Essays on Generational Economic Links Between Childhood and Adulthood

by

Connor P. Cole

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Economics)
in the University of Michigan
2021

Doctoral Committee:

Professor Martha Bailey, University of California Los Angeles, Co-Chair
Professor Charles Brown, Co-Chair
Professor James Hines
Professor Brian Jacob

Connor P. Cole

colecp@umich.edu

ORCID iD: 0000-0002-0142-1742

Dedicated to Terrence Cole

For the years of inspiration, insight and wisdom taught by word and action

“The philosophers in their long beards and short cloaks, who esteem themselves the only favourites of wisdom .... alas nature does but laugh at all their puny conjectures; they never yet made one considerable discovery, as appears in that they are unanimously agreed in no one point of the smallest moment; nothing so plain or evident but what by some or other is opposed and contradicted.” - *In Praise of Folly*

ACKNOWLEDGMENTS

In his *Autobiography,* John Stuart Mill includes the following comment about the balance between his and his collaborators' contributions to his work:

> Whoever, either now or hereafter, may think of me and of the work I have done, must never forget that it is the product not of one intellect and conscience, but of three, the least considerable of whom, and above all the least original, is the one whose name is attached to it.

In context, he is specifically thanking his wife and step-daughter for their stimulating discussions and intellectual collaborations with him, but I think the spirit of this comment rings true for any researcher's work, especially my experiences with this dissertation.

This research is indebted first and foremost to my advisers while I was in graduate school. Professor Martha Bailey collaborated with me on several projects, and taught me that daring to go down unexpected and unusual paths in research was a risk worth taking. Through our years of collaborating, I've grown better at searching for the mix of interesting data, creativity in approach, and thoroughness of execution that I admire in her work and hope to emulate in my own. Professor Charles Brown taught me the importance of knowledge of data. He is a font of institutional expertise about surveys, and I learned from him that understanding the process that creates data is as important as communicating the results found in the data. Professor James Hines, with his detailed understanding of the more labyrinthine sections of the U.S. tax code, demonstrated the importance of excellent knowledge of institutional details of policy. He also encouraged me to think more deeply about the link between theory and mechanisms with the results I've found in my research. ”Big data” empirical research is often predicated on letting the data speak for itself, but those aggregate results come from adding up individual decision-making and behavior. Puzzling through how that decision-making works is a necessary step for contextualizing results. Professor Brian Jacob showed me the importance of concision of exposition in papers with his thoughtful editing. Good empirical papers tend to have more than they show upfront, although the details are there for those who need them.

I am also thankful for the many discussions with my classmates and peers. Years of chatting with Giacomo Brusco, Benjamin Glass, Luis Baldomero Quintana, Brenden Tiempe, Arthur Sellers, and Tejaswi Velayudhan offered dozens of inspiring ideas for projects and insightful thoughts about interpreting results. I look forward to many years of collaboration in the future.

I would also like to thank Jack Carter and Joelle Abramowitz at the Michigan Federal Research Data Center. They helped me with the difficult process of securing disclosure of my results from the Census Bureau. While large datasets like the U.S. Census hold great promise for research, any person using these data must take the responsibility of data stewardship seriously to preserve public trust. In making sure that I was cautious and thorough, they ensured that my released results did not inadvertently disclose private information from the U.S. Census.

I also thank my parents Dermot Cole and Debbie Carter. Given how frequently I remember their help when I pestered them about schoolwork in grade school, I am satisfied to have produced something that I will be obligated to explain to them in turn.

Lastly, I would also like to thank my uncle Terrence Cole, to whom I have dedicated this dissertation. As a professor of history, he was a major inspiration from my earliest days in grade school, as well as one of the first people who introduced me to economics as an academic discipline. While I am saddened that he could not be here to celebrate this achievement with me, I am thankful for the many years of laughs, good advice, bad advice, and high spirits he shared with me.

While all the inspiration that may be found in this research is something I gladly share with those named here and hundreds of others unnamed, all of the mistakes are my own.

PREFACE

The following essays are loosely organized around the theme of generational economic links between childhood and adulthood. This field is an active and rapidly expanding domain in empirical research, given that there is an intuitive link between people's early life experiences and their later life outcomes. While these early life experiences are not determinative, they play an important role in physical and emotional development, as discussed in the introduction to the first chapter below. Better understanding these linkages may show how adverse outcomes in adulthood are predicted by specific circumstances in childhood, and help policy-makers better target interventions in childhood that prevent those adverse outcomes later in life.

While this link seems intuitive, quantifying it remains a difficult task for two main reasons. First, the limited time horizon of most data resources makes it difficult to measure how circumstances in childhood link up to adulthood. Second, it is often difficult to identify treatments and treated populations for causal analysis. With some exceptions, it is difficult to find clearly defined treatments that affected one group compared to another. For example, researchers have demonstrated a correlational income gradient in many metrics of child development, where children from higher income families often do better on metrics of child development. However, sources of experimental variation in income are few, so it remains an open question how much of this relationship is causal.

Because of these difficulties, future research using newly unlocked large-scale administrative data and restricted government data hold great promise. These data resources, along with increased computing power, allow researchers to calculate revealing large-scale correlations. Additionally, with these more detailed datasets, researchers can better calculate eligibility for programs with more finely grained data on place of residence and family economic conditions. For example, many benefits to families with children are means-tested. Depending on how families react to those cutoffs, they may allow opportunities for research leveraging the program eligibility rules. This dissertation's first chapter offers new evidence from one such eligibility cutoff, and offers strategies for econometrically dealing with endogenous sorting around similar cutoffs.

However, doing careful research with such resources will depend on producing quality data. As using data for these kinds of long-term analyses often requires linking one resource with another (e.g. school records with earnings records), the quality of the linking may impact empirical results. Therefore, doing careful linking is necessary to ensure these resources are used to their full potential. The second and third chapters of this dissertation explore methodological issues related to linking, showing that different linking procedures themselves may alter measured results. They also offer practical advice on how to improve the quality of linking.

In terms of future research, I believe the discontinuity in after-tax income I examine in my first chapter offers ripe variation for further examination. Any dataset that includes child date of birth, outcomes among children, and information on the economic conditions of parents allows a potential setting for looking at the effect of this income shock. Furthermore, as I suggest, tracking discontinuities in outcomes among children who were born after 2000 may be promising for researchers, as the size of the discontinuity in after-tax income at birth has only increased with time. In general, I believe more research looking at links between the conditions children face growing up and their outcomes as adults will be fruitful. Hendren and Sprung-Keyser (2020) show that, on average, interventions among children show more "bang per buck" in social outcomes than interventions among adults. These results, together with the findings in my first chapter, suggest that studying more potential interventions among children may find similarly sized effects. Future research distinguishing between the effectiveness of different types of interventions, and the relative impact at different ages among children will be of great interest to policy makers. I hope this dissertation can add to this ongoing research, and inspire further research for the future.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF TABLES

Table

LIST OF APPENDICES

Appendix

ABSTRACT

My dissertation examines the economic links between people's experiences in early and later-life. It offers new empirical evidence on the effect of income in infancy on later-life outcomes, and investigates the performance and econometric properties of the linking tools often used to create data for these long-term empirical investigations.

In my first chapter, I estimate a relationship between family income in infancy and later-life outcomes for children. Eligibility for child-related tax benefits depends on the calendar year in which a child is born. Families with children born in December are eligible for tax benefits a year earlier than families with children born a few days later in January. These differences create a discontinuity in after-tax income in infancy worth on average approximately $2,000 for families in tax year 2016. I use regression discontinuity techniques to calculate the effect of this change in after-tax income on outcomes for children and young adults in Census data. Evidence show that a $1,000 increase in after-tax income in infancy results in a 1.2 percentage point increase in the probability of a student being grade-for-age by high school, a basic indicator of academic achievement and social maturity. Effects of this income shock are larger for children from families that are more likely disadvantaged at a child's birth, including Black families, and families with low education attainment. After high school, small differences in labor force attachment, earnings and education attainment persist for the adults who experienced the income increase as children. These effects are again pronounced for Black adults and adults born in counties with low average education attainment.

In my second and third chapters, I investigate methodological problems that arise when linking data. Linking is often necessary to investigate generational economic links between childhood and adulthood.

In the second chapter, my coauthors Martha Bailey, Catherine Massey, Morgan Henderson and I review the literature in historical record linkage in the U.S., and examine the performance of widely-used automated record linking algorithms. Focusing on algorithms in current practice, our findings highlight the important effects of linking methods on data quality. We then extend our analysis to look at the consequences of these differences in data quality on inference by computing intergenerational income elasticities between fathers and sons. Many of the methods produce estimated elasticities that are statistically distinguishable from the estimated intergenerational elasticity with hand-linked data, suggesting that the linking algorithms themselves may bias inference. However, eliminating false matches renders elasticity estimates similar to each other, and statistically indistinguishable from the elasticity estimated with the hand-linked data.

In the third chapter, my coauthors Martha Bailey, Catherine Massey and I investigate two complementary strategies to address the issues we highlight in my second chapter. We investigate the use of validation variables to identify higher quality links and a regression-based weighting procedure to increase the representativeness of custom research samples. We demonstrate the potential value of these strategies using the 1850-1930 Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS). We show that, while incorrect linking rates appear low in the IPUMS-LRS, researchers can reduce error rates further using validation variables. We also show researchers can reweight linked samples to balance observed characteristics in the linked sample with those in a reference population using a simple regression-based procedure.

CHAPTER I

Effects of Family Income in Infancy on Child and Adult Outcomes: New Evidence Using Census Data and Tax Discontinuities

Researchers are finding growing evidence of sustained relationships between family economic resources in childhood and later life outcomes. Descriptive research from the U.S. shows that children from lower-income families are at higher risk of poor physical health as children (Case, Lubotsky and Paxson, 2002; Currie, 2009), more likely to perform worse in school (Michelmore and Dynarski, 2017; Reardon, 2011), and less likely to graduate high school (Stark, Noel and McFarland, 2012; Autor et al., 2019). These differences persist into adulthood, as disadvantaged children are less likely to earn college degrees (Bailey and Dynarski, 2011), more likely to have experiences in the criminal justice system, including incarceration (Chetty et al., 2019), more likely to have lower earnings (Chetty et al., 2014*a*) and more likely to have reduced longevity (Ferrie and Rolf, 2011).

The causal mechanisms underlying these relationships are an active field of study, as family income is correlated with unobservable determinants of outcomes for children. Research show that changes in permanent family income can have pronounced impacts on children from lower-income families (Akee et al., 2010; Loken, Mogstad and Wiswall, 2012; Shea, 2000; Chevalier et al., 2013; Bastian and Michelmore, 2018), although permanent income changes produced by specific transfer programs may have smaller effects (Jacob, Kapustin and Ludwig, 2014). In comparison, research on the effects of transitory changes in family income offers more mixed conclusions. Some papers find that changes in transitory family income have short-term impacts on performance of school students (Dahl and Lochner, 2012; Chetty, Friedman and Rockoff, 2011), some papers find long-term impacts (Black et al., 2014), and some papers find neither short nor long-term impacts (Cesarini et al., 2016). One critical topic left largely unaddressed in this evidence is the long-term effect of modest changes in transitory income in infancy on outcomes for children. Research suggests that conditions in infancy and early childhood may be consequential for long-term patterns of child development, so it is possible that effects could be strong at these early ages (Cunha et al., 2006; Duncan, Ludwig and Magnuson, 2011; Currie and Almond, 2011). If impacts are stronger at different ages, such a finding has consequences for transfer policy design. Most transfer policies in the U.S. are not child age-specific, and differences in impacts by age would suggest that increasing benefits at certain ages and decreasing them in others may be a low cost reform that improves outcomes for children.

This paper addresses this gap in the literature by analyzing the effect of a shock to transitory family income that happens in the first year of a child's life. If a child is born before New Year's Day, that child's family is eligible for tax benefits for that child one year earlier than if a child is born after New Year's. This discontinuity in tax policy means that the parents of children born one day earlier have larger after-tax income in the first year of a child's life. The increase in income is modest but non-trivial, worth about $2,000 on average in tax year 2016, and resulting in an average 5% increase in after-tax income. Furthermore, this increase is experienced by a broad share of families, so its effects may be analyzed and compared for families with different income levels. Note that this increase is a speeding up of the tax credit and deduction process for a child, as the families with children born in December, several years later, will be eligible for tax benefits for one year less than families of children born in January. Thus, the cost to the government of this increase in after-tax income comes from just altering the timing of the tax benefits and moving them from a child's later adolescence to infancy.

This research setting is closest to the work in Black et al. (2014) and Bastian and Michelmore (2018). Both of these papers analyze the long-term effects of income shocks from tax policy that happen early in life. Black et al. (2014) find that a $1,700 tax credit income transfer to a child's family at age 5 has effects on student achievement 10 years later. Bastian and Michelmore (2018) use implementation of state Earned Income Tax Credit programs, and conclude that increases in income in ages 0-4 have no detectable effects on high school graduation status and earnings in adulthood. This paper builds on these results with new evidence from a different research setting. Compared to Black et al. (2014), this paper looks at the effects of an income shock that happens even earlier in life, and extends analysis to effects on later life outcomes after school. Compared to Bastian and Michelmore (2018), this research looks at changes in income that reflect transitory income alone, and has more power to distinguish heterogeneous effects at different income levels.[1]

This paper calculates the effect of the shock in after-tax income around the New Year using a regression discontinuity design with date of birth as a running variable. Endogenous birth timing around the New Year is a threat to identification, and this paper accounts for this issue by omitting from the estimation process a region of observations around the New Year. This omitted region is identified using bunching estimation techniques (Chetty et al., 2011; Kleven and Waseem, 2013; Saez, 2010).

[1] The introduction of a state Earned Income Tax Credit program would impact earnings of families for years into the future and may change labor supply incentives. Hence, the results in Bastian and Michelmore (2018) are best interpreted as a mixture of changes in transitory income and permanent income.

Three assumptions are sufficient for this strategy to identify the causal effect of this boost in after-tax income on later life outcomes. First, no other treatments must coincide with the passing of the New Year. Second, the region affected by endogenous birth timing must be consistently identified using the omitted region estimation technique. Third, the evolution of an outcome must be consistently estimated using extrapolation through the omitted region.

Results show that this change in income in infancy has impacts on a child being grade-for-age by high school. Students are grade-for-age if they are in the school grade they would be in had they entered kindergarten or first grade on or before the year they were eligible to enter those grades, and if they progressed through school without ever repeating a grade.[2] Being grade-for-age is an indication that a student has met academic standards and shown social maturity in school (Xia and Kirby, 2009), so improvements in the share of students grade-for-age indicate multi-dimensional improvements in student development. Consistent with validity of the research design, there is no discontinuity in pre-school attendance and kindergarten entrance around the New Year. Children born before the New Year, who experience the increase in after-tax income, enter pre-school and kindergarten at roughly the same rate as the children born after the New Year, who do not experience it.[3] However, by the time students reach high school, students born before the New Year who experienced the increase in family income are approximately 1.1 percentage points more likely to be grade-for-age than students born after the New Year who did not. This finding is robust to a variety of checks, including restricting to students who live in their

[2] Most school systems define grade-for-age status starting from the first year a child entered kindergarten or 1st grade. As these entrance dates are not observable in Census data, this definition is the closest analogue.

[3] The claim that this result is consistent with the validity of the research design will be described in more detail later. Technically, there could be gaps that open up in this measure early on either because the grade-for-age status calculation is incorrect (which would suggest that the research set-up is flawed), or because parents want to hold back their children early on before they enter school (which would still be valid with the research design, but is more difficult to interpret). Since there is no detectable gap either way, it suggests that both possibilities have not happened.

birth state, and dividing up the sample by birth cohort to use differences in after-tax income by birth cohort to look at effects. Reinterpreting this reduced form effect as a direct effect of income, this evidence shows that an extra $1,000 in the first year of life increases the probability of the average student being grade-for-age by high school by 1.2 percentage points.

These effects of an extra $1,000 on grade-for-age status by high school are largest for groups that had lower family income at birth, including children whose mothers have a high school degree or less, and Black children. These results are consistent with the finding in Loken, Mogstad and Wiswall (2012) that the relationship between income and child outcomes is non-linear; similarly-sized increases in income have larger effects on lower-income families and smaller effects on higher-income families.

The effects of this increase in income in infancy persist after high school. Following Kling, Liebman and Katz (2007), this paper combines income, participation in the labor force, high school degree attainment and Supplemental Nutrition Assistance Program (SNAP) receipt into a single measure of economic self-sufficiency. In the years after young adults turn 19, there are suggestive but not statistically significant discontinuities in this measure in the full sample between adults who did and did not experience the income increase as infants. However, there are larger and significant discontinuities for young Black adults and adults born in counties with comparatively lower education attainment.[4] These discontinuities in outcomes last until young adults reach their mid-20s, with the discontinuities driven by differences in high school education attainment and earned income. However, these effects fade somewhat at later ages. This evidence is consistent with income in infancy having a small effect on adult outcomes that attenuates with age as young adults gather

[4] Note that looking at adults born in counties with comparably low education attainment is a slightly different subgroup than what was looked at before, children with mothers who have education attainment of a high school degree or less. A large fraction of children move away from home in their 20s, so parent education attainment cannot be defined for them. This subgroup is an imprecise proxy necessitated by data limitations.

experience in the labor force.

These results suggest that family income in infancy has effects on child development with ramifications stretching into adulthood, especially for families more likely disadvantaged at a child's birth. Furthermore, compared to some of the previous literature looking at similarly-sized income shocks at later ages, the effects on adult outcomes here are relatively large. This finding may suggest that effects of income in infancy are larger than effects from income at later ages. Overall, these findings fit within and expand on two directions of research: research into the gaps in the development of children that open up before children enter formal schooling, and research focusing on early childhood as a critical period for development. The relatively large effects measured here suggest that transfer policies aimed at families with young children may have substantial long-term benefits. As these effects come from altering the timing of tax benefits from adolescence to infancy, refocusing transfer benefits on earlier periods of life may offer a low-cost way of increasing such transfers to improve outcomes for children.

1.1 Data

The data used in this paper come from three sources: the long form sample of the 2000 Census, the 2001-2016 American Community Survey (ACS), and the Current Population Survey (CPS). The paper uses the first two resources for all reduced form regressions looking at the effect of the income shock of being born before the New Year on all outcomes. It uses the third resource to estimate the discontinuity in after-tax income for having a child born before the New Year, and to analyze general patterns of grade repetition by grade.

The long form of the 2000 Census is a household survey covering 17% of the U.S. population, or approximately 22 million U.S. households (U.S. Census Bureau, 2009). It includes a wide variety of demographic and economic data, including data on levels and sources of income, household structure, labor force participation and education attainment for respondents ages three and up. The ACS is an annual survey of households. The number of households sampled varies from year to year, but since 2011 the Census Bureau has targeted approximated 3.5 million households (U.S. Census Bureau, 2014) per year. The ACS covers many questions similar to those in the 2000 Census long form, but some question definitions are different. Appendix A covers some of the differences in definitions in more detail and describes how this paper combines the questions into single measures that can be used across years. Both the ACS and the 2000 Census long form were matched to the Numident file of the Social Security Administration using a Protected Identification Key from the Census Bureau. The Numident file offers a listed place of birth for each individual, which was coded into a county of birth by researchers at the University of Michigan.

One of the key outcomes this paper looks at is whether or not a student is grade-for-age. This research assigns grade-for-age status to a student based on four pieces of information: highest grade completed (or most recent grade enrolled), the state of birth of the child, the date of birth of the child and the date on which the household responds to the survey. Many states set explicit kindergarten and 1st grade age entrance requirements that require students to be a specific age by a certain date before being eligible to enter either kindergarten or 1st grade. Comprehensive data on these state policies for kindergarten entrance were collected by Bedard and Dhuey (2012), and they generously provided their most recent data covering 1955 to 2015. This data was compiled directly from state statutes and legislative history on school entry policies, and cross-checked against a variety of other data sources. This research assigns expected completed grades to students assuming that they entered kindergarten or 1st grade in the first year that they were eligible for those grades and then progressed through all other grades sequentially without repeating a grade. A student is grade-for-age if they have completed the most recent grade that this measure lists. Note that if students drop out of high school and do not continue on to further education, then they would be counted as not being grade-for-age.

Three complications are worth noting about this measure. First, some states do not specify statewide kindergarten entrance rules and allow local school districts to set their own cutoffs. As no clear expected grade can be assigned to these individuals without more detailed data on individual school district practices, this paper drops any individuals born in these states from any further calculation. Second, some states make the eligibility cutoff January 1st or December 31st. In the years that such cutoffs are present, children born before and after the New Year would, in addition to the difference in after-tax income, also experience the treatment of different grade eligibility rules. This paper also drops these individuals from any further calculation. Lastly, there are only a handful of grades where grade-for-age status can be reliably assigned due to the nature of the grade attainment and enrollment questions in the 2000 long form Census and 2001-2007 ACS (although grade-for-age status can be reliably assigned in the 2008-2016 ACS for all grades). This issue is described more in Appendix A. The consequence of this limitation is that grade-for-age status can only be consistently calculated in pre-school, kindergarten, 1st grade, 5th grade, 7th grade, and 9th through 11th grades.

Since this paper analyzes grade-for-age status at different grades using data from 2000 to 2016, the distribution of birth cohorts included in each calculation will differ. For example, the high school grade-for-age calculations include individuals born from 1982 to 2001, but the kindergarten enrollment calculations involve individuals born 1996 to 2011. In all, results looking at grade-for-age status include children who were born from 1982 to 2011, with the exact birth cohort of children analyzed depending on the grades looked at. To ensure that analyses of outcomes for adults continue to follow these same cohorts, this paper restricts analysis to adults who were born in 1980 and later.

Thus, the sample for analysis could broadly be described as adults and children born 1980 and later in states that had statewide kindergarten entrance cutoffs away from the New Year in the year that the student would have entered kindergarten in that state.

The CPS is a monthly sample of households in the U.S.. Although sizes of samples differ by year, the current CPS samples approximately 60,000 households per month (Bureau of Labor Statistics, 2018).

1.2 Overview of Tax Policy Relating to Children

The variation that drives this paper is the discontinuity in after-tax income for families in the first year of an infant's life depending on the birth timing of the child. There are four main child-related tax benefits that parents are eligible for: a personal exemption for a dependent, the Earned Income Tax Credit (EITC), the Child Tax Credit (CTC) and the Child and Dependent Care Credit. Parents are eligible for these tax benefits for a child starting in the tax year that a child is born. So, as Figure 1.1 shows, parents with children born in December are eligible to claim child-related tax benefits in their child's first year in life. In comparison, parents of children born a few days later in January can only claim them on tax forms starting with the next year.

Figure 1.2 estimates the average discontinuity in after-tax income for having a child born before the New Year produced by these four benefits. Without access to administrative data on tax records, it is difficult to precisely calculate the value of this discontinuity, but Figure 1.2 offers the best approximation to this calculation possible with survey data from the March CPS.[5] These estimates are in line with

[5] This paper calculates this after-tax income discontinuity by using data from the March CPS in a four year radius of a given tax year, and restricting the sample to families with at least one child three years old or younger. It then assigns the family the total income from their household of residence, and treats one of those children three years old and younger as an "infant." Finally, it computes the after-tax return for the family both with and without the "infant" three years old and

calculations from administrative data. For example, this paper estimates that the average tax benefit of having a child before the New Year was $2,150 for tax filers from 2000 to 2010. LaLumia, Sallee and Turner (2015) estimate with administrative data that the same benefit over the same time period was $2,100.

Figure 1.2 shows that this discontinuity has been steadily increasing over time, rising from about $800 in 1980 to a little over $2,000 in 2016, due primarily to increased generosity of the EITC and CTC (see Appendix B). Furthermore, the discontinuity is positive for the vast majority of families. The share of parents with no change in their tax liabilities in this calculation is around 10% prior to 1994 and falls to about 6% thereafter. These parents have zero change in tax liabilities for three reasons: either they have very low income, they have already received the maximum of relevant tax credits, or they have high incomes and high deductions. Thus, the vast majority of families experience a modest increase in after-tax income.[6]

Figure 1.2 also shows average changes in after-tax income for having a child born before the New Year for two subgroups: families where a child's mother has education attainment of a high school degree or less and Black families. These are subgroups this paper will look at later, as they have lower average income at birth than families with

younger, and the difference between the two tax returns identifies the discontinuity. Ideally, this comparison would only include parents with infants born around December and January given the fact that seasonality in the patterns of birth ensure that the characteristics of parents evolve over time (Buckles and Hungerman, 2013*b*). However, the CPS data do not identify month or quarter of birth. The use of children three years old and younger as "infants" and the use of additional years of CPS data ensure more precision and have minimal effects on point estimates. More details and robustness checks for the choices in this calculation are in Appendix A.

[6] This paper, like many papers in the EITC literature that do not have access to administrative tax data, assumes 100% take-up of tax benefits to calculate the change in after-tax income produced by these tax policies (Hoynes, Miller and Simon, 2015). Take-up rates lower than 100% would mean that the true discontinuity would be lower than the discontinuity in Figure 1.2, so Figure 1.2 is best interpreted as an upper bound. While take-up is not 100%, it is still likely high. LaLumia, Sallee and Turner (2015) find that 85% to 90% of newborns born in late December are claimed on a tax return in the 2000s. To understand how different take-up patterns might affect the discontinuity in after-tax income, Appendix A describes an exercise that adjusts Figure 1.2 for a lower bound on the estimated discontinuity. This analysis suggests that the lower bound on the discontinuity in after-tax income is at most 10% to 20% lower than the upper bound recorded in Figure 1.2. The effect of this potentially lower discontinuity in after-tax income on later results is also discussed in further detail in Appendix A.

higher education attainment and White families. As is clear, the average increases in after-tax income for these groups are similar to or slightly less than the average for all families in early years. However, they gradually increase and become equal to or larger than the average over time. The fact that these discontinuities in income are relatively large for these groups reflects the fact that the EITC and to a lesser extent the CTC are aimed at lower income families. Critical to the size of these tax benefits for these families is the fact that the EITC is a refundable tax credit and the CTC is partially refundable, meaning that individuals who have low tax obligations can actually see a positive tax return from the government.[7]

Figure 1.3 presents these changes in after-tax income as percentage increases in after-tax income. The average percent increase in after-tax income is generally larger for families where the mother has a high school degree or less and for Black families than it is for all families on average.[8] In particular, the lines rapidly diverge as the generosity of the CTC and EITC ramp up in the 1990s.

As is clear in Figure 1.1, the discontinuity in after-tax income described here in infancy does not persist into the next year.[9] In the next tax filing year parents of infants born before and after the New Year will be eligible for the same tax credits and deductions. Furthermore, parents are only eligible for these tax credits and deductions for a set number of years for a given child. Since parents of newborns born in December are eligible for tax credits and deductions a year earlier, then the parents of newborns born in January will be eligible for tax credits and deductions for one year later. For example, when children born in January turn 19, their parents

[7] The CTC was not partially refundable until tax year 2001. The CTC is partially refundable because it becomes refundable for tax filers with income over a certain threshold (Crandall-Hollick, 2016).

[8] A small share of households each year report no income, less than 5% across all years. These observations are included as a 0 percent change in after-tax income.

[9] This claim assumes that the permanent income of households is unaffected by the income shock. However, researchers have found examples where temporary income shocks result in long-term increases in earned income, presumably from parents seeking out better paying work (Black et al., 2014). This paper discusses this possibility later in the discussion section, and in Appendix B.

are still eligible for the EITC for the previous tax year. Conversely, when children born in December turn 19, their parents will not be eligible for the EITC for that tax year.[10] So, the effect of having a child born in December as opposed to January of the next year is a speeding up of the tax credit and deduction process for that child.[11]

1.3 Birth Timing Patterns

Causal analysis of the effect of this change in after-tax income needs to account for the fact that parents and doctors have some degree of control over birth timing. Doctors may deliver children using Cesarian section (C-section) surgery (32% of all births in 2017) or by inducing labor through a variety of methods, including the use of drugs (26% of all births in 2017) (Martin et al., 2018). These delivery methods can be used to alter timing of birth.

There is clear evidence of this control over birth timing in the well-known fact that fewer births happen on weekends. As is clear in Figure 1.4, there are large dips in counts of births on Saturday and Sunday. This fall on the weekends reflects a decrease in C-section surgeries, but there is a smaller but still noticeable fall in vaginal births as well (Martin et al., 2010). Figure 1.4 also shows that mothers who give birth on the weekend have slightly lower education attainment. This data alone suggest that some parents, especially parents with slightly higher education attainment, exercise some degree of control over birth timing and have specific preferences over birth timing.

[10] Parents with full-time students living at home are able to claim their children for the EITC until their children turn 24, and parents with ”permanently and totally disabled” children can claim the EITC at any age.

[11] If families have perfect foresight and perfect liquidity, then knowledge of this future change in after-tax income should attenuate the size of this discontinuity in current family income after accounting for discounting. Assuming a rate of return of 5%, then ability to borrow against future tax benefits may attenuate the current discontinuity by slightly over 40%. However, many of the lower income families with the largest after-tax increases in income are likely liquidity-constrained and hence less able to borrow against future income (Gross and Souleles, 2002). Additionally, evidence suggests that some share of families do not understand timing of how eligibility for tax benefits expires as children age (Feldman, Katuscak and Kawano, 2016). These complications likely mean that attenuation from discounting in the estimated discontinuity in family income is limited.

After regression adjusting for day of week in Figure 1.5 and taking an average of birth counts over 5 years, the distributions of births and the characteristics of births are much smoother.[12] However, there are clear disruptions in the distribution of births, especially around major holidays (including New Year's Day, Christmas and July 4th).[13] Around these days, there are always fewer births on the holidays alone, and more births on the days around them. Similar to mothers who give birth on weekends, mothers with births that occur on holidays have slightly lower average years of education than mothers with births that do not occur on holidays. However, the average years of education return to previous levels quickly in the days around a holiday. Focusing in particular around New Year's, there is a drop in births on New Year's Day, and a slightly larger drop on Christmas Day, with larger counts of births occurring before and after these holidays. Interestingly, there are relatively few births after New Year's Day compared to before, suggesting that parents and their doctors with some level of control over birth timing are more likely to move births before the New Year compared to after. This pattern may be indicative of strategic timing of births to take advantage of tax benefits, but it also may reflect other preferences over

[12] For this regression adjustment, this paper estimates the following model:

$$Y^{birthcount} = \sum_{i=1}^{6} \beta_i \mathbb{1}[d = i] + \sum_{H} \sum_{i=-5}^{5} \beta_{i_H} \mathbb{1}[d_H = i] + \epsilon \tag{1.1}$$

where the first set of indicator variables $\mathbb{1}[d = i]$ are a set of six dummy variables (excluding Monday), and the second set of indicator variables $\mathbb{1}[d_H = i]$ are 11 dummy variables for each day within 5 days of each major holiday (indexed by H). The second set of dummy variables exclude from the estimation process all days around holidays, and the first set of dummy variables indicate the average births that are observed on a given day that differ from the births observed on Monday (the omitted category variable). Then, the regression adjusted counts of births would be:

$$\hat{Y}_{adj}^{birthcount} = Y^{birthcount} - \sum_{i=1}^{6} \hat{\beta}_i \mathbb{1}[d = i] \tag{1.2}$$

[13] Within individual years there are also spikes on Memorial Day, Thanksgiving Day, and Labor Day, but those spikes are not visible in this graph as this graph averages birth counts over 5 years. While New Year's Day, Christmas and July 4th are anchored to specific days in the calendar, Memorial Day, Thanksgiving Day, and Labor Day are not, so the disruptions that happen on these days are not visible when taking an average of birth counts.

birth timing, including concerns about hospital staffing. LaLumia, Sallee and Turner (2015) find limited evidence of specifically tax-related shifting in birth-timing around the New Year, with most tax-correlated shifting concentrated in a narrow window around the New Year.[14]

1.4 Methods

Evidence in the previous section suggests that the treatment of being born before New Year's Day is not random for some children, at least within a window of New Year's Day. However, the distribution of births outside of days around New Year's appears relatively smooth, save for other holidays. Intuitively, while parents can shift births in a specific region, they may have limited desire to do so further away, either because the costs of shifting are too high, or the benefits to shifting are too low. Appendix C develops microeconomic theory foundations to justify such a way of thinking, but this general intuition inspires a regression discontinuity strategy with an omitted region (sometimes referred to as a ”doughnut regression discontinuity”).

Specifically, this paper estimates the following model:

$$Y = \beta \mathbb{1}[d < 0] + \sum_{i=1}^{c} \gamma_i^1 d^i + \sum_{i=1}^{c} \Gamma_i d^i \mathbb{1}[d < 0] + \theta \boldsymbol{X} + \epsilon \tag{1.3}$$

Where Y is some outcome, d is the distance in days to the New Year's, c is the scale of polynomial in d, $\boldsymbol{X}$ is a list of additional covariates (specifically, state fixed effects and day of week fixed effects), and the estimation process includes days in some range $[\underline{D}, \bar{D}]$ but excludes observations in an omitted range of $[\underline{d}, \bar{d}]$. Note that β is the regression discontinuity estimate that reflects the estimated drop in outcome

[14] Furthermore, LaLumia, Sallee and Turner (2015) show compelling evidence that the correlation of after-tax income and birth timing may largely reflect income tax reporting responses rather than tax-motivated shifting. Note that this result differs from Dickert-Conlin and Chandra (1999), who use data from the PSID and conclude that parents with large potential tax benefits had a high probability of altering the timing of childbirth. LaLumia, Sallee and Turner (2015) show evidence that these patterns happen primarily in a narrow window around the New Year.

Y on New Year's Day, as on that day d is 0. We can conceptualize this estimate of β as the limit of the estimated means at either side of $d = 0$, even when some region of observations is omitted in the estimation process:

$$\beta = \lim_{\epsilon_1 \uparrow 0} \mathbb{E}[Y|d = 0 + \epsilon_1, X] - \lim_{\epsilon_2 \downarrow 0} \mathbb{E}[Y|d = 0 + \epsilon_2, X] \tag{1.4}$$

Following the recommendations in the theoretical and applied literatures regarding regression discontinuity estimation, this paper adds three more features to the estimation procedure. First, it uses local linear regressions where $c = 1$ (Hahn, Todd and der Klaauw, 2001). Second, it uses a triangle kernel that weighs observations more in the regression if they are closer to the discontinuity (Fan et al., 1996). Third, it uses a variety of bandwidth choices of $[\underline{D}, \bar{D}]$ to demonstrate sensitivity of the results to the region of observations included. Demonstrating how bandwidth affects these estimates more continuously pushes the limits of disclosure of restricted data from the Census Bureau.[15]

Before discussing the sufficient conditions this paper builds up to estimate β and the validation strategies suggested by those conditions, it is useful to first review the typical sufficient conditions that would apply in this setting if there were no omitted region. First, there must be no other treatment that coincides with the passing of the New Year. Second, as described by Lee and Lemieux (2010), the joint probability of observing various values of d conditional on X and ϵ, or $f(d|X, \epsilon)$, must be continuous in d. That is, for given values of X and ϵ, the treatment as determined by the birthdate of a child is randomly determined.

To argue that this condition holds in normal settings without an omitted region, many researchers perform two tests to argue validity of the research design:

[15] There is a robust literature on optimal bandwidth selection in regression discontinuity designs (e.g. Imbens and Kalyanaraman, 2011) with the goal of minimizing expected mean squared error in estimated regression discontinuities. This paper splits the difference between the practical demands of disclosure and the theoretical recommendations by showing robustness to different choices of bandwidths.

1. Test the null hypothesis that $f(X|d)$ is continuous by testing for discontinuous changes in variables at New Year's that should not be impacted by treatment.[16]

2. Test the null hypothesis that $f(d|X)$ is smooth at the threshold. A rejection of smoothness at the treatment threshold arguably indicates precise and hence non-random control over assignment to treatment (McCrary, 2008).

Without an omitted region, both of these traditional tests are violated in this paper. Figure 1.5 shows graphical evidence of a discontinuous change in average levels of mothers' education attainment from December 31st to January 1st. Average mother's education attainment is an untreated covariate that should evolve smoothly if the first test were met. Furthermore, there is clear strategic timing of births, with more births occurring around New Year's than on New Year's. If the second test were met, this distribution would be smooth.

With an omitted region, the treatment effect can be consistently estimated under four sufficient conditions. The first two are the same as before but the third and fourth are new. First, there must be no other treatment that coincides with the passing of the New Year. Second, $f(d|X,\epsilon)$ must be continuous in d. Third, the region of manipulated birth timing must be consistently identified and dropped from analysis. Fourth, the remaining data must be sufficient to consistently estimate and extrapolate means into the omitted region. Note that the fourth condition is stronger than the conditions from Lee and Lemieux (2010) discussed above. To see why this addition requirement is necessary, suppose that $f(d|X,\epsilon)$ is continuous, but the evolution of an outcome cannot be consistently extrapolated. Then, the evolution of the outcome may behave unpredictably in the omitted region, and the estimated discontinuity may be inconsistent.

To validate this set-up, note that, if the four conditions above are met, then the first test regarding covariate smoothness described before should still be applicable.

[16] This test comes from the fact that applying Bayes' rule: $f(X,\epsilon|d) = f(d|X,\epsilon)\frac{f(X,\epsilon)}{f(d)}$.

Assuming the regression discontinuity specification is valid, there should be no discontinuities in variables that are not impacted by treatment. However, the second test is no longer applicable as a substantial share of the data is omitted, and extrapolating an estimated density into an omitted region rapidly loses power.

Using this estimation strategy depends on properly identifying the region of manipulated birth timing around the New Year. Currently, there is no standardized procedure researchers use to estimate this region. Many papers use ad hoc visual analyses of the size of the manipulated region (Barreca et al., 2011; Gauriot and Page, 2019; Almond and Doyle, 2011), but some papers suggest more regularized methods that are not applicable in this setting.[17]

This paper estimates an omitted region by applying data-driven techniques from a method widespread in the public economics bunching estimation literature (Chetty et al., 2011; Saez, 2010; Kleven and Waseem, 2013). Bunching estimation papers look at situations similar to this paper where individuals alter a running variable to take advantage of some benefit tied to that running variable. The first step of this technique estimates the length of the running variable affected by bunching. In this setting, those observations would be equivalent to the section of observations that see birth timing shifting. Thus, using this first step offers an estimate of the region of observations that should be omitted.

To apply this method, this paper uses the regression-adjusted counts of births by day from the 2000 Census for August 1989 to July 1994 graphed in Figure 1.6.[18] This

[17] Dahl, Loken and Mogstad (2014) are able to use other years where a treatment does not exist as a counterfactual to estimate the extent of the regions that are not manipulated. Hoxby and Bulman (2016) suggest a method of estimating the region using locally estimated density functions that estimate a counterfactual density. They then estimate the size of the bias in outcomes present due to sorting. In this setting, there is no counterfactual year for comparison as this discontinuity in after-tax income is always present at the New Year, and the nature of the selection process into treatment is not as clear as in Hoxby and Bulman (2016) for estimating bias.

[18] The process described here could be run for birth counts separately by year of birth, creating different omitted regions for different years of birth. This strategy would likely make the most sense with full count natality data, but given the need to weight population estimates in the Census, it seems less obvious how meaningful slight differences in birth counts are. Averaging over a number of years offers a simpler and less error-prone measure of birth counts by day.

paper follows a five step process to estimate the region of manipulated observations:

1. Visually choose an upper bound on the days that demonstrate shifted births ($\bar{d}$), following Kleven and Waseem (2013).

2. Select a lower bound ($\underline{d}$) and estimate:

$$Y_d^{birthcount} = \sum_i^c \gamma_i \cdot d^i + \sum_{i=\underline{d}}^{\bar{d}} \psi_i \cdot \mathbb{1}[d = i] + \epsilon \tag{1.5}$$

   Where the first term is a flexible polynomial of order c. Similar to Kleven and Waseem (2013), this paper uses $c = 5$, although the results are unchanged with higher order polynomials. The second term omits from the estimation process observations that fall between $\underline{d}$ and $\bar{d}$.

3. Calculate the counterfactual distribution of births implied by the estimates in step one for the days that were omitted from the estimation process in the region, $[\underline{d}, \bar{d}]$:

$$\hat{Y}_d^{birthcount} = \sum_i^c \hat{\gamma}_i \cdot d^i \tag{1.6}$$

   This counterfactual distribution of births represents the distribution of births that would be believed to exist in the absence of strategic timing of births.

4. Calculate the absolute value of the gap between the counterfactual distribution and the observed distribution of birth counts:

$$Gap_{\underline{d},\bar{d}} = \left| \sum_{\underline{d}}^{\bar{d}} \left[\hat{Y}_d^{birthcount} - Y_d^{birthcount} \right] \right| \tag{1.7}$$

5. Repeat this procedure over values of $\underline{d}$. Choose the value of $\underline{d}$ that minimizes the gap.

> Note that this choice ensures that the surplus births observed for the days before New Year's roughly equals the lost births that occur in the days on and after New Year's.[19]

Because the omitted region needs to be estimated, calculating proper standard errors for this setting means accounting for error introduced by the first step of estimating an omitted region. To do so, this paper bootstraps the estimation procedure in 2,000 replications, using a bootstrapped set of estimated cutoffs, and then applying these estimated cutoffs to bootstrapped data.

Beyond the reduced form effect identified by this discontinuity, this paper also converts these estimated effects into a direct estimated effect of \$1,000 of income. One strategy of identifying this effect is to divide the reduced form effect by the estimated change in income in Figure 1.2, and then multiply by 1,000. Letting α be the estimated increase in after-tax income, this Wald estimator would be:

$$\hat{W} = \frac{\hat{\beta}}{\hat{\alpha}} \cdot 1000 \tag{1.8}$$

This strategy is not as efficient as the two-sample two-stage least squares estimator, but that estimation procedure is not readily applicable here as the first stage was not estimated using the same regression discontinuity design (Inoue and Solon, 2010).

The delta method shows that the variation of this estimate is approximately:

$$V(\hat{W}) \approx \frac{1000^2}{\hat{\alpha}^2}\Big[V\big(\hat{\beta}\big) + \frac{\hat{\beta}^2}{\hat{\alpha}^2}V\big(\hat{\alpha}\big) - 2\frac{\hat{\beta}}{\hat{\alpha}}Cov\big(\hat{\alpha}, \hat{\beta}\big)\Big] \tag{1.9}$$

Following Angrist and Krueger (1992), this paper assumes that $\hat{\beta}$ and $\hat{\alpha}$ are inde-

[19] In some respects, this estimation process ensures that the remaining data meet a smoothness condition similar to the second validity test described above. Omitting dates that demonstrate shifted births isolates attention to births that can be modeled with the counterfactual polynomial. This process effectively finds a region of births where the density of the running variable is smooth. Of course, the density estimation process here ensures that, by design, any estimated density created with this data is smooth, but the estimation process drops observations from the analysis would not fit that smoothness.

pendent and hence the covariance term is 0.

These instrumental variables estimates should be interpreted with caution given that the increase in after-tax income, α, may be imprecisely estimated. As described in Section 3 above, the calculation in Figure 1.2 is not done with administrative tax data, and its estimation process is fundamentally different than the regression discontinuity estimation procedure for the reduced-form treatment effects. Nevertheless, if both α and β are consistently estimated, then W is also consistently estimated.

1.4.1 Estimating the Omitted Region

Figure 1.6 shows results from the density estimation procedures described in equations 1.5, 1.6 and 1.7. The horizontal lines indicate the endpoints of the region of days the procedure suggests should be omitted. Following Kleven and Waseem (2013), 9 days after the New Year appears a good endpoint for the upper region of birth dates demonstrating manipulation in birth timing. The estimation process then calculates that the lower endpoint for the omitted region is 20 days before the New Year. More days are dropped in December than January due to disruptions in birth timing around Christmas. As births shifted away from the New Year cannot be distinguished from births shifted away from Christmas, the calculation process drops all days affected by birth shifting around both holidays. This magnitude of shifting, on the order of between one to two weeks before or after a major holiday (either New Year's or Christmas), is comparable with the birth timing shifting documented elsewhere. Other papers that look at changes in birth timing to qualify for either cash or program benefits tied to birth timing of children have found similar responses (Gans and Leigh, 2009; Neugart and Ohlsson, 2013; Dahl, Loken and Mogstad, 2014). As is clear visually, the density of births appears to return to a smooth distribution outside of these dates.[20]

[20] A period of five days before and four days after Thanksgiving are also omitted from these density calculations. This omitted region was calculated using a similar process as the calculation around

1.5 Results

Having estimated the omitted region, the next step is to validate the research design. As mentioned in Section 1.4, one test for the validity of this design with this omitted region is to look for discontinuous differences in pre-treatment and untreated covariates. If the research design is valid, there should be no detectable differences except those observed at random. Table 1.1 shows the results from regression discontinuity estimates testing whether these untreated covariates for infants' families vary discontinuously.[21] All of these regression discontinuity estimates include state fixed effects, and day-of-week fixed effects. The variables analyzed include household and parent income, intensive and extensive parent labor force participation in the previous year, education attainment of parents, race of child, marital status of parents and household size.

11 out of 114 tests show significant discontinuities at the 5 percent level. This rejection rate is within the levels that would be expected with random sampling variation and independent tests if the null hypothesis of no discontinuous changes in characteristics were true. Additionally, as these tests are likely positively correlated, rates of rejection expected under this null hypothesis may be even higher. Lastly, it should be noted that most of the rejections take place within the smallest bandwidth. When bandwidths of two months or more are used, three out of 76 tests are significant. All of the point estimates discussed below will use the two month bandwidth, although

New Year's. This omission does not translate to a change in the average density depicted in Figure 1.6, as the timing of Thanksgiving (falling on the fourth Thursday in November) varies from year to year. The results estimating this estimated region are available on request.

[21] Although the results regarding outcomes for children below use pooled data from the 2001-2016 ACS and the 2000 Census, this section uses only the data from the 2000 Census for infants born 1999-2000. The Census data are better suited for looking at these questions than the ACS primarily because the 2000 Census asks for data about income types and levels in 1999 specifically, while the ACS ask about income in the "previous 12 months." This phrasing in the ACS means that, depending on the month in which families respond, they may post responses that reflect common changes in income and labor supply after birth of the newborn (Wingender and LaLumia, 2017). Hence, restricting attention to the cohort of children born 1999-2000 in the 2000 Census long form offers the cleanest test of whether characteristics differ for children born across the New Year.

other results with different bandwidths will be discussed when relevant. Hence, these results with this omitted region meet the validation test implied by the research design.

1.5.1 Effect of Family Income in Infancy on Grade-for-Age Status in School

The next step is to use this after-tax income discontinuity to examine the impact of the income discontinuity on school outcomes. The primary school outcome observable in the Census and ACS data is grade-for-age status. A student being grade-for-age is often interpreted as a basic indication of that student achieving academic and social maturity in earlier grades. Table 1.2 reports all basic results for discontinuities in grade-for-age status by grade. Figures 1.8 through 1.14 show graphical depiction of these regression discontinuities. As a reminder, all of these regression discontinuity estimates include state fixed effects and day-of-week fixed effects.

In the year that students are eligible for kindergarten, Table 1.2 and Figure 1.8 show that enrollment in kindergarten or a higher grade in the first year of kindergarten eligibility shows no discontinuity across the threshold. This result suggests that there is no detectable difference in parents delaying their child's entrance into kindergarten across the New Year. These delays are often referred to as "red-shirting."

This lack of a discontinuity in kindergarten attendance is important for contextualizing later results. This finding suggests that any subsequent detected discontinuities in grade-for-age status reflect students being retained in a grade and not kindergarten red-shirting. It is difficult to interpret the meaningfulness of changes in grade-for-age status from red-shirting. The population of students who are red-shirted do not on average have lower cognitive skills and social maturity before they enter school than children who are not red-shirted (Bassok and Reardon, 2013).[22] In contrast, repeat-

[22] Researchers often interpret parents who red-shirt children as looking to gain an advantage for their child in school by having their child enter school slightly older than the rest of the children in

ing a grade after entering school is usually interpreted as a negative signal about a student's social, emotional or academic readiness for the next grade. Students who are retained in a grade are more likely to have poorer academic performance prior to retention, lower social skills and poorer emotional adjustment. They also are more likely to display problem behaviors in class, including inattention and absenteeism (Xia and Kirby, 2009).[23] Thus, any subsequent detected changes in grade-for-age status in this setting are an indication of changes in the conditions that make students more likely to be retained within a grade.[24]

Figure 1.8 also shows an important pattern in the omitted region that is worth noting for all subsequent graphs in Figures 1.8 through 1.14. The students born right after the New Year appear to be slightly less likely to have entered kindergarten on time than the students born right before. These data were excluded from the regression discontinuity estimation process for the reasons discussed earlier regarding strategic birth timing. This drop that happens right after the New Year likely reflects both the fact that students born after the New Year did not get the income boost, and the fact that these children are negatively selected compared to the children born before the New Year. As was discussed previously regarding Figure 1.5, these children born right after the New Year come from households where mothers have, on average, slightly lower education attainment.

As children enter first grade, Table 1.2 and Figure 1.9 show that a small gap

their grade (Deming and Dynarski, 2008).

[23] Note also that students who repeat grades are more likely to be children of color from less educated and lower income households (Xia and Kirby, 2009) while red-shirted children tend to come from families with higher incomes and are more likely to be White (Bassok and Reardon, 2013).

[24] Retention policies differ across states, districts and schools, and the students that are retained in one location may not have been retained in another. As of 2018, 16 states have 3rd grade retention policies that require students to repeat a grade if those students have not reached some minimum threshold of achievement (Education Commission of the States, June 2018*b*). Even across school districts in the same state, rates of retention can vary (French, 2013), as do district policies and implementation of standards (Schwager et al., 1992). Thus, the meaningfulness of this outcome may differ from location to location, with some teachers in some states being much more willing to use it as a tool than others.

opens up in the probability of a child being grade-for-age around the New Year, with students who experience the income shock being slightly more likely to be grade-for-age than students who do not. This gap is relatively small, at around half a percentage point, and not statistically distinguishable from 0. As Figure 1.7 shows, kindergarten is one of the grades students are most likely to repeat, so a change in grade-for-age status around the New Year by this grade would not be surprising. It is worth noting that this result, unlike the other results discussed here, is relatively sensitive to the size of the omitted region. With a smaller omitted region, the gap is larger and statistically distinguishable from 0 (results available on request). These results offer suggestive evidence that a discontinuity has opened up in the share of students grade-for-age, but that discontinuity is relatively modest.[25]

These results are confirmed when looking at the share of students grade-for-age in 5th grade in Table 1.2 and Figure 1.10. As before, there is a drop in the share of students grade-for-age among the students born right after the New Year, but the estimated discontinuity reported in Table 1.2 is close to 0. This small discontinuity, coupled with the somewhat larger but still statistically insignificant discontinuity from first grade, suggest that there is at most only a modest change in the share of students grade-for-age across the New Year by this point.

Moving forward to 7th grade in Table 1.2 and Figure 1.11, a larger detectable discontinuity has opened up in the share of students grade-for-age. The regression discontinuity estimate shows that students born before the New Year see a 1.05 percentage point increase in the probability of being grade-for-age. The increase in the discontinuity here makes sense, given that Figure 1.7 shows that there is a gradual

[25] While repetition of kindergarten may represent a type of red-shirting (Deming and Dynarski, 2008), it is worth noting that the characteristics of children who repeat kindergarten are on average different than those of students who delay entrance into kindergarten. As mentioned above, children who delay entrance into kindergarten tend to be White and come from better-educated families with higher incomes than their peers who do not. The characteristics of children who repeat kindergarten tend to be similar to the characteristics of students who are held back in grades; compared to their peers they are more likely to repeat later grades, have below-average school work, and be described by their teachers as having behavioral issues (National Center for Education Statistics, 2000).

increase in retention rates from 5th grade to 7th grade. As is clear from visual inspection of Figure 1.11, this result appears somewhat sensitive to the upper bound of dates excluded, but this result is suggestive evidence of an eventual shift in grade-for-age status taking place. Table 1.2 also converts this reduced form impact into an instrumental variables estimate of the effect of $1,000 of income in infancy. These results show that $1,000 more in family income in infancy results in an 0.88 percentage point increase in the probability of a student being grade-for-age by 7th grade.

Lastly, looking at 9th, 10th and 11th grades in Table 1.2 and Figures 1.12 through 1.14, the discontinuity in the share grade-for-age appears to eventually grow in magnitude. Although there is some variation in the estimated discontinuity in grade-for-age status, the estimated discontinuity is consistently positively signed and generally significant at the 5 percent level. Furthermore, the results depicted in Figures 1.12 through 1.14 appear to become less sensitive to the upper bound on dates omitted, unlike Figure 1.11. Table 1.2 and Figure 1.14 show the average discontinuity in grade-for-age status using all high school years together. These results show that children born just before the New Year are approximately 1.13 percentage points more likely to be grade-for-age in high school. As the control mean for the share of students grade-for-age by high school is 87%, this is a meaningful shift in grade-for-age status.[26] Table 1.2 converts these reduced form results into a direct estimate of the effect of income, and shows that a $1,000 increase in income in the first year of life results in a 1.2 percentage point increase in the probability of a student being grade-for-age by high school.

[26] Changes in grade-for-age status that occur in high school are harder to interpret than changes that happen in earlier grades. Retention in high school may reflect students failing to accumulate enough credits to advance their academic standing. Hence rather than being required to repeat an entire grade, as might be the case in earlier grades, such retention may reflect students being only required to repeat one specific course (West, 2012). However, two features are worth noting of this discontinuity. First, this sort of retention, while not necessitating an additional year of schooling, indicates that a student has not met certain benchmarks, and is hence meaningful in its own right. Second, the previous results show the discontinuity in grade-for-age status evolving over time, suggesting that the discontinuity in grade-for-age status in high school reflects changes that occur both in high school and in the grades beforehand.

