselle. 2010. "How Much Does Immigration Boost Innovation?" *American Economic Journal: Macroeconomics* 2 (2): 31–56.

Iacovou, Maria. 2008. "Family Size, Birth Order, and Educational Attainment." *Marriage and Family Review* 42 (3): 35-57.

Isungset, Martin A., Mats Lillehagen, and Elisabeth Ugreninov. 2020. "One Order Fits All? Birth Order and Education in Immigrant Families." *European Sociological Review* 36 (1): 16–31.

Jensen, Robert. 2010. "The (Perceived) Returns to Education and the Demand for Schooling." *Quarterly Journal of Economics* 125 (2): 515–548.

Kantarevic, Jasmin, and Stéphane Mechoulan. 2006. "Birth Order, Educational Attainment, and Earnings: An Investigation Using the PSID." *Journal of Human Resources* 41 (4): 755-777.

Kerr, William R. 2008. "The Ethnic Composition of U.S. Inventors." Harvard Business School Working Paper 08-006.

Kerr, William R., and William F. Lincoln. 2010. "The Supply Side of Innovation: H-1B Visa Reforms and U.S. Ethnic Invention." *Journal of Labor Economics* 28 (3): 473–508.

Khanam, Rasheda, and Mohammad M. Rahman. 2007. "Child Work and Schooling in Bangladesh: The Role of Birth Order." *Journal of Biosocial Science* 39: 641-656.

Kim, Young-Joo. 2020. "Born to be More Educated? Birth Order and Schooling." *Review of Economics of the Household* 18 (2): 165–180

Kuba, Radim, Jaroslav Flegr, and Jan Havlíček. 2018. "The Effect of Birth Order on the Probability of University Enrolment." *Intelligence* 70: 61-72.

Kumar, Santosh. 2016. "The Effect of Birth Order on Schooling in India." *Applied Economics Letters* 23 (18): 1325-1328.

Lan, Xiaohuan. 2012. "Permanent Visas and Temporary Jobs: Evidence from Postdoctoral Participation of Foreign PhDs in the United States." *Journal of Policy Analysis and Management* 31 (3): 623-640

Lan, Xiaohuan. 2013. "The Effects of Green Cards on the Wages and Innovations of New PhDs." *Journal of Policy Analysis and Management* 32 (4): 807-834.

Lehmann, Jee-Yeon K., Ana Nuevo-Chiquero, and Marian Vidal-Fernandez. 2018. "The Early Origins of Birth Order Differences in Children's Outcomes and Parental Behavior." *Journal of Human Resources* 53 (1): 123-156.

Lindert, Peter H. 1977. "Sibling Position and Achievement." *Journal of Human Resources* 12 (2): 198-219.

Lleras-Muney, Adriana. 2002. "Were Compulsory Attendance and Child Labor Laws Effective: An Analysis from 1915 to 1939." *Journal of Law and Economics* 45 (2): 401-435.

Marco, Alan C., Michael Carley, Steven Jackson, and Amanda F. Myers. 2015. "The USPTO Historical Patent Data Files: Two Centuries of Innovation." SSRN Working Paper 2015-1.

Mechoulan, Stéphane, and François-Charles Wolff. 2015. "Intra-Household Allocation of Family Resources and Birth Order: Evidence from France Using Siblings Data." *Journal of Population Economics* 28: 937-964.

Monfardini, Chiara, and Sarah G. See. 2016. "Birth Order and Child Cognitive Outcomes: an Exploration of the Parental Time Mechanism." *Education Economics* 24 (5): 481-495.

Moser, Petra. 2012. "Taste-based Discrimination Evidence from a Shift in Ethnic Preferences after WWI." *Explorations in Economic History* 49 (2): 167-188.

Moser, Petra, and Shmuel San. 2020. "Immigration, Science, and Invention: Lessons from the Quota Acts." Working paper.

Moser, Petra, Alessandra Voena, and Fabian Waldinger. 2014. "German-Jewish Emigres and U.S. Invention." *American Economic Review* 104 (10): 3222–55.

Moshoeshoe, Ramaele. 2019. "Birth Order Effects on Educational Attainment: Evidence from Lesotho." *Education Economics* 27 (4): 401-424.

