The Effects of Education on Mortality:
Evidence from a Representative Sample of American Twins, Siblings, and Neighbors

ABSTRACT

Does education change people's lives in a way that delays mortality? Or is education primarily a proxy for unobserved endowments that promote longevity? Most scholars conclude that the former is true, but recent evidence based on Danish twin data calls this conclusion into question. Unfortunately, these potentially field-changing finding—that obtaining additional schooling has no independent effect on survival net of other hard-to-observe characteristics—has not yet been subject to replication outside Scandinavia. We produce the first U.S.-based estimates of the effects of education on mortality using a representative panel of male twins drawn from linked complete-count Census and death records. For comparison purposes, and to shed additional light on the roles that neighborhood, family, and genetic factors play in confounding associations between education and mortality, we also produce parallel estimates of the education-mortality relationship using data on (1) unrelated males who lived in different neighborhoods during childhood; (2) unrelated males who shared the same neighborhood growing up; and (3) non-twin siblings who shared the same family environment but whose genetic endowments vary to a greater degree.

The association between educational attainment and adult mortality in modern-day societies is well known and virtually universally observed (Elo and Preston 1996; Hummer and Hernandez 2013; Hummer and Lariscy 2011; Kitagawa and Hauser 1968; Kitagawa and Hauser 1973; Lleras-Muney 2005; Preston and Taubman 1994). Sizable educational gradients in individuals' age at and cause of death have been detected across birth cohorts, in different population groups, and in many social and institutional contexts (Hayward, Hummer and Sasson 2015). What is less clear is *why* these educational gradients exist. Are educational attainment and human survival etiologically linked, such that obtaining more schooling causes people to enjoy lower levels of mortality and longer lives? Or are the two variables related to one another because common endowments influence both, generating a spurious (or partially spurious) association?

Answering these questions is of profound scientific and policy significance. The magnitude of the association between education and mortality is large (Hummer and Lariscy 2011). If education causally affects mortality, investments in schooling could be an efficient and cost-effective means to reducing the "longevity penalty" that some groups face. On the other hand, if obtaining more schooling on its own does not cause people to live longer, then even the best-intentioned efforts to reduce mortality differentials would be of little value if education is treated as the policy lever (Hummer and Hernandez 2013). In the first scenario, education is a causal variable and could be the target of longevity-enhancing interventions. In the second, it is (at least partly) a proxy for the actual drivers of mortality.

In this paper, we use probabilistic record-linking procedures to create a set of nationally-representative longitudinal samples with linked mortality information. We then use these samples to derive causal estimates of the relationship between education and longevity. The unique data at our disposal, which we describe in more detail below, allow us to build up from a conventional covariate adjustment design to a more strenuous test of the causal relationship between education and mortality that accounts for hard-to-observe confounds at several theoretically-relevant levels of analysis. By comparing estimates obtained using different estimation strategies (and across different strategically selected subsamples of the adult population), we can (1) assess the degree to which specific early-life endowments confound the association between education and mortality and (2) evaluate variation in education effects across subgroups defined by their sociodemographic and geographic characteristics. We know of no prior work in the U.S. (or elsewhere) that has carried out such an exercise.

Background

Understanding the origins of educational gradients in mortality (and health more generally) has long been a priority on America's research and public health agenda. Causal accounts have traditionally focused on the importance of resources, skills, and knowledge (Baker et al. 2011; Mirowsky and Ross 2003; Rodgers et al. 2013; Ross and Wu 1996; Schoeni et al. 2010) acquired through education and then translated into better health behaviors and outcomes (Denney et al. 2010; Hayward et al. 2014; Hummer and Hernandez 2013). The basic conceptual model that underlies this account is summarized graphically in Figure 1, with causal arrows connecting education, *E*, to a series of mediating variables, *R* (economic and social resources), *S* (cognitive skills), and *K* (knowledge), which in turn influence health (through a variety of more proximate channels not shown) and mortality, *M*.

Associations between education and mortality could also arise if the two variables share common causes, inducing a spurious (or partially spurious) relationship between *E* and *M*

(Behrman et al. 2011). Potential confounds include people's social or economic background, their intelligence, their early-life health, and any other hard-to-observe endowments or contextual exposures that jointly predict educational attainment and survival (Hayward et al. 2014). Adding these variables—labeled *B*, *I*, *H*, and *Z*, respectively—to the causal diagram specified above, as in Figure 2, opens a series of backdoor (i.e., non-causal) paths that connect *E* to *M* ($E \leftarrow B \rightarrow M$, $E \leftarrow I \rightarrow M$, $E \leftarrow H \rightarrow M$, and $E \leftarrow Z \rightarrow M$), raising doubts about the causal nature of the association between the two variables. This concern has led some researchers to question whether education *causally* affects mortality (in the classic counterfactual sense of the word), or whether the observed association is merely the end state of a more complicated sequence of selection processes (Behrman et al. 2011; Gottfredson and Deary 2004).