While estimates of specific discontinuities are often noisy, the pattern of the evolution of the discontinuity across grades is worth noting. By 1st grade, a slight discontinuity that is statistically insignificant opens up, and by 5th grade the discontinuity is still indistinguishable from 0. While it is difficult to read much into this early pattern, it may be weak evidence of a small if undetectable gap beginning. The estimated discontinuity in grade-for-age status in 7th and 9th grade is larger, and in high school, it continues to grow. While these estimates are imprecise, they suggest a gradual increase over time in the size of the discontinuity, with perhaps the largest increases happening in grades where students are most likely to be retained.[27]

Heterogeneous Effects for Subgroups in Grade-for-Age Status Results

Tables 1.3 through 1.5 and Figures 1.15 through 1.17 break these results down further by showing how these results vary among subgroups. Here, for concision, the only grades analyzed are grades 5, 7 and then 9, 10 and 11 conjointly.[28]

Much of the previous research looking at the effects of income on outcomes for children has found non-linear impacts. Similarly sized increases in income in this research have often had larger effects for lower income families than higher income families. Ideally, to test for that non-linearity here, data would be available on the characteristics of families at birth so that families could be identified that have lower income at time of child's birth. However, without such information, identifying high impact samples depends on choosing information that retroactively could indicate high-impact groups. This paper uses two possible signifiers of high impact groups: Black students, and students with mothers who have a high school degree or less. Both of these groups have lower average income at time of child's birth because they

[27] The reasons that students are retained may differ by grade. In early grades, students are often retained on the basis of social and emotional immaturity (Xia and Kirby, 2009; Byrd and Weitzman, 1994), while in later grades retention is additionally correlated with other risk factors and grade-specific metrics of academic achievement (Peixoto et al., 2016).

[28] The use of data from high school grades conjointly is for precision. Results for individual grades are similar.

have lower average income throughout childhood (Tamborini, Kim and Sakamoto, 2015).

When comparing Black children with White children in Table 1.3 and Figure 1.16, both White and Black children have virtually no detectable discontinuity in grade-for-age status in 5th grade. For the subsequent grades, both groups show some discontinuity in the share grade-for-age around the New Year. However, in 7th grade and high school, the estimated discontinuity shows a larger point estimate for Black children. By high school, for example, the estimated discontinuity in the share grade-for-age for Black children is 1.3 percentage points, while the estimated discontinuity for White children is one percentage point. Converting these reduced form estimates into a direct effect of income in Table 1.3 shows that a $1,000 increase in family income in infancy results in a one percentage point increase for White children in the probability of being grade-for-age by high school. For Black children, the same income shock results in a 1.6 percentage point increase in grade-for-age status. It should be noted, though, that the difference between the two is significant at the 10 percent level in 7th grade and insignificant in high school. However, these tests for differences in discontinuities between White and Black children are likely imprecise given the size of the omitted region and the smaller number of Black children compared to White children. In all, these results suggest that the discontinuity is larger for Black children than White children, although the magnitude of the difference is unclear.

There are even stronger differences when comparing children born to mothers with different education attainment levels. The results in Table 1.4 and Figure 1.17 show that a large share of the estimated discontinuity in grade-for-age status in high school comes from effects on children with mothers who have lower education attainment. The discontinuity is a statistically insignificant 0.19 percentage points for children from mothers who have more than a high school degree, and 1.73 percentage points for children with mothers who have earned a high school degree or less. Furthermore, the difference between the two groups is significant at the 10 percent level among children in high school. Converting these results into a direct effect of income in Table 1.4 shows that \$1,000 of income in infancy results in a 0.17 percentage point increase in grade-for-age status for children of more educated mothers. Among children of less educated mothers, the same increase of income in infancy results in a 2.05 percentage point increase in grade-for-age status in high school.

In general, these results show that the effect of \$1,000 of income in infancy is larger for groups that are more likely to be disadvantaged at a child's birth. This result suggests that the impacts of this additional income are nonlinear, in that the benefits of increased income are stronger for families with comparatively lower incomes.

1.5.2 Robustness Checks on Grade-for-Age Status Results

Conditioning on State of Birth

This paper assigns kindergarten age eligibility cutoffs to children depending on the state in which they were born, and these cutoffs determine what the grade-for-age status of a student is. However, the appropriate state eligibility rules that children face when entering kindergarten would be those for the state the child lived in when the child was first eligible to enter kindergarten at age 5. As information on state of residence at age 5 is not available retrospectively in this data, state of birth is an imperfect proxy, and some students may have misaligned grade-for-age status.

Students will have misaligned grade-for-age status if the grade they are expected to have completed to be grade-for-age is not correct.[29] For example, if this paper's

[29] Misalignment will only happen if the child's birthdate is between both the correctly and incorrectly assigned kindergarten birthdate cutoffs. If the birthdate is after both of the cutoffs, or before, then the student would need to be in the same grade to be grade-for-age under both cutoffs, and grade-for-age status would be the same in both. Assuming that the child's birthdate is between both the correct and incorrect birthdate cutoffs, grade-for-age status is biased upwards if the incorrectly assigned birthdate cutoff is before the correct cutoff. For example, say a child is born in November in a state that had a kindergarten age-eligibility cutoff of October 1st, and moved to a state at age 5 that had an age-eligibility cutoff of December 1st. The incorrectly assigned birthdate cutoff suggests that a student should be in a grade to be grade-for-age that is lower than the grade a student would

metric of grade-for-age says that a student should have completed 8th grade to be grade-for-age, but the true grade that a student should have completed to be grade-for-age is 9th grade, then that misalignment may result in a student being improperly marked as being grade-for-age. In this setting, misaligned grade-for-age status will only bias the estimated discontinuity in grade-for-age status upwards.[30] Particularly concerning is the possibility that students may have moved from birth states to states or districts that have age-eligibility cutoffs for kindergarten that coincide with January 1st or December 31st, as this misalignment could especially bias the estimated effect upward.

One test for bias is to further restrict the sample to children who are currently residing in the same state as their state of birth. Under the assumption that students living in their state of birth did not live in another state with different age eligibility rules at age 5, these students would have correctly assigned grade-for-age status. Table 1.5 shows that effects observed among this subsample are even larger than those observed in the full sample. Notably, the control mean of students who are grade-for-age here is lower than the full sample. This pattern makes sense, as the population of students who continue to reside in their state of birth is negatively selected; families that do not engage in interstate migration are more likely to be

actually need to be in if that student were grade-for-age. Thus, even if this student were retained once, this measure will mistakenly record that student as being grade-for-age. Conversely, grade-for-age status would be biased downwards if the incorrectly assigned birthdate cutoff is after the correct cutoff. For example, suppose a child is born in November in a state that had a kindergarten age-eligibility cutoff of December 1st, and moved to a state at age 5 that had an age-eligibility cutoff of October 1st. The incorrectly assigned birthdate cutoff suggests that a student should be in a grade to be grade-for-age that is higher than the grade a student would actually need to be in if that student were grade-for-age. Thus, even if this student never skipped a grade and was never retained, this measure will mistakenly that student as not-being grade-for-age.

[30] Data in the regression discontinuities is organized by school cohort. Consider the first example in the previous footnote, where the true kindergarten eligibility age cutoff a child experienced was after the one assigned via birth state. This observation would be included in a cohort born before the New Year. As discussed in the previous footnote, that child's recorded grade-for-age status is likely biased upwards. However, the other child, who experienced a true age cutoff that was before the one assigned from the child's birth state, would not be included in a cohort before the New Year, as the first observations in that cohort would begin with the children born after the assigned birth state cutoff. Thus, misaligned grade-for-age status can only bias the estimated discontinuity upward.

less educated than families who do (Molloy, Smith and Wozniak, 2011), and previous results have already shown that effects of income on grade-for-age status are larger for less-educated families.

Thus, the findings discussed before are robust to whatever error is added from the misassignment of state of residence at age 5.

Separating Data by Birth Cohort

All of the preceding results have pooled together data across years for additional precision. However, as is clear in Figure 1.2, the size of the discontinuity in after-tax income has increased over time, so later birth cohorts see a larger discontinuity in after-tax income than earlier birth cohorts. Hence, an alternate way to use the data to explore the relationship between family income and outcomes for children is to compare the estimated discontinuity across different birth cohorts. If the relationship between after-tax income and grade-for-age status by high school is positive, then there should be increases in this discontinuity for later cohorts that saw a larger change in after-tax income for being born before the New Year.

Table 1.6 separates the sample of students in grades 9 through 11 into three different groups depending on year of birth: students born 1982-1986, 1987-1993, and 1994-2001. This combination of cohorts into years of birth reflects different eras of the EITC and CTC programs. As is clear in Figure 1.2, the average value of the discontinuity in after-tax income for having a child born before the New Year actually falls in real terms from 1982 to 1986, then begins rising from 1987 to 1993 following changes to the EITC, and then lastly increases substantially from 1994 to the early 2000s following further changes to the EITC and the introduction of the CTC.

Table 1.6 shows that an increase in the discontinuity in after-tax income by birth cohort happens alongside an increase in the estimated discontinuity in grade-for-age status by high school. Notably, the estimated discontinuity in grade-for-age status for being born before the New Year for the cohort born 1994-2001 is 60% larger than the estimated discontinuity for the cohort born 1982-1986. Since the only statistically significant change in grade-for-age status comes from the cohort of students born 1994-2001, the previous results that group all cohorts together are largely driven by children who were born in this later cohort when the EITC and CTC were most generous.

Note that this way of analyzing the data allows a check on the identifying assumption that no other treatments coincide with the passing of the New Year. If the previously observed results reflected some other treatment that occurred with the passing of the New Year, and if that other treatment remained constant, then the reduced form discontinuities in grade-for-age status across these birth cohorts should be constant. The differences across years are evidence that the previous results do not just reflect a constant New Year-specific treatment.

Interestingly, the direct effect of $1,000 on grade-for-age status by high school is relatively stable over time. Receiving $1,000 in infancy results in a 1.14 percentage point increase in grade-for-age status by high school for the 1982-1986 cohort, a 0.78 percentage point increase in the 1987-1993 cohort, and a 0.90 percentage point increase for the 1994-2001 cohort. As all these estimates have substantial standard errors on them, they are not distinguishable from each other. Hence, it is difficult to read too much into the specific pattern of results over time.

1.5.3 Effect of Income in Infancy on Outcomes in Early Adulthood

When extending analysis beyond grade-for-age status in school, the context of the treatment changes. First, there is a second discontinuity in after-tax income that happens as a child ages into adulthood. As is clear in Figure 1.1, parents of children born in December see various tax benefits expire one tax year before parents of children born in January. Research shows that the size of those tax benefits at those ages has consequences for behavior of their families, including enrollment of children in college (Manoli and Turner, 2018) and parent labor force participation (Lippold, 2019).[31]

Second, when looking at outcomes other than grade-for-age status, it is important to remember that being retained in grade is both a potential indicator of that child's progression through school but also a form of mediation that may have long-term repercussions. Research suggests that the cumulative effects of not being grade-for-age are unclear and may differ depending on the age at which retention occurs. Researchers looking at red-shirting and retention in the early grades have found that these changes may result in short-term improvements in school achievement (Datar, 2006). Researchers have also analyzed retention in later gradess related to test scores. Some researchers have found no impacts or negative impacts of retention on short-term achievement in early grades (Roderick and Nagaoka, 2005) and increases in high school dropout rates that vary by grade of retention (Jacob and Lefgren, 2009). Other researchers have found positive short-term impacts of retention on achievement and no impact on eventual high school graduation (Schwerdt, West and Winters, 2017).[32] Thus while the initial income shock treatment in infancy is clear, other compensating responses happen subsequently that may complicate interpretation of effects in adulthood.

As the discontinuity in grade-for-age status was concentrated among more likely disadvantaged households, discontinuities in outcomes in early adulthood are likely concentrated in these groups as well. However, as children age into young adulthood, many move away from their parents. Consequently, it is harder to identify children who grew up in more likely disadvantaged households as they get older. This pa-

[31] These later discontinuities in after-tax income are likely small, as the share of families that claim EITC benefits for newborns is much larger than the share of families that claim EITC benefits for older children. Appendix B discusses these patterns in more detail.

[32] The difference in these results highlights the fact that the effects of retention likely depend on other interventions related to retention.

per uses two strategies to identify these groups. First, this paper looks at outcomes among Black children. While Black children did not display consistently statistically different results in grade-for-age discontinuities than White children, Black children had larger point estimates of changes in grade-for-age status. Second, this paper looks at outcomes for children born in counties that have average mother's education attainment in the bottom quarter of the education distribution (weighted by population). Mother's education levels were a strong predictor of the discontinuity described previously, but no parent education attainment variables are observable for young adults no longer living at home. Hence, conditioning on education attainment levels in county of birth is a proxy for this group of individuals.

For relevant later life outcomes, this paper looks at high school completion rates, earned income, labor force participation, and SNAP receipt from ages 19 to 32 for children born in 1980 forward.[33] This paper follows Kling, Liebman and Katz (2007) in combining these four measures of outcomes into a single unitary measure of economic sufficiency. This single measure allows more precision in measuring effects that move in the same positive direction. To compute this measure, this paper normalizes each outcome into a z-score and adds the four z-scores with signs reflecting whether the outcome is beneficial (positive for labor force participation, earned income, and high school attainment, and negative for SNAP receipt). The normalizing mean and standard deviation for each of the z-scores come from outcomes for adults born in the month and a half after the New Year, excluding the omitted region.

Figures 1.18 through 1.20 show some of the basic variation in post-high school

[33] Age 18 is excluded here. Given the way the sample is constructed, young adults aged 18 are expected to have completed high school if they graduated on time. By definition, the previously estimated discontinuities in grade-for-age status ensure that high school graduation rates at age 18 would be different. Young adults aged 19, on the other hand would be expected to have completed high school if they graduated either on time or one year later. The results look at individuals born 1980 and later for reasons discussed earlier in the data section. This sample restriction ensures that outcomes for adults are analyzed for cohorts for which there is data from the previous section showing changes in grade-for-age status. Age 32 is an arbitrary ending age reflecting the fact that data get sparse for later ages in the 2001 to 2016 ACS when looking at adults born 1980 and later.

outcomes by age of adults. These figures show average outcomes for children born in December and January, excluding children born in the region around the New Year who are omitted in this paper. As such, they only demonstrate the underlying variation in outcomes and are not meant to be interpreted as causal impacts. As is clear, there is little detectable difference in high school graduation rates, nor in labor force attachment in the population as a whole between people born in January and December. However, there is a slightly more persistent gap in earnings, with adults born right before the New Year often earning slightly more than adults born right after the New Year. While these gaps are within the margin of error for most years, the gap varies from about $50 to $500 depending on the year. Importantly, the gap seems to attenuate or disappear in later years.

Figure 1.21 combines all four measures into a unitary measure of economic self-sufficiency for all adults. Note that, by construction, this measure has average value 0 for people born in January, but there is still a standard error on the estimate as it is an average and has sampling variation. Figure 1.21 shows that, while there is a gap of 0.04 to 0.01 standard deviations in the self-sufficiency measure in the early years, the gap disappears over time. Figures 1.22 and 1.23 show similar graphs for Black young adults and adults born in counties with comparatively low education attainment. The composite measure is recalibrated for these samples such that the measure again has average value 0 and standard deviation 1 for people born in January within this subgroup. Here, the patterns are much noisier given the smaller sample sizes, but similarly the gap varies from 0.09 to 0.01 standard deviations, and attenuates over time to low numbers by the time adults reach their late 20s and early 30s.

To formalize these comparisons, Table 1.7 computes regression discontinuities over the conjoint measure of economic self-sufficiency and each of the four outcomes separately for the full sample. Figure 1.24 shows results for discontinuities in the self-sufficiency measure. Given the small differences observed in Figures 1.21 through

1.23, it is useful to compile different ages into bins to increase precision. While the exact grouping of the bins can be somewhat arbitrary, this paper computes discontinuities for adults aged 19-22, 23-27 and 28-32 to demonstrate how patterns evolve over time. As is clear in Figure 1.21, however, there are individual outliers within these age groups that can be important for driving measured effects, so it is worthwhile to be cautious in interpreting any one given result.

Table 1.7 and Figure 1.24 show that adults aged 19-22 who experience the higher income in infancy see an estimated increase in their self-sufficiency measure of approximately 0.02 standard deviations. Converting this discontinuity into a direct effect of income, $1,000 in infancy results in a 0.03 standard deviation increase in the self-sufficiency measure. However, this gap has a wide standard error, so it is not statistically distinguishable from 0 at the 10 percent confidence level. Looking at the individual components, Table 1.7 shows that adults who experienced the income boost as children are an estimated 0.1 percentage points more likely to have completed high school off a baseline rate of 90%, and earn an estimated $8 more annually. Neither of these effects are distinguishable from 0 at the 10 percent level.

Moving to ages 23-27, young adults who experience the higher income in infancy see an estimated drop in their self-sufficiency measure of 0.02 standard deviations, again not statistically distinguishable from 0 at the 10 percent level. Converting to a direct effect of income, $1,000 in infancy results in a 0.02 standard deviation drop in the sufficiency measure. Table 1.7 shows that adults who experienced the higher income are an estimated 0.002 percentage points more likely to have completed high school, and earn an estimated $280 less annually, but again neither of these effects are distinguishable from 0.

Lastly, looking at ages 28-32, the estimated fall in the self-sufficiency measure for adults who experience the income boost is still 0.02 standard deviations, again not distinguishable from 0 at the 10 percent confidence level. Similarly, the direct effect of

$1,000 of income is -0.03 standard deviations. The adults who experienced the income shock are an estimated 0.4 percentage points less likely to have completed high school, and estimated to earn $2 less annually than adults who did not experience the income increase as infants, but again neither of these effects are distinguishable from 0.

Taking these point estimates at face value, like Figure 1.24, they suggest a weak treatment effect in early adulthood that falls over time as young adults age into their mid to late 20s, although strictly speaking no effects are distinguishable from 0.

Heterogeneous Effects by Subgroups on Outcomes in Early Adulthood

Table 1.8 and Figure 1.25 compute regression discontinuities for White and Black young adults separately. The table only reports discontinuities in the conjoint measure of self-sufficiency for concision. As most of these individual discontinuities are noisy, they should be interpreted with caution, but the high school graduation status and earned income discontinuities are referenced here for context.

White young adults who experienced the income boost as infants display a small estimated treatment in their self-sufficiency measure in ages 19-22 of 0.009 standard deviations. However, Black young adults display a much larger estimated treatment effect of 0.134 standard deviations. Both estimates are not distinguishable from 0 at the 10 percent level, but they are distinguishable from each other at the 10 percent level. Converting these reduced form results into a direct effect of income suggests that White young adults see a 0.02 standard deviation increase in their economic self-sufficiency score from $1,000 in infancy. Black young adults see a 0.18 standard deviation increase from the same sized shock. This increase in the composite score for Black young adults comes from increases in high school graduation rates. Black young adults who experienced the income boost are 2 percentage points more likely to have completed high school off a baseline high school graduation rate of 81%. While this is a large effect and distinguishable from 0 at the 1 percent level, it still has a wide standard error on it, and the effect is not sustained into later ages, so it should be interpreted with caution. Black young adults also earn $18 more annually off a mean of $6,007, but again this effect is not distinguishable from 0 at the 10 percent level.

Moving to young adults aged 23-27, White young adults who experienced the income shock display a treatment effect of -0.03 standard deviations in their self-sufficiency measure while Black young adults display a treatment effect of 0.11 standard deviations. Both estimates are not distinguishable from 0, and they are not distinguishable from each other at the 10 percent level. Converting these results into a causal effect of income suggests that a $1,000 increase in income in infancy for White children results in a decrease in their self sufficiency score of 0.05 standard deviations. For Black young adults, the same sized income shock increases their self-sufficiency score of 0.18 standard deviations. These effects among Black adults come from changes in labor force participation and earnings. Black young adults who experience the income boost are 0.5 percentage points more likely to have graduated high school off a baseline rate of 83.8%, 2 percentage points more likely to be in the labor force off a baseline rate of 69%, and earn $700 more annually off a baseline mean of $13,200. However, again, none of these indfvidual effects are distinguishable from 0 at the 10 percent level.

Note that when combining all young adults aged 19-27, the estimated treatment effect for White young adults is -0.005 standard deviations in their self-sufficiency measure. However, the estimated treatment effect for Black young adults is 0.12 standard deviations. The increase for Black young adults is statistically distinguishable from 0 at the 10 percent confidence level, and distinguishable from the treatment effect for Whites at the 5 percent level. Converting these reduced form results into a direct effect of income shows that White young adults who experienced $1,000 in after-tax income in infancy see a 0.01 standard deviation drop in their self-sufficiency score. Black young adults who experienced the same income shock see a 0.19 standard deviation increase in their self-sufficiency score.

Lastly, looking at young adults aged 28-32, the treatment effect for Whites is -0.03 standard deviations in their self-sufficiency score, and the treatment effect for Black young adults is to 0.03 standard deviations. These effects are not statistically distinguishable from 0, or distinguishable from each other at the 10 percent level. Converting to direct effects, these estimates say that for a $1,000 shock in income in infancy, White adults see a 0.02 standard deviation drop in outcomes, but Black young adults see a 0.07 standard deviation increase. Black young adults who experience the income boost are 0.6 percentage points less likely to have graduated high school off a baseline rate of 86.1%, and earn $1,227 less annually off a baseline mean of $20,500. None of these effects are distinguishable from 0 at the 10 percent level.

Overall, the treatment effects are larger for Black young adults than White young adults. Furthermore, observed treatment effects for Black young adults follow the pattern established earlier in the sample as a whole, where estimated treatment effects are largest in earlier years and appear to attenuate with time. The pattern of results here is likely more suggestive than the previous results looking at grade-for-age status. The previous results showed that White children saw an increase in the probability of being grade-for-age if they experienced the income shock as children. Taken at face value, however, some of these estimated coefficients on post-schooling outcomes for Whites suggest negative treatment effects, which would be odd given the positive effects seen on grade-for-age status earlier. The noisiness of these estimates likely reflects the fact that the sample sizes become much smaller when looking at older adults. Ultimately, what seems more instructive is that Black adults display consistently larger estimated treatment effects, and some of these treatment effects are statistically distinguishable from 0 and distinguishable from estimated treatment effects for Whites.

Table 1.9 and Figure 1.26 offer a similar exercise for young adults born in counties with average mothers' education attainment above and below the lowest quartile. Again, the table only shows effects on the composite measure of outcomes, and most of the individual discontinuities in that measure are noisy. However, as before, the high school graduation status and earned income discontinuities are referenced for context.

When looking at young adults aged 19-22, the estimated discontinuity in the self-sufficiency score for young adults born in counties with high average mothers' education attainment is 0.02 standard deviations. The estimated discontinuity for young adults born in counties with low education attainment is 0.05 standard deviations. Converting these discontinuities into a direct effect of income, young adults from counties with higher education attainment see an 0.02 standard deviation increase in their self-sufficiency score from a $1,000 shock to income in infancy. Young adults from counties with lower education attainment see an 0.06 standard deviation increase in the score from the same shock. These estimates are not statistically distinguishable from 0, or from each other at the 10 percent level. Young adults in counties with low education attainment who experience the income increase see an increase in $68 in earned income off a baseline mean of $9,074 and an 0.3 percentage point increase in the probability of having graduated high school off a baseline mean of 87.9%. None of these effects are distinguishable from 0 at the 10 percent level.

Larger effects appear when looking at young adults aged 23-27. The estimated treatment effect for young adults born in counties with high mothers' education attainment is -0.02 standard deviations, but the estimated treatment effect for young adults born in counties with low mothers' education attainment is 0.09 standard deviations. Note that these treatment effects are statistically distinguishable at the 10 percent level in the widest bandwidth. Converting these estimates into a direct effect, adults born in counties with high education attainment see a 0.04 standard deviation increase in their self-sufficiency score from a $1,000 income shock, but adults from counties with high education attainment saw a 0.12 standard deviation decrease. The young adults from counties with low education attainment who experience the income increase see a 1.0 percentage point increase in the probability of graduating high school off a baseline mean of 88.7%, and an increase of annual earned income in $679 off a baseline mean of $19,280, although again none of these effects are distinguishable from 0 at the 10 percent level.

When combining all young adults aged 19-27, the estimated treatment effect is -0.003 standard deviations for adults born in counties with higher average mothers' education attainment, and 0.07 standard deviations for adults born in counties with lower average mother's education attainment. Converting to a direct effect, adults born in counties with higher education attainment see an 0.016 standard deviation increase in their self-sufficiency score from $1,000 in income in infancy. adults born in counties with lower education attainment saw an 0.098 standard deviation decrease from the same shock.

Finally, looking at adults aged 28-32, the estimated treatment effect is 0.01 standard deviations for adults born in counties with higher average mothers' education attainment and -0.12 standard deviations for adults born in counties with lower average mothers' education attainment. Converting both in to direct effects, adults born in counties with higher average mothers' education attainment saw an 0.012 standard deviation increase in their self-sufficiency score from a $1,000 shock in infancy, but adults born in counties with lower education attainment saw a 0.20 standard deviation decrease. Young adults from counties with low education attainment who experience the income increase see a 1.4 percentage point decrease in the probability of having graduated high school off of a control mean of 90.5% and a $150 decrease in annual earned income off of a control mean of $29,120. Neither of these effects are distinguishable from 0.

These long-term effects tell a consistent story: while effects of the income increase in infancy seem to persist in terms of impacts on education attainment and earnings after turning 19, these impacts apparently attenuate with time as students age into their late 20s and early 30s. Again, as before, estimated effects are largest for groups that had lower average income at birth, specifically Black adults and adults born in counties with lower average education attainment. It is possible that the lower effects measured here at later ages reflect the fact that the cohorts analyzed in these regressions would have been born in the early 1980s when take-up of tax benefits may have been lower, and the size of the first stage jump in after-tax income in infancy more inconsistent. Future research will need to follow the current and future cohorts of graduates to see if their effects are similar to the effects measured here.

1.6 Discussion

The effects found in this research show a relationship between income in infancy and educational outcomes while in school. These estimated effects appear to persist as differences in income, education attainment and labor force attachment into early adulthood for at least some subgroups. It is difficult to directly relate these findings to other estimates. Few other papers have used such a specific, sharply defined, and relatively modest change of income in the first year of a child's life. However, some comparisons are possible to other research on the effect of family income on child outcomes.

First, the results here suggest a non-linear relationship between family income and student achievement that has been found in other settings from changes in permanent income. The effect of an additional $1,000 in transitory income in infancy on outcomes is largest for groups that likely had lower average earnings in the first year of a child's life, including Black children and children with mothers with lower education attainment. Similarly, Loken, Mogstad and Wiswall (2012) and Akee et al. (2010) find that changes in permanent family income for lower-income families have the largest impacts on outcomes for children in school and in early adulthood.

Second, this paper suggests that a $1,000 change in family income in infancy results in changes in school performance, and other papers find that similarly-sized income shocks later in a child's life also have effects on school performance. Both Chetty, Friedman and Rockoff (2011) and Dahl and Lochner (2012) find that $1,000 of contemporaneous income results in a 0.06 to 0.09 standard deviation rise in contemporaneous test scores. Black et al. (2014) find that a $1,000 income shock at age 5 results in a 0.1 to 0.6 standard deviation increase in test scores at age 15. These papers do not consider grade-for-age status, likely because there is less year-to-year variation in that measure compared to test scores. However, such changes in tests scores, especially if they happen in the lower part of the test score distribution, may have non-trivial impacts on retention. Data from Florida on test scores and retention patterns suggest that a 0.06 to 0.09 standard deviation increasein test scores correlates to a reduction in the probability of students being retained in grade 4 by 0.6 to 0.8 percentage points.[34] While this relationship from the Florida data is not causal, it is suggestive that changes in test scores from a $1,000 change in after-tax income may result in similar effects on retention as those measured in this paper.

Third, this paper finds that a $1,000 change in income in infancy results in modest long-term changes in outcomes in adulthood, and other papers show a similar relationship. Chetty, Friedman and Rockoff (2011) provide a method of linking changes in test scores to changes in future earnings. They then use these estimates to convert the impact of $1,000 in after-tax income in childhood on test scores into the impact of the income shock on later life earnings of adults. Using this method, they con-

[34] This estimate comes from the evidence reported in Schwerdt, West and Winters (2017). In Figure 2A of their paper, the authors offer average retention rates by test scores. In Appendix Figure A-2 the authors show the distribution of test scores. Shifting the distribution of test scores in the lower regions up by 0.06 to 0.09 standard deviations produces the 0.6 to 0.8 percentage point reduction in retention. Baseline retention rate in this data among all students is 1.87%.

clude that a \$1,000 increase in after-tax income when children are in later primary and high school grades results in a 0.38 to 0.57 percentage point increase in earnings as adults. Similar sized effects are present in this paper from an income shock in infancy for some subgroups. The point estimates in this paper show that a \$1,000 increase in income in infancy results in a 0.56 percentage point increase in earned income for Black young adults from ages 20-30. The same income shock results in a 0.60 percentage point increase in earned income for young adults born in counties with low average education attainment. Both estimates, it should be noted, are not distinguishable from 0 at the 10 percent level, and there are minimal effects in the population at large. Nevertheless, it is suggestive that these point estimates are within similar ranges as Chetty, Friedman and Rockoff (2011).

However, while the pattern of results in this paper fit within the pre-existing literature, the magnitudes of these estimated effects are often near or above the upper bound of previous estimates of impacts. Arguably, the larger relationships found here reflect the fact that this paper looks at the effect of family income in infancy, while other papers primarily focus on shocks to income that happen later in a child's life. To think about the context for this difference, it is necessary to look more broadly at the literature on links between experiences in childhood and later life outcomes.

A wide array of research in social science suggests that family conditions in infancy and early childhood are particularly consequential for patterns of long-term development for children. First, gaps in measured cognitive and non-cognitive abilities between children open up at early ages and are observable clearly before students enter school (Loeb and Bassok, 2007; Cunha and Heckman, 2007). Similar gaps open up in many measures of child health (Figlio et al., 2014; Case, Lubotsky and Paxson, 2002; Currie and Almond, 2011). These gaps are highly correlated with family economic resources. Second, a literature in biology suggests the existence of critical periods for development where inputs are especially important for later life outcomes (Reviewed in Cunha et al. (2006)). Lastly, research shows that some policy interventions that affect the resources available to low-income families can have both short-term consequences (Hoynes, Miller and Simon, 2015; Almond, Hoynes and Schanzenbach, 2011; Rossin-Slater, 2013) and long-term consequences for outcomes for children (Black et al., 2014; Hoynes, Schanzenbach and Almond, 2016; Aizer et al., 2016*a*; Milligan and Stabile, 2009). Those papers find effects across health, cognitive skills, non-cognitive skills, and other metrics of child development. Thus, it would not be surprising that an income shock in infancy would relate to multi-faceted improvements in outcomes for children that may have different long-term effects than income shocks later in life.

The literature on the effects of family conditions in infancy and early childhood on later life outcomes offers a few clues as to potential mechanisms. Disadvantaged families with infants are likely to be income constrained. Over the sample period included here, around 50% of Black newborns and 35% of newborns in families where the mother has a high school degree or less are in poverty. By the time those children turn 15, the shares of those families in poverty drop to 40% and 23% respectively. Releasing this constraint may have three impacts on families. First, changes in income of these families in infancy might have significant impacts on consumption patterns. Differences in income between families correlate to differences in spending patterns on children (Caucutt, Lochner and Park, 2017). Research shows that changes in income from tax credits result in changes in spending on resources that might affect child development (McGranahan and Schanzenbach, 2013). Even if parents do not spend the money directly on their children, they may spend it on goods that increase the family's earnings over time. For example, research suggests that EITC recipients use the increase in their after-tax income from the EITC to pay down debt and spend on transportation (Goodman-Bacon and McGranahan, 2008; Mendenhall et al., 2012).[35] Second, to the degree that these spending patterns might enable slightly higher labor force attachment in subsequent years, such patterns may increase the family's permanent income (Ramnath and Tong, 2017; Black et al., 2014). Third, even if consumption patterns on children and permanent income are unaffected, the simple act of loosening the family's budget constraint may have impacts on how parents interact with their children. Research has found that parental stress, parental depression, and martial conflict are all highly correlated with family income, and in turn correlated with adverse outcomes for children (Wadsworth et al., 2005; Conger et al., 1994; Gershoff et al., 2007). Thus, even small changes in the economic resources of families can have consequences for important early life experiences of children, either through changes in consumption patterns, changes in permanent income, or changes in the family environment.

Finally, note that the experiment created by the income variation in this paper has interesting consequences for policy. First, the results suggest that shifting the eligibility for child-related deductions and credits a year earlier would improve students' achievement in school. Second, the results also suggest that shifting eligibility for these tax benefits forward while removing eligibility for an additional year in adolescence may improve some outcomes in adulthood. Families with children born in January are eligible for an additional year of tax benefits after children born in December are no longer eligible. But, adults born in December, especially from groups that were more likely disadvantaged at birth, still see an increase in the self-sufficiency score as adults from the income shock in infancy. Thus, the benefits that children born in January receive from that additional year of eligibility do not completely counteract the benefits that children born in December received from that year of eligibility as infants. The cost of implementing such a policy would simply come from

[35] This research looks at spending of these recipients on average and does not specifically look at spending of parents with newborns.

altering children's age of eligibility.[36]

A full cost-benefit analysis of the effects of shifting the eligibility timeline forward is beyond the scope of this paper. Such a calculation would require taking into account all the benefits that researchers have from that additional year of eligibility (e.g. including increased college enrollment (Manoli and Turner, 2018). However, these results are suggestive that benefits geared towards families with younger children may have lasting repercussions in ways that benefits aimed at families with older children do not. Most transfer programs, including SNAP and the tax credits analyzed in this paper, do not change benefit levels in ways that relate to a child's age.[37] But, the natural experiment created by this setting suggests that increasing these transfers to families with young children may offer a cost-effective reform that would improve outcomes for children and adults.

1.7 Conclusion

This paper demonstrates compelling effects of family income in infancy on outcomes in childhood and early adulthood. Specifically, this paper shows that a \$1,000 change in family income in infancy results in a 1.2 percentage point increase in the probability of a student being grade-for-age in high school. These results are driven by large treatment effects for children more likely disadvantaged in infancy, specifically Black children and children from families with low education attainment. Small but suggestive effects on adult outcomes in earnings, labor force attachment, high school graduation status and SNAP usage persist into early adulthood, in particular among Black young adults, and adults from counties with low education attainment. As

[36] As discussed in Appendix B, the share of families that receive EITC benefits for older children is substantially lower than the share of families who receive them for newborns. So, altering the age of eligibility would also result in an increase in receipt of EITC benefits, and hence additional costs. For more on these points, turn to Appendix B.

[37] A clear exception is the Special Supplemental Nutrition Program for Women, Infants, and Children (WIC) Program which is aimed at parents with infants and children up to age 5.

the effects of an additional $1,000 in infancy are largest for children from these more likely disadvantaged groups, they suggest a non-linear relationship between changes in income and changes in child outcomes.

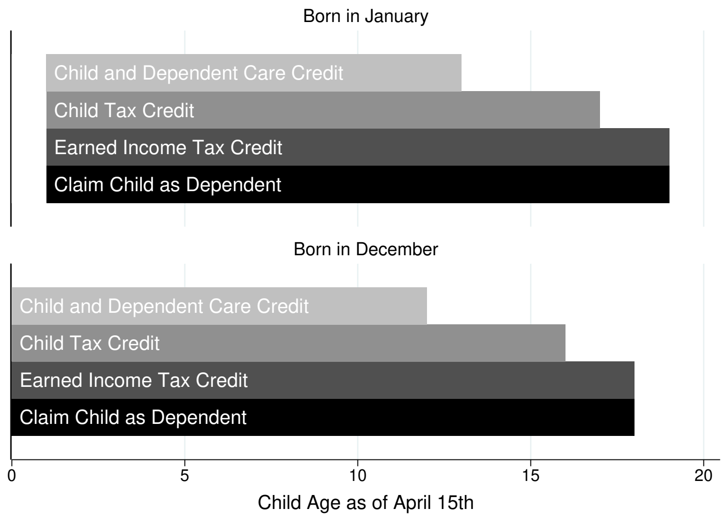
These results are on the upper end of estimated relationships between family income and outcomes for children. However, they fit in line with a broad literature suggesting that changes in family economic resources in infancy may have substantial long-term impacts on outcomes for children. This increase in income could affect children's outcomes through changing family spending patterns, improving future family earnings, or changing the home life circumstances that young children face early in life.

Furthermore, it is notable that these results come from altering timing of receipt of tax benefits from adolescence to infancy. These results may indicate that transfer programs focused on families with very young children may result in larger effects on child and adult outcomes than transfer programs aimed at families with older children. More broadly, these results suggest that altering transfer programs to be more child age-specific may a fruitful and low-cost avenue for policy reform.

In all, these results suggest that changing the resources available to low-income families can result in long-term improvements for their children. Directions for future research in this project include examining effects on siblings, and investigation into mechanisms of effects in consumption data.

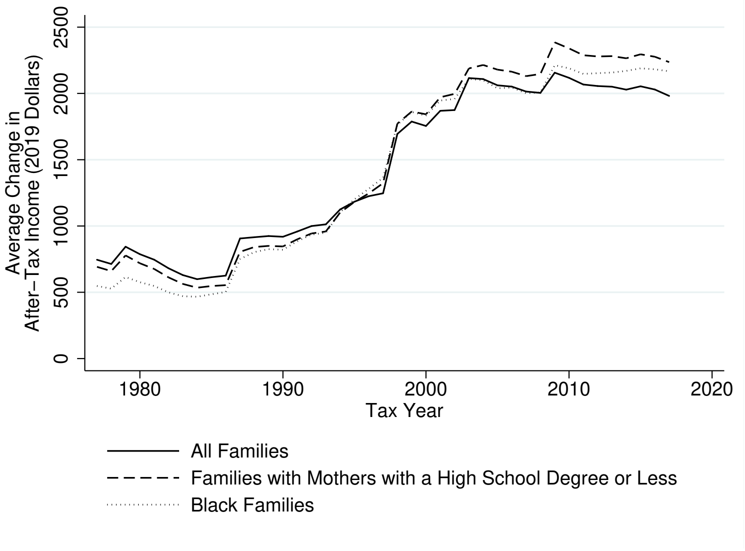
1.8 Figures and Tables

Figure 1.1: Eligibility for Child Tax Benefits for Children Born in December and January by Child Age



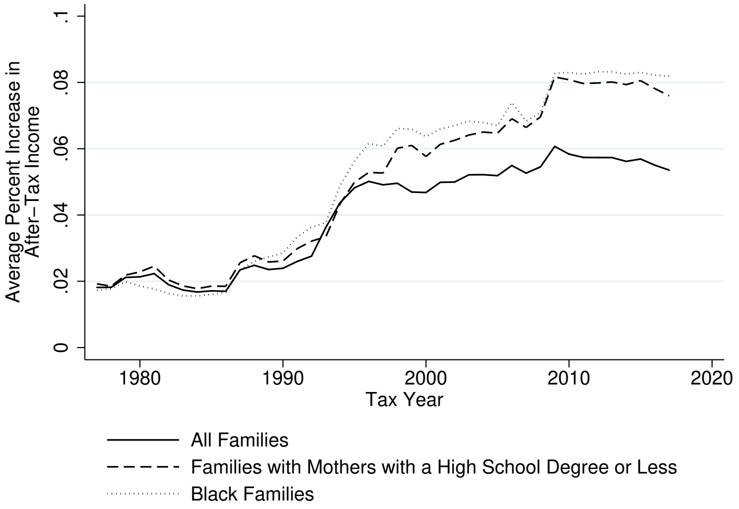
Notes: Figure depicts eligibility for tax benefits by child age and birth month. The age variable on the horizontal axis lists age as would be recorded by a family on April 15th. For example, newborns in their first year of life born in January and December would be age 0 by April 15th.

Figure 1.2: Family Tax Benefit in Infant's First Year of Life from Birth in December Compared to January



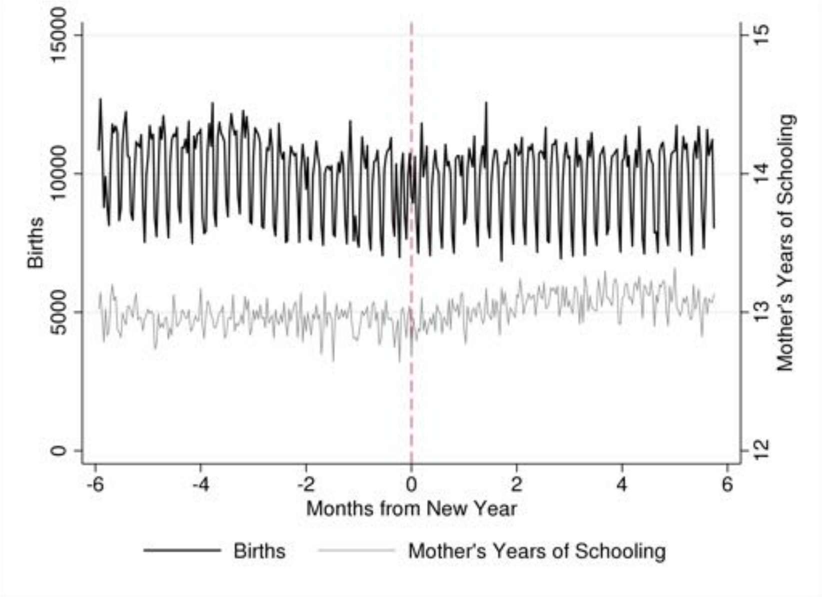
Notes: Figure depicts average estimated difference in family after-tax income in the first year of a child's life for families that have a child born in December compared to January of the next year. Incomes measured in 2019 dollars. Year variable on horizontal axis records tax year of birth. For example, the difference reported for 1986 measures the difference in after-tax income in tax year 1986 for having a child born in December 1986 compared to January 1987. Estimation process draws inspiration from Hoynes, Miller and Simon (2015) and uses the March CPS. Additional details on estimation are in the text and in Appendix A. Standard error bars here omitted for clarity, but standard errors are less than $10 for all groups and all years.

Figure 1.3: Percent Increase in Family After-Tax Income in Infant's First Year of Life from Birth in December Compared to January



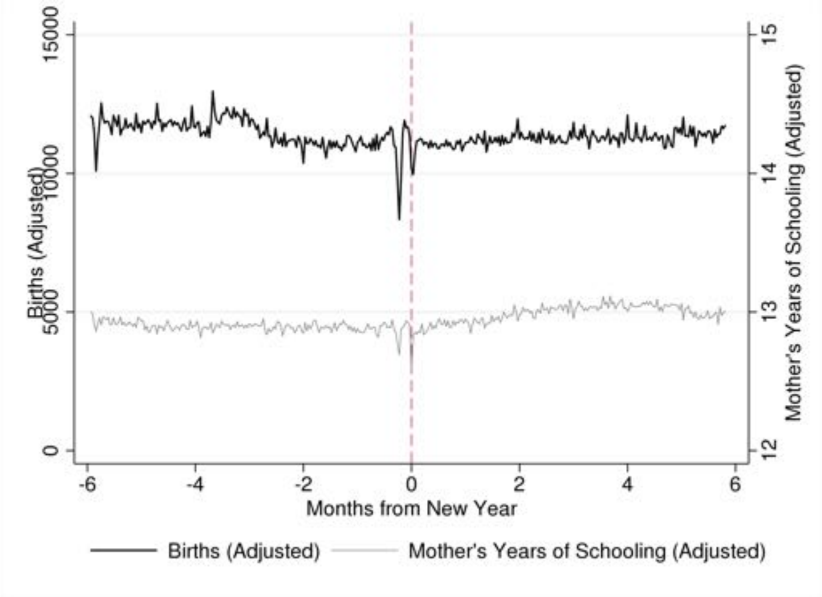
Notes: Figure depicts average percent increase in after-tax family income in the first year of a child's life for families that have a child born in December compared to January of the next year. Year variable on horizontal axis records tax year of birth. For example, the difference reported for 1986 measures the difference in after-tax income in tax year 1986 for having a child born in December 1986 compared to January 1987. Figure uses same estimation process as described in Figure 1.2, the main text, and Appendix A. Standard error bars omitted for clarity, but standard errors are less than 0.2 percentage points for all groups and all years.

Figure 1.4: Births by Day of Year - 1996 to 1997



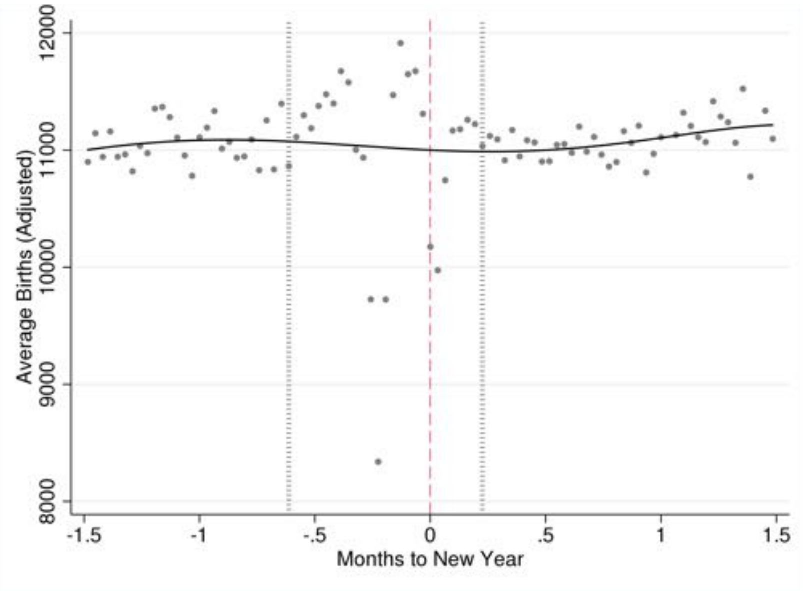
Notes: Figure depicts birth counts by day of year estimated in the 2000 Census from July 1st 1996 to June 30th 1997, centered on New Year's Day in 1997.

Figure 1.5: Births by Day of Year Adjusted by Day of Week



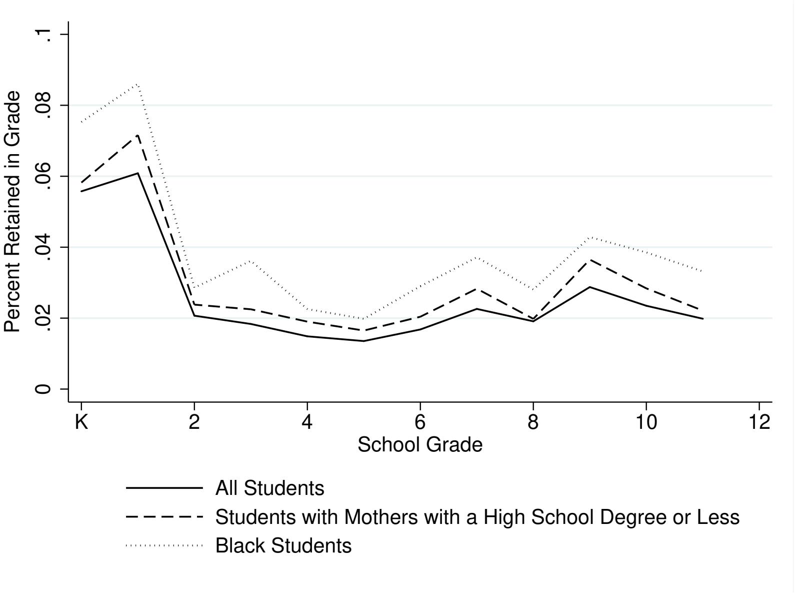
Notes: Figure depicts average births by day of year from 1989-1994 regression-adjusted for day of birth following equations 1.1 and 1.2.

Figure 1.6: Estimated Birth Timing Manipulation



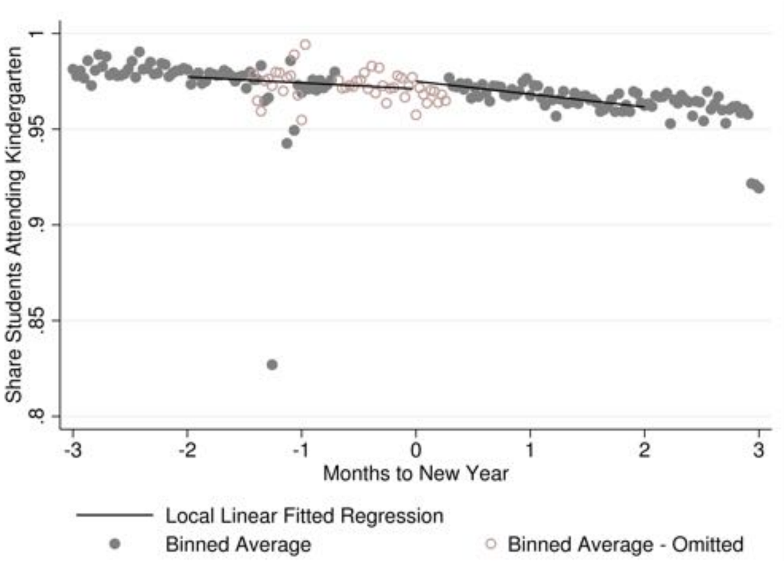
Notes: Figure depicts average births by day of year from 1989-1994 regression-adjusted for day of birth following equations 1.1 and 1.2. Vertical bars indicate manipulated region omitted from calculation. Upper bound selected visually at 9 days after the New Year. Lower bound selected through estimation process described in the text.

Figure 1.7: Average Share of Students Retained in Grade



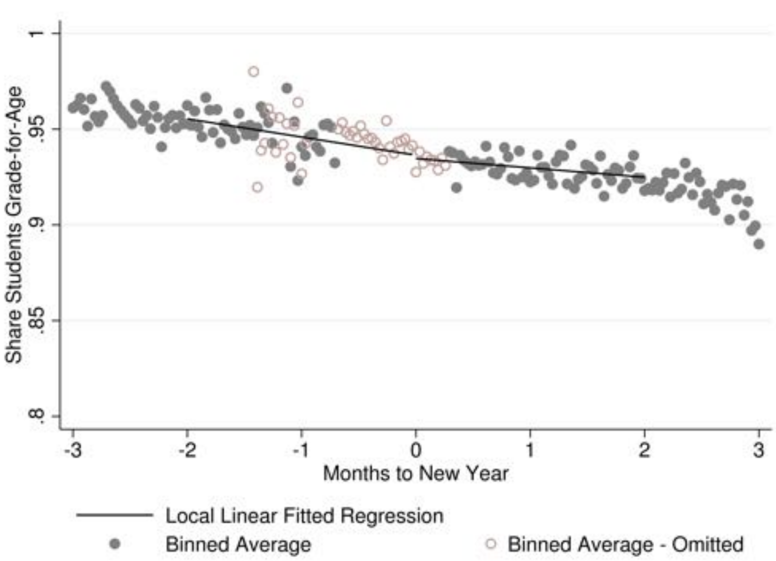
Notes: Figure depicts average share of students retained in each grade. Values estimated in the October CPS with data over the years 1990 to 2005. Standard error bars omitted for clarity, but are less than 0.1 percentage points across all groups and years.

Figure 1.8: Estimated Reduced Form Discontinuities in Grade-for-Age Status - Kindergarten



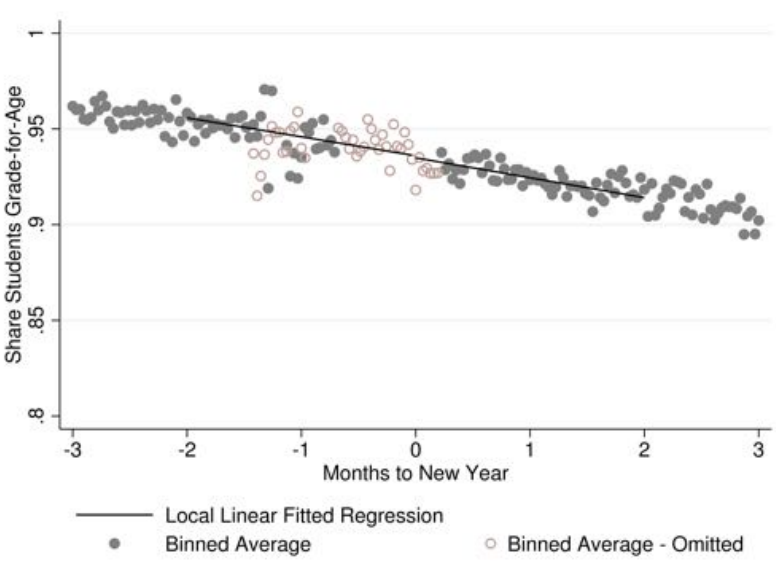
Notes: Figures depicts discontinuity in share of students attending kindergarten around the New Year. Red empty circles are data omitted from estimation process, and grey solid circles are data that could be included. The estimated line uses a bandwidth of two months around the New Year, and the solid grey circles covered by the estimated line represent data included in the estimation process. See Table 1.2 for point estimates. Regressions include fixed effects by day of week, and state of birth. Estimation process detailed in text.