Oliveira, Jaqueline. 2019."Birth Order and the Gender Gap in Educational Attainment." *Review of Economics of the Household* 17 (3): 775-803.

Oreopoulos, Philip. 2011. "Why Do Skilled Immigrants Struggle in the Labor Market? A Field Experiment with Thirteen Thousand Resumes." *American Economic Journal: Economic Policy* 3 (4): 148-171.

Orrenius, Pia, Madeline Zavodny, and Emily Kerr. 2012. "Chinese Immigrants in the U.S. Labor Market: Effects of Post-Tiananmen Immigration Policy." *International Migration Review* 46 (2): 456–482.

Peri, Giovanni, Kevin Shih, and Chad Sparber. 2015. "STEM Workers, H-1B Visas, and Productivity in U.S. Cities." *Journal of Labor Economics* 33 (S1): S1-S318.

Price, Joseph. 2008. "Parent-Child Quality Time: Does Birth Order Matter?" *The Journal of Human Resources* 43 (1): 240-265.

Rammohan, Anu, and Diane Dancer. 2008. "Gender Differences in Intrahousehold Schooling Outcomes: The Role of Sibling Characteristics and Birth-Order Effects." *Education Economics* 16 (2): 111-126.

Rubinstein, Yona, and Dror Brenner. 2014. "Pride and Prejudice: Using Ethnic-Sounding Names and Inter-Ethnic Marriages to Identify Labour Market Discrimination." *Review of Economic Studies* 81 (1): 389-425.

Ruggles, Stephen, Catherine A. Fitch, Ronald Goeken, J. David Hacker, Matt A. Nelson, Evan Roberts, Megan Schouweiler, and Matthew Sobek. 2021. "IPUMS Ancestry Full Count Data: Version 3.0." IPUMS. Minneapolis, MN (accessed April 3, 2022).

Seid, Yared, and Shiferaw Gurmu. 2015. "The Role of Birth Order in Child Labour and Schooling." *Applied Economics* 47 (49): 5262-5281.

Steinhilber, August W., and Carl J. Sokolowski. 1966. *State Law on Compulsory Attendance.* Washington, DC: U.S. Department of Health, Education, and Welfare, Office of Education.

Stephens Jr., Melvin, and Dou-Yan Yang. 2014. "Compulsory Education and the Benefits of Schooling." *The American Economic Review* 104 (6): 1777-1796.

Stuen, Eric T., Ahmed Mushfiq Mobarak, and Keith E. Maskus. 2012. "Skilled Immigration and Innovation: Evidence from Enrollment Fluctuations in U.S. Doctoral Programs." *Economic Journal* 122 (565): 1143-1176.

Stuen, Eric T., Ahmed Mushfiq Mobarak, and Keith E. Maskus. 2013. "Doctoral Students and U.S. Immigration Policy." *Science* 342 (6158): 562-563.

Su, Yaqin, Cynthia Bansak, and Cheng Cheng. 2019. "The Inter-Generational Value of a Green Card: Evidence from the CSPA of 1992." *Applied Economics Letters* 26 (20): 1713-1717.

Tan, Hui Ren. 2019. "More Is Less? The Impact of Family Size on Education Outcomes in the United States, 1850-1940." *Journal of Human Resources* 54 (4): 1154-1181.

Tenikue, Michel, and Bertrand Verheyden. 2010. "Birth Order and Schooling: Theory and Evidence from Twelve Sub-Saharan Countries." *Journal of African Economics* 19 (4): 459-495.

Wang, Weidong, Xiaohong Liu, Yongqing Dong, Yunli Bai, Shukun Wang, and Linxiu Zhang. 2020. "Son Preference, Eldest Son Preference, and Educational Attainment: Evidence From Chinese Families." *Journal of Family Issues* 41 (5): 636-666.

Weng, Qian, Xia Gao, Haoran He, and Shi Li. 2019. "Family Size, Birth Order and Educational Attainment: Evidence from China." *China Economic Review* 57: 1-23.

Zajonc, Robert B. 1976. "Family Configuration and Intelligence." *Science* 192: 227-236.

Zajonc, Robert B., and Gregory B. Markus. 1975. "Birth Order and Intellectual Development." *Psychological Review* 82 (1): 74-88.