Empirically adjudicating between these perspectives is challenging. Most work has relied on covariate adjustments to rule out possible confounders and isolate (presumably causal) effects (Kitigawa and Hauser 1973). Findings from these analyses have shown that the association between education and mortality is robust to the inclusion of several covariates, including measures of intelligence (Link et al. 2008), race (Montez et al. 2011), childhood socioeconomic status (Hayward and Gorman 2004; Luo and Waite 2005; Montez et al. 2011), and early-life health endowments (Blackwell et al. 2001; Haas 2007; Montez and Hayward 2011). If these statistical controls are enough to eliminate the threat posed by omitted variable bias (i.e., all of the backdoor paths running from *E* to *M* can be closed by conditioning on observed confounds), then the parameter of interest (the effect of education) is identified and the conditional association between education and mortality can be said to be causal.

Efforts to validate this assumption have taken several forms. One approach is to instrument education using historical information about compulsory school attendance and/or child labor laws, mimicking an experimental setup where exogenous factors sort individuals into different levels of the treatment (educational attainment). Although early proponents of this strategy observed significant (and substantively large) effects associated with education (Lleras-Muney 2005), attempts to replicate this finding have frequently failed (see, e.g., Black et al. 2015; Mazumder 2008). One reason could be the strength of the instrument: Compulsory schooling laws are, in many cases, only weakly related to variation in educational attainment (and are only relevant for the subsample of students who were induced to obtain additional schooling), making it difficult to estimate effects with sufficient precision (or to make claims about effects for students not on the margins) (Fletcher 2015). Studies that use birth registry data from outside the U.S. have sought to circumvent this problem by analyzing larger samples (Meghir et al. 2017; Lager and Torssander 2012), but findings there have been mixed as well. In some cases, researchers have observed sizable declines in mortality for birth cohorts (or jurisdictions) that were compelled to attend school for an additional year (Fischer et al. 2013; van Kippersluis et al. 2011); in other cases, they have not (Braakman 2011; Clark and Royer 2010; Lager and Torssander 2012; Meghir et al. 2018). Scholars have speculated that these discrepancies could be due, in part, to heterogeneity in the effects of education across birth cohorts and/or social and political contexts (Hayward et al. 2014), but even this hypothesis has been difficult to confirm.

As an alternative to this approach, a small but growing number of studies have turned to within-twin pair comparisons as a way to "difference out" observed and unobserved factors (e.g., *B*, *I*, and *Z*) that could confound the association between education and mortality (see, e.g., Behrman 2011; Lundborg et al. 2016; Madsen 2010). Twins—even those who end up with different levels of schooling—experience very similar social, economic, family, school, neighborhood and other environmental exposures and have identical (in the case of monozygotic or MZ twins) or similar (in the case of dizygotic or DZ twins) genes. If educational attainment is associated with mortality among pairs of twins that are concordant (or mostly concordant) with respect to these endowments, but discordant with respect to their educational attainment, the association is (conditional on several identifying assumptions) less likely to be spurious. All shared genetic and environmental exposures fall out of the model, providing (arguably) cleaner estimates of the independent effect associated with obtaining higher levels of education.

Applications of this strategy have produced intriguing, and sometimes surprising, results. Using a large population-based data set from Denmark that included just over 2,500 identical (MZ) twin pairs born between 1921 and 1950, Behrman et al. (2011) showed that the causal effect of education on mortality in Denmark *is reduced to zero* when comparing the mortality outcomes of MZ twins who are discordant on education. That the same was not true for pairs of unrelated individuals suggests that shared early-life endowments and exposures (common within pairs of identical twins but not within pairs of unrelated adults) may explain, or partially explain, the existence of educational gradients in mortality. This inference is broadly consistent with a *non*-causal interpretation of the diagram presented in Figure 2. Other researchers who have used the same Danish data have generally reached similar conclusions, observing null or attenuated effects when modeling within-twin pair differences in mortality or related health outcomes (Madsen et al. 2010; Osler, McGue and Christensen 2007).

We believe these findings are important and provocative, but we also see reasons for skepticism. First, it is unclear whether similar findings would hold if the same twin-differencing models were fit using data from outside of the distinctive Danish context (Hayward, Hummer and Sasson 2015; Lundborg, Lyttkens and Nystedt 2016). It could be the case that Denmark's set of social and educational policies render the education-mortality relationship less important than in the U.S. or other Western countries (Lundborg, Lyttkens and Nystedt 2016), where the social safety net is less secure. Until now, it has been impossible to replicate Behrman et al.'s (2011) results using U.S. data because there are no large, nationally-representative samples of U.S. twins with requisite information about education and mortality. Current U.S.-based twin data repositories (e.g., the NAS-NRC Twin Registry of WWII Military Veterans and the Minnesota Twin Registry) have been used for a variety of research on *health* gradients (Amin, Behrman and Kohler 2015), but these studies pertain to particular sub-populations (e.g., WWII veterans) or to people in particular geographic regions (e.g., Minnesota or Southern California), and only the NAS-NSF data include mortality information for most respondents.