Figure 1.9: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 1st Grade



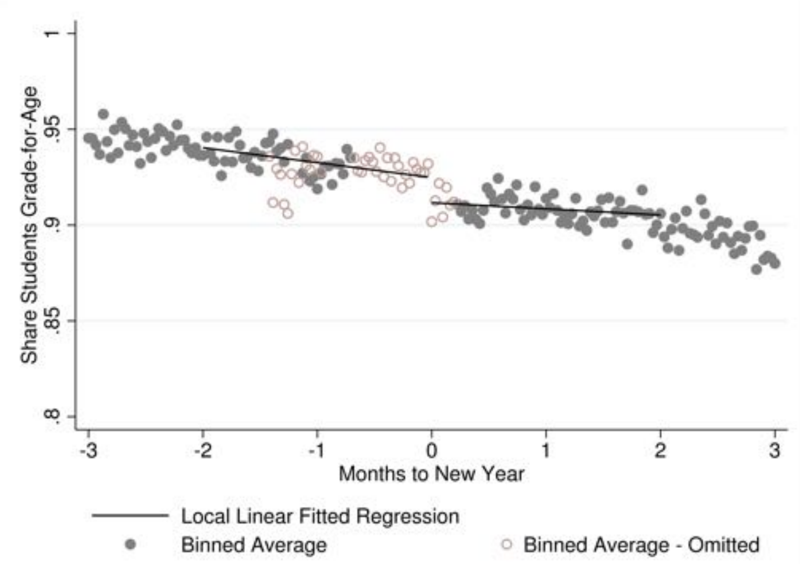
Notes: Figures depicts discontinuity in share of students grade-for-age in 1st grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.10: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 5th Grade



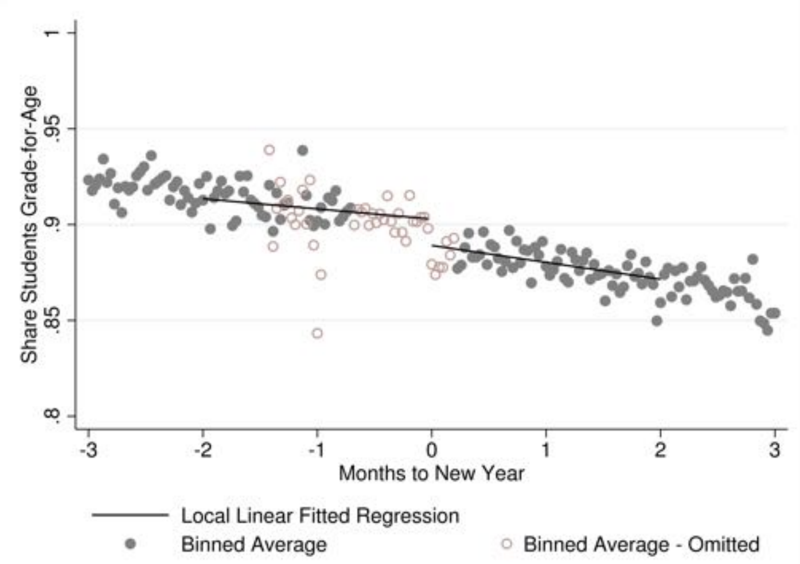
Notes: Figures depicts discontinuity in share of students grade-for-age in 5th grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.11: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 7th Grade



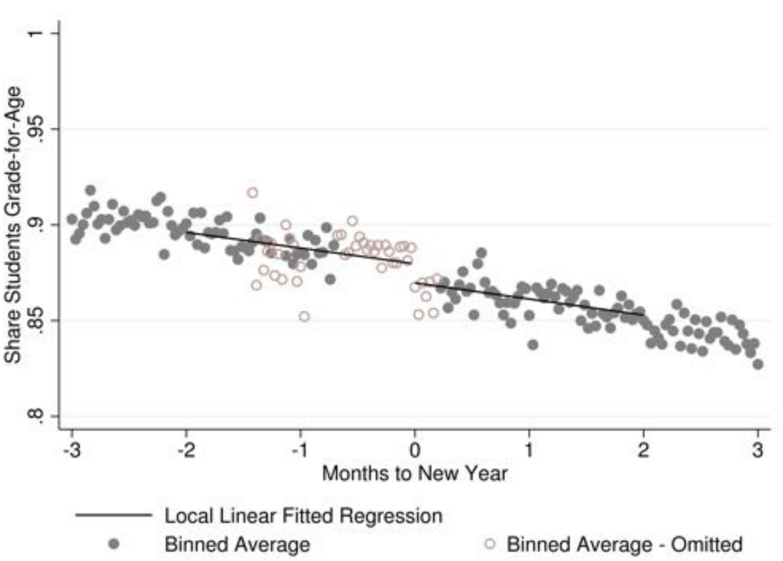
Notes: Figures depicts discontinuity in share of students grade-for-age in 7th grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.12: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 9th Grade



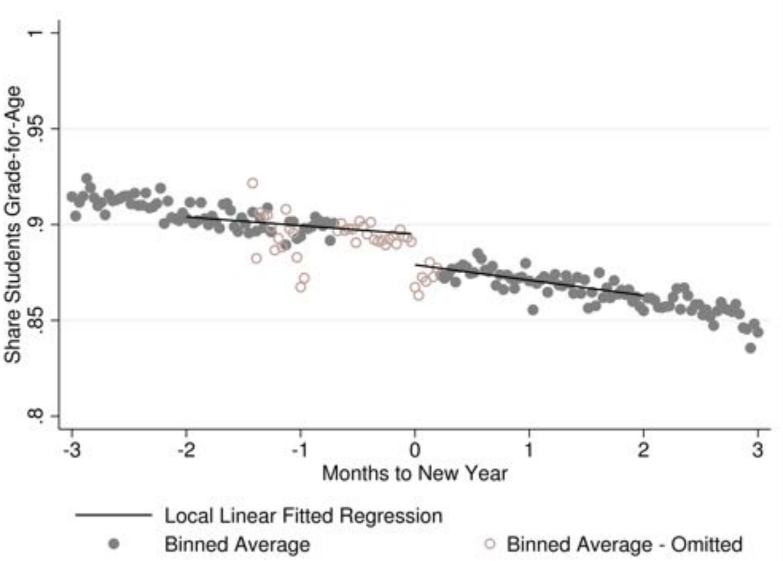
Notes: Figures depicts discontinuity in share of students grade-for-age in 9th grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.13: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 10th Grade



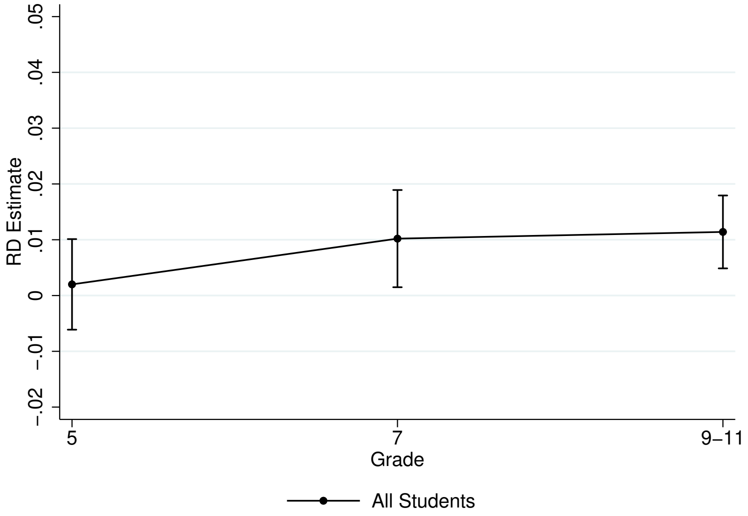
Notes: Figures depicts discontinuity in share of students grade-for-age in 10th grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.14: Estimated Reduced Form Discontinuities in Grade-for-Age Status - 9th-11th Grade



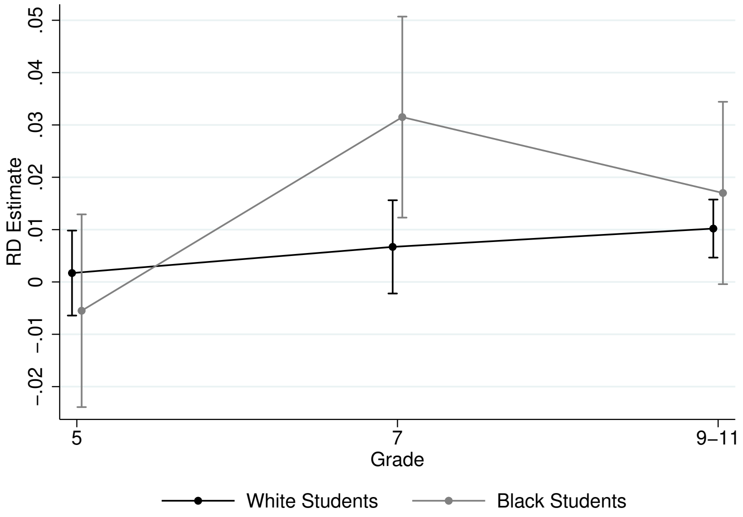
Notes: Figures depicts discontinuity in share of students grade-for-age in 9th-11th grade around the New Year. See notes to Figure 1.8 for more detail.

Figure 1.15: IV Treatment Effect of $1,000 in Infancy in Grade-for-Age Status by Grade



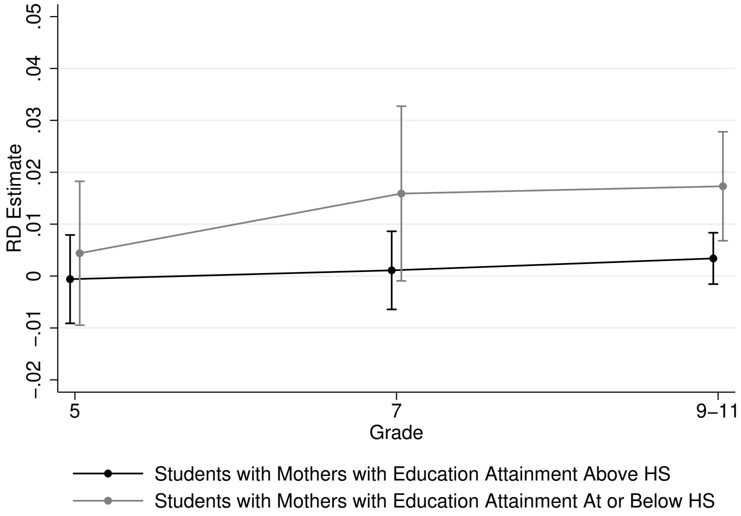
Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on grade-for-age status in 5th, 7th and 9th-11th grades recorded in Table 1.2 with a bandwidth of two months. Regressions include fixed effects by day of week, and state of birth fixed effects. Standard errors calculated with 2,000 bootstrap replications. Estimation process detailed in text.

Figure 1.16: IV Treatment Effect of $1,000 in Infancy in Grade-for-Age Status by Grade - Separated by Race



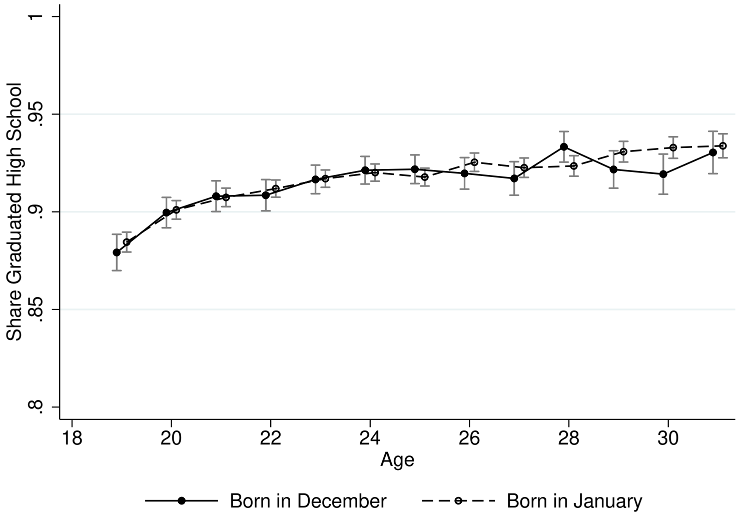
Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on grade-for-age status for White and Black children separately in 5th, 7th, and 9th-11th grades recorded in Table 1.3 with a bandwidth of two months. See additional details in Figure 1.15

Figure 1.17: IV Treatment Effect of $1,000 in Infancy in Grade-for-Age Status by Grade - Separated by Mother's Education Attainment



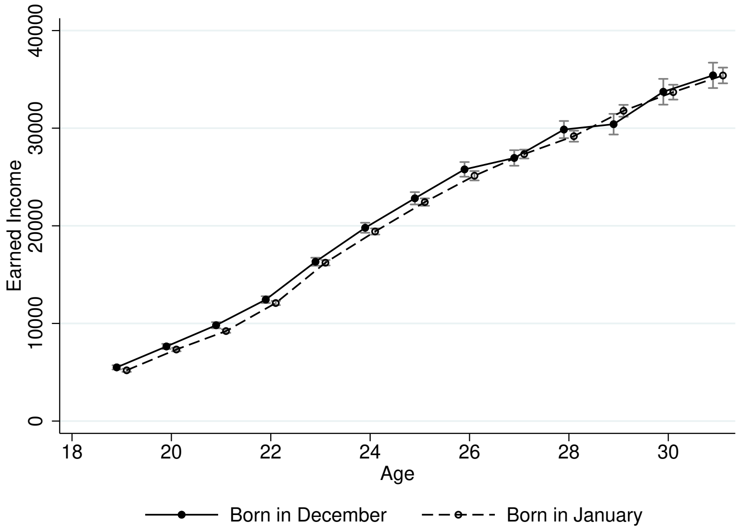
Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on grade-for-age status for children with mothers separated by education attainment in 5th, 7th and 9th-11th grades recorded in Table 1.3 with a bandwidth of two months. See additional details in Figure 1.15

Figure 1.18: Share Adults Graduated High School by Age and Birth Month



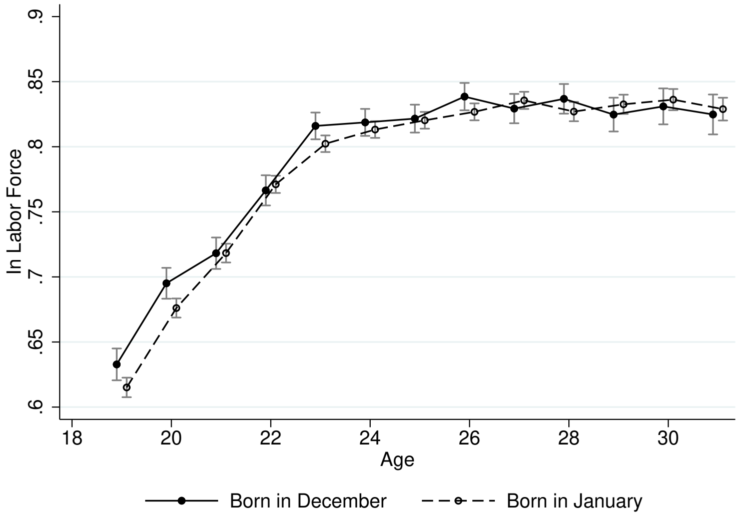
Notes: Figure depicts average share that have graduated high school in the full sample of adults by age and month of birth, omitting adults born December 11th through January 9th. ”December” births are children born from November 15th to December 10th, and ”January” births are children born from January 10th to February 15th.

Figure 1.19: Average Earned Income of Adults by Age and Birth Month



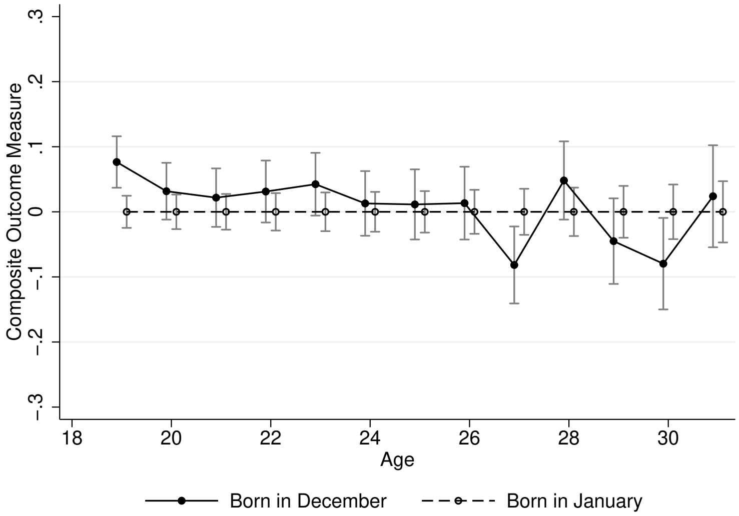
Notes: Figure depicts average earned income in the full sample of adults by age and month of birth, omitting adults born December 11th through January 9th. ”December” births are children born from November 15th to December 10th, and ”January” births are children born from January 10th to February 15th.

Figure 1.20: Share Adults in Labor Force by Age and Birth Month



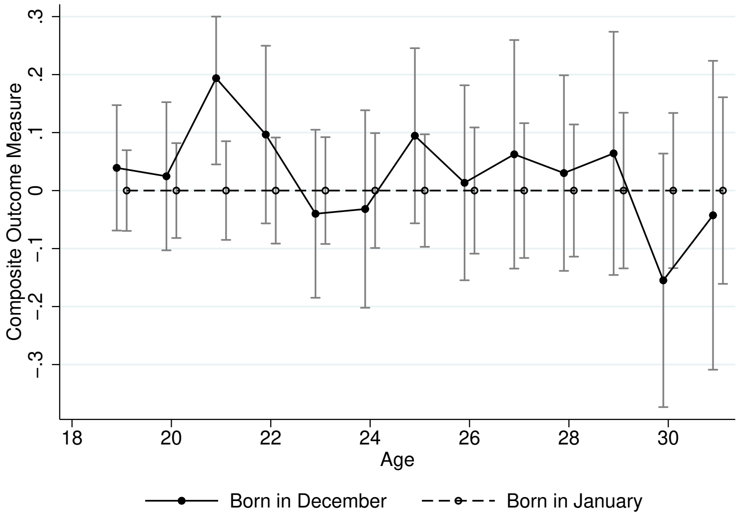
Notes: Figure depicts average share that in the labor force in the full sample of adults by age and month of birth, omitting adults born December 11th through January 9th. "December" births are children born from November 15th to December 10th, and "January" births are children born from January 10th to February 15th.

Figure 1.21: Average Composite Measure of Outcomes by Age and Birth Month



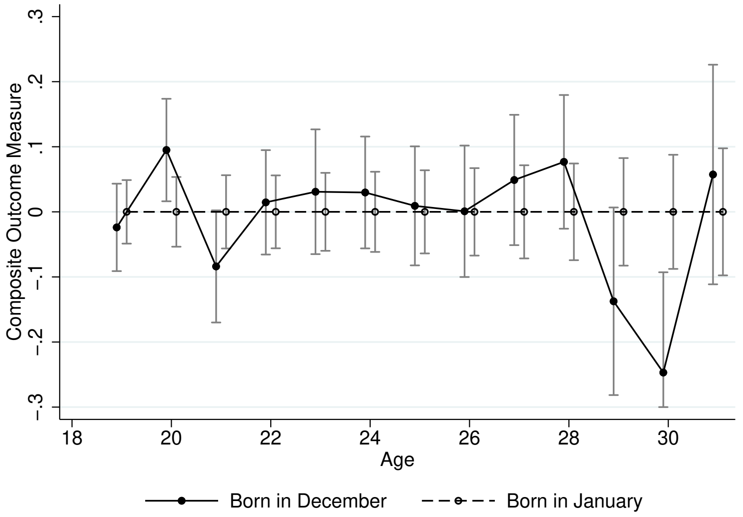
Notes: Figure depicts average trends in a composite measure of adults' outcomes by age and birth month, omitting adults born December 11th through January 9th. "December" births are children born from November 15th to December 10th, and "January" births are children born from January 10th to February 15th. The composite measure reflects labor force participation, earned income, SNAP receipt and high school graduation status. The process that creates this composite measure is described in text. Note that the measure takes on average value 0 for individuals born after New Year's Day by construction, but there is a standard error present due to sampling variation.

Figure 1.22: Average Composite Measure of Outcomes for Black Adults by Age and Birth Month



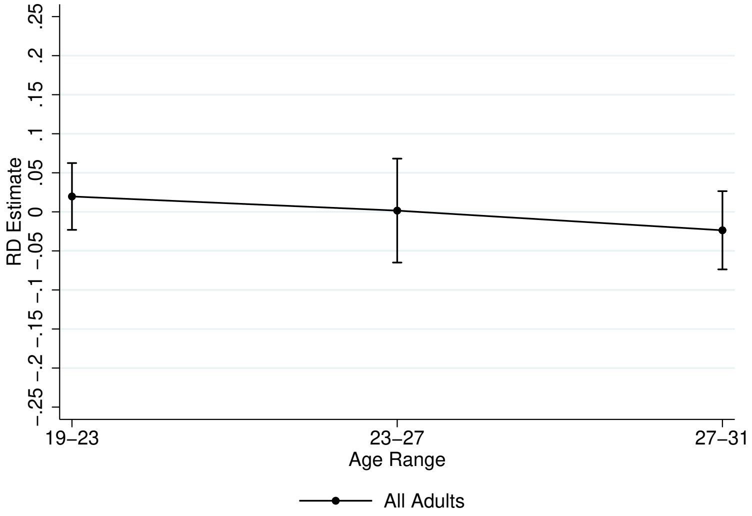
Notes: Figure depicts average trends in a composite measure of Black adults' outcomes by age and birth month, omitting adults born December 11th through January 9th. See additional details in Figure 1.21.

Figure 1.23: Average Composite Measure of Outcomes for Adults Born in Counties with Lower Education Attainment by Age and Birth Month



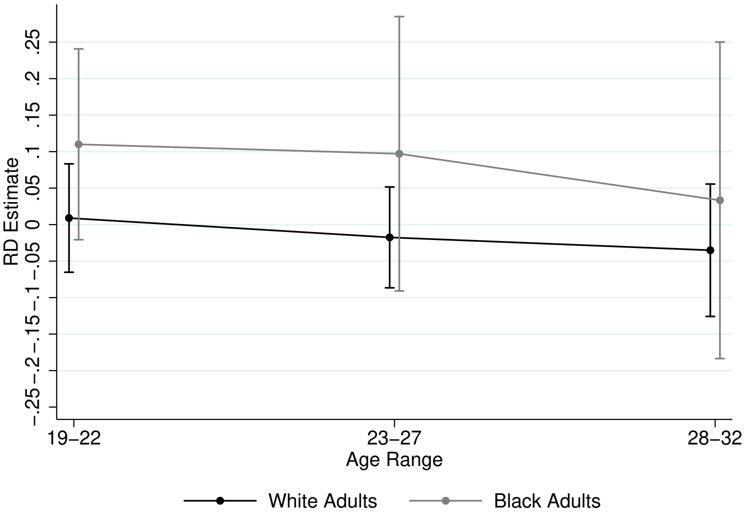
Notes: Figure depicts average trends in a composite measure of outcomes by age and birth month for adults born in counties with lower education attainment, omitting adults born December 11th through January 9th. See additional details in Figure 1.21.

Figure 1.24: Estimated IV Treatment Effect of $1,000 in Infancy on Composite Measure of Outcomes by Age Group



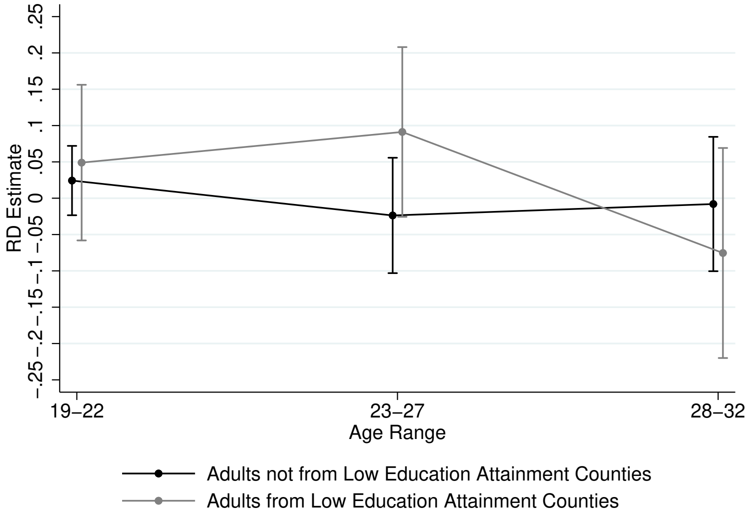
Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on composite measure of outcomes for adults aged 19-22, 23-27 and 28-32. Results recorded in Tables 1.7 with a bandwidth of two months. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation process detailed in text.

Figure 1.25: Estimated IV Treatment Effect of $1,000 in Infancy on Composite Measure of Outcomes by Age Group - Separated by Race



Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on composite measure of outcomes for White and Black adults separately aged 19-22, 23-27 and 28-32. Results recorded in Table 1.8 with a bandwidth of two months. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation process detailed in text.

Figure 1.26: Estimated IV Treatment Effect of $1,000 in Infancy on Composite Measure of Outcomes by Age Group - Separated by Birth County



Notes: Figures depicts estimated instrumental variable treatment effect of $1,000 in infancy on composite measure of outcomes for adults aged 19-22, 23-27 and 28-32. Results are separated into adults born in counties with average mothers' education attainment in the lowest quartile, and adults born into counties with average mothers' education attainment in higher quartiles. Results recorded in Table 1.9 with a bandwidth of two months. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation process detailed in text.

Table 1.1: Validating Regression Discontinuity Procedures

Outcome	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Child is White	0.725	-0.0410	-0.0227*	-0.0172*	-0.0119
	(0.001)	(0.0258)	(0.0119)	(0.0097)	(0.0340)
Child is Black	0.117	0.00140	0.00400	0.00240	0.00210
	(0.001)	(0.0129)	(0.0066)	(0.0055)	(0.0069)
Child is non-White, non-Black	0.159	0.0396**	0.0187**	0.0147*	0.00980
	(0.001)	(0.0193)	(0.0092)	(0.0077)	(0.0279)
Child State of Residence Same as Birth	0.955	-0.00430	-0.00240	-0.00480	-0.00120
	(0.001)	(0.0101)	(0.0053)	(0.0042)	(0.0045)
Total Children in Household	1.937	-0.0480	-0.0295	-0.0299	-0.0155
	(0.001)	(0.0466)	(0.0218)	(0.0197)	(0.0450)
Child Live with Both Parents	0.706	0.00980	-0.00900	-0.00560	-0.00470
	(0.001)	(0.0235)	(0.0122)	(0.0093)	(0.0147)
Child's Household Has Any Earned Income	0.807	0.0457**	0.0144	0.00600	0.00760
	(0.001)	(0.0178)	(0.0092)	(0.0077)	(0.0218)
Child's Household Has Any Other Income	0.112	-0.00510	0.000500	0.000600	0.000300
	(0.001)	(0.0130)	(0.0078)	(0.0061)	(0.0042)
Child's Household Has Any Retirement Income	0.0300	-0.00390	0.00290	0.00440	0.00150
	(0.001)	(0.0066)	(0.0039)	(0.0031)	(0.0048)

Notes: Table records estimated discontinuities in child and family covariates for a child being born before the New Year, and an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy. Results estimated using children in the 2000 Census born between 1999 and 2000. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.1: Validating Regression Discontinuity Procedures (Continued)

Outcome	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Child's Household Has Any Supplemental Income	0.0150	0.00300	0.00330	0.00310	0.00170
	(0.001)	(0.0058)	(0.0035)	(0.0030)	(0.0051)
Child's Household Has Any Welfare Income	0.0600	-0.0138	-0.00200	-0.00340	-0.00110
	(0.001)	(0.0215)	(0.0105)	(0.0081)	(0.0063)
Child's Household's Earned Income	41500	2300	474.8	79.18	249.7
	(71600)	(1700)	(950)	(800)	(859.9)
Child's Household's Other Income	469.8	-8.781	4.689	20.45	2.465
	(182.3)	(85.16)	(53.63)	(42.86)	(29.04)
Child's Household's Suppemental Income	84.83	11.92	13.89	14.39	7.309
	(23.32)	(30.15)	(19.57)	(16.46)	(22.93)
Child's Household's Total Income	42000	1600	1300	814.7	683.6
	(84000)	(1600)	(843.7)	(712.7)	(1966.)
Child's Household's Wage Income	39500	1300	70.91	-137.6	37.28
	(68500)	(1800)	(950.9)	(790.2)	(510.8)
Child's Household's Welfare Income	119.3	-72.69**	-27.29	-18.18	-14.35
	(19.20)	(35.38)	(19.45)	(15.35)	(41.51)
Maximum Age of Parents	30.72	0.142	0.103	0.0206	0.0541
	(0.002)	(0.3087)	(0.1557)	(0.1246)	(0.1724)

Notes: *Continued from page above.* Table records estimated discontinuities in child and family covariates for a child being born before the New Year, and an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy. Results estimated using children in the 2000 Census born between 1999 and 2000. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.1: Validating Regression Discontinuity Procedures (Continued)

Outcome	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Either Parent has Any Wage Income	0.880	0.0174	0.00170	-0.00160	0.000900
	(0.001)	(0.0113)	(0.0069)	(0.0055)	(0.0044)
Either Parent has Any Welfare Income	0.0480	-0.00960	-0.00240	-0.00540	-0.00130
	(0.001)	(0.0132)	(0.0080)	(0.0066)	(0.0055)
Maximum Education Attainment of Parents	13.68	0.136	-0.00610	-0.0222	-0.00320
	(0.001)	(0.1117)	(0.0629)	(0.0489)	(0.0343)
Maximum Wage Income of Parents	33000	999	1000	806	525.8
	(54500)	(1600)	(848.2)	(670.9)	(1539.)
Either Parent is in Labor Force	0.897	0.00260	-0.00300	-0.00250	-0.00160
	(0.001)	(0.0104)	(0.0068)	(0.0049)	(0.0056)
Either Parent is Married	0.808	0.0147	0.00620	0.00640	0.00320
	(0.001)	(0.0169)	(0.0104)	(0.0082)	(0.0106)
Maximum Usual Hours of Work of Parents	41.24	0.248	0.0283	-0.0144	0.0149
	(0.013)	(0.9542)	(0.5250)	(0.4227)	(0.2792)
Maximum Weeks of Work Last Year of Parents	43.04	0.842	-0.0117	-0.0482	-0.00620
	(0.013)	(0.9000)	(0.4911)	(0.4152)	(0.2588)
Either Parent Worked Last Year	0.936	0.00840	0.00260	0.00340	0.00130
	(0.001)	(0.0081)	(0.0052)	(0.0040)	(0.0046)

Notes: *Continued from page above.* Table records estimated discontinuities in child and family covariates for a child being born before the New Year, and an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy. Results estimated using children in the 2000 Census born between 1999 and 2000. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.1: Validating Regression Discontinuity Procedures (Continued)

Outcome	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Age of Mother	28.44	0.428	0.0868	0.0229	0.0457
	(0.002)	(0.3302)	(0.1772)	(0.1443)	(0.1583)
Mother Has Any Wage Income	0.681	0.0548**	0.0192	0.0111	0.0101
	(0.001)	(0.0236)	(0.0133)	(0.0109)	(0.0291)
Mother Has Any Welfare Income	0.0480	-0.00810	-0.00660	-0.00860	-0.00350
	(0.001)	(0.0122)	(0.0072)	(0.0060)	(0.0104)
Mother's Education Attainment	13.27	0.3927***	0.0721	0.0196	0.0379
	(0.001)	(0.1312)	(0.0841)	(0.0676)	(0.1152)
Mother's Wage Income	15000	2900***	1300**	850.0*	683.6
	(26500)	(1000)	(567.4)	(466.8)	(1939.)
Mother is in Labor Force	0.554	0.0302	0.00210	0.00100	0.00110
	(0.001)	(0.0295)	(0.0156)	(0.0123)	(0.0088)
Mother is Married	0.836	0.00420	0.00560	0.00840	0.00300
	(0.001)	(0.0175)	(0.0107)	(0.0079)	(0.0100)
Mother is Single Household Head	0.0770	0.00570	0.0127**	0.00730	0.00670
	(0.001)	(0.0106)	(0.0061)	(0.0052)	(0.0190)
Mother's Usual Hours of Work	25.86	1.949**	0.8732*	0.653	0.459
	(0.022)	(0.8465)	(0.4650)	(0.3950)	(1.310)

Notes: *Continued from page above.* Table records estimated discontinuities in child and family covariates for a child being born before the New Year, and an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy. Results estimated using children in the 2000 Census born between 1999 and 2000. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.1: Validating Regression Discontinuity Procedures (Continued)

Outcome	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Mother's Weeks of Work Last Year	29.36	2.157**	0.664	0.456	0.349
	(0.031)	(1.039)	(0.5903)	(0.5056)	(1.027)
Mother Worked Last Year	0.711	0.0454*	0.0179	0.0137	0.00940
	(0.001)	(0.0239)	(0.0132)	(0.0110)	(0.0272)

Notes: *Continued from page above.* Table records estimated discontinuities in child and family covariates for a child being born before the New Year, and an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy. Results estimated using children in the 2000 Census born between 1999 and 2000. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.2: Baseline Results for Regression Discontinuity Estimate of Treatment Effect on Grade-For-Age Status in School

Grade	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Pre-K	0.758	-0.00253	0.00401	0.00418	0.00238
	(0.001)	(0.01336)	(0.0081)	(0.0068)	(0.0052)
K	0.970	0.00610	-0.00230	-0.00220	-0.00150
	(0.001)	(0.0055)	(0.0025)	(0.0020)	(0.0016)
1st	0.931	0.00280	0.00520	0.00610	0.00350
	(0.001)	(0.0121)	(0.0059)	(0.0045)	(0.0039)
5th	0.915	-0.00520	-0.00180	0.00200	-0.00140
	(0.001)	(0.0083)	(0.0048)	(0.0041)	(0.0037)
7th	0.903	0.0158	0.0105*	0.0102**	0.0088*
	(0.001)	(0.0102)	(0.0057)	(0.0044)	(0.0048)
9th	0.878	0.0139**	0.0084**	0.0088***	0.0089**
	(0.001)	(0.0059)	(0.0042)	(0.0032)	(0.0044)
10th	0.864	0.00200	0.00560	0.00500	0.00630
	(0.001)	(0.0120)	(0.0066)	(0.0052)	(0.0074)
11th	0.855	0.0245***	0.0205***	0.0211***	0.0221***
	(0.001)	(0.0076)	(0.0043)	(0.0033)	(0.0047)
9th-11th	0.866	0.0123**	0.0113***	0.0114***	0.0120***
	(0.001)	(0.0059)	(0.0032)	(0.0024)	(0.0034)

Notes: Table records estimated discontinuity in grade-for-age status for a child being born before the New Year by expected grade of student for full sample. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on grade-for-age status. Results estimated using children in the 2000 Census and 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.3: Regression Discontinuity Estimates of Treatment Effect on Grade-For-Age Status in School by Race

Grade	Race	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
5th	White	0.922	0.000600	-0.00130	0.00170	-0.00100
		(0.001)	(0.0080)	(0.0049)	(0.0041)	(0.0038)
	Black	0.871	-0.0194	-0.0127	-0.00550	-0.0110
		(0.001)	(0.0188)	(0.0110)	(0.0093)	(0.0095)
	Difference		-0.0200	-0.0114	-0.00720	-0.0100
7th	White	0.912	0.00990	0.00680	0.00670	0.00560
		(0.001)	(0.0105)	(0.0059)	(0.0045)	(0.0049)
	Black	0.845	0.0218	0.0311**	0.0315***	0.0282***
		(0.001)	(0.0223)	(0.0118)	(0.0097)	(0.0107)
	Difference		0.0119	0.0244*	0.0248**	0.0226*
9th-11th	White	0.879	0.00720	0.0102***	0.0102***	0.0106***
		(0.001)	(0.0065)	(0.0036)	(0.0028)	(0.0037)
	Black	0.793	0.0207	0.0132	0.0170*	0.0160
		(0.001)	(0.0207)	(0.0111)	(0.0088)	(0.0134)
	Difference		0.0135	0.00310	0.00690	0.00540

Notes: Table records estimated discontinuity in grade-for-age status for a child being born before the New Year by expected grade of student among White and Black children. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on grade-for-age status. Results estimated using children in the 2000 Census and 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.4: Regression Discontinuity Estimates of Treatment Effect on Grade-For-Age Status in School by Mother's Education Level

Grade	Mother's Education Level	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
5th	Above High School	0.941	-0.00650	-0.00320	-0.000600	-0.00230
		(0.001)	(0.0078)	(0.0052)	(0.0043)	(0.0037)
	High School or Below	0.887	-0.00540	-0.00200	0.00440	-0.00170
		(0.001)	(0.0153)	(0.0072)	(0.0061)	(0.0060)
	Difference		0.00110	0.00130	0.00500	0.000600
7th	Above High School	0.932	-0.00120	-0.000700	0.00110	-0.000600
		(0.001)	(0.0081)	(0.0047)	(0.0038)	(0.0035)
	High School or Below	0.874	0.0207	0.0168	0.0159*	0.0153
		(0.001)	(0.0180)	(0.0107)	(0.0085)	(0.0109)
	Difference		0.0219	0.0175	0.0148	0.0159
9th-11th	Above High School	0.916	0.00350	0.00190	0.00340	0.00170
		(0.001)	(0.0058)	(0.0031)	(0.0025)	(0.0029)
	High School or Below	0.825	0.0105	0.0173**	0.0173***	0.0205**
		(0.001)	(0.0117)	(0.0067)	(0.0053)	(0.0097)
	Difference		0.00700	0.0155**	0.0139**	0.0187*

Notes: Table records estimated discontinuity in grade-for-age status for a child being born before the New Year by expected grade of student among children with different levels of mothers' education attainment. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on grade-for-age status. Results estimated using children in the 2000 Census and 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.5: Regression Discontinuity Estimate of Treatment Effect on Grade-For-Age Status in School - Children Living in Same State as Birth

Grade	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
		1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
5th	0.915	0.00150	0.00190	0.00420	0.00150
	(0.001)	(0.0093)	(0.0056)	(0.0047)	(0.0044)
7th	0.904	0.0177	0.0110*	0.0100**	0.0092*
	(0.001)	(0.0114)	(0.0063)	(0.0050)	(0.0053)
9th-11th	0.867	0.0172***	0.0129***	0.0125***	0.0138***
	(0.001)	(0.0055)	(0.0032)	(0.0025)	(0.0034)

Notes: Table records estimated discontinuity in grade-for-age status by grade of student for a child being born before the New Year among children living in the same state as birth. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on grade-for-age status. Results estimated using children in the 2000 Census and 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.6: Regression Discontinuity Estimates of Treatment Effect on Grade-For-Age Status in School by Cohort Year of Birth

Grade	Cohort Year of Birth	Control Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
9th-11th	1982-1986	0.873	0.00770	0.00740	0.00680	0.0114
		(0.001)	(0.0101)	(0.0054)	(0.0043)	(0.0084)
	1987-1993	0.856	0.00720	0.00740	0.0090*	0.00780
		(0.001)	(0.0122)	(0.0065)	(0.0051)	(0.0069)
	1994-2001	0.875	0.0244**	0.0123*	0.0114**	0.0090*
		(0.001)	(0.0121)	(0.0065)	(0.0051)	(0.0047)

Notes: Table records estimated discontinuity in grade-for-age status for a child being born before the New Year by expected grade of student among children from different birth year cohorts. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on grade-for-age status. Results estimated using children in the 2000 Census and 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.7: Baseline Results for Regression Discontinuity Estimates of Treatment Effects for Young Adults

Outcome	Age Range	Control Mean Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Composite Measure	19-27	0.0000	-0.0473	0.0028	0.0101	0.0036
		(1)	(0.0484)	(0.0261)	(0.0216)	(0.0334)
Composite Measure	19-22	0.0000	0.0481	0.0249	0.0197	0.0287
		(1)	(0.0597)	(0.0409)	(0.0336)	(0.0473)
Composite Measure	23-27	0.0000	-0.1204*	-0.0152	0.0016	-0.0201
		(1)	(0.0643)	(0.0301)	(0.0253)	(0.0397)
Composite Measure	28-32	0.0000	-0.0819	-0.0175	-0.0236	-0.0260
		(1)	(0.0748)	(0.0447)	(0.0374)	(0.0667)
Graduated High School	19-27	0.9161		0.0006	0.0012	0.0007
		(0.0006)		(0.0029)	(0.0023)	(0.0037)
Graduated High School	19-22	0.9092		0.0008	0.0011	0.0008
		(0.0009)		(0.0044)	(0.0038)	(0.0051)
Graduated High School	23-27	0.9210		0.0002	0.0011	0.0002
		(0.0007)		(0.0034)	(0.0028)	(0.0046)
Graduated High School	28-32	0.9321		-0.0047	-0.0044*	-0.0070
		(0.0009)		(0.0034)	(0.0026)	(0.0051)

Notes: Table records estimated discontinuity in adult outcomes for an adult being born before the New Year by age group for the full sample. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on adult outcomes. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.7: Baseline Results for Regression Discontinuity Estimates of Treatment Effects for Young Adults (Continued)

Outcome	Age Range	Control Mean Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Earned Income	19-27	16780		-143	-111	-182.55
		(42.6)		(182)	(155)	(232.34)
Earned Income	19-22	9582		7.5	14	8.6628
		(43)		(169)	(133)	(195.20)
Earned Income	23-27	21920		-280	-217	-369.77
		(62.7)		(292)	(244)	(385.62)
Earned Income	28-32	33100		-1.76	-376	-2.6259
		(129)		(675)	(569)	(1.0e+0)
In Labor Force	19-27	0.7763		0.0048	0.0048	0.0061
		(0.0009)		(0.0048)	(0.0037)	(0.0061)
In Labor Force	19-22	0.7238		0.0070	0.0041	0.0081
		(0.0014)		(0.0086)	(0.0064)	(0.0099)
In Labor Force	23-27	0.8138		0.0028	0.0051	0.0037
		(0.0010)		(0.0051)	(0.0039)	(0.0067)
In Labor Force	28-32	0.8234		-0.0032	-0.0010	-0.0047
		(0.0014)		(0.0081)	(0.0068)	(0.0120)

Notes: *Continued from page above.* Table records estimated discontinuity in adult outcomes for an adult being born before the New Year by age group for the full sample. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on adult outcomes. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.7: Baseline Results for Regression Discontinuity Estimates of Treatment Effects for Young Adults (Continued)

Outcome	Age Range	Control Mean Mean	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
SNAP	19-27	0.1528		0.0015	0.0004	0.0018
		(0.0007)		(0.0050)	(0.0040)	(0.0064)
SNAP	19-22	0.1480		-0.0021	-0.0020	-0.0024
		(0.0011)		(0.0091)	(0.0072)	(0.0105)
SNAP	23-27	0.1561		0.0041	0.0022	0.0054
		(0.0010)		(0.0043)	(0.0035)	(0.0057)
SNAP	28-32	0.1566		-0.0036	-0.0026	-0.0053
		(0.0013)		(0.0069)	(0.0055)	(0.0103)

Notes: *Continued from page above.* Table records estimated discontinuity in adult outcomes for an adult being born before the New Year by age group for the full sample. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on adult outcomes. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.8: Regression Discontinuity Estimate of Treatment Effects on Composite Outcomes for Young Adults by Race and Age

Outcome	Age Range	Race	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Composite Measure	19-27	White	-0.0658	-0.0121	-0.0059	-0.0145
			(0.0582)	(0.0334)	(0.0269)	(0.0404)
		Black	0.0775	0.1240*	0.0990*	0.1925*
			(0.1208)	(0.0744)	(0.0550)	(0.1155)
		Difference	0.1433	0.1361*	0.1049**	0.2071*
Composite Measure	19-22	White	0.0382	0.0206	0.0090	0.0228
			(0.0742)	(0.0440)	(0.0375)	(0.0487)
		Black	0.1101	0.1340	0.1100**	0.1794
			(0.1489)	(0.1086)	(0.0660)	(0.1454)
		Difference	0.0719	0.1134*	0.1010**	0.1566

Notes: Table records estimated discontinuity in composite measure of economic self-sufficiency for an adult being born before the New Year by age group among White and Black adults. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on the self-sufficiency measure. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.8: Regression Discontinuity Estimate of Treatment Effects on Composite Outcomes for Young Adults by Race and Age (Continued)

Outcome	Age Range	Race	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Composite Measure	23-27	White	-0.1434*	-0.0364	-0.0175	-0.0454
			(0.0788)	(0.0432)	(0.0349)	(0.0540)
		Black	-0.0148	0.1124	0.0971	0.1840
			(0.2210)	(0.1162)	(0.0949)	(0.1903)
		Difference	0.1286	0.1488	0.1146	0.2295
Composite Measure	28-32	White	-0.0257	-0.0142	-0.0351	-0.0199
			(0.1019)	(0.0567)	(0.0458)	(0.0795)
		Black	-0.1497	0.0333	0.0672	0.0664
			(0.2525)	(0.1361)	(0.1095)	(0.2714)
		Difference	-0.1240	0.0475	0.1023	0.0864

Notes: *Continued from page above.* Table records estimated discontinuity in composite measure of economic self-sufficiency for an adult being born before the New Year by age group among White and Black adults. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on the self-sufficiency measure. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

Table 1.9: Regression Discontinuity Estimate of Treatment Effects on Composite Outcomes for Young Adults by Average County Mothers' Education Attainment and Age

Outcome	Age Range	Average County Educ. Attainment of Mothers	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Composite Measure	19-27	Above Lowest Quartile	-0.0772	-0.0150	-0.0029	-0.0163
			(0.0589)	(0.0287)	(0.0241)	(0.0312)
		Below Lowest Quartile	0.0496	0.0692	0.0555	0.0987
			(0.0888)	(0.0518)	(0.0375)	(0.0739)
		Difference	0.1269	0.0842	0.0584	0.1151
Composite Measure	19-22	Above Lowest Quartile	0.0146	0.0190	0.0243	0.0192
			(0.0773)	(0.0485)	(0.0401)	(0.0492)
		Below Lowest Quartile	0.1527	0.0497	0.0049	0.0632
			(0.1206)	(0.0667)	(0.0541)	(0.0849)
		Difference	0.1381	0.0308	-0.0194	0.0440

Notes: Table records estimated discontinuity in composite measure of economic self-sufficiency for an adult being born before the New Year by age group among adults born in counties where average mothers' education is below the lowest quartile and above the lowest quartile. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on the self-sufficiency measure. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text. *Table continued in page below.*

Table 1.9: Regression Discontinuity Estimate of Treatment Effects on Composite Outcomes for Young Adults by Average County Mothers' Education Attainment and Age (Continued)

Outcome	Age Range	Average County Educ. Attainment of Mothers	Reduced Form RD Treatment Effect Estimates by Bandwidth			IV Treatment Effect of $1,000 in Infancy
			1.5 month bandwidth	2 month bandwidth	2.5 month bandwidth	2 month bandwidth
Composite Measure	23-27	Above Lowest Quartile	-0.1490**	-0.0415	-0.0237	-0.0464
			(0.0732)	(0.0321)	(0.0277)	(0.0359)
		Below Lowest Quartile	-0.0173	0.0838	0.0912	0.1240
			(0.1144)	(0.0746)	(0.0590)	(0.1103)
		Difference	0.1317	0.1253	0.1149*	0.1705
Composite Measure	28-32	Above Lowest Quartile	-0.0303	0.0102	-0.0080	0.0127
			(0.0860)	(0.0558)	(0.0467)	(0.0695)
		Below Lowest Quartile	-0.2692*	-0.1171	-0.0754	-0.1956
			(0.1490)	(0.0873)	(0.0730)	(0.1459)
		Difference	-0.2389	-0.1273	-0.0674	-0.2084

Notes: *Continued from page above.* Table records estimated discontinuity in composite measure of economic self-sufficiency for an adult being born before the New Year by age group among adults born in counties where average mothers' education is below the lowest quartile and above the lowest quartile. Table also records an instrumental variables estimate of the effect of a $1,000 increase in family income in infancy on the self-sufficiency measure. Results estimated using adults in the 2001-2016 ACS. Regressions include fixed effects by day of week, and state of birth. Standard errors calculated with 2,000 bootstrap replications. Estimation strategy described in text.

CHAPTER II

How Well Do Automated Linking Methods Perform? Lessons from US Historical Data

New large-scale linked data are revolutionizing empirical social science.[1] Record linkage is increasingly popular as a tool to create or enhance data for observational studies, randomized control trials, and lab and field experiments. Examples abound across subfields in economics, including health economics and medicine, industrial organization, development economics, criminal justice, political economy, macroeconomics, and economic history. In addition, current U.S. data infrastructure projects are linking national surveys, administrative data, and research samples to recently digitized historical records, such as the full-count 1880 (Ruggles et al. 2015, Ruggles 2006) and 1940 U.S. Censuses (the first Census to ask about education and wage income).[2] These newly available ”big data” have the potential to break new ground

[1] This chapter was written with my coauthors Martha Bailey, Catherine Massey and Morgan Henderson and published in the *Journal of Economic Literature*. Appendicies referenced in this chapter have not been included in this dissertation for concision, and are available online at https://assets.aeaweb.org/asset-server/files/13555.pdf.

[2] Many on-going initiatives link the 1940 U.S. Census to other datasets. The Census Bureau plans to link the 1940 Census to current administrative and Census data (Census Longitudinal Infrastructure Project, CLIP) and the Minnesota Population Center plans to link it to other historical censuses. The Panel Survey of Income Dynamics and the Health and Retirement Survey are linking their respondents to the 1940 Census. The Longitudinal, Intergenerational Family Electronic Micro-Database Project (LIFE-M) is linking vital records to the 1940 Census (Bailey 2018). Supplementing these public infrastructure projects, entrepreneurial researchers have also combined large datasets. See, for example, Abramitzky, Boustan, and Eriksson (2012a), Abramitzky, Boustan, and Eriksson (2013), Abramitzky, Boustan, and Eriksson (2014), Boustan, Kahn, and Rhode (2012), Mill (2013),

on old questions and open entirely novel areas of inquiry.

Machine-linking methods are critical to these projects, especially those linking U.S. Censuses. But outside of protected data enclaves, little is known about how machine algorithms influence data quality and inference, due both to false matches (Type I errors) and missed matches (Type II errors).[3] This gap in knowledge reflects the lack of ”ground truth” data. Although some diagnostic exercises are suggestive, they typically rely on selected samples (genealogy), non-U.S. samples (Goeken et al. 2017, Christen and Goiser 2007, Eriksson 2016), or rich administrative data unavailable to most researchers (Scheuren and Winkler 1993, Winkler 2006, Massey 2017, Abowd 2017). Uncertainty about the quality of machine-linked data limits their value to social science and also the development of methods to improve them.

This paper reviews the literature in historical record linkage in the U.S. and evaluates the effects of different linking algorithms on data quality. Unlike contemporary data, historical data are public and contain identifiable information, allowing us to be fully transparent about our samples and assumptions in assessing algorithm performance. Our samples include the Longitudinal Intergenerational Family Electronic Micro-database's (LIFE-M) sample of birth certificates linked to the 1940 Census (Bailey 2018) as well as a sample of Union Army veterans which the Early Indicators Project linked to the 1900 Census (Costa et al. 2017).

Even well-trained human linkers and genealogists make errors, so we also build a synthetic ground truth to validate our findings. The synthetic ground truth deliberately introduces common errors in recording, transcription and digitization of historical data. Although this synthetic ground truth is an imperfect representation of the more complicated errors in original records, the dataset's construction means that there is complete certainty about the correct links. In all cases, the synthetic

Mill and Stein (2016), Hornbeck and Naidu (2014), Aizer et al. (2016), Bleakley and Ferrie (2013, 2017, 2016), Nix and Qian (2015), Collins and Wanamaker (2014, 2015, 2016), and Eli, Salisbury, and Shertzer (2018) . This paper discusses many of the linking approaches used in these papers.

[3] ”Ground truth” is defined as data obtained by direct observation of the true link.

data produce very similar findings to the hand-linked records.

The results highlight how widely-used linking algorithms affect data quality and illustrate how different assumptions impact performance. First, we find that no linking method produces samples that are consistently representative of the linkable population, and the ways in which the data are not representative differ by algorithm. Second, widely used automated-linking algorithms produce large numbers of links that well-trained humans classify as incorrect, with rates ranging from 15 to 37 percent. Similar results in synthetic ground truth suggest that links rejected in human review are likely Type I errors. Third, false links produced by different algorithms tend to be strongly associated with baseline sample characteristics, suggesting that linking algorithms could induce systematic measurement error into analyses. In addition, the systematic measurement error varies across algorithms and records, suggesting that any bias may be difficult to predict and correct.

Our analysis also investigates how algorithm assumptions impact data quality, including phonetic name cleaning, linking more common names, and using methods to resolve ties. We find that common uses of spelling standardization in deterministic algorithms tend to increase both Type I errors, from 16 to 60 percent, as well as Type II errors. Linking more common names dramatically increases Type I errors although Type II errors fall. Lastly, including records with exact ties on name, age, and birth place (often used in conjunction with simple probability weights) increases error rates by an additional 55 to 79 percent.

After characterizing the theoretical implications of linking errors using a within-between decomposition framework, we link the same fathers and sons to the 1940 Census using different algorithms and examine the resulting estimates of intergenerational mobility. We find that some linking algorithms attenuate intergenerational income elasticity estimates by up to 20 percent. Frequently used variations in assumptions, such as including more common names and phonetic name cleaning, result in attenuation of more than 30 percent. Eliminating false matches, however, renders intergenerational income elasticities from different algorithms statistically indistinguishable. In our case study, false links appear to have a larger impact on inferences than sample composition - a finding that cautions against recent efforts to increase match rates at the expense of precision. We conclude with easy-to-implement recommendations for improving machine linking and inference with linked samples. In particular, we recommend reweighting to address sample non-representativeness and using multiple linking algorithms and supervised learning methods (with training data) to identify and reduce false links, break ties for multiple matches, and better train machine algorithms.

2.1 The Evolution of U.S. Historical Record Linkage

Record linkage has been a mainstay of social science for over 80 years. The earliest methods used painstaking manual searches to link hand-written manuscripts, and recent developments in digitized records, computational speed, and probabilistic linking techniques have expanded the possibilities for automated, or machine-based, record linkage. We briefly summarize this early literature, focusing on the components of this history that laid the groundwork for current practice.[4]

One feature important to historical and modern linking is blocking. Blocking refers to the partition of a dataset into "blocks" (or clusters of records) using a record attribute (Michelson 2006). This technique limits the number of potential matches according to the blocking attribute, thereby improving computational efficiency while (ideally) maintaining accuracy. For instance, blocking on place of birth and sex means that a linking algorithm looking for Franklin Jones born in Kentucky would only search within the set of candidate matches of men born in that state.

[4] See Ruggles, Fitch, and Roberts (2018) for a more detailed history of the findings in this early literature.

Historical linkage has always used blocking techniques to increase the feasibility of manually linking samples across manuscripts. The earliest blocking methods involved identifying a group of individuals in a particular location (e.g., township, county, or state) in one census and manually searching for the same people within the same region (the block) in the subsequent census (Malin 1935, Curti 1959, Bogue 1963, Thernstrom 1964, Guest 1987). While making manual searching feasible, this blocking strategy missed those who relocated or changed names between census years. The resulting linked samples omitted the geographically mobile population and were, therefore, unrepresentative (Ruggles 2006).

The creation of digitized state population indexes facilitated refinements in blocking.[5] In one such approach that improved on previous methods, Steckel (1988) drew a random sample of households with children at least 10 years old in a historical census. He then searched for the same household in the previous census using the birth state of the child to narrow the search. This technique was able to locate individuals who moved between the census years, but it restricted the sample of linked households to those with children surviving to age ten.

Advances in computing allowed improvements in automated matching, effectively replacing time-intensive human search with computer queries. Leveraging newly created national population indexes and Public Use Microdata Samples (PUMS), automated matching began incorporating more data elements in the linking process. An early example of this strategy was Atack, Bateman, and Gregson (1992)'s probabilistic matching software, called "PC Matchmaker." PC Matchmaker transformed names using phonetic codes and allowed for user-specified blocking and weighting schemes. Atack (2004) used this software to create a linked sample between the agricultural and population censuses between 1850 and 1880. Ferrie (1996)'s approach, which we describe in more detail in the following sections, aimed to create large, representative

[5] A state "index" is a list of individuals living in a state at a point in time.

linked U.S. Census samples and has since been embraced by the literature, forming the basis of prominent methods in use today. Before summarizing more modern methods, we present an example linking problem to illustrate common challenges in historical linking.

2.2 Current Approaches to Linking Historical Data

Matching records across sources requires choosing linking variables, also called ”features” in the computer science and statistics literatures. Modern administrative records typically have multiple, high-quality features (e.g., full legal name, Social Security Number, exact date of birth, address of residence). Outside of restricted administrative enclaves, data typically contain a limited set of noisy linking features. Historical data have the advantage of containing identifiable information, allowing transparent study of how limited data and data errors affect the quality of linked samples. Like many modern linking problems, historical data have limited information that is often measured with error.

As an example, consider the challenge of linking birth certificates to the 1940 U.S. Census. Researchers typically use ”time-invariant” features to do linking in order to minimize concerns about selection bias and non-representativeness in linked samples (Ruggles, 2006). For U.S. Census linking, these variables typically include first name, last name, age, birth state, race, and sex.[6] In practice, names may vary over time, either because Census enumerators misspelled names, the individual reported incorrectly, or the individual changed names (perhaps using a middle name or nickname in place of the given name). Goeken et al. (2017) document that in two

[6] Matching in historical settings in other countries often makes greater use of characteristics not available in U.S. data. Modalsli (2017) notes that in Norway before 1910 there is less first name variation and more flexible surname traditions than in the U.S. In addition, Norwegian censuses use 500 birthplaces (municipalities) for a population of under 2 million, whereas the U.S. Censuses identify birthplaces as 48 states and foreign countries for a much larger population (132 million residents in the 1940).

enumerations of St. Louis in the 1880 Census, nearly 46 percent of first names are not exact matches. Similarly, the Early Indicators project notes that 11.5 percent of individuals in the Oldest Old sample have a shorter first name in pension records than in the original Civil War enlistment records (Costa et al. 2017).

Similar problems arise in reported age and birth state. The recording of age in the Census tends to reflect age heaping, the common practice of rounding ages to the nearest multiple of five (A'Hearn, Baten, and Crayen 2009, Hacker 2013). In addition, birthplaces are often inaccurately recorded. Goeken et al. (2017) show that 8 percent of reported birthplaces do not match across the two enumerations of St. Louis. In addition, rates of disagreement for mothers' and fathers' birthplaces for the same individuals average 19 and 18 percent, respectively.

The digitization of hand-written manuscripts compounds errors in recording. Our comparison of two independently digitized versions of the 1940 Census by Ancestry.com and FamilySearch.org shows that 25 percent of records have different transcriptions of last name due to digitization alone.

These data quality issues are well known, and linking algorithms account for them by allowing for variation in age and name spellings. To deal with differences in age, researchers typically search over a range of ages. Researchers account for orthographic differences by using metrics such as Jaro-Winkler or Levenshtein to quantify the dissimilarity of two name strings. In some cases, researchers use phonetic string cleaning algorithms to help account for spelling differences, name Anglicization, and transcription errors. Soundex, for example, was developed in the early 20th century to help create Census links, simplifying names into phonetic codes to facilitate record searches. For example, Soundex assigns the same code (S530) to similar sounding names like "Smith," "Smyth" and "Smythe." Another cleaning algorithm, NYSIIS, the New York State Identification and Intelligence System, was developed as an improvement to Soundex in 1970. NYSIIS transforms the same root name to a common string, making names like "Wilhem" and "William" into "WALAN." While phonetically cleaned strings allow researchers to identify more candidate links, matching on them deterministically treats distinct names as the same. One implementations of NYSIIS, for instance, categorizes John and James as perfect matches (Ruggles, Fitch, and Roberts 2018).

Figure 2.1 illustrates how limited information and measurement error create challenges for matching records. The linking problem is depicted as two-dimensional scatter plot after blocking on birth state and sex, as is common in the literature. The x-axis captures the similarity between the name on record to be linked and the names of candidate links in the 1940 Census using the Jaro-Winkler similarity score, which will equal 1 if the names are identical and is less than 1 otherwise.[7] The y-axis captures the difference in the age implied by the birth certificate (which contains exact date of birth) and the reported age in the 1940 Census. A perfect match in ages occurs when the age difference is zero.

In this two-dimensional space, candidate links fall into one of four categories:

(M1) A perfect (1,0), unique match in terms of name and age similarity (Figure 2.1A).

(M2) A single, similar match that is slightly different in terms of age, name, or both (Figure 2.1B).

(M3) Many perfect (1,0) matches, leading to problems with multiple matches (Figure 2.1C).

(M4) Multiple similar matches that are slightly different in terms of age, name, or both (Figure 2.1D).

As we discuss, historical linking algorithms generally treat M1 cases as matches. However, methods differ in their treatment of candidates in the M2, M3, and M4

[7] Jaro-Winkler similarity score adapts the Jaro (1989) string score, the minimum number of single-character transpositions required to change one string into another, to up-weight differences that occur at the beginning of the string. See Winkler (2006) for an overview.

categories. To account for differences in age as in M2, researchers typically search within a band of ±3 or ±5 years. Prominent approaches to dealing with ties in categories M3 or M4 include random selection among equally likely (tied) candidates (Nix and Qian 2015), equal probability weighting of tied candidates (Bleakley and Ferrie 2016), or the use of a weighted combination of linking features to classify true matches (Feigenbaum 2016, Abramitzky, Mill, and Perez 2018). The next sections describe how commonly used linking algorithms work and ultimately classify records in cases such as M2, M3, and M4.

2.2.1 Ferrie (1996)

Ferrie's (1996) path-breaking approach links men in the 1850 U.S. Census to men who were 10 years and older in the 1860 U.S. Census. Ferrie (1996) begins by selecting a sample of uncommon names from the 1850 Census.[8] To correct for minor orthographic differences (category M2 above), his algorithm transforms last names using NYSIIS codes and also truncates the untransformed first name at the fourth letter. The algorithm then links his sample to the 1860 Census and eliminates candidate links that were not born in the same state and not living with the same family. The algorithm keeps all candidate links within a ±5 year difference in age (or "age band") and, if more than two links remain, chooses the link with the smallest age difference. At the end of this process, the algorithm drops cases where two individuals from 1850 link to the same observation in 1860.[9] This process produces a linked sample of 4,938 men - 9 percent of the male population in 1850, and 19 percent of the population of men with uncommon names. Ferrie has used different approaches in more recent work, including smaller age ranges, different ways of parameterizing name dissimilarities like SPEDIS, or altered restrictions on common names. More

[8] Ferrie (1996) searched for 25,586 men in the 1860 Census whose surname and first name appeared ten or fewer times in 1850.

[9] Ferrie (1996) does not specify a process for multiple match disambiguation; in his linking from 1850-1860, there were no ties after minimizing the difference in age.

than 20 years later, Ferrie's approach has become the foundation for much of the historical linking literature.

Two features of this algorithm are especially worth noting. First, the decision to make links among observations with uncommon names reduces both the computational burden and the number of candidate matches of the M3 variety. Consequently, this method never attempts to link common names like "John Smith." Second, the decision to use NYSIIS and truncate first name reduces problems associated with minor orthographic differences, but it may also increase ties of the M3 variety and, therefore, the number of records the algorithm will not link. The independent effects of both of these choices are considered in our subsequent analysis.

2.2.2 Abramitzky, Boustan, and Eriksson (2012 and 2014)

Abramitzky, Boustan, and Eriksson's (2012, 2014) "Iterative Method" scale up Ferrie (1996) to use the full-count census. This procedure relaxes Ferrie's (1996) uncommon name restriction to the extent that the combination of name and age provide distinctive information. Summarized in a detailed web appendix, Abramitzky, Boustan, and Eriksson (2012b) select a sample of boys ages 3 to 15 with unique name-age combinations in the 1865 Norwegian Census, standardize first and last names using NYSIIS codes, and look for exact, unique matches in U.S. and Norwegian Censuses. For the observations in the 1865 Census without an exact, unique link (M1), the algorithm then searches for a name match within a ±1 age band and, if there is no match in this band, the algorithm searches within a ±2 age band. The algorithm does not link a record if more than one candidate match exists within an age band. The algorithm ultimately links a sample of 2,613 migrants and 17,833 non-migrants from a primary sample of 71,644 individuals for a match rate of 29 percent. Abramitzky, Boustan, and Eriksson (2014) use the same procedure to link men ages 18 to 35 with unique age-name combinations from the 1900 U.S. Census to the 1910 and 1920 Censuses, producing a sample of 20,225 immigrant and 1,650 native-born men for match rates of 12 percent and 16 percent. The authors provide the most recent version of their code for our analysis, which has been used for record linkage in a number of high profile papers.[10] A variation on this approach is also reported in Abramitzky, Boustan, and Eriksson (2014)'s appendix as a robustness check. Similar to Ferrie's (1996) uncommon name restriction, this robustness check requires that names be unique within a five-year age band (a ±2 year difference).

Two differences to Ferrie (1996) are worth noting. First, Abramitzky, Boustan and Erickson (2012, 2014) link more common names, while Ferrie's (1996) algorithm does not. Second, Abramitzky, Boustan and Erickson (2012, 2014) use a narrower age band than Ferrie (1996).

2.2.3 Feigenbaum (2016) and IPUMS (2015)

A common feature of Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014) is that they search among identical, phonetically cleaned names for a match that uniquely has the minimum age difference. This restriction reduces computational

[10] Many other papers have used variations on Ferrie (1996) and Abramitzky et al. (2014). These variations are similar in that they require matches to match completely on a cleaned name variable. For example, Abramitzky et al. (2013) use the Abramitzky et al. (2014) algorithm to match men aged 3 to 15 in the 1865 Norwegian Census to the 1880 U.S. and Norwegian Censuses, and match 26 percent of records unique by name and birth year from the 1865 Census. Boustan et al. (2012) link the IPUMS sample of the 1920 U.S. Census to the 1930 U.S. Census using a link uniqueness age band that functions similar to an uncommon names restriction, and match 24 percent of men unique by name, age and birthplace in their 1920 U.S. sample. This same dataset was used in Hornbeck and Naidu (2014). Collins and Wanamaker (2015) use a variation on the Ferrie (1996) method with an alternate name uniqueness requirement to match southern men younger than 40 in the 1910 Census to the 1930 Census. They match 24 percent of their records from the 1910 Census. Other papers use name similarity measures. Eli, Salisbury, and Shertzer (2018) match Civil War recruitment records from Kentucky to U.S. Censuses before and after the Civil War using a variation of Ferrie (1996) without an uncommon names restriction, and impose an additional restriction on Jaro-Winkler string dissimilarity after generating candidate matches with NYSIIS. They match 30 percent of selected records of recruits from the 1860 U.S. Census to the 1880 U.S. Census. Aizer et al. (2016) match state mother's pension records to other data sources, including the Social Security Death Master File, using a variation on Ferrie (1996) without an uncommon names restriction and Soundex, but allow some additional matches to differ in exact name but have low SPEDIS and Levenstein dissimilarity measures. Aizer et al. (2016) match 48 percent of their records to the Social Security Death Master File, but this high match rate reflects the fact that they can use exact date of birth to match their observations.

burden, but comes at the cost of excluding very similar (though not exact) names with exact or very close age matches. New methods in probabilistic linking relax these assumptions and allow machine models to weight different kinds of disagreements in names and ages.[11] The key insight is that the best link may not exactly match on name (or phonetically cleaned name) or age as in Figure 2.1B and Figure 2.1D but it may dominate other candidates when simultaneously considering both age and name differences. One class of machine-learning algorithms are known as supervised learning methods and use ”training data” to classify matches. Training data may be a subset of data coded by humans (sometimes genealogists and sometimes by others) or result from the observation of true links (called ground truth). If the training data are ground truth and the model is well specified, the computer will learn how to classify links to approximate this truth. However, if the training data are of limited quality, the computer algorithm will replicate these incorrect decisions. Another potential limitation is that if the training data have little in common with the records to be linked, the supervised-learning algorithm will have unpredictable performance. Consequently, the advantage and disadvantage of supervised-learning algorithms is that they depend heavily on the quality of the training data and its similarity with the data to be linked.

The Minnesota Population Center (MPC) uses a supervised learning method to create the Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS), a set of links between the 1850 to 1930 one-percent samples and the 1880 Census (Ruggles et al. 2015). Using clerically reviewed data, MPC trains a Support Vector Machine (SVM) using features of matches that, after specifying a few tuning parameter choices, classifies links as true or false (Goeken et al. 2011). Illustrating the conservative nature of this approach, MPC produced final match rates of 12 percent for native-born whites, 3 percent for foreign-born whites, and 6

[11] See Mullainathan and Spiess (2017) for a useful primer on machine learning for economists.

percent for African Americans for the 1870-1880 links.[12] Unfortunately, this model is proprietary and we cannot use it in our analysis.

In a similar spirit, Feigenbaum (2016) uses a supervised-learning technique to link the 1915 Iowa state Census to the 1940 U.S. Census. After creating training data by hand, he estimates a probit model to quantify the joint importance of different record feature in determining a link, including name Jaro-Winkler scores, differences in age, indicators for Soundex matches of first or last name, indicators for matches in letters or names, and indicators for matching truncated first or last names. He then tunes his model so that a link is only chosen if its probability of being a match is sufficiently high and sufficiently greater than the second-best candidate's match probability (if a second-best candidate exists). These cutoffs are derived from training data to assess the Type I and Type II errors of different choices. Feigenbaum (2016) achieves a match rate of 57 percent.[13]

2.2.4 Abramitzky, Mill, and Perez (2018)

An alternative to supervised learning is unsupervised learning, an approach which evaluates the quality of different links without training data. These algorithms depend on using observed patterns in the data to classify the data by quality of potential link. Similar to other deterministic methods like Ferrie (1996) and Abramitzky, Boustan,

[12] Researchers have used the IPUMS-LRS for a variety of research questions, including the economic effects of racial fluidity (Saperstein and Gullickson 2013), long-term differences in black and white women's labor-force participation (Boustan and Collins 2014), and intergenerational coresidency (Ruggles 2011).

[13] We focus on Feigenbaum (2016) in this analysis because it was developed for U.S. data, and is transparent and easy to replicate. Many other researchers have also incorporated probabilistic and machine learning. Mill (2013) and Mill and Stein (2016) use an expectation maximization method. Similar to the IPUMS-LRS, Wisselgren et al. (2014) use a support vector machine and link the 1890 Swedish Census to the 1900 Swedish Census in a few select parishes for match rates ranging from 25 to 72 percent. Antonie et al. (2014) link across historical Canadian census data and achieve linkage rates from 17.5 percent (Quebec) to 25.5 percent (New Brunswick). Other work uses Ancestry.com's search algorithm to link records. Bailey et al. (2011) link records of lynching to the 1900 to 1930 U.S. Censuses using the Ancestry.com's algorithm, linking 45 percent of their lynching records. Collins and Wanamaker (2014) and (2015) search Ancestry using Soundex for names as well as age and place of birth for white and black men ages 0 to 40 resident in Southern states in the 1910 Census. They match 19 and 24 percent of men, respectively, to a unique person in 1930.

and Eriksson (2014), an advantage of unsupervised learning is that it creates links without training data. Moreover, if training data are error ridden or too different from the dataset of interest, the lack of reliance on them is a feature. Not relying on human decisions may also be a limitation, because the algorithm's performance depends on difficult-to-validate modelling assumptions. Abramitzky, Mill, and Perez (2018) use unsupervised learning in the form of the expectation-maximization algorithm (Fellegi and Sunter 1969, Winkler 2006, Dempster, Laird, and Rubin 1977) to link Censuses in the U.S. and Norway. Building on Mill (2013), they fit a mixture model that allows for conditionally independent multinomial probabilities of specific age distances and discretized Jaro-Winkler scores for two distributions. They then fit this model with observed data using the expectation-maximization algorithm. Then, using their results, they calculate the estimated probability that a given potential match is a correct match conditional on the Jaro-Winkler score and age distance of the match. Like Feigenbaum (2016), they create a final set of matches by applying cut-offs to these estimated probabilities so that links are chosen that reach a sufficiently high estimated probability that is sufficiently greater than the second-best candidate match. However, unlike Feigenbaum (2016), these cutoffs are not guided by training data. The approach is not completely automated, because it requires the user to define tuning parameters, including discretization thresholds of Jaro-Winkler scores and probability cut-offs for classifying a link.[14] With regards to the latter, using lower cut-offs will create more matches but potentially include more marginal matches less likely to be correct. Conversely, higher cut-offs will create fewer matches but create matches that have a higher estimated probability of being correct. To address this trade-off, Abramitzky, Mill, and Perez (2018) use two cut-offs, a more conservative and less conservative choice. Using these two cut-offs for their algorithm, they achieve

[14] These choices may be consequential. For example, setting a high Jaro-Winkler similarity threshold corresponding to (0.92,1] assigns the same estimated match probability to a pair with first names, Katherine/Catherine, as to a pair with an exact match on first name, all else equal.

match rates of 5 percent and 15 percent in their Census data. Our analysis implements these cutoffs and considers the effects of alternate cut-offs in an online appendix.

In summary, existing linking methods involve a variety of modelling choices with unknown effects on data quality. Which set of assumptions should researchers use in different contexts? What are the implications of different assumptions for error rates? The next sections answers these questions by presenting a systematic comparison of methods in different records.

2.3 Data and Metrics of Automated Method Performance

Our analysis considers four different linking algorithms: Ferrie (1996); Abramitzky (Abramitzky, Boustan, and Eriksson 2014); regression-based, supervised learning (Feigenbaum 2016); and unsupervised machine learning (Abramitzky, Mill, and Perez 2018). Detailed web appendices, published articles, and posted code make replicating these methods straightforward. Ferrie (1996) and Feigenbaum (2016) describe their methods step by step, which we implement exactly.[15] We present Feigenbaum (2016) using both his regression coefficients for the Iowa Census-1940 training data (labeled "Iowa coef.") as well as coefficients estimated using hand-linked samples (called "Estimated coef."; see Online Appendix A for details and coefficient estimates). To implement Abramitzky, Boustan, and Eriksson (2014) and Abramitzky, Mill, and Perez (2018), we use the code provided by the authors (see Online Appendix A) and report two cutoff implementations per the latter's recommendation, "less conservative" and "more conservative." For interested readers, we created a public Stata ado-file that implements these methods and the variations we consider in this paper (Bailey and Cole 2019).

[15] Unlike Ferrie (1996), we do not limit links based on family continuity. In addition, we treat records with multiple matches after the last step as having no link, although Ferrie reports having none of these instances and, therefore, does not indicate how he would have dealt with them.

2.3.1 Hand-linked and Synthetic Data

We examine the performance of each algorithm in two high-quality, hand-linked historical samples: the LIFE-M sample of birth certificates linked to the 1940 Census (Bailey 2018) and the Early Indicators Project's genealogically linked sample of Union Army veterans (Costa et al. 2017).

The LIFE-M sample is based on a random draw from birth certificates from Ohio and North Carolina. These birth certificates are then linked to siblings' birth certificates using parents' names. We exclude girls because they typically changed their name at marriage in this era, making them hard to find as adults in the Census (see Online Appendix B). The LIFE-M sample consists of 42,869 boys born from 1881 to 1940, 24,408 of whom were born in North Carolina and 18,461 born in Ohio.

The LIFE-M sample of boys is then linked to the 1940 full-count U.S. Census using a semi-automated process, making use of both computer programming and human input. Our linking variables include first, middle (when available), and last name, birth state, and age. We do not use race, because it is not available on all birth certificates (see section 2.5.2 for an analysis of this limitation).

After cleaning and standardizing the data, we use bi-gram matching on name and age similarity within a birth state to generate candidate links (Wasi 2014).[16] Each candidate is independently reviewed by two "data trainers" who choose a correct link (or no link) from the set of candidates. If the two trainers agree, we treat their choice (link or no link) as the truth. In cases where the two trainers disagree, the records are independently re-reviewed by three new trainers to resolve these discrepancies (see section 2.3.2).[17] LIFE-M data trainers are instructed to reject links if they are not

[16] We generate a set of candidate links using "reclink.ado," an algorithm that uses bi-gram comparisons of name strings. We also block on the first letter of last names to reduce computation time.

[17] Data trainers participate in a rigorous orientation process where they receive detailed feedback on their accuracy relative to an answer key. They continue this process for 10 to 20 hours per week until their matches agree with the truth dataset 95 percent of the time. After completing this orientation, trainers become part of the larger team that conducts independent clerical review.

completely certain the links are correct.

LIFE-M trainers also work under a senior data trainer and receive multiple rounds of feedback across approximately 30 hours of work. Before trainers are allowed to work on the LIFE-M team, their decisions must reach a 0.95 correlation with a truth dataset.[18] After trainers achieve this threshold, they begin receiving training batches from an automated distribution system, which guarantees that links are reviewed initially by two different trainers and that discrepancies are reviewed by three additional trainers. This automated system also distributes audit batches at least once per week to provide weekly feedback to trainers about their accuracy. Trainers meet weekly to discuss their mistakes, difficult cases, and learn about historical-contextual factors affecting the quality of the data.

The Family History and Technology Lab at Brigham Young University (BYU) performed two independent quality checks of the LIFE-M links. First, BYU research assistants used genealogical methods and multiple data sources to hand link a random sample of 543 of the 18,461 Ohio boys, 241 of which had been linked by LIFE-M. The BYU team had no knowledge of LIFE-M's links. Among links made by both LIFE-M and BYU, BYU agreed with LIFE-M matches 93.4 percent of the time (16/241 matches were discordant). Second, BYU compared 1,043 LIFE-M links to those already on the FamilySearch.org "Tree." (FamilySearch.org tree links are created by genealogists and users of FamilySearch.org, who are independent of the LIFE-M process.) For 1,043 birth certificates linked to the 1940 Census by LIFE-M and FamilySearch.org users, the LIFE-M links agreed with FamilySearch.org users 96.7 percent of the time. A link-weighted average of the two exercises implies that LIFE-M's false link rate is around 3.9 percent. To account for potential errors in the LIFE-M data, we additionally require all links that differ from the LIFE-M sample

[18] This truth dataset has been vetted by multiple individuals for accuracy. The cases for this truth dataset are selected to test the trainers' knowledge and decision-making for a variety of linking problems.

to be re-reviewed using the "police line-up" process described in section 2.3.2

Our second sample is the Oldest Old sample of Union Army veterans from the Early Indicators Project. Costa et al. (2017) created this sample of 2,076 individuals at least 95 years old linked to the 1900 complete-count U.S. Census using genealogical methods and a rich set of supplementary information. These veterans tended to report complete and accurate information to ensure they would receive their army pensions and benefits. Moreover, sources such as gravestone databases, obituaries, newspaper accounts, veterans associations and pension files allow multiple cross-validation exercises, ultimately resulting in a high match rate of 90 percent among men confirmed to live beyond the 1900 Census. The Early Indicators Project scores matches on a scale of 1 to 4 to indicate their confidence in a match. We use 1,887 matches coded as the highest quality (1 and 2) as the hand-linked sample. Importantly, we do not use all possible records for which matches were attempted.

Because these hand-linked data may contain errors, we validate our conclusions by building a third sample: a synthetic ground truth. This synthetic ground truth adds noise to true links to mimic common errors in historical data while ensuring complete certainty about correct and incorrect links. That is, this synthetic dataset characterizes the performance of each matching algorithm relative to an objective truth, which shares important commonalities with the LIFE-M sample.

We construct the synthetic ground truth in two steps. First, we take all of our Ohio and North Carolina born boys' birth certificates, randomly drop 10 percent to reflect mortality and emigration and drop another 5 percent to reflect under-enumeration.[19] Using the LIFE-M records as a basis allows us to retain sample name characteristics

[19] Based on life tables from 1939 to 1941, we calculate that 8 percent of our sample should be un-linkable due to death prior to 1940 (National Office of Vital Statistics 1948). Moreover, Census analyses estimate that around 5.4 percent of individuals were missed in 1940 (West and Robinson 1999). This calculation leaves some scope (about 1.5 percentage points) for emigration, which reflects the fact that we think emigration for native-born boys would have been much lower than for those born abroad. To the extent that our approximation of emigration is too low, the actual Type II errors should be adjusted accordingly.

(e.g., ethnic origin and other conventions and name commonness). To account for orthographic differences in enumeration or transcription errors, we add noise to names and ages to reflect age heaping and transcription or digitization errors (Goeken et al. 2017, Hacker 2013, 2010).[20] One limitation of this approach is that the true error structure in names and ages is unknown, so our decisions about how to simulate error may be simplistic and incomplete.

The resulting synthetic truth dataset is a noisy version of the truth for 85 percent of the Ohio and North Carolina boys. Then, we append to a random sample of boys from the 1940 Census who were born in Michigan, Indiana, Tennessee and South Carolina. Because these states neighbor Ohio and North Carolina, these individuals are incorrect links by construction. We chose these states because they share regional naming conventions and have similar demographic and economic characteristics. The size of our random sample of boys from neighboring states ensures that our set of candidates for each state has the same number of observations as in the LIFE-M linking exercise: 3,133,982 boys from the relevant age ranges born in Michigan and Indiana for Ohio and 1,904,592 boys born in Tennessee or South Carolina for North Carolina. When linking to this synthetic dataset, we emulate the common process of blocking on birthplace and consider only the synthetic Ohio data as candidate matches for the Ohio boys and only the synthetic North Carolina data as candidate matches for the North Carolina boys.

[20] To mimic age-heaping, 25 percent of ages are rounded to the closest multiple of 5. We introduce orthographic and transcription errors as follows. In 10 percent of cases, the first and middle names are transposed (if a middle name exists) and, in 5 percent of cases, the first and last names are transposed. In 5 percent of cases each, the first character of the first name or last name is randomly changed. In 5 percent of cases, each second character of the first name or last name is randomly changed. In 5 percent of cases, each third character of the first or last name is randomly changed. In 5 percent of cases each, we add a repeated letter o first names (e.g., ”James” to ”Jamees”) or last names. In 5 percent of cases each, a random letter is dropped or two letters are transposed in the first or last name (e.g., ”Matthew” to ”Mathew” or ”William” to ”Willaim”). In 5 percent of cases, we replace the first name with an initial. In 50 percent of cases, we drop middle names (resulting in the same share of observations having middle names as is observed in the 1940 Census).

2.3.2 Performance Criteria

We use four main criteria to measure performance. The first two are almost universally reported in papers using linked samples.

(1) Match rate: We calculate the match rate as the share of records that were linked of the sample that we attempted to link. Even if matching were perfect, this rate is expected to be less than 100 percent due to emigration and mortality. Notably, emigration and mortality are not expected to have different impacts by method, so they should not impact the relative performance of methods.

(2) Representativeness: We compare characteristics for the linked sample to the same characteristics for the unlinked sample. in a multivariate, linear probability model with Huber-White standard errors (Huber 1967, White 1980). A heteroskedasticity-robust Wald test of model significance tests the null hypothesis that the covariates are jointly related to successful linkage.[21] This straightforward, single summary metric and the regression coefficients describe the extent the extent of non-representativeness as well as the subgroups that are under-represented.

These measures alone are inadequate to assess link quality. This fact is easily illustrated in an example. Consider a matching algorithm that randomly links individuals between two datasets. This algorithm would perform very well in terms of the first two criteria, because the entire sample would be matched and identical to the baseline sample in observed characteristics (and, therefore, representative). Few researchers, however, would want to work with these data, because - with large enough datasets - the incidence of false links would approach 100 percent.

We, therefore, use two more criteria to assess link performance (Abowd and Vilhuber 2005, Kim and Chambers 2012).

[21] We implement this in Stata by multiplying the F-statistic reported in Stata following a regression with robust standard errors by the relevant degrees of freedom parameter. Note that this test could be very conservative in the sense that it would reject the null hypothesis due to one variable's significance in the regression and does not weight for the ”importance” of different covariates.

(3) False link rate (Type I error rate):[22] We compare links for each automated method to a measure of the truth. We treat the high-quality, hand-linked Early Indicators dataset as the "truth," given that genealogists have used multiple data sources to confirm each link. In the synthetic data, we know the true link, so we code differences in links between an algorithm and the synthetic data as Type I errors.

For the LIFE-M data, we subject discrepancies between the hand-links and the algorithm to an additional blind review. Similar to a "police line-up," two reviewers independently review the LIFE-M link (made by hand), the link made by the automated method, and close candidate links. Reviewers may choose to code any of these links as correct or incorrect. This process gives the links from the hand-match and the automated method an equal shot at being chosen to avoid preferential treatment. For the LIFE-M data, only links that are rejected in clerical review as part of the police line-up are treated as Type I errors. This analysis may understate the true Type I error rate if the hand-links are incorrect and agree with the automated method.

(4) False negative rate (Type II error rate):[23] This metric captures the fraction of true links that are not found, or 1 - Match Rate*(1 - Type I Error Rate). With this definition, the false negative rate can never be zero, because mortality and emigration mean that many individuals cannot be linked even with perfect data.[24]

[22] Computer scientists focus on precision, or 1-T1 error rate presented here.

[23] Computer science focuses on a similar statistic, "recall." This measure is defined as the number of true links found by the algorithm divided by number of linkable observations, or those linked by the data trainers.

[24] Note also that, if the marginal link is more likely to be incorrect, an increase in the match rate within a specific algorithm has a weakly negative effect on Type II error rates and a weakly positive effect on Type I error rates. If the marginal link is wrong, then the Type II error rate would not change but the Type I error rate would increase. However, if the marginal link is correct, the Type II error rate would fall and the Type I error rate would decrease.

2.4 The Performance of Prominent Automated Matching Methods

Because a central focus of a growing literature is linking to the newly available 1940 Census, we begin our analysis linking birth certificates to the 1940 Census. We then corroborate our findings using our synthetic ground truth and the Oldest Old sample from the Early Indicators Project.

2.4.1 Evaluating Algorithms Using the LIFE-M Data

Figure 2.2 compares the performance of selected, prominent automated linking methods to the hand-linked LIFE-M data, where each of these methods uses the same information to create links - name, age, and birth state. The length of each bar represents the match rate, computed as the share of the baseline sample of 42,869 boys who were successfully matched to the 1940 complete count Census. LIFE-M hand-review matched 45 percent of the baseline sample. Ferrie's (1996) method matched 28 percent of the baseline sample, and Abramitzky, Boustan, and Eriksson (2014) achieve a higher link rate of 42 percent. This result makes sense because Abramitzky, Boustan, and Eriksson (2014) do not impose Ferrie's (1996) uncommon name restriction. Feigenbaum's (2016) regression-based machine learning method matches 52 percent of the baseline sample both when using Iowa coefficients and when we estimate the coefficients using a random sample of the LIFE-M links. Abramitzky, Mill, and Perez (2018)'s expectation-maximization method links 46 percent of the sample when using less conservative cutoffs and 28 percent of the sample with more conservative cutoffs.

Across the board, these match rates are higher than in the original studies. For instance, the Ferrie (1996) method matches 28 percent of the LIFE-M data versus his published figure of 9 percent of all men between 1850 and 1860 Censuses. Similarly, Abramitzky, Boustan, and Eriksson (2014) link 40 percent of the LIFE-M sample, whereas the same method links only 29 percent in Abramitzky, Boustan, and Eriksson (2012b) and 16 percent of native-born men in Abramitzky, Boustan, and Eriksson (2014). These higher match rates likely reflect two factors: the LIFE-M boys are on average 24 years old in the 1940 Census, so mortality and outmigration are likely lower for them than in other studies. In addition, birth certificate data quality is higher compared to other sources. Birth certificates (1) contain a complete and correct full name, often including middle names omitted in the Census; (2) record the exact date of birth rather than age in years;[25] and (3) capture the birth state by construction (it is issued by the birth state and so should not have reporting error like the Census).

Figure 2.2 also summarizes the share of links that human reviewers rejected in a blinded review using the "police line-up" method. These rejected links are presented in two ways. First, the share of the entire sample determined to be wrong for each method is displayed in red. For less than 2 percent of original sample, trainers reversed LIFE-M decisions upon re-review in favor of the link chosen by one of the automated methods. Consistent with genealogical validation by BYU, these reversals are rare. Second, the column on the far right in Figure 2.2 presents that share of links that were rejected by human reviewers (the estimated Type I error). We compute this share by dividing the share of the total sample that is incorrect by the match rate. Because the LIFE-M match rate is 45 percent, this implies a Type I error rate of 4 percent (approximately 0.017/0.445). As shown in section 2.6, the implications of measurement error for inference is linked to the share of incorrect links, so our discussion focuses on this second metric.

Relative to clerical review, the share of false links for automated methods is higher across the board. The lowest Type I error rate occurs in the more conservative version of Abramitzky, Mill, and Perez (2018) at 15 percent. Ferrie's (1996) method

[25] Massey (2017) shows that decreasing the noise in age results in higher match rates and lower Type I error rates.

of selecting uncommon names achieves the second lowest Type I error rate at 25 percent. These error rates are consistent with Massey (2017) who uses contemporary administrative data linked by Social Security Number as the ground truth. She finds that methods similar to Ferrie (1996) are associated with 19 to 23 percent Type I error rates.. Abramitzky, Boustan, and Eriksson (2014)'s refinement of Ferrie (1996) increases match rates to 40 percent, but only half of the added links appear to be correct, and the Type I error rate increases to 32 percent. Feigenbaum's (2016) supervised, regression-based machine learning model produces a Type I error rate of 34 percent when using the Iowa coefficients, and the Type I error rate decreases to 29 percent when estimated using LIFE-M data. Finally, Abramitzky, Mill, and Perez (2018)'s less conservative cut-off results in the highest error rate at 37 percent. The difference between the conservative and less conservative versions of Abramitzky, Mill, and Perez (2018) highlights the sensitivity of performance to the tuning parameters.

In terms of missed links, Ferrie (1996) correctly linked the lowest share of the sample without error at 21 percent, and Feigenbaum (2016)'s algorithm estimated with the LIFE-M data correctly linked the largest share of the sample without error at 37 percent. It is worth noting that Feigenbaum (2016) and Abramitzky, Mill, and Perez (2018) allow for a variety of different choices of sample restrictions within their linking methods that alter the trade-off between Type I and Type II errors in their matches. We implemented versions of these methods that reflected how they were implemented in each. Our Online Appendix Figures A1 and A5-A7 show how altering these restrictions impacts results in both cases.

Table 2.2 describes the representativeness of the linked sample. Because birth certificates do not contain socio-demographic measures found in the Census (race, age, or incomes of the parents), we regress a binary dependent variable (1= linked records) on a variety of covariates from the birth certificates. These variables include the individual's exact date of birth;[26] the number of siblings in the family; the number of characters in the infants' (boys'), mothers', and fathers' names - a characteristic which is strongly positively correlated with years of schooling and income from wages in the 1940 Census; and the share of family records with a misspelled mother's or father's name, which we expect to be negatively correlated with years of schooling and income (Aizer et al. 2016).[27] Table 2.2 presents the Wald-statistic for tests of whether these covariates are jointly associated with an observation being linked (p-value beneath). If a representative set of birth certificates were linked, then these characteristics would not be jointly related to whether an observation was linked. However, Wald-statistics for the joint test of the association of these characteristics with linking show a persistent association. For all methods, including LIFE-M's clerical review, we reject representativeness at the 1-percent level.

The signs and magnitudes of the regression results provide clues about the individuals easier to link (see the full set of regression results in Online Appendix C). Many automated methods are more likely to link boys with higher incidence of misspelled father's last name, and more likely to link boys with a longer mother's name. All methods except Feigenbaum (2016) with estimated coefficients are more likely to link children with longer names. Based on the correlation of name length with wage income in the 1940 Census, this finding indicates that linked records are drawn from more affluent families. Some methods are more likely to link individuals with more siblings, while other methods are more likely to link individuals with fewer siblings. In short, even though no linking algorithm generates representative samples, different algorithms yield samples that are non-representative in different ways.

Finally, Table 2.3 tests for the systematic correlation of links rejected in hand

[26] Exact day of birth (1-366, due to leap years) is as close to a continuous measure as we can get in historical records, and season of birth is strongly correlated with socio-economic characteristics in modern data (Buckles and Hungerman 2013).

[27] We measure misspellings in father and mothers' names as the number of name spellings in the birth certificates of all siblings that differ from the modal spelling divided by the total number of children in a family

review with baseline characteristics. The method is identical to what is presented in Table 2.2 but that the dependent variable is equal to 1 if the link was rejected in a blind review. If the rejected links are systematically related to baseline characteristics, this suggests that the algorithm introduces systematic measurement error in variables of interest. Column 1 of Table 2.3 reports the heteroskedasticity-robust Wald-statistic (p-value beneath) by method for the LIFE-M data (see the full set of regression results are in Online Appendix D). For each algorithm, we reject the null hypothesis that errors in linking are unrelated to baseline characteristics at the 1-percent level. False links are significantly negatively associated with the length of a mother's name and length of a father's name in nearly all samples, suggesting that being falsely linked is negatively associated with affluence. Patterns across other variables are more varied. For example, the number of siblings is positively associated with the probability that a link is incorrect for the Feigenbaum (2016) algorithm, but the number of siblings is negatively associated with the probability that a link is incorrect in the Abramitzky, Boustan, and Eriksson (2014) sample. In short, different algorithms appear to induce different types of systematic measurement error.

2.4.2 Evaluating Algorithms Using the Synthetic Ground Truth and Early Indicators Samples

One critique of these findings is that human errors survive even the blind review process. This could lead the incidence of Type I errors in the LIFE-M analysis to be too high or too low relative to the truth. To address this potential issue, we reevaluate algorithm performance in synthetic data (where the truth is known). Because this objective truth is not influenced by human reviewers at all, this exercise validates those obtained from human review. We also evaluate the same algorithms using the Early Indicators, data, proving a complimentary perspective using a sample that was linked by genealogists and is known to be highly accurate.

For both the synthetic and Early Indicators data, Table 2.1 compares the match rates and error rates for each prominent algorithm. Recall, for the synthetic data, a perfect match rate is 85 percent, because 15 percent of the original links are absent by design. Patterns in match rates across methods are slightly higher in the synthetic data but generally within a few percentage points of the LIFE-M match rates, with the exception of Feigenbaum (2016) and Abramitzky, Mill, and Perez (2018). Notably, both methods perform substantially better in the synthetic data than in the LIFE-M data with match rates of 56 and 57 percent for Feigenbaum (2016) with the Iowa and estimated coefficients and 52 and 32 percent for Abramitzky, Mill, and Perez (2018) with the less and more conservative cutoffs. The match rates for Early Indicators' veterans linked to the 1900 complete count Census are generally higher than in the LIFE-M sample, which reflects the fact that all individuals in these data are known to be linkable. The LIFE-M data, however, contains both individuals who can be linked and those who cannot.

Table 2.1 also shows that patterns of error rates in the synthetic and Early Indicators data are similar to those in LIFE-M. Importantly, the best performing algorithms in LIFE-M continue to perform the best in the synthetic and Early Indicators data. Figure 2.2 describes patterns of error rates graphically across algorithms and datasets. In most cases, the error rates are slightly lower in the synthetic data and Early Indicators data relative to the LIFE-M records, ranging from 11 to 33 percent. Because there was no hand linking involved in producing the synthetic data, similar error levels suggest that the results of the LIFE-M hand-linking process and blind review reflect true errors in the automated linking algorithms. Larger reductions in error rates for machine-learning algorithms like Feigenbaum (2016) and Abramitzky, Mill, and Perez (2018) suggest that these methods may be effective at detecting the simple errors we simulated. Because the Early Indicators data contain only individuals who have been successfully linked by genealogists, Type I error rates are lower, ranging from 10 to 24 percent versus 15 to 37 percent in the LIFE-M data. The fact that the patterns of error rates are the similar in all datasets, however, provides strong support for the notion that prominent machine linking methods in current practice make considerable errors.

The findings for representativeness in the synthetic and Early Indicators data are also similar, with Table 2.2 suggesting that the linked samples are unrepresentative. For the synthetic data, this exercise allows a particularly strong test of the hypothesis that the non-representativeness of linked samples reflects the linking algorithm per se. Because we randomly dropped 15 percent of individuals, non-random attrition due to differential death, enumeration, or emigration is ruled out by construction. The only reason that the linked synthetic sample would not be representative is that the methods link certain groups more systematically than others. Consistent with this hypothesis, the Wald-statistics and p-values in column 2 reject representativeness for all methods in the synthetic data at the 1-percent level. Most methods are less likely to link individuals with more siblings. In the Early Indicators data, nearly all methods are more likely to link individuals with U.S.-born mothers; some methods are more likely to link individuals with longer first or last names, while others exhibit the reverse correlation.[28] (See Online Appendix C for the full set of regression results.)

Table 2.3 underscores the finding that false are systematically related to baseline sample characteristics. For all methods in both the synthetic and Early Indicators data, we reject the null hypothesis that false links are unrelated to baseline covariates - a pattern that may complicate inference by introducing systematic measurement error. (See Online Appendix D for the full set of regression results.)

[28] For the synthetic dataset, we use the same covariates as in the LIFE-M data when considering representativeness. For the Early Indicators data, we use continuous variables in age and length of first and last names and dummy variables for speaks English, owns a farm, currently married, foreign born, day of birth by year, literacy, and foreign born status of parents.

2.4.3 Summary of Findings

Prominent algorithms yield widely varying results - even using the same data and linking variables. In the LIFE-M data, match rates range from 28 to 52 percent, while the share of links rejected in the police line-up ranges from 15 to 37 percent and the associated Type II error rate ranges from 63 to 79 percent. A synthetic ground truth dataset confirms these patterns and also suggests that machine-learning algorithms like Feigenbaum (2016) and Abramitzky, Mill, and Perez (2018) are effective at detecting and correcting for synthetic errors, which speaks to their potential value in improving the quality of linked data. An equally important finding is that error rates vary across datasets - even when links are created using the same algorithm and linking variables. The share of links rejected by humans in the Early Indicators data is slightly lower and the match rates are higher, possibly owing to the fact that the data consist of individuals selected on having been linked by genealogists (i.e., living in the U.S. and also less likely to have changed a name or its spelling). However, error rates likely differ across the datasets due to differences in data quality that are difficult to easily measure and diagnose. This variation across datasets cautions against generalizing this paper's findings and recommends that researchers examine their linked samples for clues about error rates, representativeness, and systematic measurement error.

2.5 How Variations in Algorithms Alter Method Performance

Understanding what drives differences in algorithm performance is key to improving existing methods and current practice. This section considers how the performance of these algorithms changes when varying key features of their set-ups. First, we examine the role of different phonetic name cleaning strategies for algorithms that require agreement in cleaned names. Second, we extend the Ferrie (1996) algorithm to include more common names or eliminate them using a narrow age-band as in Abramitzky, Boustan, and Eriksson (2014)'s robustness test. Third, we examine equal weighting of exact ties (i.e., multiple, exact matches). A final section examines the robustness of these findings across all methods to using middle names, information on race, and extensions to population-to-population linking.

2.5.1 Phonetic Name Cleaning, Common Names, and Ties

Phonetic string cleaning algorithms account for orthographic differences that could lead a true match to be missed, such as minor spelling differences, name Anglicization, and transcription/digitization errors. Figure 2.3 and Table 2.4 show how the performance of the Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014) algorithms vary with three types of phonetic name cleaning: no cleaning (labeled "Name"), Soundex (labeled "SDX"), and NYSIIS. Interestingly, although name cleaning is intended to increase match rates, it can also decrease match rates if it increases ties (by removing meaningful spelling variations). This interaction is important for the Ferrie (1996) method, which matches between 20 and 33 percent of baseline sample depending on the phonetic name cleaning used. Because this cleaning creates more common name strings, and Ferrie's algorithm restricts the sample to uncommon names, the algorithm discards more links due to ties when cleaning is used: the match rate falling from 33 (Name) to 28 (NYSIIS), to 20 percent (Soundex). Because Abramitzky, Boustan, and Eriksson (2014) does not restrict to uncommon names, it does not show reductions.

The likelihood of Type I errors increases with the use of phonetic name cleaning in Ferrie (1996) and Abramitzky, Boustan, and Eriksson (2014). As shown in Table 2.4 and Figure 2.3, using NYSIIS rather than uncleaned names increases Type I error rates by an average of 18 percent, ranging from as little as 5 to as much as 25 percent across datasets. Using Soundex rather than uncleaned names increases

Type I error rates by an average of 36 percent, ranging from 14 to as much as 64 percent. These increases occur because, in addition to orthographic and transcription errors, phonetic codes may remove meaningful spelling variation. For example, both Soundex and NYSIIS would code "Meyer" and "Moore" as the same name, whereas reviewers tend to treat these as different names. Counterintuitively, the increase in error rates induced by phonetic name cleaning universally decreases the share of the sample that is correctly linked.

Another modification to Ferrie (1996) is to link more common names. Recall that, for computational reasons, Ferrie (1996) discarded matches if there were 10 or more candidate matches, regardless of age differences. We relax this assumption and include records with 10 or more candidates when we find links (labeled "Ferrie 1996 + common names"). Table 2.4 and Figure 2.3 show that including common names results in significantly higher match rates, including higher shares of true links than in the original method. The results are almost identical to those of Abramitzky, Boustan, and Eriksson (2014), which is because this method's key deviation from Ferrie's (1996) is attempting to link more common names. The inclusion of common names roughly doubles the share of incorrect links but it also decreases Type II error rates.

Similar in spirit to the common names restriction, Abramitzky, Boustan, and Eriksson (2014) implement a robustness check that discards links if a name tie occurs within a two-year age band. As reported in Figure 2.3 under Abramitzky, Boustan, and Eriksson 2014 (NYSIIS, Robustness), this restriction lowers the match rate to 24 percent but it also halves the share of the sample that is incorrectly linked, from 14 percent to 6 percent, and changes the Type I error rate from 32 percent to 23 percent. This robustness check is very similar to the uncommon names restriction and, therefore, performs almost identically to Ferrie (NYSIIS) in terms of match rates and Type I errors. Relatedly, the Abramitzky, Mill, and Perez (2018) algorithm with more conservative cut-offs is also similar to an uncommon names restriction. By requiring a high threshold of the probability of a candidate match being a correct match, and requiring the second-best options for the observations in the match to be much lower, this restriction ensures that the links made have no close analogues either due to differences in spelling or age. This unsupervised approach slightly outperforms Ferrie (NYSIIS) in terms of Type I and Type II error rates.

A third variation on prominent algorithms relates to how ties are handled (e.g., cases like M3 in Figure 2.1C and M4 in Figure 2.1D). Ties are very common in contexts with limited information (such as matching between U.S. Censuses). If one could break exact ties or use ties in the analysis, researchers could match the majority of the sample, raising match rates substantially. For the Ferrie (1996) algorithm, using both common names and ties raises the match rates from 20 to 33 percent to 69 to 86 percent.

Two main approaches to using exact ties have been suggested by the literature. First, the statistics literature offers an alternative to tie-breaking by using probability weighting. For instance, one could use a weight that is the conditional probability that the match is correct (Scheuren and Winkler 1993, Lahiri and Larsen 2005). In the absence of other data features, this suggestion simplifies to weighting by $\frac{1}{J_r}$, where J_r is the number of exact ties for record r.[29] Nix and Qian (2015)'s random selection among ties is similar in spirit. Their process draws one of the candidate matches with probability $\frac{1}{J_r}$. Importantly, simple probability weighting and random selection among ties have the same expected performance in certain contexts such as those we consider later for our case study. [30] We label results that include ties as "Ferrie 1996

[29] This simple probability weighting differs from Lahiri and Larsen (2005), because match probabilities vary in their data due to a specifically defined data generating process, and they are able to trim candidate links with lower match probabilities. We do not assume a specific data generating process. Furthermore trimming is not possible when all records are equally tied.

[30] We explain this result later in more detail, but the intuition is straightforward. Let N be the number of observations, M be the number of primary records with multiple exact ties as their best matches, J_r the number of ties for a primary record r$= 1, 2, ..., M$. Assuming that one of the ties is the correct link, the expected number of false matches for records with ties is $\sum_{r=1}^{M} \frac{J_r - 1}{J_r} = M - \sum_{r=1}^{M} \frac{1}{J_r}$

+ ties." Table 2.4 shows that including ties may dramatically increase match rates, but Figure 2.3 shows that this substantially increases the share of observations that are incorrectly matched in every sample. This makes sense, because at most one of the candidate links can be correct. For instance, if there are ten candidate "John Smith" links and only one of these is the correct link, nine out of ten of these links are incorrect. Notably, the Type I error rate is higher in the Early Indicators data, reflects the fact that they are selected upon being successfully linked and have fewer close ties.

Figure 2.5 describes the mixed progress in historical automated linking since 1996. As the literature has moved from the use of Ferrie's (1996) uncommon name sample and increased match rates, some methods have also increased Type I errors (and decreased precision). The hope of researchers using these methods is that, on net, they are increasing the share of true links in their sample as well as sample representativeness. However, for the synthetic and Early Indicators data, the pattern of Type I and Type II errors suggest that there may be scope to improve in both dimensions by leveraging the strengths of different algorithms.[31] Similar to the findings for prominent methods, each of these variations produces samples that are unrepresentative (Table 2.5). Moreover, these variations produce false links that are systematically related to baseline sample characteristics in all datasets (Table 2.6).

for both random selection and simple probability weighting. As the number of multiples increases for a given record, the probability weight on a false match gets smaller as does the weight on the true match. The results from probability weighting may differ slightly due to sampling variation.

[31] Of course, the level of Type II errors in the LIFE-M and synthetic data is overstated, because some infants did not survive until the 1940 Census, emigrated, or were missed by enumerators in the 1940 Census. We estimate that these factors likely account for around 15 percent of missed links (see footnote 16). A linking method that linked all LIFE-M or synthetic data individuals correctly would locate at (0.15, 0), missing only the 15 percent individuals who are unlinkable and making no errors. Because these sources of attrition affect all methods equally, these factors do not influence our comparisons across methods.

2.5.2 Robustness: Middle Names, Race, and Population-to-Population Linking

How much should we expect these results to change with the addition of information commonly available in historical datasets? A first robustness check considers how the addition of middle name or race could reduce Type I error rates. For middle names, we examine a subsample of cases where middle name or initial was available for both the birth certificate and the linked Census record. Then, we calculate the number of false links that would have been eliminated had the automated method required that middle initial match for all potential matches after running a matching algorithm. We apply this restriction ex post, but it would also be possible to include middle name agreement in the matching process as a feature considered by an algorithm in the process of making matches.

Table 2.7 shows that the availability of middle initials may reduce match rates but also the reduce rate of false matches. For comparison, columns 1 and 2 reprint the information on match rates and Type I error rates from Table 2.4. Column 3 shows the share of matched observations that have information on middle initial in the birth certificate and the Census record, which range from a quarter to a third of matches. Column 4 reports the share of matches that have discordant middle initials among the subset of matches that have middle names in both records, ranging from 20 percent to 57 percent. Column 5 reports the Type I error rate among the matches with discordant middle initials. What is clear is that the Type I error rate is always above 87 percent within this subset. Presumably, these error rates are high because disagreements among middle initials mattered to the trainers considering these observations when making matches. Finally, columns 6 and 7 recalculate the match rate and Type I error rates after dropping observations that have discordant middle initials. Match rates tend to drop by several percentage points, but Type I error rates drop by more. For example, Ferrie (1996) with NYSIIS drops from a match rate of 28 percent and an associated Type I error rate of 25 percent to a match rate of 26 percent and an associated Type I error rate of 20 percent. Type II error rates are nearly unchanged despite the drop in match rates, with all changes in Type II error rates never exceeding one percentage point. This result suggests that the addition of more information contained in middle names can substantially reduce Type I linking errors with minimal changes in Type II errors, at least among observations that have middle initials in the LIFE-M data.

A second robustness check compares race indicated on the 1940 Census for LIFE-M linked records to the race reported on an individual link.[32] Column 3 of Table 2.8 shows the share of linked birth certificates that would not have been erroneously linked by an algorithm that blocked on race. Interestingly, only a small share of incorrect links have discordant races, ranging from 0 to 5 percent. When we omit incorrect links with discordant races in column 5, we find that the match rate drops slightly and the Type I error rates decrease by no more than two percentage points across methods. This is consistent with Massey (2017), who finds that errors in linking the 2005 Current Population Survey to the Numident only decreased by 0.07 percentage points when blocking on race. In contrast to using middle initial, race does not appear to add much information to reduce errors. Note, however, that including race may result in better link disambiguation among links that appear to be close substitutes.

A third robustness check examines to the implications of linking a sample (rather than a population) to a Census. The critical difference in these two settings is that an observation could appear to be unique in a sample while having duplicates or near-duplicates in the population. This is important because many algorithms drop a match if (1) it occurs more than once in the set of records to be linked or (2) more than one observation links to the same record. Both are more likely to occur for a

[32] To the extent that some individuals ”pass” for other races, this robustness check may eliminate true links (Nix and Qian 2015, Mill 2013). Note that race is only observable for the observations that LIFE-M successfully linked, as we infer race from the 1940 Census given that it is not reported on birth certificates.

population than a sample. Sample-to-population linkage may, therefore, result in a higher share of incorrect matches than population-to-population linkage.

To quantify the importance of linking a sample to a population as we do here, we compare our results to matching the universe (e.g., the population) of birth certificates to the 1940 Census. First, we match the universe of Ohio and North Carolina birth certificates to the 1940 Census using each automated method, including adjustments described above. Then, we isolate attention to the subset of records in the LIFE-M sample to assess performance. Since the LIFE-M sample is a random subsample of birth certificates, we expect the results to generalize to the population.

Figure 2.6 displays the results, with the horizontal axis depicting Type I error rates in matching between the sample and the 1940 Census and the vertical axis displaying the results matching the population of birth certificates to the 1940 Census. Results along the 45-degree line indicate perfect agreement in the two procedures. As expected, false link rates fall below the 45-degree line for all methods that use the post-linking adjustments, suggesting that Type I error rates are somewhat higher for sample-to-population linkage. The Type I error rate in Ferrie (1996) with exact names is 20 percent for a sample but 17 percent for the population; for Abramitzky, Boustan, and Eriksson (2014) with exact names, the Type I error rate is 25 percent for the sample versus 20 percent for the population. In short, population-to-population linking may reduce errors but the improvements are not large, as no method achieves a Type I error rate lower than 15 percent.

2.5.3 Summary of Findings

Variations on machine-linking algorithms may improve or worsen performance. Deterministic algorithms that clean names using Soundex or NYSIIS perform worse than using raw name strings. Similarly, trying to link common names (especially in conjunction with phonetic name algorithms) tends to increase error rates and the share of the initial sample correctly linked. Tie breaking or weighting ties equally could dramatically increase sample sizes but may have the unintended effect of using more incorrect matches in analyses. Including middle names as a linking criterion appears to have large effects on Type I error rates. However, using race information or using population-to-population linking appears to alter data quality only modestly.

2.6 How Automated Methods Affect Inferences

Our final analysis explores the consequences of Type I and Type II errors for inferences about historical rates of intergenerational mobility. Following the intergenerational literature (Solon 1999, Black and Devereux 2011), we consider the following benchmark specification,

$$\log(y) = \pi \log(x) + \epsilon \tag{2.1}$$

where the dependent and independent variables have been rescaled to capture only individual deviations from population means. The dependent variable, $\log(y)$, refers to the log of son's wage income in adulthood in the 1940 Census. The key independent variable, $\log(x)$, refers to the parent's log wage income in the 1940 Census. Within this framework, we interpret π as the intergenerational income elasticity. The magnitude of π is an important indicator of the role that parents' wage incomes play in determining their children's wage income. Intergenerational mobility is measured as 1- π, which is often regarded as a metric of economic opportunity. Our analysis uses the LIFE-M sample of 19,486 boys (43 percent of the 45,442 that were linked to the 1940 Census) and samples linked using different automated methods to estimate intergenerational mobility. Unlike other analyses using the Census and Panel Survey of Income Dynamics, we must link fathers from birth certificates to the 1940 Census to obtain their income information. Links for fathers are obtained using only the LIFE-M clerical review method, so that father links remain constant in all regressions. By using the same links for fathers and different methods to link sons, our analysis describes differences in the estimates that are driven by differences in methods used to link sons.

2.6.1 How Type I Errors Affect Inferences

Different kinds of Type I errors could have different implications for inferences about intergenerational mobility. Within the regression framework in equation (2.1), measurement error in son's income (the dependent variable in the regression) that is uncorrelated with father's income will still allow us to estimate π consistently using OLS, though the estimates will be less precise. However, measurement error on the right-hand side in father's income (the independent variable in the regression) is more consequential. At first glance, considering measurement error in father's income seems counter to our problem of using different linking methods to link sons. Note, however, that linking a boy to the wrong man in 1940 is equivalent to assigning the wrong father's income to that man.

Our conceptual framework for thinking about linking-induced measurement error is similar to Horowitz and Manski (1995). We assume that a linking method, ℓ, induces Type I error in matches by erroneously linking a father to a son (we do not derive bounds here, but that is a useful avenue for future research). The presence of this measurement error allows us to divide the sample into two groups, g: one for which the links are correct, denoted with a $*$, and another for which the link is imputed (or incorrectly classified), i. Following Greene (2008) and Stephens and Unayama (2017), we decompose the OLS estimate of π for a sample linked with method, l, into the sum of within and between covariance for the correct, $*$, and imputed groups, i. b denotes the between component. Let $s_{xy}^{\ell *} + s_{xy}^{\ell i} = \sum_g s_{xy}^{\ell g} =$

$\sum_g \sum_k \big(\log(x_{kg}) - \overline{\log(x_{kg})}\big)\big(\log(y_{kg}) - \overline{\log(y_{kg})}\big)$ and let $s^{\ell b}_{xy} = \sum_g N_g\big(\overline{\log(x_{kg})} - \overline{\overline{\log(x_{kg})}}\big)\big(\overline{\log(y_{kg})} - \overline{\overline{\log(y_{kg})}}\big)$ where group means are defined with a single bar and overall means are defined by two bars, such that:

$$\hat{\pi}^{\ell} = \frac{s^{\ell}_{xy}}{s^{\ell}_{xx}} = \frac{s^{\ell *}_{xy} + s^{\ell i}_{xy} + s^{\ell b}_{xy}}{s^{\ell}_{xx}} = \frac{s^{\ell *}_{xx}}{s^{\ell}_{xx}}\hat{\pi}^{\ell *} + \frac{s^{\ell i}_{xx}}{s^{\ell}_{xx}}\hat{\pi}^{\ell i} + \frac{s^{\ell b}_{xx}}{s^{\ell}_{xx}}\hat{\pi}^{\ell b} \tag{2.2}$$

Equation (2.2) shows that an OLS estimator converges in probability to a weighted average of the plim for the correct links, $\pi^{\ell *}$, imputed links, $\pi^{\ell i}$, and the between group term ($*$ vs. i) $\pi^{\ell b}$, where the weights on each term reflect the share of variance due to each component, θ:

$$\text{plim } \hat{\pi}^{\ell} = \theta^{\ell *}\text{plim } \hat{\pi}^{\ell *} + \theta^{\ell i}\text{plim } \hat{\pi}^{\ell i} + \theta^{\ell b}\text{plim } \hat{\pi}^{\ell b} \tag{2.3}$$

The between group component can be thought of as the ”selection” term. In some cases, we expect that the plim of the between term is zero (e.g., if the means of son’s income or father’s income are the same for the imputed and correctly linked groups). This pattern could happen in practice if errors (e.g., enumeration or transcription error) randomly assign records to these groups. Initially, we assume this term is zero to simplify exposition but later relax this assumption. Note also that, if the variances of father income are equal across all groups, the weights θ become the share of the sample in each category. Now, consider the probability limit of the two remaining non-weight terms, $\hat{\pi}^{\ell *}$ and $\hat{\pi}^{\ell i}$. The first term represents the elasticity for the linked subsample, plim $\hat{\pi}^{\ell *} = \pi$. The second term is an estimated elasticity for the imputed observations. If we assume $\text{cov}\big(\epsilon, \log(x^{\ell i})\big) = 0$, then

$$
\begin{aligned}
\text{plim } \hat{\pi}^{\ell i} =& \frac{\text{cov}\big(\log(y^*), \log(x^{\ell i})\big)}{\text{var}\big(\log(x^{\ell i})\big)} = \frac{\text{cov}\big(\pi \log(x^*) + \epsilon, \log(x^{\ell i})\big)}{\text{var}\big(\log(x^{\ell i})\big)} \\
=& \pi \frac{\text{cov}\big(\log(x^*), \log(x^{\ell i})\big)}{\text{var}\big(\log(x^{\ell i})\big)}
\end{aligned}
\tag{2.4}
$$

If the imputed father's income is the same as the true father's income, $\log(x^*) = \log(x^{\ell i})$, then plim $\hat{\pi}^{\ell i}$=plim $\hat{\pi}^i$. However, if $\frac{\text{cov}\big(\log(x^*),\log(x^{\ell i})\big)}{\text{var}\big(\log(x^{\ell i})\big)} \neq 1$, then plim $\hat{\pi}^{\ell i} \neq \pi$ and the degree of the inconsistency depends on the relationship between the true and imputed father's income.

There are several special cases of interest. First, suppose that there is no relationship between the true father's income and the imputed father log income, or that $\frac{\text{cov}\big(\log(x^*),\log(x^{\ell i})\big)}{\text{var}\big(\log(x^{\ell i})\big)} = 0$. Then, the plim $\hat{\pi}^{\ell i} = 0$ and the estimator is inconsistent in proportion to the share of imputed links, plim $\hat{\pi}^{\ell} = \theta^{\ell *}\pi$. Second, consider the case where imputed father's income equals the true father's income plus noise, or $\log(x^{\ell i}) = \log(x^*) + u$. Under the assumptions of the classical errors in variables model (plim($u\epsilon$) $= 0$, plim $\big(u \log(x^*)\big) = 0$, and plim $\big(u \log(y)\big) = 0$), then plim $\hat{\pi}^{\ell i} = \theta^{\ell i} \frac{\sigma_x^2}{\sigma_x^2 + \sigma_u^2}\pi$. Moreover, plim $\hat{\pi}^{\ell} = (1 - \theta^{\ell i})\pi + \theta^{\ell i}\frac{\sigma_x^2}{\sigma_x^2+\sigma_u^2}\pi$.

Third, it is well known that non-classical measurement error for the imputed fathers could lead to under or over-statement of the parameter of interest, plim $\hat{\pi}^{\ell i} > \pi$ or plim $\hat{\pi}^{\ell i} < \pi$.

As a final exercise, consider the effect of Type I errors on inference using exact ties. Consider a setting where N is the total number of records that one wishes to link and for $M \leq N$ of these records, $r = 1, 2, ...M$, there are J_r candidate matches that are tied. For instance, if the first record with ties involves 30 potential matches for a John Smith, age 40, then for $r = 1$, $J_1 = 30$. A second record, however, may only have 4 ties, so $r = 2$ and $J_2 = 4$. Assume that there is one correct link among the ties for record, r, indexed by $j = 1$, $\log(y)$, and imputed (but incorrect links)

$\log(y_j), j = 2, ...J_r$. From the researcher's perspective, the correct link is unknown and the probability that any one of the ties is correct is $\frac{1}{J_r}$.

Assume that one of these records would be selected at random to use in the analysis. By the same logic as in equation (2.2), a regression estimate of the intergenerational income elasticity can be decomposed into a variance-weighted sum of elasticities for three groups of observations - correct, unique links, denoted $*$; a correct link from the ties, denoted $j = 1$; incorrect links from the ties, denoted $j > 1$, and a "selection term" (which we assume is zero). Therefore, the estimated elasticity will be $\hat{\pi}^\ell = \frac{s^*_{xx}}{s_{xx}}\hat{\pi}^{*\ell} + \frac{s^{j=1}_{xx}}{s_{xx}}\hat{\pi}^{j=1,\ell} + \frac{s^{j>1,\ell}_{xx}}{s_{xx}}\big(\sum_{j=2}^{J_r}\frac{s^{j,\ell}_{xx}}{s^{j>1,\ell}_{xx}}\hat{\pi}^{j\ell}\big)$ and plim$\big(\sum_{j=2}^{J_r}\frac{s^{j,\ell}_{xx}}{s^{j>1,\ell}_{xx}}\hat{\pi}^{j\ell}\big) = \sum_{j=2}^{J_r}\lambda_j\text{plim }\pi^{j\ell}$

Therefore, the estimate of the intergenerational income elasticity using random selection to break ties is,

$$\text{plim }\hat{\pi}^\ell = \pi\Big[\theta^{*j} + \theta^{j=1,\ell} + \theta^{j>1,\ell}\big(\sum_{j=2}^{J_r}\lambda_j\psi_{j\ell}\big)\Big] \tag{2.5}$$

Note that this estimator is inconsistent if $\psi_{j\ell} = \frac{\text{cov}\Big(\log(x_1),\log(x_j)\Big)}{\sigma^2_{x_j}} < 1$.[33]

The degree of inconsistency is, again, determined by how much information is in the incorrect ties. If the weights, θ and λ, simplify to the expected shares of observations in each group (as they would if variances were equal across all groups as we note above), the degree of inconsistency is also related to the share of all records with exact ties, $\frac{M}{N}$ (implicit in the weight), as well as the number of multiples for each record, J_r.

These conclusions are identical if the elasticity is estimated with a probability-weighted estimator where the weight is the probability that any exact multiple in a set of exact multiples is the true match, or $\frac{1}{J_r}$. The probability limit of the estimator will be the same, although the performance of these methods may diverge in smaller

[33] Note that $\frac{\text{cov}\Big(\log(y_j),\log(x_1)\Big)}{\sigma^2_{x_j}} = \frac{\text{cov}\Big(\pi\log(x_j)+\epsilon,\log(x_1)\Big)}{\sigma^2_{x_j}} = \frac{\text{cov}\Big(\log(x_j),\log(x_1)\Big)}{\sigma^2_{x_j}}\pi = \text{plim }(\pi^j)$.

samples.[34] This result is intuitive because the same share of imputed observations would be present using probability weighting or random selection for exact ties. In summary, the presence of imputed links - either through random selection or probability weighting - will generally lead to inconsistency, with the degree of inconsistency increasing in the number of records with ties, the number of exact ties for a given record, as well as the relationship between imputed observation and the truth. After examining the role of Type II errors, we examine the quantitative importance of these errors in a case study.

2.6.2 How Type II Errors Affect Inferences

Social scientists are accustomed to working with small representative samples. As long as links are representative of the underlying population, higher Type II error rates should only reduce precision. Across linking methods and datasets, however, this paper finds evidence that samples of links are not representative. If Type II errors result in the selective representation of different groups and the relationship of interest is heterogeneous across these groups, Type II errors may also lead to inconsistent estimates of population parameters in linked samples.

Heterogeneity in intergenerational income elasticities is believed to exist for many reasons. For instance, researchers have concluded that intergenerational income elasticities are larger for blacks than whites (Duncan 1968, Margo 2016) and that patterns of mobility are substantially different for farmers compared to non-farmers (Hout and Guest 2013, Xie and Killewald 2013). If one group is over-represented in the linked data, this will bias inferences about the historical rate of the population's intergen-

[34] Reducing the influence of observations with less information is why some statisticians recommend truncating lower probability links, where presumably the covariance between the income of the father for the imputed link and the true link is small (Scheuren and Winkler 1993, Lahiri and Larsen 2005). Although Lahiri and Larsen (2005) propose an exactly unbiased estimator of π, this result only holds when the estimated link probability is uncorrelated with father's income and where an exact data generating process for links is estimated. But this result breaks down in many historical settings, because the distribution of matching variables (name, age, and birth place) are correlated with outcomes and, often, a parent's socioeconomic status (see Online Appendix C and D.

erational mobility.

To make this point concretely, assume that the two groups in equation (2.3) are high mobility, h, and low mobility, l (rather than correctly and imputed links). Denote the intergenerational income elasticities of these groups as π^h and π^l (where $\pi^h \leq \pi^l$), and the share of the variation attributable to each group is θ^h and θ^l, respectively. Finally, assume that there are no errors in linking. Therefore, following the logic of equation (2.2), the regression estimate of the population elasticity parameter for a given linking method, ℓ, is,

$$\text{plim } \hat{\pi}^{\ell} = \theta^{\ell h}\text{plim } \hat{\pi}^{\ell h} + \theta^{\ell l}\text{plim } \hat{\pi}^{\ell l} + \theta^{\ell b}\text{plim } \hat{\pi}^{\ell b} \tag{2.6}$$

The inconsistency of the probability limit in equation (2.6) depends upon several factors. First, if $\pi^h = \pi^l$ and the means for both groups of fathers and sons are the same, the selection term is 0 and having a non-representative sample will not affect inference. Having a representative sample matters only to the extent that the relationship of interest varies across those groups or the group's characteristics differ. Second, if $\pi^h \neq \pi^l$ (and the group means are the same), Type II errors that effectively decrease the share of variation attributable to one group will lead to an inconsistent estimate of the population intergenerational elasticity parameter. Suppose that a linking method introduces Type II errors, which effectively decreases the variation attributable to observations representing the low mobility group. (In the extreme, high rates of Type II errors could imply that none of the total variation is attributable to low mobility group.) These Type II errors would result in an elasticity estimate that puts lower weight on the low-mobility group, resulting in a lower estimated elasticity. Third, if $\pi^h = \pi^l$ but the group means are different, then the selection term will not be 0 and inferences will be affected in an ambiguous way. Both heterogeneity and selection, of course, may vary greatly across samples. The following case study examines the combined implications of non-representativeness (through heterogeneity and selection) using inverse propensity weights to adjust for differences in observed characteristics (DiNardo, Fortin, and Lemieux 1996, Heckman et al. 1998).

2.6.3 Results: Intergenerational Elasticity Estimates from the 1940 Census

Different linking methods could have large effects on intergenerational income elasticity estimates through their influence on both Type I and Type II error rates. Figure 2.7A reports estimates of the intergenerational elasticity using samples of sons linked using different methods. For the LIFE-M links, we estimate an income elasticity of 0.24 between fathers and sons. Consistent with lifecycle bias and transitory income fluctuations attenuating our estimates, this estimate is lower than modern estimates.[35] These biases, however, should not affect our comparisons across different linking methods for the same set of records.

Several important patterns emerge. First, higher Type I errors in matching tend to be associated with smaller intergenerational elasticities. Consistent with attenuation described in equations (4) and (5), estimates using linking samples with higher Type I error rates tend to be smaller. Using NYSIIS and Soundex tends to increase Type I error rates and produce smaller estimates than using the reported name. Moreover, weighting ties results in Type I error rates ranging from 50 to 67 percent and yields

[35] For instance, Chetty et al. (2014) estimates 0.33, which is itself smaller than estimates for the same period using survey data (Mazumder 2015). Life-cycle bias may attenuate the estimated intergenerational elasticity regardless of matching method (Mazumder 2005, Haider and Solon 2006, Black and Devereux 2011, Mazumder 2015). In addition, wage income observed in the 1940 Census is an imperfect measure of permanent income, and we expect the single year observation of income for both generations can generate downward bias in estimated elasticities due to the importance of transitory income (Solon 1992, Zimmerman 1992, Mazumder 2005, Ward 2019). On the other hand, the absence of farm and self-employed income in 1940 may lead this analysis to overstate mobility by excluding father-son pairs of farmers - an occupation that tends to be highly persistent across generations (Hout and Guest 2013, Xie and Killewald 2013). However, lifecycle bias and transitory income fluctuations should have similar effects for all methods.

intergenerational income elasticity estimates of 0.19 to 0.11. However, Type I error is not the only factor determining bias. It is notable that the more conservative Abramitzky, Mill, and Perez (2018), the method with the lowest Type I error rates, yields an intergenerational income elasticity that is 20 percent smaller and statistically different than the true coefficient. Conversely, a method with a comparatively high Type I error rate such as Feigenbaum (2016) achieves an estimated intergenerational income elasticity that is statistically indistinguishable from the LIFE-M elasticity. These results may reflect the fact that sample composition or a more systematic correlations of the errors with certain characteristics impact the coefficient. Online Appendix Figures A1 through A10 consider alternate implementations of Feigenbaum (2016) and Abramitzky et al. (2018). In general, the implementations that place more weight on precision (minimizing Type I errors) achieve estimated elasticities that are closer to the estimate with the hand-linked data. However, the decrease in Type I error is accompanied by a decrease in match rates and an increase in standard errors.

To examine the role of non-representativeness, we use inverse propensity-score weights to reweight the linked sample to have the characteristics of the LIFE-M birth certificate sample (Bailey, Cole, and Massey 2019).[36] Figure 2.7B shows that the reweighted intergenerational income elasticities tend to be slightly smaller in magnitude than the unweighted Figure 2.7A estimates. This result may stem from the modest over-representation of larger, less-mobile families in the linked sample. The attenuation of the coefficient for the more conservative Abramitzky, Mill, and Perez (2018) is cut in half, however, by using weights. For this case study, however, the effects of non-representativeness (as measured by the changes induced by reweighting)

[36] To construct these weights, we first run a probit model of link status (for each method) on covariates, X, which include an indicator variable for presence of middle name, length of first, middle, and last name, polynomials in day of birth, polynomials in age, an index for first name commonness, an index for last name commonness, number of siblings, an indicator variable for presence of siblings, and the length of own name as well as father's and mother's names. We then use the estimated propensity of being linked, $P_i(L_i = 1|X_i)$, for each method and reweight observations by $\frac{1-P_i(L_i=1|X_i)}{P_i(L_i=1|X_i)} \frac{q}{1-q}$, where q is the share of records that are linked. Distributions of inverse propensity score weights are plotted in Online Appendix E

on observed characteristics appear modest in comparison to the role of errors in linking. Of course, one might also choose to re-weight the sample to resemble the 1940 Census. Online Appendix E shows that these results are nearly identical to results presented here.

While estimates using machine-linked samples appear attenuated relative to LIFE-M, the attenuation is not always as severe as one might expect with random error. For instance, Ferrie (1996) with name results in a 20 percent Type I error rate but the intergenerational elasticity estimate obtained from these links is one percentage point different from the LIFE-M estimate. If the selection term in equation (3) were zero, and the signal to noise ratio in equation (4) were zero, one would expect to estimate 0.19 (=0.80*0.24). Therefore, one might think that fathers' incomes for imputed links positively covaries with the truth or that the Ferrie (1996) is positively selected on immobility. For tie-breaking methods, however, the attenuation appears more consistent with random error. For instance, Ferrie (1996) with common names and ties and Soundex shows a 69 percent Type I error rate and the intergenerational elasticity estimate is 0.11 in Figure 2.7A.

Figure 2.7C and Figure 2.7D directly examine the effects of incorrect matches by plotting $\hat{\pi}^*$, or the estimated elasticity for the "true" links (plotted as o with 95-percent confidence intervals) and $\hat{\pi}^i$, or the estimated elasticity for the "false" links (plotted as x) from separate regressions. Without the incorrect links, the estimates of the intergenerational income elasticity are very similar across groups at around 0.23 without weights (Figure 2.7C) and 0.23 with inverse propensity-score weights (Figure 2.7D). The comparability of unweighted estimates is especially striking given how different in size and representativeness the samples are. For instance, the number of true links varies from around 482 for Ferrie 1996 (Soundex) to 1,600 when using exact ties Ferrie 1996 (Soundex), but the unweighted intergenerational income elasticities are estimated to be 0.22 and 0.23, respectively. Consistent with Type I errors introducing attenuation, π^i tends to be smaller than π^* across methods. And, consistent with the observations about the magnitudes above, the unweighted estimated intergenerational elasticities for the imputed links for Ferrie 1996 (Name) are 0.15 and only 0.05 for Abramitzky, Boustan, and Eriksson (2014) (Soundex) and 0.05 for Ferrie 1996 with common names and ties (Soundex) - a statistical zero in the latter two cases. On the other hand, the correlation of incorrect links for Feigenbaum (2016) is very high, which shows how the regression-based classification system selects links with a very high correlation to the true link in this setting - even when incorrect. In short, the inclusion of imputed links appears to have large effects on OLS estimates of the intergenerational income elasticities, biasing them toward zero in most cases. After purging incorrect links, reweighting the linked sample to resemble the set of birth certificates has a minimal effect on the estimates.

In summary, the lower attenuation for some methods reflects the fact that the sample of correct links is selected on having lower intergenerational mobility (i.e., a higher elasticity), pulling the point estimate up, while the measurement error tends to pull the estimates down (see Figures 2.7C and 2.7D). Notably, the bias in estimates from some linking algorithms is as large as transitory income bias or life-cycle bias, which may each further attenuate intergenerational elasticities by more than 20 percent (Solon 1999, Haider and Solon 2006, Mazumder 2018). This case study suggests that researchers should carefully consider the role of measurement error due to linking algorithms, as their influence on inference could be significant.

We have discussed estimated elasticities here to explore the consequences of errors in linking, and we do not intend to offer these estimates as an unbiased estimate of the true elasticity in the early 20th Century due to issues with transitory income bias and life-cycle bias in our data. However, with strong assumptions, we can relate our estimates to the existing literature estimating intergenerational elasticities. If we assume the independence of transitory income bias and life-cycle bias, and assume both factors attenuate elasticities by approximately 20 percent given the ages at which we observe sons and fathers (Haider and Solon 2006), the true LIFE-M estimate of the elasticity could be around 0.36 for the early 20th century (over 56 percent or 1/(0.64) larger than observed). This is smaller than recent estimates (Mazumder 2018), suggesting that the US was much more fluid historically than in the later 20th century (Feigenbaum 2016, Ferrie and Long 2013). However, our estimate is based on fathers and sons with wage income from Ohio and North Carolina, and to the degree that the patterns of mobility present for this sample are not representative of the country as a whole, these findings may not generalize (Ward 2019).

2.7 Lessons for Historical Record Linking

New large-scale linked data hold the potential to shift the knowledge frontier, increasing the urgency for developing reliable linking methods. Using different U.S. samples, this paper documents how linking algorithms and resulting errors could have large effects on scientific conclusions and policy inferences. Not only are linked samples not representative, but existing algorithms yield high rates of false matches. Moreover, the incidence of false matches are systematically related to baseline sample characteristics, suggesting that linking-induced measurement error may introduce complicated forms of bias into analyses. Our case study shows that linking algorithms may severely attenuate estimates of intergenerational income elasticities.

The variability in our estimates across datasets implies that it is difficult to diagnose how much linking assumptions matter for different research questions using different records. Nevertheless, our results suggest that reducing false matches and choosing methods that generate false matches more highly correlated with the truth are crucial for improving inferences with linked data - even when reducing Type I errors increases Type II errors.

An easy remedy when linking richer data is to use more information - especially continuous variables or those with many values (e.g., Social Security Numbers or exact dates of birth). In addition, higher quality information (e.g., administrative records rather than individual reports) will result in lower error rates than we document. For contexts with limited linking variables which are measured with error, systematic clerical review (e.g., LIFE-M) and genealogical methods (e.g., Early Indicators) generally attain lower error rates than machine algorithms. Because these methods are cost prohibitive for most projects, we draw on our findings to recommend several easy-to-implement and lower-cost changes to current practice.

First, we recommend careful examination of a sample of links resulting from automated algorithms. Applying close scrutiny to a sample of links allows researchers to diagnose and potentially remedy systematic problems with machine-linking algorithms arising for specific records or in a particular historical context. In fact, many of the links coded as incorrect in clerical review are easy to identify as such. These cases can be used to improve machine-linking algorithms.

Second, caution is advised in linking phonetically cleaned names in deterministic algorithms or in commonly occurring name-age combinations. Phonetic cleaning tends to remove meaningful variation in names that allows algorithms to make better links. Eliminating commonly occurring name-age combinations, like Ferrie's (1996) approach of only linking uncommon names or Abramitzky, Boustan, and Eriksson (2014)'s robustness check using unique name-age combinations in a five-year window, substantially reduces the incidence of false matches. Together with reweighting, these restrictions achieve results in our case study that are statistically indistinguishable from hand-linked data. In contrast, weighting name-age ties equally by the inverse of their empirical frequency incorporates information from a large number of false links and results in substantial attenuation. In addition, researchers may incorporate more information in the linking process to break ties and distinguish true links from close alternatives. One such example in historical data is middle name or middle initial.

Third, using even a small sample of clerically reviewed data to train a machine-learning algorithm (or applying the results of another researcher's model based on similar training data) can improve the quality of linked samples. Notably, even when these machine-methods make incorrect links, the correlation of these links with the truth appears to be much higher than for other algorithms in our setting. These errors, therefore, have less impact on inference. An additional feature of some machine-linking algorithms is that they allow researchers to choose the importance of Type I and Type II errors, balancing the trade-off to fit a particular application. Although Feigenbaum (2016) and Abramitzky, Mill, and Perez (2018) choose a specific penalty for Type I and Type II errors, different parameter choices can drive Type I error rates lower while linking much of the sample correctly.

A fourth strategy for reducing Type I errors is to combine multiple methods and use the intersection of the links across sets - a form of ensemble machine learning in the spirit of "bagging" or "boosting." By construction, requiring links to be classified as such by more than one algorithm should tend to decrease match rates. But, to the extent that different methods make errors for different reasons, taking the set of common links helps avoid idiosyncratic reasons for errors. We illustrate the value of this approach in Figure 2.8 for our example of intergenerational elasticity, where we plot the Type I and Type II errors associated with the 131,071 possible combinations ($2^{17} - 1$) of the 17 algorithms in this paper for each dataset. Overall, combining methods drives down Type I error rates and increases Type II error rates. For example, when using LIFE-M data, combining two methods like Ferrie (1996) and Feigenbaum (2016) drives the Type I error rate to 10 percent - a substantial improvement over error rates for either method individually. Combining 12 methods achieves error rates as low as 6 percent, which is almost as precise as hand-linking.

Using combinations of methods may also improve inference. As shown in Figure 2.9A, across all combinations of methods, unweighted intergenerational elasticity estimates range from 0.11 to 0.24 (circle markers) and inverse-propensity-score reweighted estimates range from 0.13 to 0.24 (square markers). Based on an unweighted linear regression, a 10 percentage point increase in the Type I error rate tends to decrease the elasticity by 0.028, whereas this number is 0.015 in the weighted regression. Interestingly, in both weighted and unweighted cases, the mean over all combinations yields the value to the elasticity obtained in the hand-linked LIFE-M sample. As in our intergenerational elasticity example using single methods, Figure 2.9B shows that eliminating the incorrect links yields an average intergenerational elasticity nearly identical to the hand-linked sample (0.22) whereas the average intergenerational elasticity estimates for the incorrect links are less than half that value (0.096). These findings hold even when considering only the most prominent matching algorithms.

Finally, after limiting the role of linking errors, we recommend using multiple record features to assess and improve sample representativeness. Survey methods for constructing weights and allocating values are easy to implement and have well-documented properties. Making greater use of common record features such as name length or other socio-demographic information also allows researchers to use survey research methods or, as is more common in economics (and used in this paper), construct inverse-propensity weights to reduce sample selection and improve representativeness in observed characteristics (see Bailey, Cole, and Massey (2019) for a simple how-to guide). Ideally, this reweighting also improves balance in terms of unobserved characteristics, but there is no way to test this claim. A close examination of what is referred to as the common support assumption also informs researchers about where more time-intensive genealogical or clerical review methods may increase the representation of hard-to-link groups.

Many discussions of inference with linked data implicitly or explicitly assume that the match rate is just as important to inference as match quality. Our findings suggest that the quality of inferences with linked data may be improved by putting less emphasis on increasing sample sizes (which in our analysis tend to be associated with higher rates of false matches) and more emphasis on increasing the share of correct links. That is, social scientists wishing to conduct inference on linked data might increase the weight they place on decreasing Type I error rates over increasing sample sizes (decreasing Type II error rates). In the parlance of machine learning, this would involve weighting precision more heavily. Indeed, modern surveys such as the Panel Survey of Income Dynamics and the National Longitudinal Survey demonstrate that much can be learned from high-quality small samples with summary statistics and weights to describe and adjust for non-representativeness. Ultimately, increasing sample sizes for difficult-to-link subgroups (such as individuals with common names) will not likely be achieved without more data or higher quality record features to disambiguate similar records. More research to uncover data to describe the groups underrepresented in linked samples will serve both to broaden knowledge about them and improve the ability of modern machine learning methods to link them.

2.8 Figures and Tables

Figure 2.1: Examples of Common Linking Problems in Historical Samples

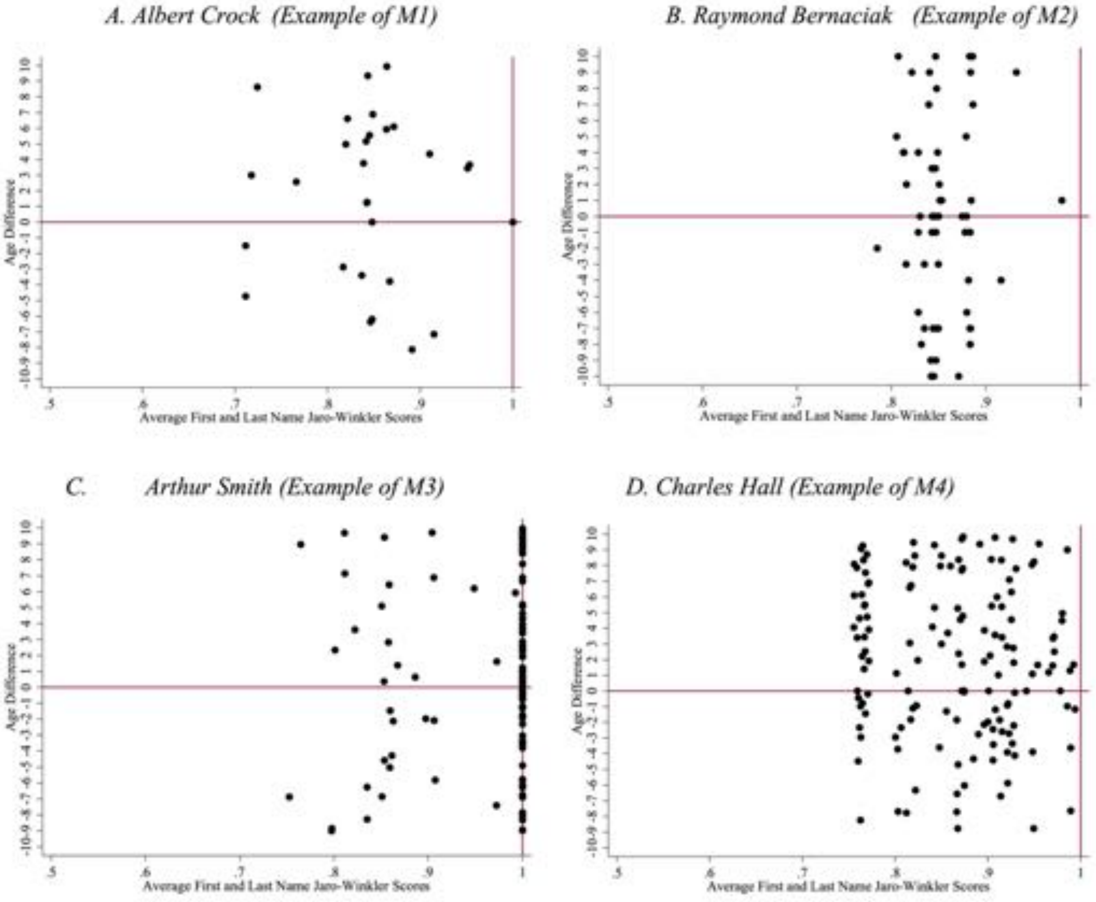
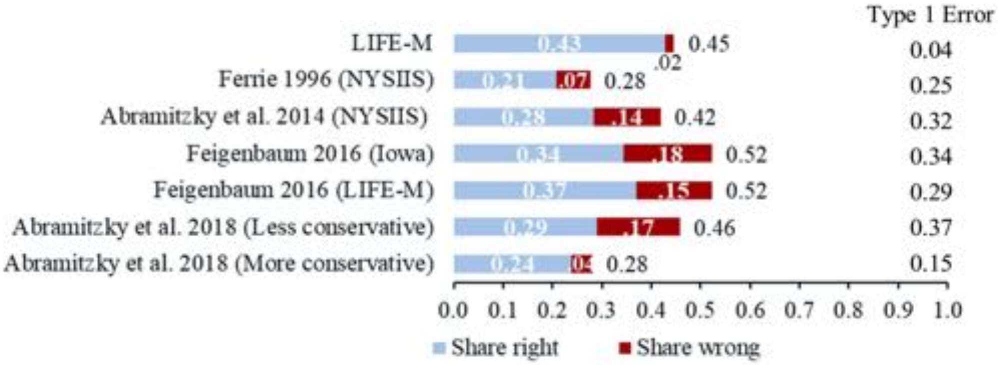
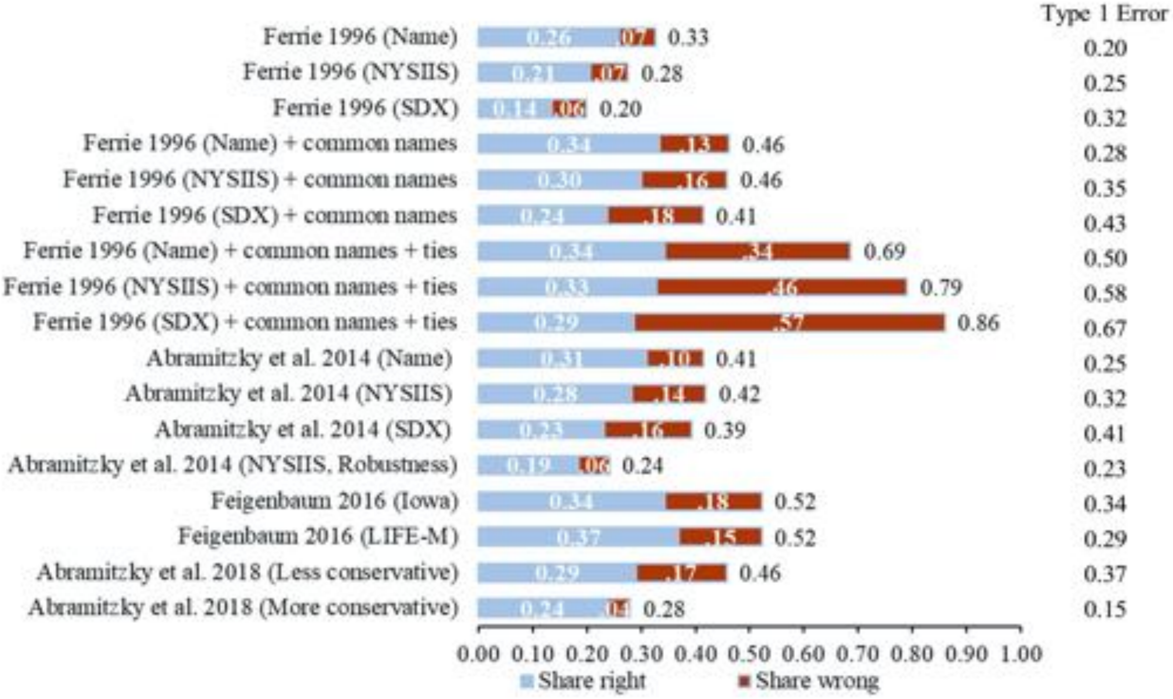


Figure 2.2: Match Rates and False Links for LIFE-M Hand-Linked Data and Selected Automated Linking Methods



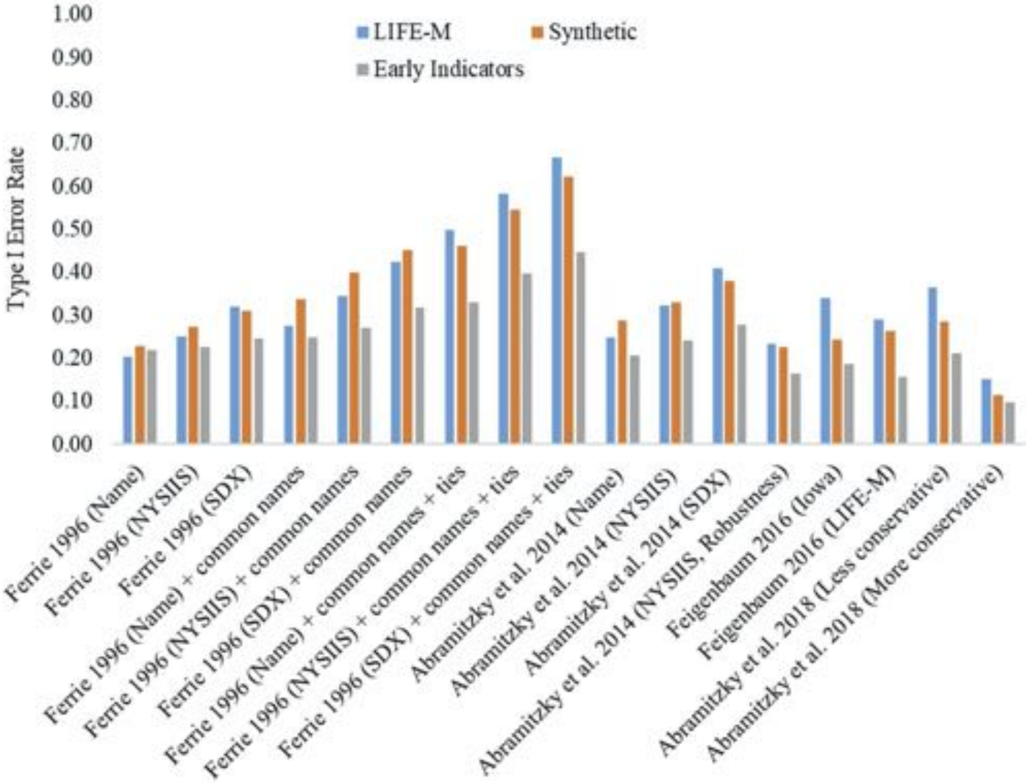
Notes: The bars show the performance of different algorithms linking LIFE-M boys to the 1940 Census. See text for details and Table 2.1 for numerical estimates. As a reminder, LIFE-M represents handlinked data before police batch review, and Ferrie (1996) NYSIIS, Abramitzky et al. 2014 (NYSIIS), Feigenbaum 2016 (LIFE-M), Abramitzky et al. 2018 (Less conservative) and Abramitzky et al. 2018 (More conservative) reflect the methods used to produce the main results from these papers.

Figure 2.3: Match Rates and False Links for Common Variations on Automated Linking Methods



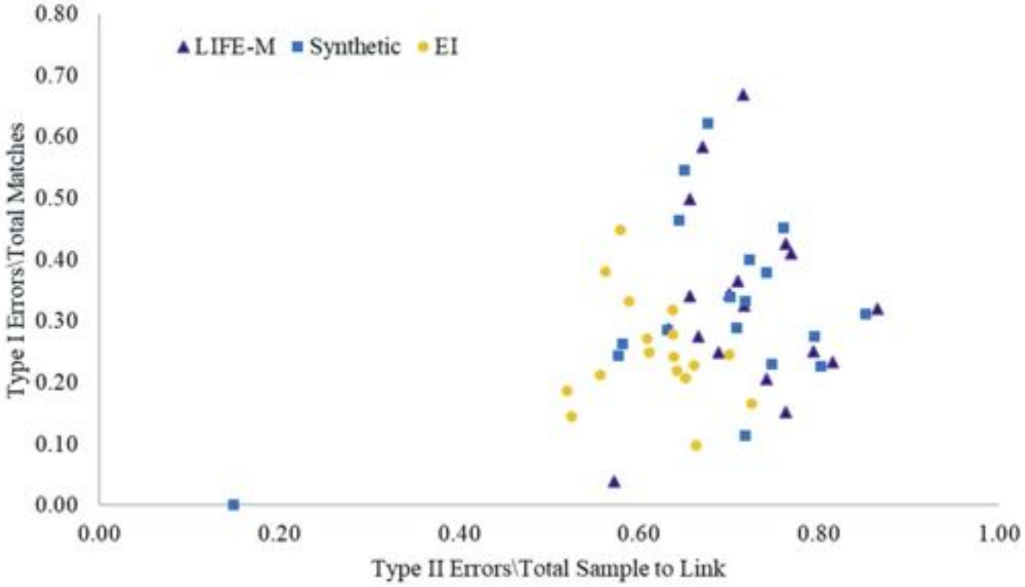
Notes: See Figure 2.2 notes.

Figure 2.4: Share of Incorrect Links (Type I Error Rate) by Method and Dataset



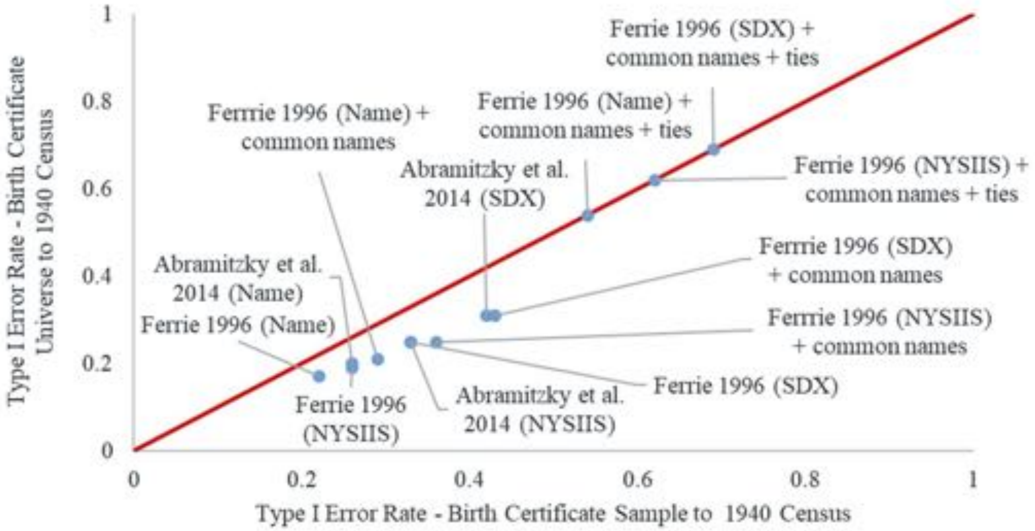
Notes: See Figure 2.2 notes and Table 2.1 for numerical estimates.

Figure 2.5: Type I vs. Type II Error Rates by Method and Dataset



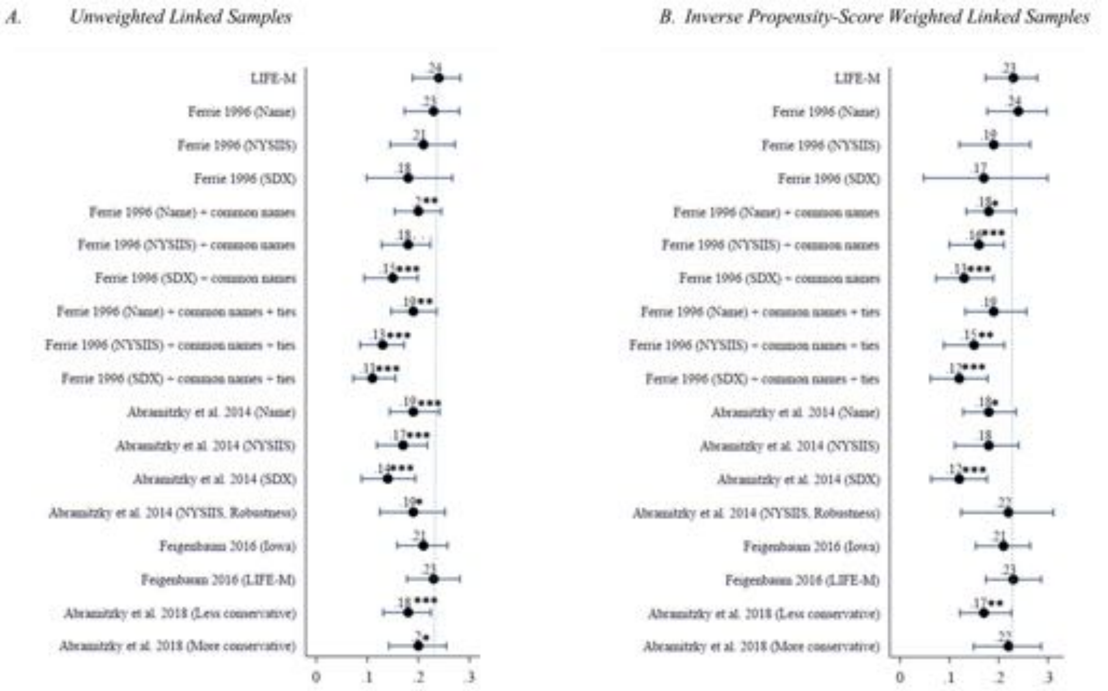
Notes: Points plot Type I and Type II error rates using different algorithms and data in Table 2.2.

Figure 2.6: A Comparison of Method Performance in Sample-to-Population and Population-to-Population Linking



Notes: The y-axis plots the Type I error rate implied by linking birth certificates for all boys in the same cohorts as the Ohio and North Carolina LIFE-M sample to the 1940 Census using different automated methods. The x-axis plots the Type I error rate implied by linking the LIFE-M sample of boys in the Ohio and North Carolina birth certificates to the 1940 Census using different automated methods.

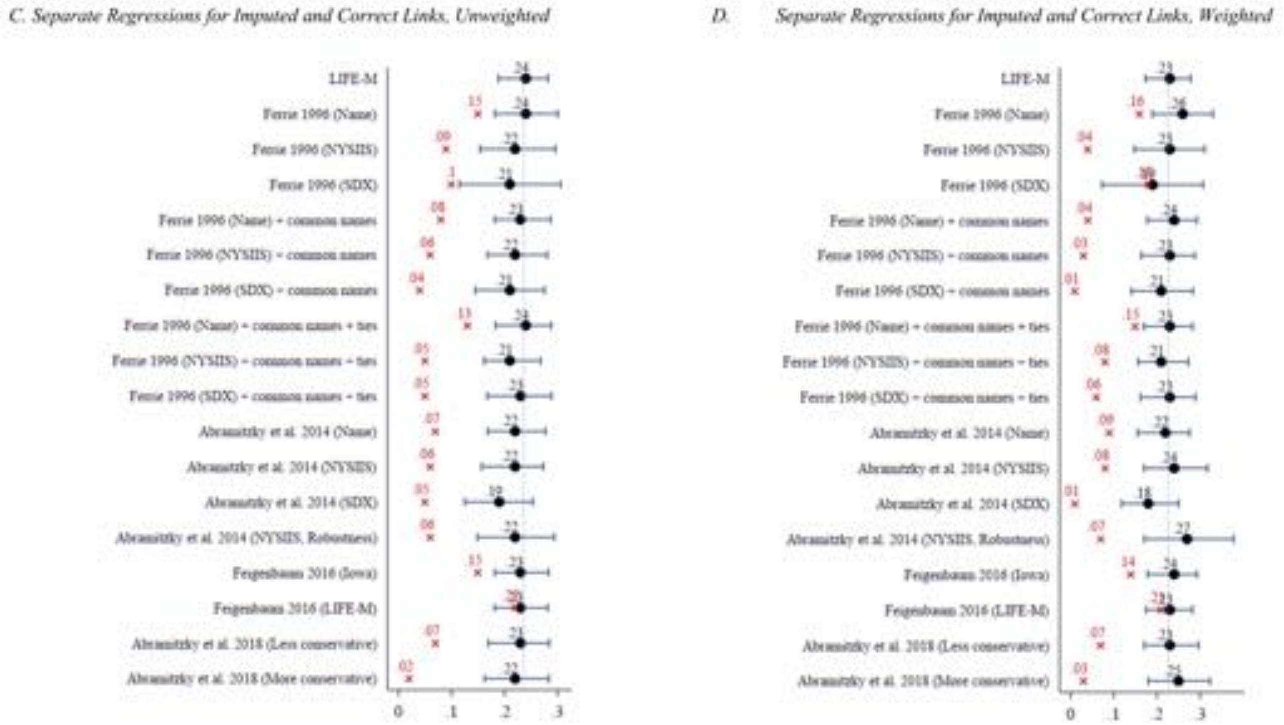
Figure 2.7: Intergenerational Income Elasticity Estimates



Notes: Differences in estimates reflect the incidence of Type I and Type II errors. The sample sizes of father-son pairs are lower than when matching sons only, because not all linked sons had income from wages and fathers who were also linked who also had income from wages. Sample sizes are 1,834 for LIFE-M, 1,313 for Ferrie 1996 (Name), 1,064 for Ferrie 1996 (NYSIIS), 708 for Ferrie 1996 (Soundex), 1,751 for Ferrie 1996 (Name) + common names, 1,702 for Ferrie 1996 (NYSIIS) + common names, 1,466 for Ferrie 1996 (SDX) + common names, 2,354 for Ferrie 1996 (Name) + common names + ties, 2,648 for Ferrie 1996 (NYSIIS) + common names + ties, 2,875 for Ferrie 1996 (SDX) + common names + ties, 1,610 for Abramitzky et al. 2014 (Name), 1,600 for Abramitzky et al. 2014 (NYSIIS, Robustness), 1,412 for Abramitzky et al. 2014 (SDX), 999 for Abramitzky et a. 2014 (NYSIIS, Robustness) 1,955 for Feigenbaum 2016 (Iowa), 1,855 for Feigenbaum 2016 (LIFE-M), 1774 Abramitzky et al. 2018 (Less conservative), 1206 Abramitzky et al. 2018 (More conservative).

Notes and figure continued in page below

Figure 2.7: Intergenerational Income Elasticity Estimates (Continued)

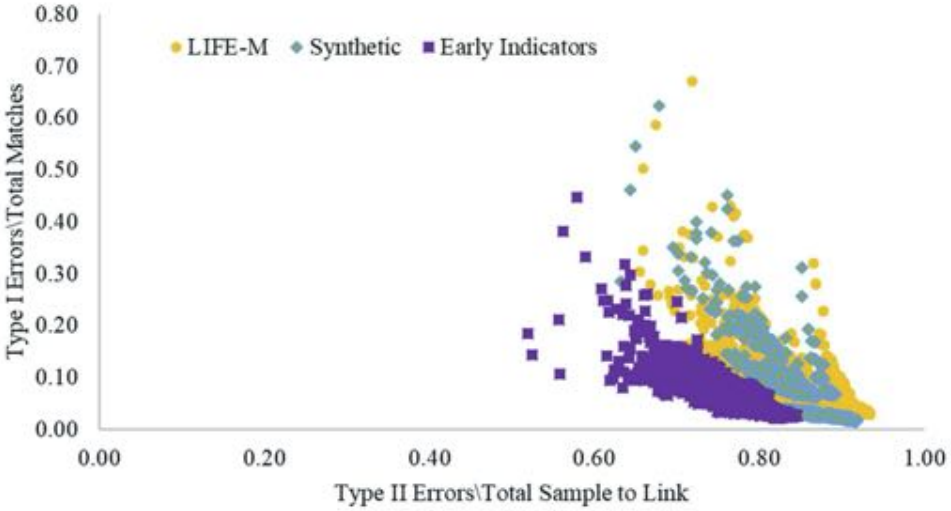


C. Separate Regressions for Imputed and Correct Links, Unweighted

D. Separate Regressions for Imputed and Correct Links, Weighted

Notes: *Continued from page above.* Reweighted estimates were reweighted to represent the LIFE-M sample of birth certificates linked to the 1940 Census. Weighting variables include day of year measured from one to 365, polynomials in age, first and last name commonness indexes and the interaction of the two, a dummy variable for presence of siblings, polynomials in the number of siblings, polynomials in the length of child, mother, and father name, and state fixed effects.. * indicates that the estimate is statistically different from the LIFE-M estimate at the 10-percent, ** at the 5-percent, and *** at the 1-percent levels.

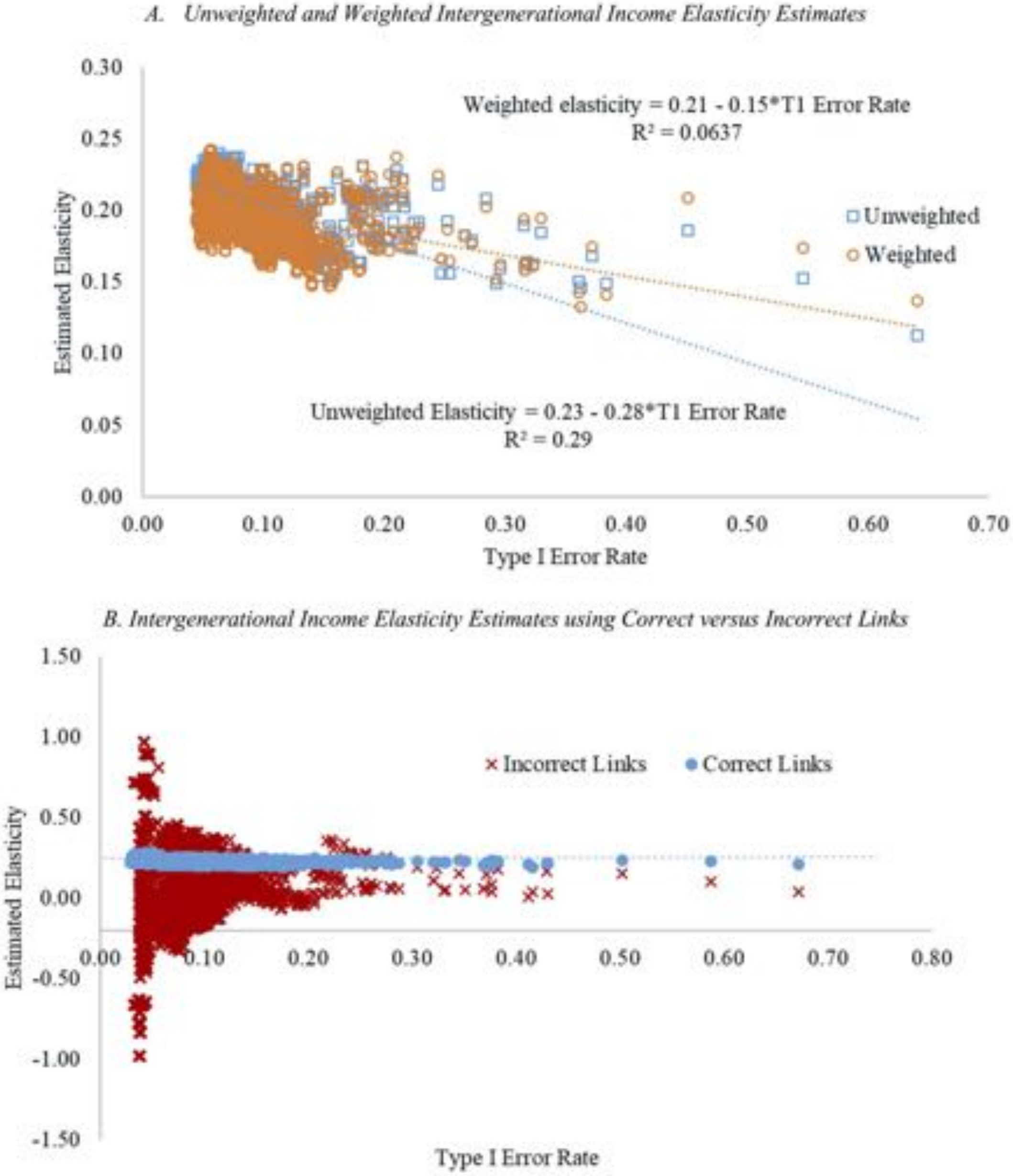
Figure 2.8: Type I vs. Type II Error Rates for Different Combinations of Methods and Dataset



Notes: Each point represents the Type I and Type II error rate for 131,071 different combinations of the 17 methods considered in this paper by dataset.

s

Figure 2.9: Intergenerational Income Elasticity Estimates across Method Combinations



A. Unweighted and Weighted Intergenerational Income Elasticity Estimates

B. Intergenerational Income Elasticity Estimates using Correct versus Incorrect Links

Notes: Each point represents intergenerational elasticity estimate plotted against the Type I error rate for one of the 131,071 different combinations of the 17 methods considered in this paper. Panel A pools all links and panel B plots estimates separately for correct and incorrect links. See also Figure 2.6 notes.

Table 2.1: Summary of Performance of Prominent Linking Methods, by Algorithm and Dataset

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Hand-links or Synthetic	0.45	0.85	1	0.04	0	0	0.57	0.15	0
Ferrie 1996	0.28	0.28	0.44	0.25	0.27	0.23	0.79	0.79	0.66
Abramitzky et al. 2014	0.42	0.42	0.48	0.32	0.33	0.24	0.72	0.72	0.64
Feigenbaum 2016 (Iowa coefficients)	0.52	0.56	0.59	0.34	0.24	0.19	0.66	0.58	0.52
Feigenbaum 2016 (Estimated coefficients)	0.52	0.57	0.57	0.29	0.26	0.14	0.63	0.58	0.52
Abramitzky et al. 2018 (Less conservative)	0.46	0.52	0.56	0.37	0.29	0.21	0.71	0.63	0.56
Abramitzky et al. 2018 (More conservative)	0.28	0.32	0.37	0.15	0.11	0.1	0.76	0.72	0.66

Notes: EI stands for the Early Indicators data. Each estimate in the table is for a match rate, Type I error rate, or Type II error rate as described in text. These estimates are depicted in graphical form in Figures 2.2, 2.3 and 2.4.

Table 2.2: Representativeness of Links Created by Prominent Linking Methods, by Algorithm and Dataset

	LIFE-M	Synthetic Data	EI
Ferrie 1996 (NYSIIS)	445.9	277.5	38.2
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (NYSIIS)	457	387.2	12.7
p-value	*0*	*0*	*0.24*
Feigenbaum 2016 (Iowa coef.)	195.7	34.9	50
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Estimated coef.)	334.9	62.2	44
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (Less conservative)	788.3	485	46.6
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (More conservative)	1350	673	51.4
p-value	*0*	*0*	*0*
Observations	42,869	42,869	1,785

Notes: Each estimate is a heteroskedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for linked record) for samples described in the text. Relevant p-values are reported in italics. The covariates included in the LIFE-M sample and synthetic data are age, number of siblings, length of names of individuals and parents, fraction of siblings with misspelled parents' names, and an observation coming from Ohio. The covariates included in the Early Indicators data are age, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. These sample sizes are slightly smaller due to missing values. See appendices for full regression results.

Table 2.3: Randomness of False Links Created from Prominent Linking Methods, by Algorithm and Dataset

	LIFE-M	Synthetic Data	EI
Ferrie 1996 (NYSIIS)	242.9	35.1	26.2
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (NYSIIS)	500.9	64.3	39.4
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Iowa coef.)	1806	448	38.9
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Estimated coef.)	1559	802	19.4
p-value	*0*	*0*	*0.03*
Abramitzky et al. 2018 (Less conservative)	559.3	112.9	43
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (More conservative)	139.8	17.4	18.3
p-value	*0*	*0.03*	*0.05*

Notes: Each estimate is a heteroskedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for falsely linked record) for samples described in the text. Relevant p-values are reported in italics. The covariates included in the LIFE-M sample and synthetic data are age, number of siblings, length of names of individuals and parents, fraction of siblings with misspelled parents' names, and an observation coming from Ohio. The covariates included in the Early Indicators data are age, currently married, foreign born, day of birth by year, literacy, length of first and last names, and foreign born status of parents. See appendices for full regression results.

Table 2.4: Summary of Algorithm Performance When Varying Assumptions

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Ferrie 1996 (Name)	0.33	0.33	0.46	0.2	0.23	0.22	0.74	0.75	0.64
Ferrie 1996 (NYSIIS)	0.28	0.28	0.44	0.25	0.27	0.23	0.79	0.79	0.66
Ferrie 1996 (SDX)	0.2	0.22	0.4	0.32	0.31	0.25	0.86	0.85	0.7
Ferrie 1996 (Name) + common names	0.46	0.45	0.52	0.28	0.34	0.25	0.66	0.7	0.61
Ferrie 1996 (NYSIIS) + common names	0.46	0.46	0.54	0.35	0.4	0.27	0.7	0.72	0.61
Ferrie 1996 (SDX) + common names	0.41	0.44	0.53	0.43	0.45	0.32	0.76	0.76	0.64
Ferrie 1996 (Name) + common names + ties	0.69	0.66	0.62	0.5	0.46	0.33	0.66	0.64	0.59
Ferrie 1996 (NYSIIS) + common names + ties	0.79	0.77	0.71	0.58	0.55	0.4	0.67	0.65	0.57
Ferrie 1996 (SDX) + common names + ties	0.86	0.86	0.76	0.67	0.62	0.45	0.71	0.68	0.58

Notes: See Table 2.1 notes. *Notes and table continued in page below.*

Table 2.4: Summary of Algorithm Performance When Varying Assumptions (Continued)

	A. Match Rates			B. Type I Error Rate (False Links)			C. Type II Error Rate (Missed links)		
	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI	LIFE-M	Synthetic	EI
Abramitzky et al. 2014 (Name)	0.41	0.41	0.44	0.25	0.29	0.21	0.69	0.71	0.65
Abramitzky et al. 2014 (NYSIIS)	0.42	0.42	0.48	0.32	0.33	0.24	0.72	0.72	0.64
Abramitzky et al. 2014 (SDX)	0.39	0.42	0.5	0.41	0.38	0.28	0.77	0.74	0.64
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.26	0.33	0.23	0.23	0.17	0.81	0.8	0.72
Feigenbaum 2016 (Iowa coef.)	0.52	0.56	0.59	0.34	0.24	0.19	0.66	0.58	0.52
Feigenbaum 2016 (LIFE-M coef.)	0.52	0.57	0.57	0.29	0.26	0.16	0.63	0.58	0.52
Abramitzky et al. 2018 (Less conservative)	0.46	0.52	0.56	0.37	0.29	0.21	0.71	0.63	0.56
Abramitzky et al. 2018 (More conservative)	0.28	0.32	0.37	0.15	0.11	0.1	0.76	0.72	0.66

Notes: *Continued from page above.* See Table 2.1 notes.

Table 2.5: Representativeness of Links When Varying Algorithm Assumptions

	LIFE-M	Synthetic Data	EI
Ferrie 1996 (Name)	688.3	390.7	47.5
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS)	445.9	277.5	38.2
p-value	*0*	*0*	*0*
Ferrie 1996 (SDX)	130.6	71.1	57
p-value	*0*	*0*	*0*
Ferrie 1996 (Name) + common names	412.3	378.1	35.5
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS) + common names	402.6	447	16
p-value	*0*	*0.1*	*0.1*
Ferrie 1996 (SDX) + common names	208.8	310.6	25.6
p-value	*0*	*0*	*0*
Ferrie 1996 (Name) + common names + exact ties	178.8	452.1	75.6
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS) + common names + exact ties	148.2	363.9	69.9
p-value	*0*	*0*	*0*
Ferrie 1996 (SDX) + common names + exact ties	104.7	174.7	43.6
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (Name)	454.6	271.9	32.3
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (NYSIIS)	457	387.2	12.7
p-value	*0*	*0*	*0.24*
Abramitzky et al. 2014 (SDX)	255.7	257.8	17.1
p-value	*0*	*0*	*0.07*
Abramitzky et al. 2014 (NYSIIS, Robustness)	568.6	397.7	31.2
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Iowa coef.)	195.7	34.9	50
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Estimated coef.)	334.9	62.2	44
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (Less conservative)	788.3	485	46.6
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (More conservative)	1350	673	51.4
p-value	*0*	*0*	*0*
Observations	42,869	42,869	1,785

Notes: See Table 2.2 notes.

Table 2.6: Randomness of False Links When Varying Algorithm Assumptions

	LIFE-M	Synthetic Data	EI
Ferrie 1996 (Name)	468.3	79.3	45.8
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS)	242.9	35.1	26.2
p-value	*0*	*0*	*0*
Ferrie 1996 (SDX)	81.5	0.9	17.5
p-value	*0*	*0.99*	*0.06*
Ferrie 1996 (Name) + common names	772.1	115.3	48.6
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS) + common names	429	64.4	39.2
p-value	*0*	*0*	*0*
Ferrie 1996 (SDX) + common names	157.7	17.4	32.4
p-value	*0*	*0.03*	*0*
Ferrie 1996 (Name) + common names + exact ties	1859	466.2	60.1
p-value	*0*	*0*	*0*
Ferrie 1996 (NYSIIS) + common names + exact ties	1163	249.6	92.3
p-value	*0*	*0*	*0*
Ferrie 1996 (SDX) + common names + exact ties	457.8	55.4	61.7
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (Name)	744.4	100.2	54.3
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (NYSIIS)	500.9	64.3	39.4
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (SDX)	223.2	41.7	28.6
p-value	*0*	*0*	*0*
Abramitzky et al. 2014 (NYSIIS, Robustness)	239	24.3	32.4
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Iowa coef.)	1806	448	38.9
p-value	*0*	*0*	*0*
Feigenbaum 2016 (Estimated coef.)	1559	802	19.4
p-value	*0*	*0*	*0.03*
Abramitzky et al. 2018 (Less conservative)	559.3	112.9	43
p-value	*0*	*0*	*0*
Abramitzky et al. 2018 (More conservative)	139.8	17.4	18.3
p-value	*0*	*0.03*	*0.05*

Notes: See Table 2.3 notes.

Table 2.7: How Middle Initials Could Reduce Errors in Linking in LIFE-M Data

	(1) Table 4 Match Rate	(2) Table 4 Type I Error Rate	(3) Share Matches with Middle Initials for Both Records	(4) Share of (3) with Discordant Middle Initials	(5) Type I Error Rate in (4)	(6) Revised Match Rate	(7) Revised Type I Error Rate
Ferrie 1996 (Name)	0.33	0.2	0.28	0.26	0.9	0.3	0.15
Ferrie 1996 (NYSIIS)	0.28	0.25	0.26	0.27	0.91	0.26	0.2
Ferrie 1996 (SDX)	0.2	0.32	0.24	0.3	0.93	0.19	0.27
Ferrie 1996 (Name) + common names	0.46	0.28	0.29	0.35	0.94	0.42	0.2
Ferrie 1996 (NYSIIS) + common names	0.46	0.35	0.27	0.37	0.95	0.41	0.28
Ferrie 1996 (SDX) + common names	0.41	0.43	0.26	0.41	0.96	0.37	0.36
Ferrie 1996 (Name) + common names + exact ties	0.69	0.5	0.3	0.44	0.97	0.6	0.43
Ferrie 1996 (NYSIIS) + common names + exact ties	0.79	0.58	0.29	0.5	0.98	0.68	0.52
Ferrie 1996 (SDX) + common names + exact ties	0.86	0.67	0.28	0.57	0.98	0.72	0.61

Notes: This table uses the LIFE-M data to evaluate changes in algorithm match rates and Type I error rates with the addition of middle initials. Column 7 computes match rates after dropping links with discordant middle initials. Column 8 computes revised Type I error rates by dropping links with discordant middle initials. See text for details. *Table continued in page below.*

Table 2.7: How Middle Initials Could Reduce Errors in Linking in LIFE-M Data (Continued)

	(1) Table 4 Match Rate	(2) Table 4 Type I Error Rate	(3) Share Matches with Middle Initials for Both Records	(4) Share of (3) with Discordant Middle Initials	(5) Type I Error Rate in (4)	(6) Revised Match Rate	(7) Revised Type I Error Rate
Abramitzky et al. 2014 (Name)	0.41	0.25	0.3	0.31	0.94	0.38	0.18
Abramitzky et al. 2014 (NYSIIS)	0.42	0.32	0.27	0.33	0.94	0.38	0.26
Abramitzky et al. 2014 (SDX)	0.39	0.41	0.27	0.39	0.96	0.35	0.35
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.23	0.26	0.26	0.89	0.23	0.18
Feigenbaum 2016 (Iowa)	0.52	0.34	0.31	0.23	0.93	0.48	0.3
Feigenbaum 2016 (LIFEM)	0.52	0.29	0.33	0.23	0.93	0.48	0.24
Abramitzky et al. 2018 (Less conservative)	0.46	0.37	0.29	0.34	0.95	0.41	0.3
Abramitzky et al. 2018 (More conservative)	0.28	0.15	0.29	0.2	0.87	0.26	0.11

Notes: *Continued from page above.* This table uses the LIFE-M data to evaluate changes in algorithm match rates and Type I error rates with the addition of middle initials. Column 7 computes match rates after dropping links with discordant middle initials. Column 8 computes revised Type I error rates by dropping links with discordant middle initials. See text for details.

Table 2.8: How Using Race Could Reduce Errors in Linking in LIFE-M Data

	(1) Table 4 Match Rate	(2) Table 4 Type I Error Rate	(3) Share Matches - 1940 Race Variables Different than LIFE-M	(4) Revised Match Rate	(5) Revised Type I Error Rate
Ferrie 1996 (Name)	0.33	0.2	0	0.33	0.2
Ferrie 1996 (NYSIIS)	0.28	0.25	0.01	0.28	0.25
Ferrie 1996 (SDX)	0.2	0.32	0.01	0.2	0.31
Ferrie 1996 (Name) + common names	0.46	0.28	0.01	0.46	0.27
Ferrie 1996 (NYSIIS) + common names	0.46	0.35	0.01	0.46	0.34
Ferrie 1996 (SDX) + common names	0.41	0.43	0.02	0.41	0.42
Ferrie 1996 (Name) + common names + exact ties	0.69	0.5	0.01	0.69	0.49
Ferrie 1996 (NYSIIS) + common names + exact	0.79	0.58	0.03	0.79	0.58
Ferrie 1996 (SDX) + common names + exact ties	0.86	0.67	0.06	0.86	0.66

Notes: This table uses the LIFE-M to evaluate changes in linking rates with the addition of race. Column 4 computes match rates after dropping links with discordant race. Column 5 computes revised Type I error rates by dropping links with discordant race. See text for details. *Table continued in page below.*

Table 2.8: How Using Race Could Reduce Errors in Linking in LIFE-M Data (Continued)

	(1)	(2)	(3)	(4)	(5)
	Table 4 Match Rate	Table 4 Type I Error Rate	Share Matches - 1940 Race Variables Different than LIFE-M	Revised Match Rate	Revised Type I Error Rate
Abramitzky et al. 2014 (Name)	0.41	0.25	0.01	0.41	0.25
Abramitzky et al. 2014 (NYSIIS)	0.42	0.32	0.01	0.42	0.32
Abramitzky et al. 2014 (SDX)	0.39	0.41	0.02	0.39	0.4
Abramitzky et al. 2014 (NYSIIS, Robustness)	0.24	0.23	0.01	0.24	0.23
Feigenbaum 2016 (Iowa)	0.52	0.34	0.01	0.52	0.34
Feigenbaum 2016 (LIFEM)	0.52	0.29	0	0.52	0.29
Abramitzky et al. 2018 (Less conservative)	0.46	0.37	0.02	0.46	0.36
Abramitzky et al. 2018 (More conservative)	0.28	0.15	0	0.28	0.15

Notes: *Continued from page above.* This table uses the LIFE-M to evaluate changes in linking rates with the addition of race. Column 4 computes match rates after dropping links with discordant race. Column 5 computes revised Type I error rates by dropping links with discordant race. See text for details.

CHAPTER III

Simple Strategies for Improving Inference with Linked Data: a Case Study of the 1850-1930 IPUMS Linked Representative Historical Samples

Until recently, the dearth of longitudinal or intergenerational U.S. data for the late 19th and 20th centuries limited the study of important social, economic, demographic, and health questions.[1] Much of the existing work on these questions has instead used cross-sectional or aggregated data - data that answer some questions but that often leave the mechanisms for both observed effects and policy generalizability unclear.[2]

Large-scale linked data are allowing researchers to break new ground on older questions and open entirely novel areas of inquiry.[3] New work, however, suggests

[1] This chapter was written with my coauthors Martha Bailey and Catherine Massey and published in *Historical Methods*. Appendicies referenced in this chapter have not been included in this dissertation for concision, and are available online at https://www.tandfonline.com/doi/suppl/10.1080/01615440.2019.1630343?scroll=top.

[2] See, for instance, early - life public health initiatives (Alsan & Goldin, 2015; Cutler & Miller, 2005), exposures to environmental pollutants (Clay, Lewis, & Severnini, 2016) and animal diseases (Rhode & Olmstead, 2015), and access to medicines (Bleakley, 2007). Other examples include the long-run effects of exposure to human capital initiatives through Rosenwald schools (Mazumder & Aaronson, 2011).

[3] On-going and proposed projects are linking national surveys, administrative data, and research samples to recently digitized historical records, such as the full-count 1880 (Ruggles, 2006; Ruggles, Genadek, Grover, & Sobek, 2015) and 1940 U.S. Censuses (the first U.S. census to ask about education and wage income) and newly available administrative sources. The Census Bureau plans to link the 1940 Census to current administrative and census data (Census Longitudinal Infrastructure Project, CLIP) and the Minnesota Population Center plans to link it to other historical censuses. The Panel Survey of Income Dynamics (PSID) and the Health and Retirement Survey (HRS) are

that the prevalence of false links and missed matches in historical U.S. linked data may limit the contributions of this research. Bailey, Cole, Henderson, and Massey (2019) show that commonly used methods consistently produce non-representative samples and high rates of false matches (or Type I errors), ranging from 15 to 37 percent, and higher rates of missed matches (or Type II errors), ranging from 63 to 79 percent, depending on the linking algorithm used. In addition, false matches do not occur at random; they are systematically predicted by baseline characteristics, suggesting that machine linking algorithms may introduce complicated forms of bias into analyses. To this point, Bailey et al.'s (2019) case study of linking birth certificates to the 1940 Census shows that - for the same set of records - prominent linking algorithms attenuate intergenerational income elasticity estimates by up to 20 percent. In that setting, Bailey et al. (2019) show that false links generate a critical part of this bias, and eliminating Type I errors from matches produces estimates that are indistinguishable from estimates of elasticities in data linked by hand.

This paper proposes two practical and complementary methods that aim to address these concerns and improve inference with linked data, regardless of the linking method used to create the data. First, we suggest using ”validation variables” - variables that include information on the likelihood that a link is correct and information that was not used in the original linking process. Validation variables can help identify subsets of lower quality links for greater scrutiny. Second, we recommend creating custom weights for linked samples to improve their representativeness. These weights mitigate the biases that arise from low linking rates (high Type II errors) as well as the biases introduced by restricting samples with validation variables. We demonstrate

linking their respondents to the 1940 Census. The Longitudinal, Intergenerational Family Electronic Micro-Database Project (LIFE-M) is linking vital records to the 1940 Census (Bailey, Anderson, Karimova, & Massey, 2016). Supplementing these public infrastructure projects, entrepreneurial researchers have also combined large datasets. See, for example, Abramitzky, Platt Boustan, and Eriksson (2012, 2013, 2014), Boustan, Kahn, and Rhode (2012), Hornbeck and Naidu (2014); Mill (2013); Mill and Stein (2016), Aizer, Eli, Ferrie, and Lleras-Muney (2016), Bleakley and Ferrie (2014; 2016; 2013), Nix and Qian (2015), Collins and Wanamaker (2016), and Eli, Salisbury, and Shertzer (2016).

how researchers can create these weights using inverse-propensity score reweighting. Although neither of these methods is new, they have rarely been applied individually or together in empirical papers using linked historical data.

This paper illustrates the value of these two strategies using the 1850-1930 Integrated Public Use Microdata Series Linked Representative Samples (IPUMS-LRS), a well-known and frequently used dataset in historical research. In section 3.1, we review the linking and weighting methodology used to create the IPUMS-LRS dataset, emphasizing the components of its construction that are relevant to our later analysis. In section 3.2, we demonstrate two examples of validation variables: name commonness (which can be used in almost all historical samples) and parent birthplace disagreement (which is specific to the IPUMS-LRS). Using a new hand-linked dataset, we show that both validation variables produce subsamples with fewer observations that human reviewers code as incorrect. In section 3.3, we show how generating custom weights can improve the representativeness of the IPUMS-LRS, even relative to the provided weights available in the linked data. In contexts where weights are not available, analyzing representativeness and generating custom weights are even more important. The value of these strategies for the IPUMS-LRS - a highly curated dataset - demonstrates their potential to improve research with other linked datasets.

3.1 A Brief Overview of the IPUMS-LRS

The IPUMS-LRS consist of roughly 500,000 individuals for seven pairs of years: 1850-1880, 1860-1880, 1870-1880, 1880-1900, 1880-1910, 1880-1920, and 1880-1930 (the 1890 Census was excluded, because most of the original manuscripts were destroyed in a fire). These samples were created by the Minnesota Population Center (MPC), which linked the full-count 1880 Census (which was digitized by the Church of Jesus Christ of Latter-Day Saints) to the one-percent samples of the 1850, 1920 and 1930 Censuses, the 1.2 percent samples of 1860, 1870 and 1900 Censuses, and the 1.4 percent sample of the 1910 Census (Ruggles, 2006). Our analysis focuses on linked men from these samples.

To link men from one Census to the 1880 Census, the MPC produced a cross product of individuals across the two Censuses (e.g. 1850 and 1880). Using the Freely Extensible Biomedical Record Linkage software (FEBRL), the MPC kept each potential match from the cross product if the two observations had names that met a string similarity threshold, shared the same birthplace (state or country), and had ages that fell in a specified window.[4] They then trained a support vector machine (SVM) classifier using a set of hand-matched Census data, and applied the SVM to the non-training data in the cross product. Using these results, they kept all potential matches that had a predicted match probability that exceeded a match "quality" threshold and dropped all matches that had multiple potential links to 1880.

The MPC used two strategies to create representative samples. First, like many modern linking projects, they linked observations using theoretically time-invariant characteristics such as name, age, and birthplace rather than characteristics like place of residence, occupation, and family structure that may change over time. The use of these time-invariant characteristics limits selection bias in creation of links (Ruggles, 2006). For instance, linking individuals by using information on state of residence could make the sample much less geographically mobile than the population of interest to researchers.

Second, because different population subgroups might have different likelihoods of being linked, the MPC created weights to balance the representation of observed characteristics for the "linkable" population. Linkable men are those who were alive in both years and resided in the U.S. and could, therefore, be enumerated by the Census in both years. To determine the population of linkable men, the MPC took the final

[4] FEBRL is a record linking software developed by the ANU Data Mining Group and the Centre for Epidemiology and Research in the New South Wales Department of Health. See Christen and Churches (2005) for more information.

year Census and dropped men younger than the gap in years between Censuses (e.g. for 1880 in the 1860-1880 data, they drop everyone 20 years and younger). Because Census data do not specify when foreign-born men immigrated to the U.S., the MPC estimated the share of these foreign-born men who were present in the first year using life tables.

The MPC created weights for the linkable population using an iterative process. To start, the MPC assigned each observation a weight that was the inverse match rate for the relevant birth and race group (with denominators described by the linkable population). Then, they applied these weights and calculated weighted inverse match rates for other covariates, including relationships to head of household, individual birthplaces, 5-year age groups, and categories for size of place and occupation. They used these new inverse match rates to iteratively alter the weights until arriving at a final weight.

The IPUMS-LRS weights were designed to allow researchers to adjust the characteristics of the linked sample to resemble a simple random sample from the linkable population and, therefore, make inferences about this population's characteristics. The MPC is careful to note the potential limitations of these weights, saying "researchers must decide whether the constructed weights are appropriate for their specific samples" (Goeken, Huynh, Lynch, & Vick, 2011).

3.2 Validation Variable as a Method to Improve Match Quality

The first method that we suggest for improving inference in historical linked data is to use one (or many) "validation variables." A validation variable is a variable that is correlated with whether a link is correct but was not used deterministically in the linking process. Consequently, a researcher can condition on a validation variable to obtain a subsample with a smaller Type I error rate. Additionally, researchers can use validation variables to examine the links where the validation variable fails (i.e., links that are expected to have a higher Type I error rate) to investigate the performance of their algorithm by applying more scrutiny to a subset of more questionable records.

To motivate the purpose and practice of validation variables, we first lay out some basic theory. Consider the full dataset of links observed, L_i, and let whether or not a given link is correct be described by the following function:

$$C_i = f(Y(X_i), X_i, Z_i) \tag{3.1}$$

where C_i is an indicator variable equal to 1 if the link is correct, $Y(X_i))$ is the impact of the linking algorithm, which considers the information in X_i, and lastly Z_i, or variables that were not included in the linking process. Note that X_i impacts C_i through the process of the linking function and independently of the linking function, A validation variable, V_i, is a variable that satisfies the following properties:

1). $cov(C_i, V_i, L_i = 1) > 0$

2). $Var(V_i|L_i = 1) \neq 0$

The first condition assures that the validation variable contains relevant information on whether the links are correct. The second condition ensures that the validation variable varies after conditioning on the observed links, which means that the validation variable is adding information beyond what is in the linking algorithm. If the validation variable agrees with all linking decisions, this condition will not be met. Note that a validation variable could be either a variable that was not included in the linking process (e.g. Z_i) or a variable that was included in the linking process but is used differently than it was in the linking process (e.g. X_i). Good validation variables may be more or less difficult to find depending on the linking setting, but our next section provides several examples hiding in plain sight.

3.2.1 Examples of Validation Variables

We use two different validation variables to demonstrate how these variables may reduce incorrect links: name commonness and disagreements in parents' place of birth. We chose these two variables because the first is available in almost all historical linking contexts, but the second is specific to the IPUMS-LRS. Here we describe these variables and offer intuition for why they might be effective as validation variables.

Our first example of a validation variable, name commonness, is a broadly applicable validation variable. Name commonness is available in many linking situations and is intuitively correlated with whether a link is correct. More common names, for example "John Smith," have more possible matches than less common names. Therefore, measurement error in other features (age or birthplace) may lead an algorithm to select an incorrect match more frequently for more common names. Observations with uncommon names, on the other hand, have fewer potential matches available, so measurement error in other linking variables are less likely to cause an algorithm to choose an incorrect link. Bailey et al. (2019) provide empirical support for this intuition and show that eliminating more common names from the linking process significantly reduces incorrect links, or Type I errors, in some algorithms.

Some papers use name commonness restrictions in the matching process or as a robustness check, implicitly treating it as a validation variable. Abramitzky et al. (2012, 2014) use such a strategy, verifying that their results from their main dataset hold for links that have name-birth place combinations that are unique in a two-year age band. For our exercise, we similarly create a validation variable equal to 1 if a name-birthplace combination has only one observation within a two-year band of the individual's name.[5] The validation variable would be equal to zero for very common names and equal to one for less common names. As an example, the validation variable for "John Smith" born in Ohio aged 30 in the 1880 Census would be equal to zero, if multiple "John Smiths" ages 28 to 32 born in Ohio appeared in the 1880 Census.

Our second validation variable, parent birthplace disagreement, is specific to the IPUMS-LRS. When matching the 1850, 1860 and 1870 Census samples to the 1880 full count Census, the MPC did not include parent birthplaces in the linking process.[6] If parent birthplaces are correctly recorded for an individual in the Census, they should be consistent over time. Although some parent birthplaces may be measured with error (Goeken, Lynch, Lee, Wellington, & Magnuson, 2017), limiting attention to matches that agree in parent birthplaces would intuitively tend to select matches that are more likely to be correct.[7]

3.2.2 Examining the Effectiveness of Validation Variables

Bailey et al. (2019) recommend that researchers create training data (hand-links) for some of their observations in order to document the performance of their algorithm and similarly defend their choice of validation variables. We follow this advice and link a subsample of the 1850-1880 IPUMS-LRS to directly examine the quality of

[5] We are performing this restriction on the data ex post as we only have access to the finished IPUMS-LRS matches. However, Abramitzky et al. (2012, 2014) as described in Bailey et al. (2019), perform this restriction before engaging their matching algorithm.

[6] The MPC did use parental birthplace when linking the 1900, 1910, 1920 and 1930 Census samples to the 1880 full count Census.

[7] Data quality issues prior to 1880 are the reason that the MPC did not use this variable in the matching process for 1850-1870. For these years, parent birthplaces can only be inferred from individuals living at home with their parents. Furthermore, relationships within a household in those years are not listed by Census takers, and need to be inferred from the order in which individuals are listed in the Census and the ages of individuals. In Online Appendix I, we demonstrate that, although parent birthplace is clearly measured with error, patterns of parental birthplace disagreement between individuals living at home with their parents and those not living at home are similar in the years after 1880. Therefore, assuming that the imputed household relationships are accurate in the years prior to 1880, this evidence suggests that parent birthplace disagreement patterns for children living at home might be similar to parent birthplace disagreements for people who are not living at home with their parents.

the IPUMS-LRS and the performance of our validation variables. To link these data, we randomly selected 653 IPUMS-LRS linked men who were aged 0 to 25 and living at home with their parents in 1850. An experienced group of genealogical linkers at the Family History and Technology Lab at Brigham Young University (BYU) then linked these observations by hand to the 1880 full count Census, without knowledge of the IPUMS-LRS links. The team at BYU used all the information available to the MPC and used additional information available to them through Ancestry.com and FamilySearch.org's databases. For the purpose of our exercise, we treat BYU's links as the truth and use these links to examine the performance of our validation variables.[8]

Table 3.1 summarizes the differences between the 1850-1880 IPUMS-LRS links and BYU's links.[9] The resulting share of links rejected by hand linkers is 10.0 percent, which is higher than the Type I error rate estimated by the MPC but is still low relative to machine-linked datasets analyzed in Bailey et al. (2019). Seventy percent of the differences come from cases where BYU determined that there was not enough data to reliably state a link. This outcome often occurred when a record had several possible matches, and genealogists were unsure about which possible match was correct. The remaining 30 percent of differences come from matches where BYU identified a link that disagreed with the IPUMS-LRS link.

Columns 3 and 4 of Table 3.1 examine the usefulness of our first validation variable, keeping only records that are unique for a given name, birthplace and age within a two-year band.[10] The first row under column 3 shows that 627 out of the 653 links considered by BYU make this cut using exact names. Given that many linking papers

[8] Online Appendix I provides more indirect evidence to demonstrate the relevance of parent birthplace disagreement as a validation variable without using hand-linked data.

[9] It is worth noting that hand-linked data are not "true" matches. Human error in matching may also produce false matches or fail to capture all 'true' matches. Given the dearth of longitudinal historical data, we have no direct test of the effectiveness of matching by hand.

[10] For completeness, we also considered other age bands, including a one-year and three-year age band in addition to the two-year age band in Table 3.1. The larger the band, the more observations tend to be dropped from consideration, but the Type I error rate also falls.

use phonetic cleaning to alter names for matching, columns 5 and 7 summarize the number of links that are unique in terms of NYSIIS or Soundex cleaned name and age combinations for the same age band.[11] The results show that requiring uniqueness of first and last name within the two-year age-radius lowers the rate of disagreement with hand linkers slightly, by 4 to 16 percent (0.4 to 1.6 percentage points on a base of 10.0 percentage points) depending on the name cleaning used. The drop in disagreements is likely small in part due to the fact that error rates are lower in the IPUMS-LRS data than in many other linked data. In other datasets, Bailey et al. (2019) show that a similar restriction in the Abramitzky et al. (2012, 2014) algorithm reduces rates of disagreement with hand linkers by as much as 10 percentage points.

Table 3.2 repeats this exercise using the validation variable for parent birthplace disagreement. As was the case for common names, genealogists are more likely to disagree with IPUMS-LRS links when parent birthplaces disagree. Dropping observations with a disagreement in father's birthplace drops the discrepancies with genealogists by 20 percent, a reduction of 2.0 points relative to a base of 10.0 percentage points. Dropping observations with a disagreement in mother's birthplace reduces disagreements by 18 percent, a reduction of 1.8 percentage points, and dropping observations with a disagreement in both mother and father birthplaces drops the error rate by 16 percent, a reduction of 1.6 percentage points.

If one takes records linked by genealogists as the truth, both sets of results suggest that conditioning on validation variables could reduce incorrect links. As a final test, we further probe the strength of the relationship between our validation variables and the determination by linkers that a link is incorrect. Specifically, we regress BYU's determination that a link is incorrect on our two validation variables as well as other data characteristics measuring a match's quality, including differences in age, own

[11] Researchers use name cleaning algorithms to adjust exact names for errors in transcription, recording and changes in phonetic spelling. For more background on these algorithms, see Bailey et al. (2019).

birthplace, and differences in recorded name using Jaro-Winkler similarity scores. This regression tests whether the validation variable contains information beyond that already present in these other features of the matches.

Table 3.3 shows the results from this regression using validation variables for name commonness and parent birthplace disagreement. Columns 1 and 4 show the unadjusted difference in error rates between observations that meet the validation variable and those that fail, demonstrating that the validation variables predict disagreements. Columns 2 and 5 show the correlation between the validation variables - after adjusting for the similarity of the individual's first and last name, difference in expected age, and own birthplace disagreement. Records with a higher similarity in first and last names or a smaller difference in expected age are negatively associated with BYU's determination that the link is incorrect, which is consistent with these record features partially determining matches. However, the inclusion of these covariates barely alters the partial correlation of the validation variables with link correctness. Similarly, the correlation between the validation variables with the likelihood of a link being judged incorrect by a reviewer is nearly unchanged by the inclusion in columns 3 and 6 of additional covariates, including indicator variables for living in an urban area, being in school, being born abroad, having a mother born abroad, having a father born abroad, residence on a farm, race, and Census region of residence. Across specifications, our validation variables remain a sizable and statistically significant predictor of the IPUMS-LRS link agreeing with hand-linked records.

Overall, our findings suggest the value of using a validation variable to diagnose and potentially increase link quality. Even though name commonness and discrepant parent birthplaces are noisy determinants of link quality, they appear to help diagnose errors and select higher quality links without having to examine the entirety of a dataset by hand. Here we have only considered two validation variables, and other validation variables may be more or less effective in other settings depending on the matching process that produced the linked data. When selecting and implementing validation variables, researchers should consider the strength of the correlation of a validation variable with whether links are incorrect, and the effect of restricting on a validation variable on missed matches, called Type II errors. For instance, imposing restrictions on name commonness using exact names produces a limited decrease in Type I errors, but match rates drop non-trivially, resulting in increases in Type II errors. This limited decrease in Type I error likely reflects the fact that the MPC considered some variation of name commonness in their linking. On the other hand, imposing restrictions on name commonness using NYSIIS- and Soundex-cleaned names produces a larger drop in Type I errors and also a larger increase in Type II errors, because these cleaned variables contain different information than that which was used in the algorithm. Thus, name commonness in our setting is more similar to an X_i variable, using the terminology of the linking example before: some part of this information was included in the MPC algorithm, but using a different part of the information still impacts incorrect link rates.

Parental birthplace was not explicitly used in the MPC's linking process for the 1850 Census data and is, therefore, more similar to the Z_i variable in our framework. We see a large drop in Type I errors from using this information as a validation variable, with the drop again potentially reflecting that the information from this validation variable was not captured by the other variables in X_i. Thus, selecting validation variables relies on knowledge of how the sample was initially constructed, and researchers will want to balance improvements in link quality from drops in Type I error rates against (sometimes) non-trivial increases in Type II errors.

3.3 Increasing the Representativeness of Linked Samples

Validation variables can help purge samples of lower quality links, but their effect on Type II errors raises concerns about sample representativeness. This concern motivates a second and complementary strategy for improving inference with linked samples: generating customized weights for the analytic sample. Generating custom weights may be important even in high quality linked data that contain weights (such as the IPUMS-LRS), as problems with representativeness may occur when researchers select certain subsamples for which weights do not balance covariates or because the relevant covariates were not used in the creation of weights. Consequently, weights may not create representative samples (Andrews & Oster, 2017; Angrist & Pischke, 2009; Caliendo & Kopeinig, 2008; Solon, Haider, & Wooldridge, 2015).

There are many ways to generate customized weights. Here, we document a simple, two-part procedure. First, we recommend that researchers document the degree to which their linked data are representative of the reference population using a regression test. Note, this investigation can be implemented in a manner similar to balance tests in randomized control trials (Duflo, Glennerster, & Kremer, 2007). Some papers currently do this check by reporting means of covariates of interest for the linked population and the reference population in the style of a covariate balance test. While this approach is valid, a regression provides a more concise joint test of representativeness. Second, we recommend that researchers construct and use custom weights using inverse propensity-score matching and report weighted results alongside unweighted results. While applying custom weighting may be especially important when using restrictions like validation variables, this strategy can also be used with nearly all historical linked data, as most historical linked samples have problems with non-representativeness.

Testing the representativeness of linked data requires establishing the relevant population for comparison - the reference population of interest. Consider a linking setting like IPUMS-LRS where links are between two Census years. The reference population would be the set of individuals who were alive and present in the U.S. in the earlier year and was still alive and present in the U.S. in the later year. That is, some of the observations present in the earlier year would not be linkable to the later year due to mortality and migration. Some of the observations in the later year would not be linkable to the earlier year if they had not been born yet, or if they had immigrated into the U.S. between the Censuses. Depending on the research questions, either year could be used for testing representativeness, so researchers would need to decide which is the relevant reference population for their analysis.

When testing representativeness in the IPUMS-LRS samples, we follow the MPC and identify the reference population as the individuals alive in the second Census: the 1880 full count Census for the 1850-1880, 1860-1880, and 1870-1880 samples. We examine the 1910 Census for the 1880-1910 sample, 1920 for the 1880-1920 sample, and 1930 for the 1880-1930 sample. Following the MPC, we identify as the reference population the potentially linkable individuals within this Census who would have been alive in the previous year by dropping all individuals who (given their reported age) would not have been alive in the earlier Census year (e.g. men younger than 30 in the 1880 Census in the 1850-1880 IPUMS-LRS). Unlike the MPC, we make two further restrictions on the sample of links to simplify our analysis. First, we drop from consideration all men born outside the U.S. The MPC included these individuals and created weights for them using life tables to account for the fact that some of the foreign-born men present in the later year may have immigrated into the U.S. between the two Census years. For simplicity, we avoid these adjustments by isolating attention to U.S. born men. Second, we drop all non-white men from our analyses. The MPC included these individuals, but given issues with counting African-American men in the 1850 and 1860 Censuses, we wanted to limit attention to men who could have been counted in the previous Census.[12] Thus, for our analysis, we restrict attention to matches within the population of white U.S.-born men present in the final year of the Census. Note that here we are not imposing any restrictions

[12] In 1850 and 1860, African-American slaves were enumerated separately under a slave schedule.

related to our validation variables - we are considering the representativeness of the IPUMS-LRS data overall.

3.3.1 A Simple Regression Test of Representativeness

Our representativeness test uses a simple regression method proposed in Bailey et al. (2019). Specifically, we recommend that researchers take the reference population data, create a dummy variable equal to 1 if an observation is linked, and then regress the dummy variable on a series of covariates describing the reference population. If using a linear probability model, we recommend researchers use Huber-White standard errors to account for the fact that errors of a linear probability model are heteroskedastic (Huber 1967, White 1980). Our representativeness test-statistic is a heteroscedasticity-robust Wald test of joint significance of the covariates. Under the null hypothesis of representativeness of the linked sample, there should be no relationship between the covariates and the likelihood an observation is linked.

The advantage of this test over variable-by-variable balance test of means is that it accounts for the correlations among covariates and the joint relationship of the group of covariates with the likelihood of being linked, aggregating all information in the relevant covariates into a single test statistic. Furthermore, the magnitudes of the regression coefficients conveniently quantify which characteristics are more or less likely to result in a linked observation after controlling for other record characteristics. Note, however, that this technique is only a diagnostic test of the null hypothesis of representativeness, and rejecting the null hypothesis is not an indication that inference estimates are necessarily biased for two reasons. First, statistical significance of differences in covariates does not imply scientific significance, as magnitude of the bias may be slight (McCloskey, 2005). Moreover, if the relationship of interest (e.g. job mobility) is homogeneous for all groups in the population, selecting a non-representative sample would not bias estimates.

Table 3.4 summarizes the results of the representativeness tests for all of the IPUMS-LRS samples. Since the MPC provides weights to adjust for the non-representativeness of linked data, we compare the sample characteristics using both unweighted and weighted data. The first two columns present the results of a regression of a binary dependent variable (=1 if the observation is linked) on a subset of the covariates that the MPC used to construct their weights. These include 11 binary variables for relationship of an individual to the head of household (e.g. spouse, child, etc.); eight binary variables for birthplaces by region (e.g. Northeast, Mid-Atlantic); and up to 14 binary variables for the size of the place the individual currently lives in (see table notes for details). For the unweighted results in column 1, the p-values show that the Wald test easily rejects the null-hypothesis of representativeness. After we apply IPUMS-LRS weights in column 2, we fail to reject representativeness at the 5-percent level in this subset of characteristics for three samples, which suggest the IPUMS-LRS weights largely work as intended. However, for the other four samples, applying the weights results in p-values that reject representativeness at conventional levels of significance.

Columns 4 and 5 consider the entirety of the covariates that the MPC used in their weighting procedure (all previous variables from columns 1 and 2 as well as binary variables for five-year age groups and four categories for occupations) both with and without weights. Unsurprisingly, we reject representativeness in the unweighted samples at the 1-percent level in all cases. After we apply IPUMS-LRS weights in column 5, we fail to reject representativeness for this full set of weighting covariates at the 5-percent level for the 1850-1880, 1870-1880, 1880-1910, and 1880-1920 samples.

Finally, columns 7 and 8 consider all variables that were used by the MPC to calculate weights and additional variables that were not. These additional variables include binary variables for whether or not a man lives with his parents, whether that man's parents were born in the U.S., the region of the country that man lives in, his marital status, farm status, the number of co-resident siblings, and an indicator variable for whether or not an individual lives in the same state as birth. In both weighted and unweighted samples across all years, the p-values show we reject representativeness at the 1-percent level for each sample. This result is less surprising, as the IPUMS-LRS weights might only be expected to achieve balance in covariates used to create these samples.[13] Similarly, in other settings, weights may not create representative samples for every research question or purpose and may not work well when isolating attention to specific subgroups (Caliendo and Kopeinig 2008, Angrist and Pischke 2009, Solon et al. 2015, Andrews and Oster 2017).

Looking beneath the test of statistical significance, this lack of representativeness may have consequences for inference. For brevity, Table 3.5 presents a subset of estimates for the 1860-1880 sample from the regressions underlying Table 3.4. We report the full set of regression results for all samples in Online Appendix IV for the interested reader. As a complement to these findings, Table 3.6 presents more standard mean comparisons for a subset of covariates in the 1860-1880 sample (the full set of mean comparisons for all samples are reported in Online Appendix III). The IPUMS-LRS weights improve representativeness with respect to some variables, especially those used in the construction of the weights, including age categories, size of place categories and current location of residence categories. As one might expect, however, the weights do little to balance the representation of characteristics that were not included in their construction. Moreover, some categories that were included in the weighting process remain unbalanced. For example, some IPUMS-LRS samples after applying weights over-represent heads of household while others underrepresent them. These patterns could be important for inference for a variety

[13] It is worth noting that these findings hold up in more traditional t-tests as well. Notably, we reject the null hypothesis of equality of means among the variables not included by the MPC roughly 63 percent of the time across all samples. See Online Appendix III for the full set of results. Note also that if the weights addressed all issues with representativeness of the data that there should not be these issues with other variables.

of research questions on family structure, particularly those relating to structure of intergenerational co-residing families (Ruggles, 2011).

In terms of migration and nativity outcomes, the weighted IPUMS-LRS often produce unrepresentative samples of Census region of residence and parental birthplaces. The weighted IPUMS-LRS samples over-represent individuals from the Northeast in five of six samples, including the 1860-1880 data reported in Table 3.4. All samples underrepresent U.S.-born children with foreign-born parents - a finding that could affect inferences about U.S. immigration from Asia (Hatton, 2011) and Europe (Abramitzky et al., 2012). Furthermore, all IPUMS-LRS samples, including the 1860-1880 sample shown in Tables 3.4 and 3.5, over-represent individuals living in the same state as where they were born. Living in the same state as birth increases the probability of being linked among U.S.-born white men by 4 to 6 percentage points across all samples after applying IPUMS-LRS weights. This suggests that the linked IPUMS-LRS sample appears less geographically mobile, which could affect inferences about intergenerational occupational mobility, occupation selection, and generational household structure.

Thus, overall, even in datasets like the IPUMS-LRS that have weights that work as intended for adjusting the covariates that were included in the weighting process, these weights may not be effective when considering different subsamples of the data, or other covariates that were not included in the weighting process. This lack of representativeness may create biases in inference from over or under-representation of specific groups if heterogeneous effects are present (Bailey et al. 2019).

3.3.2 Creating Weights Customized to a Sample or Question of Interest

If non-representativeness or imbalance in certain characteristics is a concern, researchers should report weighted results that adjust for that imbalance in addition to traditional unweighted estimates. If weights are not available, or the weights do not adjust sufficiently for non-representativeness, then researchers may construct their own using an application-specific inverse propensity (IP) score reweighting technique.

This approach requires that (1) the propensity of being linked is properly specified and can be consistently estimated (often described as unconfoundedness assumption) and that (2) the distribution of the propensity of being linked spans the same support as the reference population (often described as a common support assumption). It is impossible to test assumption (1) directly and it could be violated in a linking situation where the probability of being linked depends on unobservable features of an observation that are correlated with the variables included in the weight estimation process. However, theory can guide the selection of variables for (1). Assumption (2), on the other hand, can be tested directly by examining the estimated link propensities of linked records and the reference population.

This method can be implemented using the following steps:

1). Append the data for the linked sample to the population which the researcher wants the reweighted sample to represent.

2). Create a dependent variable, L_i, equal to 1 for each observation, i, in the linked sample and 0 for each observation in the reference population. Using this dependent variable, estimate a probit model on covariates of record characteristics, X_i (for instance, the variables used in columns 7-9 in Table 3.4).

3). Using the results from the probit, predict the conditional probability of being linked, $P(L_i = 1|X_i)$, for each observation.

4). To check assumption (2) regarding common support, plot the probabilities of being linked for the linked and unlinked observations. The overlap in the two distributions provides information on which individuals can be compared. Also, Crump, Hotz, Imbens, and Mitnik (2009) recommend trimming extreme probabilities, which is another easy-to-implement strategy for improving inference.

5). Using the predicted probabilities, researchers may calculate weights as $W_i = (1 - P(L_i = 1|X_i))/(P(L_i = 1|X_i)) * q/(1-q)$, where q is the share of records that are linked. If a certain set of characteristics is underrepresented in the linked sample relative to the population of interest, this weight will increase the influence of this particular observation. The second component normalizes these probabilities to fit the size of the linked and unlinked samples.

We implement this procedure for each sample using the covariates in columns 7-9 in Table 3.4 and find evidence that the common support assumption holds. Intuitively, the common support assumption requires that there is sufficient overlap in the characteristics of links and the reference population, as summarized by the propensity score, so that the former can be reweighted to look like the latter.

Applying these weights to the IPUMS-LRS samples makes a meaningful difference in our representativeness calculations. Although only a handful of means in Online Appendix III remain statistically significant after reweighting, column 3 of Table 3.4 shows that coefficient estimates from the regression are very close to zero for a large number of covariates in the 1860-1880 IPUMS-LRS. This finding is substantively different from the unweighted (column 1) and IPUMS-LRS weighted results (column 2). Moreover, columns 3, 6 and 9 of Table 3.4 show that we fail to reject representativeness for all of the IPUMS-LRS samples (p-values very close to one) after applying IP-weights for these covariates of interest. Of course, if we omit certain variables when constructing the IP-weights and then test for representativeness in these same variables after applying IP-weights, we also tend to reject representativeness, just as we did when considering the MPC's weights with variables they had not included in the reweighting process. It is important, therefore, that researchers specify the propensity score equation in step 3 with covariates to achieve balance in characteristics relevant for answering a particular research question.

Although we have only been considering the overall representativeness of the

IPUMS-LRS data, we find the same results regarding lack of representativeness of linked data and effectiveness of IP weights after imposing restrictions using our two validation variables. We omit those results here for brevity.

Lastly, it is important to note that, even though this reweighting procedure produces a sample very similar in observed characteristics, the resulting data may still be unbalanced in terms of unobserved characteristics, and reweighting will only accurately address bias from non-representativeness if the unconfoundedness assumption described earlier holds. That is, reweighting's effectiveness ultimately depends on the assumptions specified earlier, although the hope is that reweighting at least mitigates the problem of non-representativeness of linked data (DiNardo, Fortin, & Lemieux, 1996; Heckman, 1979).

3.4 Recommendations and Conclusions

Many important questions relate to how individuals, families, and communities changed over time, and new linked samples are critical in facilitating new research on these questions. As documented in Bailey et al. (2019), measurement error induced by linking algorithms may have substantial implications for inference. In light of this evidence, this paper suggests two complementary strategies to improve inference with linked samples.

First, we recommend using a validation variable that is correlated with link quality and not deterministically used in the linking process in order to improve inferences. These two conditions imply that the validation variable will contain additional information about link quality. These variables allow researchers to perform robustness tests by purging links more likely to be incorrect from their analysis samples without the high cost of hand linkage. For our case study using the 1850-1880 IPUMS-LRS, we use name uniqueness and parental birthplaces to identify a set of links more likely to be correct. Although both of these variables are noisy indicators of linking errors, regression evidence demonstrates that name commonness and discordance in both parents' birthplaces are nevertheless powerful predictors of incorrect links - even in a high quality sample like the IPUMS-LRS. Purging samples of links with common names reduces the error rate in the pre-1880 IPUMS by up to 15 percent, and dropping observations with discordant parent birthplaces, reduces the error rate by up to 20 percent. We have only examined two examples of variables but other contexts may lead to other potential validation variables.

Limiting samples by purging potentially false links may also increase problems with non-representativeness, an issue with almost all linked data. This problem leads us to suggest a second, complementary strategy for improving inferences with linked records. Like many surveys and historical samples, the IPUMS-LRS (even with weights) are not generally representative of the reference population of potentially linkable individuals. However, applications of inverse probability weighting can substantially improve representativeness. To this end, we describe a simple inverse propensity score reweighting approach similar to that proposed by DiNardo et al. (1996) and demonstrate its effectiveness for the IPUMS-LRS. This method is easily adaptable to various applications and will generally produce representative samples catered to specific research objectives under the assumptions we specify. A close examination of the value of these weights also informs researchers about where more time-intensive genealogical or clerical review methods may increase the representation of hard-to-link groups. Used in combination with validation variables, custom reweighting may help improve inference with linked data.

3.5 Figures and Tables

Table 3.1: Name uniqueness in 1850-1880 IPUMS-LRS and linking errors from comparison to genealogically linked sample

	All IPUMS Observations		Uniqueness in Exact Name in Two Year Radius		Uniqueness in NYSIIS Name in Two Year Radius		Uniqueness in Soundex Name in Two Year Radius	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Total IPUMS Observations	653	100.00%	627	100.00%	573	100.00%	489	100.00%
Total IPUMS-LRS Correct	588	90.05%	567	90.43%	525	91.62%	446	91.21%
Total IPUMS-LRS Incorrect	65	9.95%	60	9.57%	48	8.38%	43	8.79%
A) Matched by BYU	19	2.91%	16	2.55%	13	2.27%	12	2.45%
B) Not Matched by BYU	46	7.04%	44	7.02%	35	6.11%	31	6.34%

Notes: This table uses a hand-linked sample of the 1850-1880 censuses produced by the BYU Family History and Technology Lab. When IPUMS-LRS agrees with BYU, we call the link "correct." When IPUMS-LRS differs from BYU, we call the link "incorrect." Incorrect links can be further divided into links where both IPUMS-LRS and BYU link an observation but they choose different links, and links where IPUMS-LRS linked an observation but BYU did not.

Table 3.2: Parent birthplace disagreements in 1850-1880 IPUMS-LRS and linking errors from comparison to genealogically linked sample

	All IPUMS Observations		Observations without Father Birthplace Disagreement		Observations without Mother Birthplace Disagreement		Observations without Father and Mother Birthplace Disagreement	
	Count	Percent	Count	Percent	Count	Percent	Count	Percent
Total IPUMS Observations	653	100.00%	514	100.00%	475	100.00%	547	100.00%
Total IPUMS-LRS Correct	588	90.05%	473	92.02%	436	91.79%	501	91.59%
Total IPUMS-LRS Incorrect	65	9.95%	41	7.98%	39	8.21%	46	8.41%
A) Matched by BYU	19	2.91%	13	2.53%	13	2.74%	15	2.74%
B) Not Matched by BYU	46	7.04%	28	5.45%	26	5.47%	31	5.67%

Notes: This table uses a hand-linked sample of the 1850-1880 censuses produced by the BYU Family History and Technology Lab. When IPUMS-LRS agrees with BYU, we call the link ”correct.” When IPUMS-LRS differs from BYU, we call the link ”incorrect.” Incorrect links can be further divided into links where both IPUMS-LRS and BYU link an observation but they choose different links, and links where IPUMS-LRS linked an observation but BYU did not.

Table 3.3: Regression-adjusted measurement of validation variable correlation with linking errors in 1850-1880 IPUMS-LRS

Covariates	Dependent Variable: 1=Incorrect Link					
	(1)	(2)	(3)	(4)	(5)	(6)
Uniqueness of NYSIIS Name and	-0.13**	-0.12**	-0.11**			
Birthplace in 2-Year Age Radius	-0.047	-0.046	-0.048			
No Disagrement in Both Father				-0.13**	-0.13**	-0.12**
and Mother Birthplaces				-0.05	-0.051	-0.051
Jaro Winkler - Own Last Name		-1.08*	-1.04*		-1.24**	-1.09*
		-0.581	-0.596		-0.588	-0.59
Jaro Winkler - Own First Name		-0.23*	-0.21		-0.27**	-0.24*
		-0.127	-0.137		-0.13	-0.137
Difference in Expected Age		0.06**	0.06**		0.06**	0.06**
		-0.024	-0.025		-0.025	-0.025
Own Birthplaces Disagree		0.14	0.21		0.14	0.22
		-0.226	-0.229		-0.228	-0.23
Constant	0.21***	1.46**	1.18*	0.21***	1.66**	1.09*
	-0.046	-0.6	-0.625	-0.049	-0.628	-0.615
Additional covariates	N	N	Y	N	N	Y
R-squared	0.006	0.01	0.022	0.011	0.051	0.088

Notes: The regression results are obtained from regressing a binary dependent variable (=1 if a link is incorrect, 0 if the link is correct) on the indicated covariates in the male 1850-1880 Census links from the IPUMS-LRS. Additional covariates are all measured in the 1850 Census, and include an indicator variable for living in an urban area, being in school, being white, being born abroad, mother born abroad, father born abroad, farm status, and region fixed effects. In columns 1-3, regressions use 653 observations, and regressions in columns 4-6 use 618 observations; the difference in the number of observations in regressions reflects the fact that columns 4-6 require that a man be living at home with both parents. Heterskedasticity-robust standard errors are reported beneath each estimate, and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***).

Table 3.4: Regression test of representativeness of IPUMS-LRS

	Restricted IPUMS Covariates			All IPUMS Covariates			Al IPUMS Covariates and Other Covariates		
Years Matched	Unweighted	IPUMS Weighted	BCM Weighted	Unweighted	IPUMS Weighted	BCM Weighted	Unweighted	IPUMS Weighted	BCM Weighted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1850-1880	1287	13678	5	1529	15043	14	1714	1094	20
	0	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*
1860-1880	1819	92	8	2149	100	19	2503	332	24
	0	*0*	*1*	*0*	*0*	*1*	*0*	*0*	*1*
1870-1880	3122	60	7	3560	71	22	4301	540	33
	0	*0*	*1*	*0*	*0.03*	*1*	*0*	*0*	*1*
1880-1900	1970	25	2	2603	79	9	3275	355	13
	0	*0.73*	*1*	*0*	*0*	*1*	*0*	*0*	*1*
Covariates	C,B,H	C,B,H	C,B,H	C,B,H,O	C,B,H,O	C,B,H,O	C,B,H,O	C,B,H,O	C,B,H,O
Included				A	A	A	A,R,C,X	A,R,C,X	A,R,C,X

Notes: Each estimate is a heteroscedasticity-robust Wald-test from a separate regression of a binary dependent variable (=1 for linked record) for samples described in the text. the associated p-values is printed beneath in italics. ”IP Weights” refers to inverse-propensity score weighted estimates. Covariate abbreviations are as follows. C denotes dummy variables for size of local city (under 1,000 or unincorporated; 1,000 to 2,499; 2,500 to 3,999; 4,000 to 4,999; 5,000 to 9,999; 10,000 to 24,999; 25,000 to 49,999; 50,000 to 74,999; 75,000 to 99,999, 100,000 to 199,999; 200,000 to 299,999; 300,000 to 499,999; 500,000 to 599,999; 600,000 to 749,999; 750,000 to 999,999; 1 million to 1.99 million and 2 million and up). B denotes dummy variables for birth location (Northeast, Mid-Atlantic Region, East North Central Region, West North Central Region, South Atlantic Region, East South Central Region, West South Central Region, Mountain Region, and born outside U.S.). *Notes and table continued in page below.*

Table 3.4: Regression test of representativeness of IPUMS-LRS (Continued)

	Restricted IPUMS Covariates			All IPUMS Covariates			Al IPUMS Covariates and Other Covariates		
Years Matched	Unweighted	IPUMS Weighted	BCM Weighted	Unweighted	IPUMS Weighted	BCM Weighted	Unweighted	IPUMS Weighted	BCM Weighted
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1880-1910	1512	46	18	1939	52	20	2355	293	25
	0	*0.03*	*0.95*	*0*	*0.26*	*1*	*0*	*0*	*1*
1880-1920	961	44	23	1190	53	27	1390	204	29
	0	*0.05*	*0.8*	*0*	*0.17*	*0.98*	*0*	*0*	*1*
1880-1930	772	130	18	937	335	18	1070	455	18
	0	*0*	*0.96*	*0*	*0*	*1*	*0*	*0*	*1*
Covariates Included	C,B,H	C,B,H	C,B,H	C,B,H,O A	C,B,H,O A	C,B,H,O A	C,B,H,O A,R,C,X	C,B,H,O A,R,C,X	C,B,H,O A,R,C,X

Notes: *Continued from page above.* H denotes dummy variables for relationship of individual to household head (head/householder, spouse, child, child-in-law, parent, parent-in-law, sibling, sibling-in-law, grandchild, other relatives, parent friend or visitor). O denotes dummy variables for occupation (white collar occupation, farming occupation, semi-skilled occupation, unskilled occupation), and A denotes age category variables (dummy variables for five-year categories of ages). R denotes dummies for region of residence (Northeast, Midwest, West), E is a set of dummy variables for whether an individual lives with his mother, lives with his father, or lives with both parents. X is a set of dummy variables for whether an individual's father was born abroad, mother was born abroad, marital status, or farm status and whether they were living in the same state as birth. It also includes number of siblings in the household. Weights for BCM are described in text.

Table 3.5: Regression estimates of representativeness of 1860-1880 IPUMS-LRS

	Unweighted	IPUMS	Unweighted	IPUMS	Unweighted	IPUMS	IP
		Weighted		Weighted		Weighted	Weighted
Covariates	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Born in Northeast	0.05***	-0.05	0.04**	-0.06	0.01	-0.10**	-0.01
	-0.016	-0.045	-0.016	-0.045	-0.016	-0.045	-0.043
Born in Mid-Atlantic	-0.06***	-0.05	-0.06***	-0.05	-0.08***	-0.09**	-0.01
Region	-0.015	-0.044	-0.015	-0.044	-0.016	-0.043	-0.042
Born in East North	-0.04**	-0.04	-0.04**	-0.05	-0.05***	-0.06	-0.01
Central Region	-0.015	-0.043	-0.016	-0.044	-0.016	-0.042	-0.04
Born in West North	-0.02	-0.04	-0.02	-0.04	-0.03**	-0.05	-0.01
Central Region	-0.016	-0.044	-0.016	-0.044	-0.016	-0.042	-0.041
Born in South Atlantic	-0.02	-0.07	-0.02	-0.08*	-0.03*	-0.08*	-0.01
Region	-0.016	-0.044	-0.016	-0.045	-0.016	-0.044	-0.043
Born in East South	-0.05***	-0.06	-0.05***	-0.07	-0.06***	-0.07*	-0.01
Central Region	-0.016	-0.044	-0.016	-0.045	-0.016	-0.044	-0.042
Born in West South	-0.01	-0.03	-0.01	-0.04	-0.02	-0.04	-0.01
Central Region	-0.017	-0.047	-0.017	-0.047	-0.018	-0.046	-0.044
Born in Mountain	-0.06***	-0.03	-0.05**	-0.03	-0.05**	-0.02	-0.02
Region	-0.021	-0.091	-0.021	-0.092	-0.021	-0.094	-0.089
Relationship to head:	0.05***	-0.02	0.04***	-0.03***	0.03***	-0.06***	0
head/householder	-0.003	-0.012	-0.003	-0.013	-0.004	-0.015	-0.015

Notes: The regression results are obtained from regressing a binary dependent variable (=1 if a record is linked, 0 if in the linkable population) on the indicated covariates (N = 8,673,750). Standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). *Table continued in page below.*

Table 3.5: Regression estimates of representativeness of 1860-1880 IPUMS-LRS (Continued)

	Unweighted	IPUMS	Unweighted	IPUMS	Unweighted	IPUMS	IP
		Weighted		Weighted		Weighted	Weighted
Covariates	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Relationship to head:	-0.01	-0.32***	-0.02	-0.33***	-0.03	-0.35***	-0.05
spouse	-0.026	-0.092	-0.025	-0.091	-0.025	-0.09	-0.147
Relationship to head:	0.04***	0.02	0.05***	0.02	0.03***	-0.02	0.01
child	-0.003	-0.014	-0.003	-0.014	-0.007	-0.024	-0.024
Relationship to head:	-0.01	-0.32***	-0.02	-0.33***	-0.03	-0.35***	-0.05
spouse	-0.026	-0.092	-0.025	-0.091	-0.025	-0.09	-0.147
Relationship to head:	0.04***	0.02	0.05***	0.02	0.03***	-0.02	0.01
child	-0.003	-0.014	-0.003	-0.014	-0.007	-0.024	-0.024
Relationship to head:	0.02**	-0.06	0.02*	-0.06*	0.01	-0.09**	0.01
child-in-law	-0.01	-0.036	-0.01	-0.037	-0.011	-0.038	-0.038
Relationship to head:	0.09***	-0.04	0.01	-0.06	0.01	-0.07**	0.01
parent	-0.014	-0.033	-0.015	-0.035	-0.015	-0.035	-0.034
Relationship to head:	0.08***	-0.04	0.01	-0.06	0	-0.08*	0.01
parent-in-law	-0.018	-0.042	-0.018	-0.043	-0.018	-0.043	-0.044
Relationship to head:	0.02***	-0.06**	0.02***	-0.07**	0.02**	-0.07***	0.01
sibling	-0.007	-0.027	-0.007	-0.027	-0.008	-0.027	-0.027
Relationship to head:	0.02*	-0.08*	0.02*	-0.08*	0.02	-0.09**	0
sibling-in-law	-0.011	-0.041	-0.011	-0.041	-0.011	-0.041	-0.041

Notes: *Continued from page above.* The regression results are obtained from regressing a binary dependent variable (=1 if a record is linked, 0 if in the linkable population) on the indicated covariates (N = 8,673,750). Standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). *Table continued in page below.*

Table 3.5: Regression estimates of representativeness of 1860-1880 IPUMS-LRS (Continued)

	Unweighted	IPUMS	Unweighted	IPUMS	Unweighted	IPUMS	IP
		Weighted		Weighted		Weighted	Weighted
Covariates	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Lives with mother					0	0	0
					-0.006	-0.018	-0.018
Lives with father					0	0.03	0.01
					-0.009	-0.027	-0.027
Lives with both parents					0.02	0.03	-0.02
					-0.009	-0.028	-0.028
Father: born abroad					-0.02***	-0.05***	0
					-0.005	-0.018	-0.018
Mother: born abroad					-0.02***	-0.08***	0
					-0.005	-0.019	-0.019
Lives in Northeast					0.02***	0.05**	0
					-0.005	-0.023	-0.023
Lives in Midwest					0.01***	0.02	0
					-0.004	-0.019	-0.02
Lives in West					0	0	0
					-0.006	-0.036	-0.037
Number of siblings					0	0	0
					-0.001	-0.003	-0.003

Notes: *Continued from page above.* The regression results are obtained from regressing a binary dependent variable (=1 if a record is linked, 0 if in the linkable population) on the indicated covariates (N = 8,673,750). Standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). *Table continued in page below.*

Table 3.5: Regression estimates of representativeness of 1860-1880 IPUMS-LRS (Continued)

	Unweighted	IPUMS	Unweighted	IPUMS	Unweighted	IPUMS	IP
		Weighted		Weighted		Weighted	Weighted
Covariates	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Lives in same state					0.02***	0.05***	0
as birth					-0.002	-0.01	-0.01
Constant	0.08***	0.56***	0.18***	0.51***	0.18***	0.51***	0.46***
	-0.019	-0.053	-0.053	-0.102	-0.053	-0.11	-0.112
Observations	97123	97123	97123	97123	97123	97123	97123
R-squared	0.017	0.002	0.022	0.003	0.025	0.012	0
Wald Statistic	1631	59.2	2029	67.7	2400	255	5.1
Prob ≥ F	0	0	0	0.03	0	0	1

Notes: *Continued from page above.* The regression results are obtained from regressing a binary dependent variable (=1 if a record is linked, 0 if in the linkable population) on the indicated covariates (N = 8,673,750). Standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***).

Table 3.6: T-Tests of means in the 1860-1880 IPUMS-LRS and the linkable population

	Unweighted	IPUMS Weighted	IP Weighted
Variables			
Age	3.787***	-0.088	0.161
	-0.173	-0.166	-0.164
Born in Northeast	0.107***	0.001	0.001
	-0.004	-0.007	-0.007
Born in Mid-Atlantic Region	-0.093***	0.002	-0.005
	-0.004	-0.008	-0.008
Born in East North Central Region	-0.021***	0.008	0.001
	-0.004	-0.007	-0.007
Born in West North Central Region	0	0.003	0
	-0.002	-0.003	-0.002
Born in South Atlantic Region	0.024***	-0.012*	0.002
	-0.004	-0.006	-0.006
Born in East South Central Region	-0.018***	-0.006	0
	-0.003	-0.005	-0.005
Born in West South Central Region	0.003*	0.002	0
	-0.002	-0.002	-0.002
Born in Mountain Region	-0.001*	0	0
	-0.001	0	-0.001
Born in Pacific Region	0	0.001	0
	-0.001	-0.001	-0.001
Relationship to head: head/householder	0.066***	-0.021***	0.001
	-0.005	-0.006	-0.006
Relationship to head: spouse	-0.000**	-0.001***	0
	0	0	0
Relationship to head: child	-0.003	0.023***	0
	-0.004	-0.005	-0.005
Relationship to head: child-in-law	-0.002**	-0.001	0
	-0.001	-0.001	-0.001
Relationship to head: parent	0.004***	-0.001	0
	-0.001	-0.001	-0.001
Relationship to head: parent-in-law	0.002**	0	0
	-0.001	-0.001	-0.001

Notes: A selected set of mean comparisons shows the difference between the means of the linked IPUMS-LRS and the linkable population without IPUMS-LRS weights in column (1); standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). Columns (2) and (3) present the same statistics using IPUMS-LRS and IP weights. The Online Appendix presents the full set of mean comparisons for 1860-1880 and all other IPUMS-LRS years. *Table continued in page below.*

Table 3.6: T-Tests of means in the 1860-1880 IPUMS-LRS and the linkable population (Continued)

Variables	Unweighted	IPUMS Weighted	IP Weighted
Relationship to head: sibling	-0.003***	-0.003**	0
	-0.001	-0.001	-0.002
Relationship to head: sibling-in-law	-0.002**	-0.002*	0
	-0.001	-0.001	-0.001
Relationship to head: grandchild	0	0	0
	0	-0.001	0
Relation to head: other	-0.061***	0.006	-0.002
	-0.003	-0.006	-0.005
In white collar occupation	0.013***	0	0
	-0.004	-0.005	-0.005
In farming occupation	0.063***	-0.001	0.002
	-0.005	-0.007	-0.007
In semi-skilled occupation	-0.031***	0.001	0
	-0.004	-0.006	-0.006
In unskilled occupation	-0.053***	-0.003	-0.002
	-0.004	-0.006	-0.005
In other or N/A occupation	0.009***	0.004	0
	-0.003	-0.003	-0.003
Lives with mother	-0.002	0.022***	-0.001
	-0.004	-0.005	-0.005
Lives with father	0.005	0.027***	0
	-0.004	-0.005	-0.004
Lives with both parents	0.006	0.025***	0
	-0.004	-0.005	-0.004
Father: born abroad	-0.063***	-0.040***	-0.001
	-0.003	-0.005	-0.006
Mother: born abroad	-0.061***	-0.041***	-0.001
	-0.003	-0.005	-0.006
Lives in Northeast	0.042***	0.023**	-0.004
	-0.005	-0.011	-0.011
Lives in Midwest	-0.030***	-0.004	0
	-0.005	-0.011	-0.011

Notes: *Continued from page above.* A selected set of mean comparisons shows the difference between the means of the linked IPUMS-LRS and the linkable population without IPUMS-LRS weights in column (1); standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). Columns (2) and (3) present the same statistics using IPUMS-LRS and IP weights. The Online Appendix presents the full set of mean comparisons for 1860-1880 and all other IPUMS-LRS years. *Table continued in page below.*

Table 3.6: T-Tests of means in the 1860-1880 IPUMS-LRS and the linkable population (Continued)

Variables	Unweighted	IPUMS Weighted	IP Weighted
Lives in West	-0.014***	-0.005**	0.003
	-0.002	-0.005	-0.005
Lives in South	0.002	-0.014	0.002
	-0.005	-0.01	-0.01
Currently married	0.058***	-0.008	0.002
	-0.005	-0.006	-0.006
Farm status	0.054***	0.01	0.003
	-0.005	-0.008	-0.008
Number of siblings in household	-0.031**	0.052***	-0.001
	-0.015	-0.02	-0.018
Living in same state as birth	0.044***	0.050***	-0.003
	-0.005	-0.008	-0.009

Notes: *Continued from page above.* A selected set of mean comparisons shows the difference between the means of the linked IPUMS-LRS and the linkable population without IPUMS-LRS weights in column (1); standard errors are reported beneath and stars indicate conventional levels of statistical significance, e.g., 10-percent (*), 5-percent (**), and 1-percent (***). Columns (2) and (3) present the same statistics using IPUMS-LRS and IP weights. The Online Appendix presents the full set of mean comparisons for 1860-1880 and all other IPUMS-LRS years.

APPENDICES

APPENDIX A

Additional Detail on Variables and Data

This paper uses the 2000 long-form Census and the 2001-2016 ACS to estimate causal regression discontinuities. These estimates identify the effect of an increase in family income from being born before the New Year on later-life outcomes for children. It also uses the CPS to estimate the size of the family's discontinuity in after-tax income from having a child born before the New Year. This appendix discusses data quality issues associated with these two data sources sequentially.

Assigning Grade-for-Age Status in the 2000 Census and 2001-2016 ACS

As described in the text, this paper assigns grade-for-age status to students based on four pieces of information: a child's highest grade completed or current grade enrolled, the state of birth of the child, the year and date of birth of the child, and the day on which households respond to the survey. Many states set explicit Kindergarten and 1st grade age entrance requirements that require students to be a specific age by a certain date before being eligible to enter either Kindergarten or 1st grade in that state. Comprehensive data on these state policies were collected by Bedard and Dhuey (2012) and they generously provided their most recent data covering 1955 to 2015. Using this data, this paper assigns expected completed grades to students assuming that they entered Kindergarten or first grade in the first year that they were eligible for those grades and then progressed through all other grades sequentially without repeating a grade. A student is grade-for-age for the purposes of this research if they have completed the most recent grade that this measure records a student as having completed.[1]

Four complications are worth noting about this measure. First, some states do not specify statewide Kindergarten entrance rules and allow local school districts to specify their own rules. No clear expected grade can be assigned to these individuals without more detailed data on individual school district practices. Consequently, this paper drops any individuals born in these states from any further calculation involving either outcomes for children or outcomes for adults.

Second, some states make the eligibility cutoff January 1st or December 31st. In the years that such cutoffs are present, children born before and after the New Year would, in addition to the treatment described, also experience the treatment of different grade eligibility rules. This paper also drops these individuals from any further calculation.

Third, there are only a handful of grades where grade-for-age status can be reliably assigned due to the nature of the grade attainment and enrollment questions in the 2000 Census and 2001-2007 ACS. The 2008-2016 ACS allow respondents to mark grade completion and grade attendance in all primary and secondary grades. However, the 2000 Census and 2001-2007 ACS only allow respondents to list whether their children have completed Nursery School/Preschool through 4th grade, 5th grade through 6th grade, 7th grade through 8th grade, and 9th, 10th, 11th and 12th grades. These same surveys only allow respondents to list whether their children have recently attended Nursery School/Preschool, Kindergarten, 1st through 4th grade, 5th grade through 8th grade, and 9th grade through 12th grade. Therefore, the best grades to

[1] As noted in the paper, most school systems define grade-for-age status starting from the first year a child enters Kindergarten or 1st grade. As these entrance dates are not observable in Census data, this definition is the closest analogue.

measure grade-for-age status would be grades where students would be expected to have completed or be currently attending a grade where the student's family could have listed completion or attendance of a prior grade. These grades would be pre-Kindergarten, Kindergarten, 1st, 5th, 7th, 9th, 10th and 11th grades. To see why, for example, 6th grade cannot be included, note that whether or not a student has completed 5th or 6th grade cannot be distinguished from that student's information in the 2000 Census and the 2001-2007 ACS. Note that the recent grade completed question can be used to determine grade-for-age status for 5th, 7th, 9th, 10th and 11th grades. The recent grade enrolled question can be used to calculate enrollment status for pre-Kindergarten and Kindergarten, and grade-for-age status in 1st grade.

Fourth, the response day of a household will affect the most recent grade a student may have completed or attended. In both the Census and the ACS, the education attainment question asks for the highest grade completed by a respondent and most recent grade enrolled. Thus, the date of response to an individual survey matters for determining the most recent grade a student has completed or recently attended.

The effect of date of response differs between the most recent grade enrolled and the most recent grade completed questions. Consider first how date of response will affect completed grades, which are used to calculate grade-for-age status in 5th grade and up. Suppose a student is in 5th grade in March 2001. If that family were responding to the ACS in that month, that family would list that student as having completed the fourth grade. However, suppose the student progressed to the next grade, the school year ended in May, and the family responded to the ACS in June. Then, that family would list that student as having completed the 5th grade. To account for this issue, this paper assumes that households responding to surveys between January 1st and April 10th will still have their children enrolled in the grade that they would have enrolled in at the beginning of the school year. Thus, these children will be recorded as having finished the previous grade they completed before enrolling in their current grade. This paper also assumes that households that respond to surveys between July 1st and December 31st will either have completed the previous grade (if the student passed and is grade-for-age) or will only have completed the grade before that (if the student was retained and is not grade-for age). As grade-for-age status cannot be ascertained reliably for the intervening months, this paper drops individuals who respond in those months from consideration for all calculations.[2] To ensure that post-schooling outcomes look at similarly structured cohorts as well, this paper also omits responses from these months when looking at outcomes for adults.

Date of response affects the ways families answer the question regarding the most recent grade enrolled in a slightly different manner. The most recent grade enrolled question is used to calculate enrollment status for students in pre-Kindergarten and Kindergarten, and grade-for-age status in 1st grade. Suppose a student is in Kindergarten in March 2001, and the family responded to the ACS in that month. That family would list that student as being enrolled in Kindergarten. Now suppose the student progressed to the next grade and the school year ended in May. If the family responded to the ACS in June, that family would still list that student as having most recently attended Kindergarten. If the households respond by October, however, it is likely that the next school year has begun, and the family would list that student as having been most recently enrolled in 1st grade. To account for this issue, this paper assumes that households responding to surveys between January 1st and April 10th will still have their children enrolled in the grade that they would have enrolled

[2] Since almost all states allow districts to set school calendar start and end dates (Education Commission of the States, April 2018*a*), there is substantial variation in the dates at which the school year ends for students in the U.S.. Ideally, the April 10th date would be the latest possible date before any school district has ended the school year and the July 1st date would be the earliest possible date after any school district has ended the school year. Although national data for all districts is not available on school start and end dates, Florida collects data on these dates for its school districts. In Florida, all school districts start school in August to September, and end the school year in May to June (Florida Department of Education, 2020). A sample of large school districts surveyed by Pew indicates that most school districts start school in August to September as well (Desilver, 2019). Hence, the sampling restrictions by date of response used in this paper fit with the limited data available.

in at the beginning of the school year. This paper also assumes that households that respond to surveys between September 30th and December 31st will either be enrolled in the next grade (if the student passed and is grade-for-age) or will still be enrolled in the same grade (if the student was retained and is not grade-for age). As grade-for-age status cannot be ascertained reliably for the intervening months when using the current grade enrolled question, this paper drops individuals who respond in those months from these calculations. Again, note that this specific adjustment only happens when looking at enrollment in pre-Kindergarten and Kindergarten and grade-for-age status in 1st grade.[3]

These sampling restrictions are necessary to ensure accurate assignment of grade-for-age status, but they may introduce bias related to response dates. If different types of households are more likely to respond to the survey at different times, then restricting attention to individuals who respond in specific months may bias the sample. If these sample restrictions change the sample in ways that do not vary across the New Year, it would mean that the treatment effect measured by the discontinuity is a local treatment effect for the population created by the sampling restrictions. If the sample restrictions change the sample in ways that vary across the New Year, it could bias the estimated treatment effect in complex ways that make any treatment effects measured harder to interpret.

The bias introduced in the ACS data by these sampling restrictions by date of response is likely small. As mentioned in the text, the ACS samples households throughout the year, with the vast majority of households assigned a sampling date

[3] Note that this set-up is similar to the previous adjustment when looking at grade-for-age status by grade completed, but omits slightly more data from the summer months. It is possible to assign families who respond in these summer months to a grade-for-age calculation with the most recent grade enrolled variable. Families who respond in the summer would presumably list their children as having been most recently enrolled in the grade that their student completed in the early spring. However, the previously described restrictions on response dates are used throughout the paper when looking at adults. Hence, omitting these months from the calculation keeps data sampling decisions as similar as possible among all calculations.

in the year at random (U.S. Census Bureau, 2019).[4] Hence, children born before and after the New Year are sampled at similar rates at different times across the year, and restricting attention to households sampled in particular months should not bias the composition of the sample of observations. The effect of this sampling restriction on the 2000 U.S. Census data is more complicated. The vast majority of responses to the 2000 Census happened in March through the end of April (Stackhouse and Brady, 2003*a*). Hence, most responses would have been sent in by April 10th. However, the households that respond later are more likely to be harder to reach, and more likely to be larger than households that respond earlier (Stackhouse and Brady, 2003*b*). These factors may correlate with family disadvantage, meaning that dropping responses in the summer months drops observations from families that are more likely disadvantaged.

One check on the potential bias of this sampling feature of the 2000 Census data is to drop this data from calculations. Table 6 offers a version of such a check. This table separates the data by birth cohorts when looking at grade-for-age status by high school. The 2000 Census data would not be included in the regression discontinuity calculations looking at children born 1987-1993 or 1994-2001, as the children born in these cohorts were not in high school in 2000. As is clear, the measured discontinuities in grade-for-age status for the cohorts born after 1987 are in the same range or larger than those for the birth cohort born before. Thus, the bias introduced by this sampling feature of the 2000 Census is likely minor.

One further issue with household response dates worth noting is how date of response affects enrollment rates in nursery school and pre-Kindergarten. Other school grades are nearly always organized by regular school calendars. So, the previously mentioned omissions of households by month of response result in data that reflect the average likelihood of a child being grade-for-age within that grade. However,

[4] Exceptions include households in rural Alaska and areas with high concentrations of Native Americans.

with children in pre-Kindergarten, there are many different enrollment policies across states, districts and local private care providers. The diversity of programs and program structures ensures that more children tend to be enrolled in pre-Kindergarten programs for months closer to the beginning of the next school year. The 2000 Census responses happen primarily in the later spring months before the lead-up to the next school year. Hence, the children in the 2000 Census are more likely to be enrolled in pre-Kindergarten than if these children were surveyed in the previous fall. While including the 2000 Census data does not impact the significance of discontinuities in enrollment across the New Year, it does increase average enrollment levels in pre-Kindergarten. Thus, this paper restricts attention to individuals in the ACS 2001-2016 for this calculation. The average in this data offers a more accurate estimate of average likelihood of being enrolled in nursery school or pre-Kindergarten in the year prior to Kindergarten enrollment.

Estimating the Discontinuity in After-Tax Income using CPS Data

As described in the text, this paper uses the March CPS to estimate the ize of the discontinuity in after-tax income for a family for having a child born in December rather than January. The estimation process draws inspiration from Hoynes, Miller and Simon (2015). The sample for the estimation process are parents with at least one infant under three who are in the March CPS in a four year radius for the year after the tax year. So for example, when calculating the discontinuity for the 1986 tax year, this paper uses all parents with at least one infant under three in a four year radius of the 1987 March CPS (1983 to 1991). Note that the central year in the data included is the year after the relevant tax year. The CPS March income data reflect income from the previous calendar year, which is the relevant year for computing taxes for the tax year. Parents with an infant under three are treated as having at least one infant under one who could have been born in January or December. The inclusion of other survey years and other child ages in the data is only to increase precision when calculating effects for smaller and more likely disadvantaged groups. A later part of this section investigates potential bias introduced by this choice.

Using this sample, this paper calculates tax obligations for having a child born in December by summing income measures at the family level and calculating the total state and federal tax burden using TAXSIM assuming that the family with the infant under three is the relevant tax filing unit.

This paper calculates tax obligations for having a child born in January using the same data with the same income measures but reducing the number of dependents under the age of 13 by one (as if the infant were born after December and hence not claimed on that year's tax return). The tax discontinuity is then the difference between the two calculated tax obligations. The percent change in after-tax income is this change divided by the after-tax income calculated for that family assuming the child was born in January. Families with no reported income are included in all calculations, but they comprise a small share of households over all years, and are included as a 0 increase in income and a 0 percent change in income.

Appendix Figure A.1 shows a check on the potential for bias from including parents with slightly older children and other years of survey data in the calculation. This figure shows the average estimated discontinuity when using only parents with infants under 1 and responses in the current tax year, and compares it to the results in Figure 2. As is clear, the measure is somewhat noisier, reflecting the smaller sample sizes, but the evolution of the discontinuity is similar over time, with the average gap between the two measures being $44. Note that using just the individuals with newborns who were born during the tax year results in a larger estimated increase in after-tax income. This difference is because families with older children are less likely to be in poverty, and hence usually have smaller CTC and EITC tax credits. However, the bias is relatively small across all years. Thus, it is likely the case that the other estimated discontinuities in Figure 2 are only slightly biased downwards by including families with older children and other tax years of data.

This paper, like many papers in the EITC literature that do not have access to administrative tax data, assumes 100% take-up of tax benefits to calculate the change in after-tax income produced by these tax policies (Hoynes, Miller and Simon, 2015). While take-up is not 100%, it is still likely high. LaLumia, Sallee and Turner (2015) find that 85% to 90% of newborns born in late December are claimed on a tax return in the 2000s. Of the remaining 15% to 10% of children who do not appear on tax returns, 5 percentage points are children whose parents do file tax returns but do not claim their newborn on that year's tax return, a phenomenon driven by low-income parents. Thus, likely 10 to 5 percentage points of the remaining share of newborns not claimed on taxes likely come from parents who are not required to file tax returns.

While the data in LaLumia, Sallee and Turner (2015) do not allow a strict calculation about take-up rates, a separate literature on take-up of the EITC suggests that, conditional on eligibility, take-up of the EITC is substantial. Among eligible families with children, Scholz (1994) estimates EITC take-up in 1990 of 80% to 86%, and U.S. Government Accountability Office (2001) find EITC take-up in 1999 is 86%. A large share of the families who do not claim EITC benefits are families not required to file taxes. For example, Blumenthal, Erard and Ho (2005) suggest that take-up of the eligible population of parents that are required to file taxes is 90% to 95%. Note furthermore that these take-up rates consider families with all ages of children, but the relevant take-up rate of interest for this paper would be take-up among families with newborns. Research shows that take-up of benefits among families with newborns is especially large. For example, twice as many newborns appear in tax returns as 11 year-olds (Dowd and Horowitz, 2011).

Take-up of child-related tax benefits like the EITC is likely high for three reasons. First, the IRS has taken steps to ensure low income households claim EITC benefits.

Prior to 1991, the IRS had a policy of offering the EITC to tax filers they deemed eligible even if they failed to claim it (U.S. Government Accountability Office, 1993). After 1991, the IRS switched to mailing tax filers who they concluded might be eligible to remind them of the availability of tax benefits (U.S. Government Accountability Office, 1993). Second, private tax preparers encourage low-income filers to file for the EITC since the tax preparers can claim a fraction of the tax return as compensation (Blumenthal, Erard and Ho, 2005). These arrangements have likely boosted outreach to low income eligible tax payers. Third, as the size of the credit has increased, so has the willingness of families to file to claim it (Blumenthal, Erard and Ho, 2005).

Without administrative data, it is impossible to come up with a precise understanding of how differential take-up might affect the estimated discontinuity in after-tax income used in this paper. Any decrease in take-up would by definition lower the estimated discontinuity. As such, Figure 2 in the paper is best understood as an upper bound on the size of the discontinuity in after-tax income.

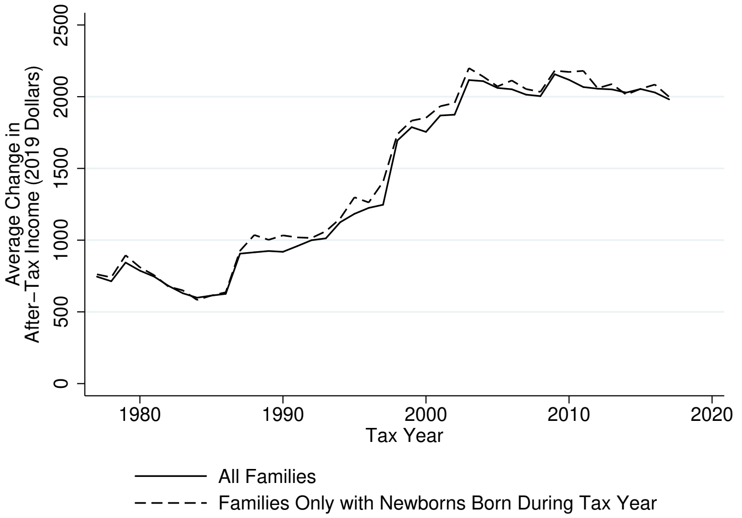
A descriptive exercise with the CPS data offers a lower bound. For each year, assume that 10% of newborns are not claimed in tax returns, and assume that these newborns come from families with either zero AGI, or families with the largest possible increases in after-tax income among the families not required to file taxes. Assume an additional 5% of newborns are also not claimed in tax returns, and assume that these newborns come from families who are legally required to file taxes and have the largest possible increases in after-tax income among this population. These percentages follow the results in LaLumia, Sallee and Turner (2015) above, where 10% of newborns were not claimed on taxes because their parents did not file taxes, and an additional 5% were not claimed even through the families filed tax returns. Note that because this adjustment drops observations from the population of filers who see large changes in after-tax income, it maximizes the drop in the estimated discontinuity that comes from this adjustment.

Appendix Figure A.2 below compares the results from this exercise to the estimated discontinuity reported in the paper in Figure 2. As is clear, this process adjusts the estimated discontinuity to be somewhere from 10% to 20% lower depending on the year. The estimated EITC take-up rate in the CPS data after applying these adjustments is 70% to 75%, which is lower than the take-up estimates listed above. Hence, this lower bound is conservative.

This paper does not do similar exercises like Appendix Figure A.2 for the two subgroups analyzed in the paper, children born to families with lower education attainment and Black children. Doing a calculation like Appendix Figure A.2 for these groups would require taking a clear stand on where the newborns not claimed on tax returns come from and their distribution among different demographics. It is not clear how to do such an exercise with available data. It is likely the case that a larger proportional share of these newborns come from families with low education attainment and Black families, as they likely have lower average income at time of a child's birth, and are hence more likely to not be required to file taxes. Hence, the percentage drops could be larger for these groups.

If the true discontinuity in after-tax income across the New Year is lower than was reported in the paper, then that would alter the instrumental variables estimates of the direct effect of income in infancy on later-life outcomes. A lower discontinuity in after-tax income would suggest that the real size of the estimated coefficient in the first stage is smaller, which would suggest that the instrumental variables estimates should be larger (as the denominator α in equation 8, would be lower). The effect of this drop on each instrumental variable estimate would depend on the years included, as the gap in the first stage differs by year. However, as the maximum gap between the upper bound and lower bound in after-tax income in Appendix Figure A.2 is 20%, that would suggest that instrumental variables estimates in the paper could be at most 25% higher.[5]

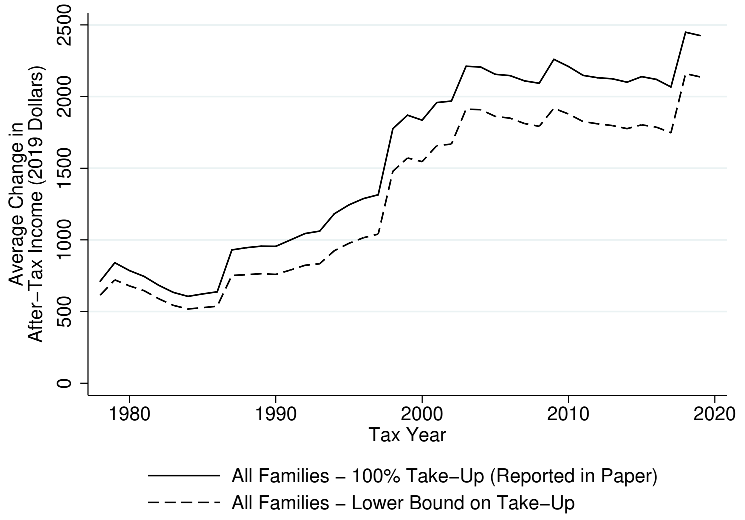
Figure A.1: Robustness of Estimated Average Increase in After-Tax Income from Having Newborn in December Compared to January Under Alternate Samples (2019 Dollars)



Notes: Figure depicts average increase in after-tax income for all families. The solid line is the average increase depicted in the paper. The dotted line uses an alternate subsample of the data, restricting attention to families with children aged 0 in the relevant March CPS year and using only CPS data from the relevant year. Details in the text. Standard error bars omitted for clarity, but standard errors are less than $100 for both lines and for all years.

[5] Note: $\frac{1}{0.8} = 1.25$

Figure A.2: Bounding Exercise for Estimated Average Increase in After-Tax Income from Having Newborn in December Compared to January (2019 Dollars)



Notes: Figure depicts average estimated discontinuity in after-tax income for families for having a child born in December compared to January of the next year by tax year of birth in 2019 dollars. The solid line is the average increase depicted in the paper, and assumes 100% take-up of eligible benefits. The dotted line is a robustness exercise that offers a lower bound on the estimated average increase in family income. Details in the text. Standard error bars omitted for clarity, but standard errors are less than $50 for both lines and for all years.

APPENDIX B

Tax Policies Related to Children

As discussed in the paper, the discontinuity depicted in Figures 2 and 3 reflects four main child-related tax benefits that depend on timing of birth: personal exemptions for a dependent, the EITC, the CTC and the Child and Dependent Care Credit. These four tax benefits have changed substantially over time, but eligibility for them in the first year of a child's life has always been determined by calendar year of birth, with children first eligible for them in the first tax year that they are born.

For all years in the data in Figure 2, parents may claim infant dependents as a personal exemption for a reduction in their taxable income. In tax year 2017, if a parent has a taxable income greater than 0 after applying other deductions, and if that parent has an infant born in December 2017, that parent could reduce their taxable income by up to $4,050. The value of this change in their tax obligations depends on their marginal tax rate. However, it is important to note that this benefit is not refundable, meaning that the additional benefit of the deduction can only reduce a parent's tax obligations to 0. Hence, it provides limited benefits to families that already have low tax obligations.

Starting in 1975, parents were also eligible to claim EITC benefits for infant dependents. This program, over time, has substantially increased the discontinuity in after-tax income from claiming an infant on a tax return. The EITC offers households with earned income above 0 a benefit that gradually increases in income until it reaches a maximum level and eventually phases out to 0. Importantly, this benefit is refundable, meaning that it can both reduce tax obligations and result in a tax refund where a parent receives a refund for the difference between tax obligations and the size of the EITC credit. Following its enactment, the real value of the EITC declined from 1975 to 1986 as the credit was not adjusted annually for inflation (Crandall-Hollick, 2018*b*). Legislative changes since 1987 have gradually made the size of the EITC credit more generous. This increase has happened through both raising the maximum benefit in real dollars, and increasing the number of children for whom tax filers can claim an EITC benefit.[1]

Since 1998, parents with infants who have incomes below a certain level are also eligible for the Child Tax Credit (CTC). Similar to the EITC, the child tax credit is partially refundable, and gradually phases out for tax filers with sufficiently high incomes.

Technically, there is a fourth infant-related tax credit that parents are eligible for if they have an infant born before December 31st of a tax year: the Child and Dependent Care Credit. Given the lack of information on child care expenses in the CPS, this credit is omitted from consideration here, although it would on average increase the size of the discontinuity in after-tax income.[2]

[1] One notable change from 1986 complicating analysis of take-up in this data is the fact that, beginning in tax year 1987, tax filers were required to list the Social Security Number for exemptions for dependents that they claimed. It is well-known that this requirement resulted in a drop of the number of dependents claimed from 77 million in tax year 1986 to 70 million in tax year 1987. Thus, it is possible that there is not as sharp a discontinuity in claiming of dependents around the New Year in years prior to 1987. Parents with children born after the New Year in those earlier years may be claiming them inappropriately regardless of timing of birth. There is no way to accommodate this issue in this data when calculating the increase in after-tax income in Figure 2. This issue would complicate analysis of results because it would suggest that the discontinuity in after-tax income is potentially less sharp in earlier years. However, it should be noted that Table 6 looks at grade-for-age status of high schoolers, and separates the data into children born before and after 1987. As is clear, the measured change in grade-for-age status for being born before the New Year is larger for the cohorts of children born after 1987. Whatever the take-up issues created by this specific policy change in 1987, the same basic causal results are observed for cohorts born afterward.

[2] The average size of this credit among tax filers who claim it is smaller than credits from the EITC and CTC. The average value of the credit is usually \$500 to \$600 as opposed to over \$1,000.

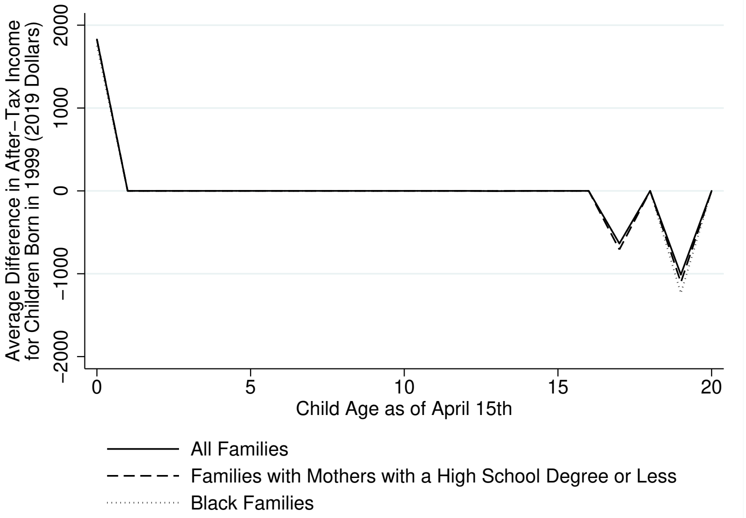
As depicted in Figure 1, eligibility for tax benefits phases out over time as children age. As a result, there are later discontinuities in after-tax income that occur as children reach various ages. For example, as shown in Figure 1, in the calendar year in which children born in December turn 17, their families are no longer eligible to claim the Child Tax Credit for them. However, families with children born in January are still eligible to claim the Child Tax Credit for their children in that tax year.

Appendix Figure B.1 offers an indication of how these changes in eligibility impact after-tax income for families as their children age. This figure looks at the evolution of the gap in after-tax income by child age for the cohort of families with children born in December 1999 or January 2000. This gap is estimated in the March CPS using the procedures discussed earlier in Appendix A. As is clear, when children are infants, families with December births see the increase in after-tax income depicted in Figure 2. In the next year, however, all families are eligible for the tax credits, so the difference disappears.[3] When the children born in December turn 17, however, their families are no longer eligible for the Child Tax Credit for them, so the families with children born in January see slightly larger after-tax incomes. When these children turn 18, there is again no difference in their after-tax income as both groups are eligible for the same tax benefits. However, when these children turn 19, the families with children born in January see slightly larger after-tax incomes, as they are eligible to claim the EITC for these children and the families with children born in December are not.

It is concentrated among middle and upper-middle income taxpayers, and is claimed by only 13 percent of taxpayers with children. Hence, its impact on after-tax income for the tax discontinuity studied here is likely comparatively small. (Crandall-Hollick, 2018*a*)

[3] This estimation strategy cannot account for changes in income that might happen because of responses to the income shock in infancy. Black et al. (2014) show that a modest shock of a slightly larger size than the shock considered in this work resulted in a long-term change in labor force participation of mothers. If similar dynamics happen here, then there may be a non-zero difference in income in the years after children are infants. This possibility is a direction for future work described in the conclusion.

Figure B.1: Difference in After-Tax Income for December and January Births by Age of Child for Children born in December 1999 compared to January 2000 (2019 Dollars)



Notes: Figure depicts average estimated difference in family after-tax income by child age for families that have a child born in December 1999 compared to January of 2000. Incomes measured in 2019 dollars. Age variable on the horizontal axis lists age as would be recorded by a family on April 15th. For example, newborns in their first year of life born in January and December would be age 0 by April 15th. Estimation process draws inspiration from Hoynes, Miller and Simon (2015) and uses the March CPS. Additional details on estimation are in the text and in Appendix A. Standard error bars here omitted for clarity, but standard errors are less than $10 for all groups and all years.

APPENDIX C

Theoretical Foundations of Birth Shifting

To better understand the choices families make about birth timing and the meaning of the discontinuity described earlier, it is necessary to think about the incentives families face when considering timing births around the New Year. This appendix offers theoretical foundations for two features of the intuition underlying the empirical method. First, there is a limit on how far birth timing is moved by families as, outside of a region around the New Year, there is less incentive to engage in strategic birth-timing. Second, omitting data around the New Year restricts attention to a sample that can identify the theoretical effect of the change in treatment across the threshold.

Consider the following one period family utility optimization problem:

$$\begin{aligned} \max_{d,C,F,L} \quad & V(\delta C, F, L) - f(d-d') - \eta \mathbb{1}[d=0] \\ & w.r.t \quad p_C C + p_F F = wL + \mathbb{1}[d<0]T(wL, d<0) + \mathbb{1}[d \geq 0]T(wL, d \geq 0) + I \end{aligned}$$

Assume that:

$$\begin{aligned}
&\frac{\partial V}{\partial C} > 0, \frac{\partial V}{\partial F} > 0, \frac{\partial V}{\partial L} < 0 \\
&\frac{\partial T}{\partial L}_{d<0} > 0, \frac{\partial T}{\partial L}_{d\geq 0} > 0, \frac{\partial^2 T}{\partial L^2} = 0 \\
&V \text{ is concave}
\end{aligned}$$

In the first equation, C is spending on a newborn, δ is a multiplier on C drawn from a distribution (where higher levels of δ indicate high marginal utility of investments in C), F is spending on the rest of the family, L is a unitary measure of labor for the household, d is the realized date of birth (centered such that $d = 0$ is New Year's day) and d' is the date of birth that would happen without a parent altering the timing of birth, and $f(d - d')$ is a cost function that reaches a minimum when $d = d'$. This term reflects the fact that altering the exact date of birth of a child away from the expected due date, either by Cesearian section or induced labor, is costly to a family in terms of consequences to an infant and a mother's health. Given the relatively smooth distribution of births outside of holidays depicted in Figure 5, assume that d' is randomly assigned. The final term, η is a utility cost to being born on the New Year independent of tax benefits.

$T(wL)$ is an equation representing tax obligations, but the tax schedule differs in this first year depending on whether a child is born before or after New Year's Day. So, there are two separate functions T if d is less than or greater than 0. Assume that, for each level of wL, the after-tax income of having a child before the New Year is greater than having a child after the New Year, or $T(wL, d < 0) > T(wL, d \geq 0)$. Assume that the tax schedule is linear for simplicity. I is a fixed endowment.

Lastly, suppose that the family optimization problem proceeds in the following order:

1). A family chooses L given a certain prior on d', $g(d')$;

2). d' is realized;

3). A family chooses C, F and d to maximize utility with respect to the budget constraint.

Note that the later timing of choices over C, F and d compared to earlier decisions over L reflects the fact that changes in real economic behavior, such as labor supply, are more difficult for births that might happen close to the New Year. Further away from the New Year, there may be more opportunities to alter economic activity after a child's birth.

A critical piece of the family's optimization problem that will determine their decisions is the shape of the cost function for altering birth timing, f. Consider three possibilities:

Case 1: $f(d - d') = \infty$ if $d - d' \neq 0$

Suppose that $f(d - d')$ is infinite for every value except $f(0)$, and keep w, p_C, p_F and $g(d')$ the same for all families. Then, the infinite utility cost associated with altering birth timing means that a family would have no desire to alter birth timing, and families would be randomly assigned on either side of New Year's Day depending on their assignment of d'. In such a scenario, L would be constant for everyone with the same δ, and the additional shock to income given by being bumped into a different tax bracket would be a pure income shock that would both impact investments in C and F. Thus, a simple comparison of people born before and after the New Year will identify the effect of the income boost.

This outcome is depicted in a simulated example in Appendix Figure C.1. Note that the counts of births are relatively smooth, as is average δ. The lack of variation in both variables reflects the fact that no selection across the New Year occurs in this setting.

Case 2: $f'(d-d')=0$ **and** $f \geq 0$

Suppose that $f'(d-d')=0$, and keep w, p_C, p_F and $g(d')$ the same for all families. Then, the lack of a utility cost that varies with d means that families' decisions about birth timing is unaffected by the assignment of d'.

In such a scenario, families' choice of L and d would depend on their value of δ and the value of f. Families that have $d < 0$ would not have any incentive to shift birth timing, as there is no tax benefit to doing so. Among the families that have $d \geq 0$, families with higher δ would be more willing to shift birth timing. They would more highly value the marginal utility of an additional dollar of expenditure on their newborn, and hence would value more highly the value of the tax benefit from being born before the New Year. Importantly, though, families' choices over d would not change depending on d', as the costs to altering birth-timing are constant. Note that the selection here ensures that the families with births before the New Year are different than families with births after the New Year.

This model has important implications for what happens near the discontinuity. First, unlike the infinite cost setting before, actual observed birthdays d will not be randomly distributed, and a larger mass of individuals will move from the days after New Year's Day to the day right before New Year's Day. Second, comparing spending patterns of individuals right before the New Year to spending patterns of individuals born on New Year's day is no longer indicative of the pure income effect of increasing a family's economic resources. The individuals born after the New Year will include people with comparatively low values of δ, indicating that their spending on their infants will be comparatively lower, and the individuals born before the New Year will include people with comparatively higher values of δ, indicating that their spending on their infants will be comparatively higher. Thus, a comparison of their spending will both indicate the pure effect of the increase in after-tax income, but also the difference in the distribution of δ that comes from the people selecting to have births before the New Year having higher marginal utility of spending on children. These differences would mean that a naive comparison of spending on children at the New Year would offer a biased upwards treatment effect.

This outcome is depicted in a simulated example in Appendix Figure C.2. For this graph, assume that each family has a function f that is a constant draw from some distribution. In this situation, there are an abnormally large number of births that happen on the day before the New Year, reflecting shifting of births from families that would have otherwise had births after the New Year. Technically, in this setting, families would be indifferent between scheduling births on the day before New Year's or on any other day before New Year's. As is clear, there are permanently lower births after New Year's, reflecting the fact that families' decisions to alter birth timing is unrelated to d. Furthermore, the average δ of births that happen the day before the New Year is noticeably higher than the days around it, reflecting the fact that the families that move to schedule a birth before New Year's Day have higher δ. Conversely, the children who are born after New Year's have lower average δ.

Case 3: $f(d - d')$ is convex

Suppose alternatively that f is convex, and keep w, p_C, p_F and $g(d')$ the same for families. As in case 2, families assigned births d' that are before New Year's Day see no benefit from altering their birth timing as the tax benefits to having a child before the New Year are always larger. So they will continue to select d' as a child's birth date. However, families with $d' \geq 0$ will choose $d = -1$ as long as the utility they achieve from having their birth before the New Year is larger than that they would have if they timed their births after the New Year. That is, as long as:

$$V(\delta C_{-1}, F_{-1}, L) - f(-1 - d') > V(\delta C_{d'}, F_{d'}, L) - \eta \mathbb{1}[d' = 0]$$

Where C_{-1}, F_{-1}, $C_{d'}$, $F_{d'}$ represent consumption choices such that budget sets balance at either $d = -1$ or $d = d'$. As in case 2, families' choice of L and d would depend on their value of δ and the value of f. Taking L and d as given, note that, for any given level of δ , the convex cost in d' means that there is some maximum date past which individuals will not move the timing of their birth. Furthermore, note that for each level of d', the individuals who move the timing of their birth will have larger values of δ, indicating a larger marginal utility of spending on children.

As in case 2, there is selection into birth timing around the New Year. However, for each level of δ, there is some birthdate d' such that no family would move timing of the birth. Thus, dropping birthdates that appear affected by birth shifting and restricting attention to days away from the New Year gives a sample unaffected by the bias created by the uneven distribution of δ. A comparison of spending between these restricted samples would identify, again, the pure income effect of the change in resources on investments in children.

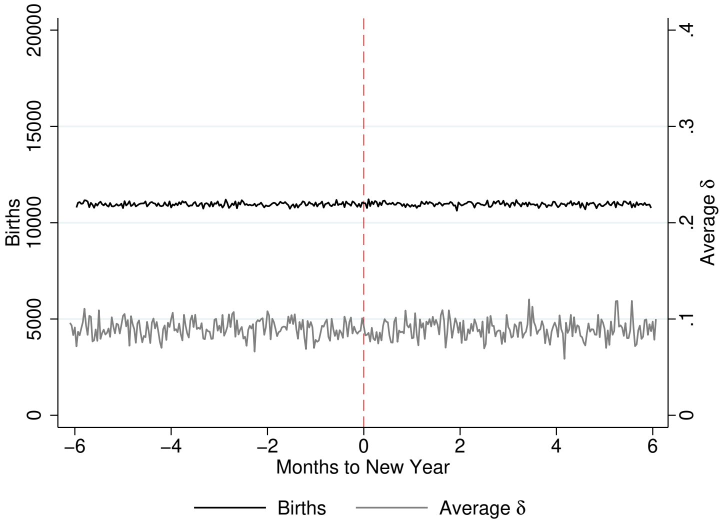
This outcome is depicted in a simulated example in Appendix Figure C.2. Note that there is a massive spike in births on the day before New Year's Day, as this would be the least costly day for families to move timing of birth to.

Some complications of how families perceive the discontinuity are important. First, the analysis in this paper focuses less on immediate spending on children then on intermediate and longer-term outcomes for children, which can be thought of as demonstrating the long-term consequences of that spending. The discussion section at the end touches on how similar income shocks tend to be spent by families in other settings, but there are none directly comparable to the shock in this paper.

Second, the size of the discontinuity in resources will depend on how families understand the tax system. As discussed in the text, this income shock is technically a speeding up of the tax benefits related to children, as families that have children born in December are eligible for the tax benefits one year before families with children born in January, but then their eligibility expires one year earlier as well. If families fully understand this feature of how the system works, then the shock to their spending might be smaller in the short-run, as they could borrow against future earnings (hence increasing I in the model above). As discussed in the text, there is evidence that some share of families misunderstand the timing of how benefits expire in the tax system. Furthermore, the families that benefit from these transfers, especially less educated families, are likely credit constrained, and thus less able to borrow against future income. Both of these features of this setting mean that families with children born in January have limited ability to borrow against future earnings.

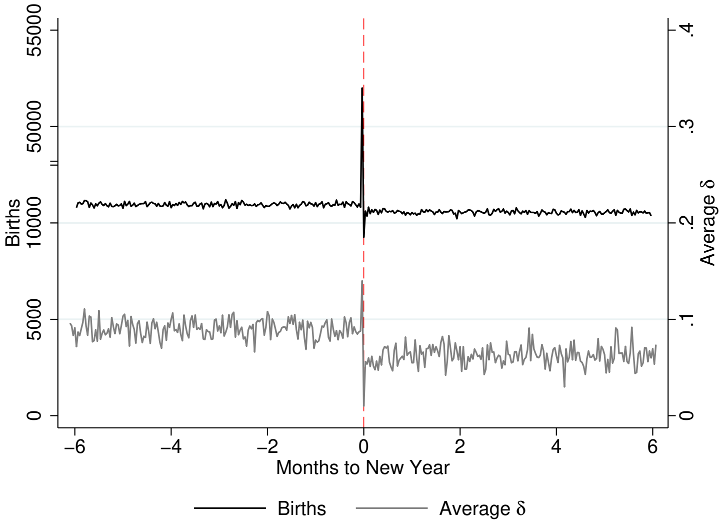
Thus, this setting shows that basic microeconomic theory and simple assumptions about the optimization process can explain the basic intuition motivating the empirical approach in this paper. First, there is limited birth shifting outside of a window around holiday. Second, omitting the data that demonstrate shifting ensures that a comparison of people born after and born before the New Year identifies the effect of the increase in after-tax income.

Figure C.1: Simulation Of Births by Day of Year Under Case 1 for f



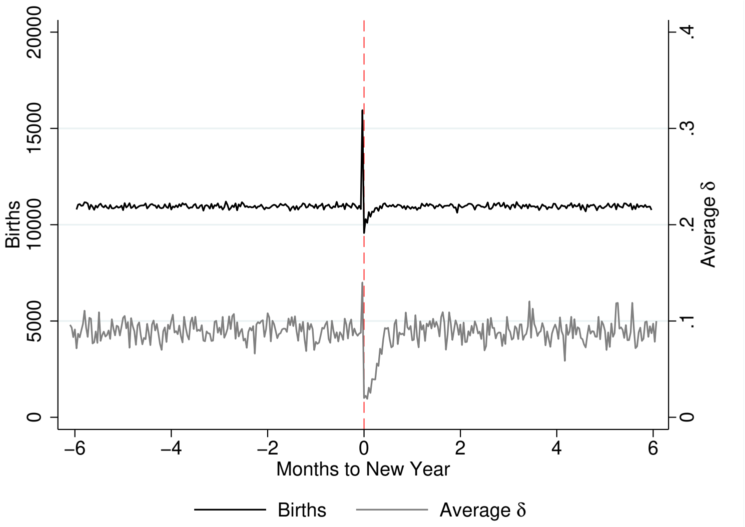
Notes: Graph shows simulated distribution of births by day of year under case 1 for f described above, where $f(d - d') = \infty$ if $d - d' \neq 0$.

Figure C.2: Simulation Of Births by Day of Year Under Case 2 for f



Notes: Graph shows simulated distribution of births by day of year under case 2 for f described above, where $f'(d - d') = 0$ and $f \geq 0$.

Figure C.3: Simulation Of Births by Day of Year Under Case 3 for f



Notes: Graph shows simulated distribution of births by day of year under case 3 for f described above, where $f(d-d')$ is convex.

BIBLIOGRAPHY

BIBLIOGRAPHY

Abowd, John M. 2017. *Large-scale Data Linkage from Multiple Sources: Methodology and Research Challenges.* NBER Summer Institute Methods Lecture.

Abowd, John M., and Lars Vilhuber. 2005. "The Sensitivity of Economic Statistics to Coding Errors in Personal Identifiers." *Journal of Business and Economic Statistics*, 23(2): 133–165.

Abramitzky, Ran, Leah Boustan, and Katherine Eriksson. 2012*a*. "Web Appendix: Europe's tired, poor, huddled masses: Self-selection and economic outcomes in the age of mass migration."

Abramitzky, Ran, Leah 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. 2012*b*. "Europe's Tired, Poor, and Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review*, 102.

Abramitzky, Ran, Roy Mill, and Santiago Pérez. 2018. "Linking Individuals Across Historical Sources: a Fully Automated Approach." *National Bureau of Economic Research Working Paper Series*, 1(24324).

Abramitzky, R., L. Platt Boustan, and K. Eriksson. 2012. "Europe's Tired, Poor, and Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration." *American Economic Review*, 102(5): 1832–1856.

Abramitzky, R., L. Platt Boustan, and K. Eriksson. 2013. "Have the Poor Always been Less Likely to Migrate? Evidence from Inheritance Practices during the Age of Mass Migration." *Journal of Development Economics*, 102: 2–14.

A'Hearn, Brian, Jörg Baten, and Dorothee Crayen. 2009. "Quantifying Quantitative Literacy: Age Heaping and the History of Human Capital." *The Journal of Economic History*, 69(3): 783–808.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney. 2016*a*. "The Long-Run Impact of Cash Transfers to Poor Families." *American Economic Review*, 106(4): 935–71.

Aizer, Anna, Shari Eli, Joseph Ferrie, and Adriana Lleras-Muney. 2016*b*. "The Long Term Impact of Cash Transfers to Poor Families." *American Economic Review*, 106(4): 935–971.

Akee, Randall K. Q., William E. Copeland, Gordon Keeler, Adrian Angold, and E. Jane Costello. 2010. "Parents' Incomes and Children's Outcomes: A Quasi-experiment Using Transfer Payments from Casino Profits." *American Economic Journal: Applied Economics*, 2(1): 86–115.

Almond, Douglas, and Joseph J. Doyle. 2011. "After Midnight: A Regression Discontinuity Design in Length of Postpartum Hospital Stays." *American Economic Journal: Economic Policy*, 3(3): 1–34.

Almond, Douglas, Hilary W. Hoynes, and Diane W. Schanzenbach. 2011. "Inside the War on Poverty: The Impact of Food Stamps on Birth Outcomes." *The Review of Economics and Statistics*, 93(2): 387–403.

Alsan, M., and C. Goldin. 2015. "Watersheds in Infant Mortality: The Role of Effective Water and Sewage Infrastructure, 1880 to 1915." *NBER Working Paper*, 21263.

Andrews, I., and E. Oster. 2017. "Weighting for External Validity." *NBER Working Paper*, 23826.

Angrist, J.D., and J.-S. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton, NJ:Princeton University Press.

Angrist, Joshua D., and Alan B. Krueger. 1992. "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples." *Journal of the American Statistical Association*, 87(418): 328–336.

Antonie, Luiza, Kris Inwood, Daniel J. Lizotte, and J.Andrew Ross. 2014. "Tracking People Over Time in 19th Century Canada for Longitudinal Analysis." *Machine Learning*, 95(1): 129–146.

Atack, Jeremy. 2004. "A Nineteenth-Century Resource for Agricultural History Research in the Twenty-First Century." *Agricultural History*, 78(4): 389–412.

Atack, Jeremy, Fred Bateman, and Mary Eschelbach Gregson. 1992. "Matchmaker, Matchmaker, Make Me a Match." *Historical Methods*, 25(2): 53–65.

Autor, David, David Figlio, Krzysztof Karbownik, Jeffrey Roth, and Melanie Wasserman. 2019. "Family Disadvantage and the Gender Gap in Behavioral and Educational Outcomes." *American Economic Journal: Applied Economics*, 11(3): 338–81.

Bailey, Amy Kate, Stewart E. Tolnay, E.M. Beck, and Jennifer D. Laird. 2011. "Targeting Lynch Victims: Social Marginality or Status Transgressions?" *American Sociological Review*, 76(3): 412–436.

Bailey, Martha, and Susan Dynarski. 2011. "Gains and Gaps: Changing Inequality in U.S. College Entry and Completion." In *Whither Opportunity? Rising Inequality, Schools, and Children's Life Chances.* , ed. G. J. Duncan and R. J. Murnane, Chapter 6, 117–132. New York, NY:Russell Sage Foundation.

Bailey, Martha, Connor Cole, and Catherine G. Massey. 2019. *Simple Strategies for Improving Inference with Linked Data: A Case Study of the 1850-1930 IPUMS Linked Representative Historical Samples.* University of Michigan Working Paper.

Bailey, Martha J. 2018. *Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database.* University of Michigan Working Paper.

Bailey, Martha J., and Connor Cole. 2019. "Autolink.ado." accessed 2019-06-13.

Bailey, M.J., C. Cole, M. Henderson, and C.G. Massey. 2019. "How Well Do Automated Linking Methods Perform in Historical Samples? Evidence from New Ground Truth." *NBER Working Paper*, 24019.

Bailey, M.J., S. Anderson, A. Karimova, and C.G. Massey. 2016. "Creating LIFE-M: The Longitudinal, Intergenerational Family Electronic Micro-Database." Retrieved from.

Barreca, Alan, Melanie Guldi, Jason Lindo, and Glen R. Waddell. 2011. "Saving Babies? Revisiting the Effect of Very Low Birth Weight Classification." *The Quarterly Journal of Economics*, 126(4): 2117–2123.

Bassok, Daphna, and Sean F. Reardon. 2013. ""Academic Redshirting" in Kindergarten: Prevalence, Patterns, and Implications." *Educational Evaluation and Policy Analysis*, 35(3): 283–297.

Bastian, Jacob, and Katherine Michelmore. 2018. "The Long-Term Impact of the Earned Income Tax Credit on Children's Education and Employment Outcomes." *Journal of Labor Economics*, 36(4): 1127–1163.

Bedard, Kelly, and Elizabeth Dhuey. 2012. "School-Entry Policies and Skill Accumulation Across Directly and Indirectly Affected Individuals." *Journal of Human Resources*, 47(3): 643–683.

Black, Sandra E., and Paul J. Devereux. 2011. "Recent Developments in Intergenerational Mobility." In *Handbook of Labor Economics.* , ed. Card David and Ashenfelter Orley, 1487–1541. Amsterdam:Elsevier.

Black, Sandra, Paul Devereux, Katrine V. Loken, and Kjell G Salvanes. 2014. "Care or Cash? The Effect of Child Care Subsidies on Student Performance." *The Review of Economics and Statistics*, 96(5): 824–837.

Bleakley, H. 2007. "Disease and Development Evidence from Hookworm Eradication in the American South." *Quarterly Journal of Economics*, 122(1): 73–117.

Bleakley, H., and J. Ferrie. 2014. "Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run." Retrieved from.

Bleakley, H., and J. Ferrie. 2016. "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations." *Quarterly Journal of Economics*, 131(3): 1455–1495.

Bleakley, H., and J.P. Ferrie. 2013. "Up from Poverty? The 1832 Cherokee Land Lottery and the Long-run Distribution of Wealth." *NBER Working Paper*, 19175.

Bleakley, Hoyt, and Joseph Ferrie. 2017. "Land Opening on the Georgia Frontier and the Coase Theorem in the Short- and Long- Run."

Blumenthal, Marsha, Brian Erard, and Chih-Chin Ho. 2005. "Participation and Compliance With the Earned Income Tax Credit." *National Tax Journal*, 58(2): 189–213.

Bogue, A. 1963. *From Prairie to Corn Belt: Farming on the Illinois and Iowa Prairies in the Nineteenth Century.* Chicago:University of Chicago Press.

Boustan, Leah Platt, and William Collins. 2014. "The Origins and Persistence of Black-White Differences in Women's Labor Force Participation from the Civil War to the Present." In *Human Capital and History: The American Record.* , ed. Leah Boustan, Carola Frydman and Robert A. Margo. Chicago, IL:University of Chicago Press.

Boustan, Leah Platt, Matthew E. Kahn, and Paul W. Rhode. 2012. "Moving to Higher Ground: Migration Response to Natural Disasters in the Early Twentieth Century." *American Economic Review: Papers and Proceedings*, 102(3): 238–244.

Buckles, Karey S., and Daniel M. Hungerman. 2013*a*. "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics*, 95: 3 711–724.

Buckles, Kasey S., and Daniel M. Hungerman. 2013*b*. "Season of Birth and Later Outcomes: Old Questions, New Answers." *The Review of Economics and Statistics*, 95(3): 711–724.

Bulman, George, Robert Fairlie, Sarena Goodman, and Adam Isen. 2017. "Parental Resources and College Attendance: Evidence from Lottery Winnings." NBER Working Paper 22679, Cambridge, MA.

Bureau of Labor Statistics. 2018. "Current Population Survey: Handbook of Methods." Bureau of Labor Statistics, Washington, DC.

Byrd, Robert S., and Michael L. Weitzman. 1994. "Predictors of Early Grade Retention Among Children in the United States." *Pediatrics*, 93(3): 481–487.

Caliendo, M., and S. Kopeinig. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching." *Journal of Economic Surveys*, 22(1): 31–72.

Case, Anne, Darren Lubotsky, and Christina Paxson. 2002. "Economic Status and Health in Childhood: The Origins of the Gradient." *American Economic Review*, 92(5): 1308–1334.

Caucutt, Elizabeth M., Lance Lochner, and Youngmin Park. 2017. "Correlation, Consumption, Confusion, or Constraints: Why Do Poor Children Perform so Poorly?" *The Scandinavian Journal of Economics*, 119(1): 102–147.

Cesarini, David, Erik Lindqvist, Robert Ostling, and Bjorn Wallace. 2016. "Wealth, Health, and Child Development: Evidence from Administrative Data on Swedish Lottery Players." *The Quarterly Journal of Economics*, 131(2): 687–738.

Chetty, Raj, John N. Friedman, and Jonah Rockoff. 2011. "New Evidence on the Long-Term Impacts of Tax Credits." Working Paper. Accessed October 11th, 2020. Available: https://www.irs.gov/pub/irs-soi/11rpchettyfriedmanrockoff.pdf.

Chetty, Raj, John N. Friedman, Tore Olsen, and Luigi Pistaferri. 2011. "Adjustment Costs, Firm Responses, and Micro vs. Macro Labor Supply Elasticities: Evidence from Danish Tax Records." *The Quarterly Journal of Economics*, 126(2): 749–804.

Chetty, Raj, Nathaniel Hendren, Maggie R. Jones, and Sonya R. Porter. 2019. "Race and Economic Opportunity in the United States: an Intergenerational Perspective." *The Quarterly Journal of Economics*, 135(2): 711–783.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, and Emmanuel Saez. 2014*a*. "Where is the Land of Opportunity? The Geography of Intergenerational Mobility in the United States." *The Quarterly Journal of Economics*, 129(4): 1553–1623.

Chetty, Raj, Nathaniel Hendren, Patrick Kline, Emmanuel Saez, and Nicholas Turner. 2014*b*. "Is the United States Still a Land of Opportunity? Recent Trends in Intergenerational Mobility." *American Economic Review*, 104(5): 141–47.

Chevalier, Arnaud, Colm Harmon, Vincent O'Sullivan, and Ian Walker. 2013. "The Impact of Parental Income and Education on the Schooling of Their Children." *IZA Journal of Labor Economics*, 2(8).

Christen, P., and T. Churches. 2005. "Febrl - Freely extensible biomedical record linkage." Retrieved from.

Christen, Peter, and Karl Goiser. 2007. "Quality and Complexity Measures for Data Linkage and Deduplication."

Clay, K., J. Lewis, and E. Severnini. 2016. "Canary in a Cola Mine: Infant Mortality." In *Property Values, and Tradeoffs Associated with Mid-20th Century Air Pollution.* 1–61.

Collins, William J., and Marianne H. Wanamaker. 2014. "Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data." *American Economic Journal: Applied Economics*, 6(1): 220–252.

Collins, William J., and Marianne H. Wanamaker. 2015. "The Great Migration in Black and White: New Evidence on the Selection and Sorting of Southern Migrants." *Journal of Economic History*, 75(4): 947–992.

Collins, William J., and Marianne H. Wanamaker. 2016. "Up from Slavery? African American Intergenerational Economic Mobility Since 1880." *NBER Working Paper*, 23395.

Conger, Rand D., Xiaojia Ge, Glen H. Elder, Frederick O. Lorenz, and Ronald L. Simons. 1994. "Economic Stress, Coercive Family Process, and Developmental Problems of Adolescents." *Child Development*, 65(2): 541–561.

Costa, Dora L., Heather DeSomer, Eric Hanss, Christopher Roudiez, Sven E. Wilson, and Noelle Yetter. 2017. "Union Army Veterans, All Grown Up." *Historical Methods*, 50(2): 79–95.

Crandall-Hollick, Margot L. 2016. "The Child Tax Credit: Current Law and Legislative History." Congressional Research Service Report R41873, Washington, DC.

Crandall-Hollick, Margot L. 2018*a*. "Child and Dependent Care Tax Benefits: How They Work and Who Receives Them." Congressional Research Service CRS Report R44993, Washington, DC.

Crandall-Hollick, Margot L. 2018*b*. "The Earned Income Tax Credit (EITC): A Brief Legislative History." Congressional Research Service Report R44825, Washington, DC.

Crump, R.K., V.Joseph Hotz, Guido W. Imbens, and O.A. Mitnik. 2009. "Dealing with Limited Overlap in Estimation of Average Treatment Effects." *Biometrika*, 96(1): 187–199.

Cunha, Flavio, and James Heckman. 2007. "The Technology of Skill Formation." *American Economic Review*, 97(2): 31–47.

Cunha, Flavio, James J. Heckman, Lance Lochner, and Dimitriy V. Masterov. 2006. “Interpreting the Evidence on Life Cycle Skill Formation.” In . Vol. 1 of *Handbook of the Economics of Education*, , ed. E. Hanushek and F. Welch, 697–812. Elsevier.

Currie, Janet. 2009. “Healthy, Wealthy, and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Capital Development.” *Journal of Economic Literature*, 47(1): 87–122.

Currie, Janet, and Douglas Almond. 2011. “Human Capital Development Before Age Five.” *Handbook of Labor Economics*, , ed. David Card and Orley Ashenfelter Vol. 4, 1315–1486. Elsevier.

Curti, Merle. 1959. *The Making of an American Community: A Case Study of Democracy in a Frontier County.* Stanford:Stanford University Press.

Cutler, D.M., and G. Miller. 2005. “The Role of Public Health Improvements in Health Advances: The 20th Century United States.” *Demography*, 42(1): 1–22.

Dahl, Gordon B., and Lance Lochner. 2012. “The Impact of Family Income on Child Achievement: Evidence from the Earned Income Tax Credit.” *American Economic Review*, 102(5): 1927–56.

Dahl, Gordon B., Katrine V. Loken, and Magne Mogstad. 2014. “Peer Effects in Program Participation.” *American Economic Review*, 104(7): 2049–74.

Datar, Ashlesha. 2006. “Does Delaying Kindergarten Entrance Give Children a Head Start?” *Economics of Education Review*, 25(1): 43–62.

Deming, David, and Susan Dynarski. 2008. “The Lengthening of Childhood.” *Journal of Economic Perspectives*, 22(3): 71–92.

Dempster, A.P., N.M. Laird, and D.B. Rubin. 1977. “Maximum Likelihood from Incomplete Data via the EM Algorithm.” *Journal of the Royal Statistical Society. Series B (Methodological*, 39(1): 1–38.

Desilver, Drew. 2019. “”Back to school” Means Anytime from Late July to After Labor Day, Depending on Where in the U.S. You Live.” Pew Research Center, Washington, DC. Accessed October 11th, 2020. Available: https://www.pewresearch.org/fact-tank/2019/08/14/back-to-school-dates-u-s/.

Dickert-Conlin, Stacy, and Amitabh Chandra. 1999. “Taxes and the Timing of Birth.” *Journal of Political Economy*, 107(1): 161–177.

DiNardo, J., N.M. Fortin, and T. Lemieux. 1996. “Labor Market Institutions and the Distribution of Wages, 1973-1992: A Semiparametric Approach.” *Econometrica*, 64(5): 1001–1044.

Dowd, Tim, and John B. Horowitz. 2011. "Income Mobility and the Earned Income Tax Credit." *Public Finance Review*, 39(5): 619–652.

Duflo, E., R. Glennerster, and M. Kremer. 2007. "Chapter 61 Using Randomization in Development Economics Research: A Toolkit." In *Handbook of Development Economics*. Vol. 4, , ed. T.P. Schultz and J.A. Strauss, 3895–3962. Elsevier.

Duncan, Greg J., Jens Ludwig, and Katherine A. Magnuson. 2011. "Child Development." *Targeting Investments in Children: Fighting Poverty When Resources are Limited*, , ed. Phillip B. Levine and David J. Zimmerman, 27–58. University of Chicago Press.

Duncan, Otis Dudley. 1968. "Patterns of Occupational Mobility among Negro Men." *Demography*, 5(1): 11–22.

Education Commission of the States. April 2018*a*. "State Comparison: School Start/Finish." Accessed October 11th, 2020. Available: http://ecs.force.com/mbdata/mbquestci?rep=IT1804.

Education Commission of the States. June 2018*b*. "State Kindergarten Through Third-Grade Policies: Is There a Third Grade Retention Policy?" Accessed October 11th, 2020. Available: http://ecs.force.com/mbdata/MBQuest2RTanw?rep=KK3Q1818.

Eli, Shari, Laura Salisbury, and Allison Shertzer. 2018. "Ideology and Migration after the American Civil War." *Journal of Economic History*, 78(3).

Eli, S., L. Salisbury, and A. Shertzer. 2016. "Migration in Response to Civil Conflict: Evidence from the Border of the American Civil War." In *NBER Working Paper 22591.*

Eriksson, Björn. 2016. "The Missing Links: Data Quality and Bias to Estimates of Social Mobility." Accessed September 15,.

Fan, Jianqing, Irene Gijbels, Tien-Chung Hu, and Li-Shan Huang. 1996. "A Study of Variable Bandwidth Selection for Local Polynomial Regression." *Statistica Sinica*, 6(1): 113–127.

Feigenbaum, James J. 2016. "A Machine Learning Approach to Census Record Linking." Accessed March 28,.

Feldman, Naomi E., Peter Katuscak, and Laura Kawano. 2016. "Taxpayer Confusion: Evidence from the Child Tax Credit." *American Economic Review*, 106(3): 807–35.

Fellegi, Ivan P., and Alan B. Sunter. 1969. "A Theory for Record Linkage." *Journal of the American Statistical Association*, 64.

Ferrie, Joseph, and Karen Rolf. 2011. "Socioeconomic Status in Childhood and Health After Age 70: A New Longitudinal Analysis for the U.S., 1895-2005." *Explorations in Economic History*, 48(4): 445–460.

Ferrie, Joseph P. 1996. "A New Sample of Males Linked from the 1850 Public Use Micro Sample of the Federal Census of Population to the 1860 Federal Census Manuscript Schedules." *Historical Methods*, 29(4): 141–156.

Figlio, David, Jonathan Guryan, Krzysztof Karbownik, and Jeffrey Roth. 2014. "The Effects of Poor Neonatal Health on Children's Cognitive Development." *American Economic Review*, 104(12): 3921–55.

Florida Department of Education. 2020. "School District Start and End Dates, 2005-06 through 2012-13." Accessed October 11th, 2020. Available: http://www.fldoe.org/core/fileparse.php/7584/urlt/0086559-startenddates.xls.

French, Ron. 2013. "Michigan's 13,000 "Redshirt" Kindergartners." Bridge: Michigan, Lansing, MI. Accessed October 11th, 2020. Avaialble: https://www.bridgemi.com/talent-education/michigans-13000-redshirt-kindergartners.

Gans, Joshua, and Andrew Leigh. 2009. "Born on the First of July: An (Un)natural Experiment in Birth Timing." *Journal of Public Economics*, 93(1-2): 246–263.

Gauriot, Romain, and Lionel Page. 2019. "Does Success Breed Success? a Quasi-Experiment on Strategic Momentum in Dynamic Contests." *The Economic Journal*, 129(624): 3107–3136.

Gershoff, Elizabeth T., J. Lawrence Aber, C. Cybele Raver, and Mary Clare Lennon. 2007. "Income Is Not Enough: Incorporating Material Hardship Into Models of Income Associations With Parenting and Child Development." *Child Development*, 78(1): 70–95.

Goeken, R., L. Huynh, T.A. Lynch, and R. Vick. 2011. "New Methods of Census Record Linking." *Historical Methods*, 44(1): 7–14.

Goeken, R., T. Lynch, Y.N. Lee, J. Wellington, and D. Magnuson. 2017. "Evaluating the Accuracy of Linked U." In *S. Census Data: A Household Approach.* Retrieved from.

Goodman-Bacon, Andrew, and Leslie McGranahan. 2008. "How do EITC Recipients Spend Their Refunds?" *Economic Perspectives*, 32(QII): 17–32.

Greene, William H. 2008. *Econometric Analysis.* . 6th ed., New York:Pearson.

Gross, David B., and Nicholas S. Souleles. 2002. "Do Liquidity Constraints and Interest Rates Matter for Consumer Behavior? Evidence from Credit Card Data." *The Quarterly Journal of Economics*, 117(1): 149–185.

Gubbels, Jeanne, Claudia E. van der Put, and Mark Assink. 2019. "The Effectiveness of Parent Training Programs for Child Maltreatment and Their Components: A Meta-Analysis." *International Journal of Environmental Research and Public Health*, 16(13): 2404.

Guest, A.M. 1987. "Notes from the National Panel Study: Linkage and Migration in the Late Nineteenth Century." *Historical Methods*, 20(2): 63–77.

Hacker, J.David. 2010. "Decennial Life Tables for the White Population of the United States, 1790-1900." *Historical Methods*, 43(3): 45–79.

Hacker, J.David. 2013. "New Estimates of Census Coverage in the United States, 1850-1930." *Social Science History*, 37(1): 71–101.

Hahn, Jinyong, Petra Todd, and Wilbert Van der Klaauw. 2001. "Identification and Estimation of Treatment Effects with a Regression-Discontinuity Design." *Econometrica*, 69(1): 201–209.

Haider, Steven J., and Gary Solon. 2006. "Life-Cycle Variation in the Association between Current and Lifetime Earnings." *American Economic Review*, 96.

Hatton, T. 2011. "The Cliometrics of International Migration: A Survey." In *Economics and History: Surveys in Cliometrics.* , ed. L. Oxley, 187–216. London:Wiley-Blackwell.

Heckman, James J., Hidehiko Ichimura, Jeff Smith, and Petra Todd. 1998. "Characterizing Selection Bias Using Experimental Data." *Econometrica*, 66.

Heckman, J.J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica*, 47(1): 153–161.

Hornbeck, Richard, and Suresh Naidu. 2014. "When the Levee Breaks: Black Migration and Economic Development in the American South." *American Economic Review*, 104(3): 963–990.

Horowitz, Joel L., and Charles F. Manski. 1995. "Identification and Robustness with Contaminated and Corrupted Data." *Econometrica*, 63(2): 281–302.

Hout, Michael, and Avery M. Guest. 2013. "Intergenerational Occupational Mobility in Great Britain and the United States since 1850: Comment." *American Economic Review*, 103.

Hoxby, Caroline M., and George B. Bulman. 2016. "The Effects of the Tax Deduction for Postsecondary Tuition: Implications for Structuring Tax-Based Aid." *Economics of Education Review*, 51: 23–60. Access to Higher Education.

Hoynes, Hilary, Diane W. Schanzenbach, and Douglas Almond. 2016. "Long-Run Impacts of Childhood Access to the Safety Net." *American Economic Review*, 106(4): 903–34.

Hoynes, Hilary, Doug Miller, and David Simon. 2015. "Income, the Earned Income Tax Credit, and Infant Health." *American Economic Journal: Economic Policy*, 7(1): 172–211.

Huber, P.J. 1967. "The behavior of maximum likelihood estimates under nonstandard conditions." Vol. 1, 221–233.

Imbens, Guido, and Karthik Kalyanaraman. 2011. "Optimal Bandwidth Choice for the Regression Discontinuity Estimator." *The Review of Economic Studies*, 79(3): 933–959.

Inoue, Atsushi, and Gary Solon. 2010. "Two-Sample Instrumental Variables Estimators." *The Review of Economics and Statistics*, 92(3): 557–561.

Jacob, Brian A., and Lars Lefgren. 2009. "The Effect of Grade Retention on High School Completion." *American Economic Journal: Applied Economics*, 1(3): 33–58.

Jacob, Brian A., Max Kapustin, and Jens Ludwig. 2014. " The Impact of Housing Assistance on Child Outcomes: Evidence from a Randomized Housing Lottery." *The Quarterly Journal of Economics*, 130(1): 465–506.

Jaro, Matthew A. 1989. "Advances in Record Linking Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association*, 84(406): 414–420.

Kim, Gunky, and Raymond Chambers. 2012. "Regression Analysis under Probabilistic Multi-Linkage." *Statistica Neerlandica*, 66(1): 64–79.

Kleven, Henrik J., and Mazhar Waseem. 2013. "Using Notches to Uncover Optimization Frictions and Structural Elasticities: Theory and Evidence from Pakistan." *The Quarterly Journal of Economics*, 128(2): 669–723.

Kling, Jeffrey R, Jeffrey B Liebman, and Lawrence F Katz. 2007. "Experimental Analysis of Neighborhood Effects." *Econometrica*, 75(1): 83–119.

Lahiri, P., and Michael D. Larsen. 2005. "Regression Analysis with Linked Data." *Journal of the American Statistical Association*, 100(469): 222–230.

LaLumia, Sara, James M. Sallee, and Nicholas Turner. 2015. "New Evidence on Taxes and the Timing of Birth." *American Economic Journal: Economic Policy*, 7(2): 258–93.

Lavy, Victor, Giulia Lotti, and Zizhong Yan. 2020. "Empowering Mothers and Enhancing Early Childhood Investment: Effect on Adults Outcomes and Children Cognitive and Non-Cognitive Skills." *Journal of Human Resources*, 55(3).

Lee, David S., and David Card. 2008. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics*, 142(2): 655–674.

Lee, David S., and Thomas Lemieux. 2010. "Regression Discontinuity Designs in Economics." *Journal of Economic Literature*, 48(2): 281–355.

Lippold, Kye. 2019. "The Effects of the Child Tax Credit on Labor Supply." Working Paper. Accessed October 11th, 2020. Available: http://acsweb.ucsd.edu/ klippold/pdfs/Lippold_CTC_Paper.pdf.

Loeb, Susanna, and Daphna Bassok. 2007. "Early Childhood and the Achievement Gap." *Handbook of Research in Education Finance and Policy*, , ed. H.F. Ladd and E.B. Fiske, 517–534. Routledge Press.

Loken, Katrine. 2010. "Family Income and Children's Education: Using the Norwegian oil Boom As a Natural Experiment." *Labour Economics*, 17: 118–129.

Loken, Katrine V., Magne Mogstad, and Matthew Wiswall. 2012. "What Linear Estimators Miss: The Effects of Family Income on Child Outcomes." *American Economic Journal: Applied Economics*, 4(2): 1–35.

Malin, J. 1935. "The Turnover of Farm Pouplation in Kansas." *Kansas Historical*, 20: 339–372.

Manoli, Day, and Nicholas Turner. 2018. "Cash-on-Hand and College Enrollment: Evidence from Population Tax Data and the Earned Income Tax Credit." *American Economic Journal: Economic Policy*, 10(2): 242–71.

Margo, Robert A. 2016. "Obama, Katrina, and the Persistence of Racial Inequality." *Journal of Economic History*, 76(2): 301–341.

Martin, Joyce A., Brady E. Hamilton, Michelle J.K. Osterman, Anne K. Driscoll, and Patrick Drake. 2018. "Births: Final Data for 2017." Division of Vital Statistics Reports, 67(8). National Center for Health Statistics, Hyattsville, MD.

Martin, Joyce A., Brady E. Hamilton, Paul D. Sutton, Stephanie J. Ventura, T.J. Mathews, Sharon Kirmeyer, and Michelle J.K. Osterman. 2010. "Births: Final Data for 2007." Division of Vital Statistics Reports, 58(24). National Center for Health Statistics, Hyattsville, MD.

Massey, Catherine G. 2017. "Playing with matches: An assessment of accuracy in linked historical data." *Historical Methods: A Journal of Quantitative and Interdisciplinary History:1-15.*

Mayer, Susan E., Ariel Kalil, Philip Oreopoulos, and Sebastian Gallegos. 2019. "Using Behavioral Insights to Increase Parental Engagement: The Parents and Children Together Intervention." *Journal of Human Resources*, 54(4): 900–925.

Mazumder, B., and D. Aaronson. 2011. "The Impact of Rosenwald Schools on Black Achievement." *Journal of Political Economy*, 119(5): 821–888.

Mazumder, Bhashkar. 2005. "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data." *Review of Economics and Statistics*, 87(2): 235–255.

Mazumder, Bhashkar. 2015. *Estimating the Intergenerational Elasticity and Rank Association in the U.S.: Overcoming the Current Limitations of Tax Data.* Federal Reserve Bank of Chicago Working Paper.

Mazumder, Bhashkar. 2018. "Intergenerational Mobility in the United States: What We Have Learned from the PSID." *Annals of the American Academy of Political and Social Science*, 680(1): 213–234.

McCloskey, D. 2005. "The Trouble with Mathematics and Statistics in Economics." *History of Economic Ideas*, XIII, 3: 85–102.

McCrary, Justin. 2008. "Manipulation of the Running Variable in the Regression Discontinuity Design: A Density Test." *Journal of Econometrics*, 142(2): 698–714.

McGranahan, Leslie, and Diane W. Schanzenbach. 2013. "The Earned Income Tax Credit and Food Consumption Patterns." Chicago Federal Reserve WP 2013-14, Chicago, IL.

Mendenhall, Ruby, Kathryn Edin, Susan Crowley, Jennifer Sykes, Laura Tach, Katrin Kriz, and Jeffrey R. Kling. 2012. "The Role of Earned Income Tax Credit in the Budgets of Low-Income Households." *Social Service Review*, 86(3): 367–400.

Michalopoulos, Charles, Kristen Faucetta, Carolyn J. Hill, Ximena A. Portilla, Lori Burrell, Helen Lee, Anne Duggan, and Virginia Knox. 2019. "Impacts on Family Outcomes of Evidence-Based Early Childhood Home Visiting: Results from the Mother and Infant Home Visiting Program Evaluation." Office of Planning, Research, and Evaluation, Administration for Children and Families, U.S. Department of Health OPRE Report 2019-07, Washington, DC.

Michelmore, Katherine, and Susan Dynarski. 2017. "The Gap Within the Gap: Using Longitudinal Data to Understand Income Differences in Educational Outcomes." *AERA Open*, 3(1).

Michelson, M., and C.A. Knoblock. 2006. "Learning Blocking Schemes for Record Linkage."

Miller, Cynthia, Rhiannon Miller, Nandita Verma, Nadine Dechausay, Edith Yang, Timothy Rudd andJonathan Rodriguez, and Sylvie Honig. 2016. "Effects of a Modified Conditional Cash Transfer Program in Two American Cities: Findings from Family Rewards 2.0." MDRC, Washington, DC.

Milligan, Kevin, and Mark Stabile. 2009. "Child Benefits, Maternal Employment, and Children's Health: Evidence from Canadian Child Benefit Expansions." *American Economic Review*, 99(2): 128–32.

Mill, R. 2013. "Record Linkage across Historical Datasets. Inequality and Discrimination in Historical and Modern Labor Markets." Retrieved from.

Mill, R., and L.C. Stein. 2016. "Race, Skin Color, and Economic Outcomes in Early Twentieth-Century America." Retrieved from.

Modalsli, Jorgen. 2017. "Intergenerational Mobility in Norway, 1865-2011." *The Scandanavian Journal of Economics*, 119(1): 34–71.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States." *Journal of Economic Perspectives*, 25(3): 173–96.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspectives*, 31(2): 87–106.

National Center for Education Statistics. 2000. "Children Who Enter Kindergarten Late or Repeat Kindergarten: Their Characteristics and Later School Performance." U.S. Department of Education: Office of Educational Research and Improvement NCES Report 2000-039, Washington, DC.

National Office of Vital Statistics. 1948.

Neugart, Michael, and Henry Ohlsson. 2013. "Economic incentives and the Timing of Births: Evidence from the German Parental Benefit Reform of 2007." *Journal of Population Economics*, 26(1): 87–108.

Nix, Emily, and Nancy Qian. 2015. "The Fluidity of Race: 'Passing' in the United States, 1880-1940."

Peixoto, Francisco, Vera Monteiro, Lourdes Mata, Cristina Sanches, Joana Pipa, and Leandro S. Almeida. 2016. ""To be or not to be Retained ... That's the Question!" Retention, Self-esteem, Self-concept, Achievement Goals, and Grades." *Frontiers in Psychology*, 7: 1550.

Ramnath, Shanthi P., and Patricia K. Tong. 2017. "The Persistent Reduction in Poverty from Filing a Tax Return." *American Economic Journal: Economic Policy*, 9(4): 367–94.

Ratcliffe, Caroline. 2019. "Child Poverty and Adult Success." Urban Institute, Washington, DC. Accessed October 11th, 2020. Available: https://www.urban.org/sites/default/files/publication/65766/2000369-Child-Poverty-and-Adult-Success.pdf.

Reardon, Sean F. 2011. "The Widening Academic Achievement Gap Between the Rich and the Poor: New Evidence and Possible Explanations." *Whither opportunity? Rising Inequality, Schools, and Children's Life Chances*, , ed. G. J. Duncan and R. J. Murnane, 91–116. Russell Sage Foundation.

Rhode, P., and A.L. Olmstead. 2015. *Arresting Contagion: Science, Policy and Conflicts over Animal Disease Control.* Cambridge, MA:Harvard Univ. Press.

Roderick, Melissa, and Jenny Nagaoka. 2005. "Retention Under Chicago's High-Stakes Testing Program: Helpful, Harmful, or Harmless?" *Educational Evaluation and Policy Analysis*, 27(4): 309–340.

Rossin-Slater, Maya. 2013. "WIC in Your Neighborhood: New Evidence on the Impacts of Geographic Access to Clinics." *Journal of Public Economics*, 102: 51–69.

Ruggles, S. 2006. "Linking Historical Censuses: A New Approach." *History and Computing*, 14(1-2): 213–224.

Ruggles, S., K. Genadek, J. Grover, and M. Sobek. 2015. *Integrated Public Use Microdata Series (Version 6.0) [Machine-Readable database.* Minneapolis:University of Minnesota.

Ruggles, Steven. 2011. "Intergenerational Coredisence and Family Transitions in the United States, 1850-1880." *Journal of Marriage and the Family*, 73(1): 138–148.

Ruggles, Steven, Catherine A. Fitch, and Evan Roberts. 2018. "Historical Census Record Linkage." *Annual Review of Sociology*, 44.

Saez, Emmanuel. 2010. "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy*, 2(3): 180–212.

Saperstein, Aliya, and Aaron Gullickson. 2013. "A Mulatto Escape Hatch? Examining Evidence of U.S. Racial and Social Mobility in the Jim Crow Era." *Demography*, 50.

Scheuren, Fritz, and William E. Winkler. 1993. "Regression analysis of data files that are computer matched." *Survey methodology*, 19(1): 39–58.

Scholz, John Karl. 1994. "The Earned Income Tax Credit: Participation, Compliance, and Anti-Poverty Effectiveness." *National Tax Journal*, 47(1): 63–87.

Schwager, Mahna T., Douglas E. Mitchell, Tedi K. Mitchell, and Jeffrey B. Hecht. 1992. "How School District Policy Influences Grade Level Retention in Elementary Schools." *Educational Evaluation and Policy Analysis*, 14(4): 421–438.

Schwerdt, Guido, Martin R. West, and Marcus A. Winters. 2017. "The Effects of Test-Based Retention on Student Outcomes over Time: Regression Discontinuity Evidence from Florida." *Journal of Public Economics*, 152(C): 154–169.

Shea, John. 2000. "Does Parents' Money Matter?" *Journal of Public Economics*, 77(2): 155–184.

Solon, Gary. 1992. "Intergenerational Income Mobility in the United States." *American Economic Review*, 82(3): 393–408.

Solon, Gary. 1999. "Intergenerational Mobility in the Labor Market." In *Handbook of Labor Economics.* , ed. Orley Ashenfelter and David Card, 1761–1800. Amsterdam:Elsevier.

Solon, G., S.J. Haider, and Wooldridge. n.d..

Stackhouse, Herbert F., and Sarah Brady. 2003*a*. *Census 2000 Evaluation A.7.a: Census 2000 Mail Response Rates.* Vol. 1, Washington, DC:U.S. Census Bureau.

Stackhouse, Herbert F., and Sarah Brady. 2003*b*. *Census 2000 Evaluation A.7.b: Census 2000 Mail Return Rates.* Vol. 1, Washington, DC:U.S. Census Bureau.

Stark, Patrick, Amber M. Noel, and Joel McFarland. 2012. "Trends in High School Dropout and Completion Rates in the United States: 1972-2015." National Center for Education Statistics NCES Report 2015-015, Washington D.C.

Steckel, R. 1988. "Census Matching and Migration: A Research Strategy." *Historical Methods*, 21: 52–60.

Stephens, Jr., Melvin, and Takashi Unayama. 2017.

Tamborini, Christopher, ChangHwan Kim, and Arthur Sakamoto. 2015. "Education and Lifetime Earnings in the United States." *Demography*, 52: 1383–1407.

Thernstrom, S. 1964. *Poverty and Progress: Social Mobility in a Nineteenth Century City.* Cambridge:Harvard University Press.

U.S. Census Bureau. 2009. *U.S. Census Bureau, History: 2000 Census of Population and Housing.* Vol. 1, Washington, DC.

U.S. Census Bureau. 2014. *American Community Survey Design and Methodology.* Washington, DC.

U.S. Census Bureau. 2019. "American Community Survey: Accuracy of Data (2019)." U.S. Census Bureau, Washington, DC. Accessed October 11th, 2020. Available: https://www2.census.gov/programs-surveys/acs/tech_docs/accuracy/ACS_Accuracy_of_Data_2019.pdf.

U.S. Government Accountability Office. 1993. "Earned Income Tax Credit: Design and Administration Could Be Improved." Government Accountability Office GAO/GGD-93-146, Washington, DC.

U.S. Government Accountability Office. 2001. "Earned Income Tax Credit Eligibility and Participation." Government Accountability Office GAO-02-290R, Washington, DC.

Wadsworth, Martha E., Tali Raviv, Bruce E. Compas, and Jennifer K. Connor-Smith. 2005. “Parent and Adolescent Responses to Poverty-Related Stress: Tests of Mediated and Moderated Coping Models.” *Journal of Child and Family Studies*, 14: 283–298.

Ward, Zachary. 2019. “Intergenerational Mobility in American History: Accounting for Race and Measurement Error.”

West, Kirsten K., and J.Gregory Robinson. 1999. “What do we know about the Undercount of Children?” U.S.” *Census Bureau Population Division Working Paper*, 39.

West, Martin R. 2012. “Is Retaining Students in the Early Grades Self-Defeating?” Brookings Institution, Washington, D.C. Accessed October 11th, 2020. Avaialble: https://www.brookings.edu/research/is-retaining-students-in-the-early-grades-self-defeating/.

White, H. 1980. “A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity.” *Econometrica*, 48: 817–830.

Wingender, Philippe, and Sara LaLumia. 2017. “Income Effects on Maternal Labor Supply: Evidence from Child-Related Tax Benefits.” *National Tax Journal*, 70(1): 11–52.

Winkler, William E. 2006. “Overview of Record Linkage and Current Research Directions.” *Research Report Series*. Statistics 2006 (2).

Wisselgren, Maria J., Soren Edvinsson, Mats Berggren, and Maria Larsson. 2014. “Testing Methods of Record Linkage on Swedish Censuses.” *Historical Methods*, 47(3): 138–151.

Xia, Nailing, and Sheila Nataraj Kirby. 2009. “Retaining Students in Grade: A Literature Review of the Effects of Retention on Students’ Academic and Nonacademic Outcomes.” RAND Corporation, Santa Monica, CA.

Xie, Yue, and Alexandra Killewald. 2013. “Intergenerational Occupational Mobility in Britain and the U.S. since 1850: Comment.” *American Economic Review*, 103.

Zimmerman, David J. 1992. “Regression Toward Mediocrity in Economic Stature.” *American Economic Review*, 82(3): 409–429.