Second, it is not clear that effects estimated using twin-differencing models pertain equally within all population subgroups. Prior twin-based estimates of the effects of education on health and mortality can be thought of as estimates of the *average* treatment effect (or ATE). A recent study by Heyward et al. (Hayward, Hummer and Sasson 2015), however, suggests that the association between education and mortality may be stronger for certain segments of the population; in particular, for men, for whites, and for younger adults. Under these conditions, estimates of the ATE of education on mortality, which represent a weighted average of all group-specific estimates, could mask meaningful variation within the adult population. This in turn suggests that, even if Behrman et al.'s (2011) findings hold *on average* across the U.S. population, they may not hold within specific subgroups of individuals.

Finally, it is not known *why* twin-based studies have produced results that diverge from findings obtained using more conventional covariate adjustment designs and/or alternative identification strategies. In their conclusion, Behrman and colleagues (2011) write that education may serve as "a marker for parental family and individual-specific endowments that are uncontrolled in the usual estimates" (p. 1367), but they are unable to provide additional information about what those endowments might be. Because twins share the same (or most of the same) genetic, family, neighborhood, and school characteristics, finely grained analyses of individual confounds (i.e., the additional variables included in Figure 2) are generally not feasible. While this does not negate the overall contribution of their research, it does lead to a coarse assessment of the underlying causal model.

RESEARCH DESIGN

Data and Measures

We address two of the above listed issues using a unique and (until now) untapped data resource: the digitized complete-count U.S. Censuses for 1920 and 1940. With support from NSF and NIH, and in collaboration with Ancestry.com, the Minnesota Population Center is finishing work on complete-count versions of the 1850-1940 census files (Ruggles 2014). From the 1920 U.S. Census, we have drawn samples of males aged 0-4, representing four different groups: (1) 100,000 randomly selected males ("random"); (2) 50,000 randomly selected males, paired with the next enumerated age-appropriate male, conditional on different households but the same enumeration district ("neighbors"); (3) 50,000 singleton males from households with at least two age-appropriate singleton brothers, each paired with one randomly selected brother ("siblings");

and (4) all same-sex male multiple births ("twins").[1] Using probabilistic record linkage methods, these samples were linked to the 1940 U.S. Census, from which information on educational attainment was obtained. Information on the individual's age at death was then obtained by linking the 1920-1940 U.S. Census samples to death records in the Social Security Death Master File (SSDMF). In doing so, we are able to produce nationally-representative samples that are sufficiently large to identify even modestly-sized education effects, allowing for an original and sophisticated assessment of the magnitude of the causal effects of education on men's mortality, variation in those effects, and the specific factors (if any) that confound their estimation.

Below we describe in greater detail our strategy for identifying male panel members in the 1920 U.S. Census; for linking panel members across the 1920 U.S. Census, the 1940 U.S. Census, and mortality records; and for estimating he effects of education on mortality.

Linking Records

From the 1920 U.S. Census, we selected the "random" sample by taking a random draw of

[1] We will not be able to distinguish MZ from DZ twins in our analyses, but a publication from the period in question (Hamlett 1934) estimates that among same-sex twin pairs, 50% will be MZ. Contemporary twin databases, such as the Swedish and Danish twin registry (Skytthe et al 2011, Ljungqvist et al 1998), however, suggest substantially lower shares of MZ twins among same-sex twin pairs, at around 35%. These discrepancies could reflect differences across time and context in the composition of twin births. Either way, our inability to differentiate between MZ and DZ twins will tend to bias our results *in favor* of inferring effects of education on mortality, due to our inability to completely account for unobserved genetic confounds.

100,000 individuals from the universe of males who at the time of the census were between the ages of 0 and 4. We selected the “neighbor” sample by initially drawing a sample of 50,000 males from the same number of different households; we then selected the next enumerated age-appropriate male, provided that he was residing in the same enumeration district but in a different household as the index individual. Due to some index individuals being in the last household for a given enumeration district, the sample amounts to 99,545 individuals. The “sibling” sample was created by first selecting households with at least two singleton males between the ages of 0 and 4. In order to ascertain that the siblings indeed were full siblings, we restricted the sample to individuals with unambiguous links to the same mother and father. From these households, 50,000 were again randomly selected, with a limit of one individual per household. We then randomly selected one brother per index individual, resulting in a sample of 100,000 individuals in 50,000 brother pairs. Finally, for “twins,” we selected all multiple births in the designated birth cohorts, with the stipulation that the cases have (1) unambiguous links to the same mother and father (thus not being stepchildren) and (2) the same recorded age in years at the time of enumeration. While the census does contain the individual’s age in years as well as month for the birth cohorts we included, the amount of missing values on the latter variable was substantial. We are able to recover exact birthdays from death records, however, which we use in auxiliary analyses to establish the robustness of our results; these analyses are described in more detail below. Our initial sample of twins includes approximately 81,000 individuals, belonging to over 40,000 multiple birth pairs. Across all samples, we have restricted the population to individuals born in the United States.

Linking 1920 to 1940 U.S. Census Records

Linking cases from our four sub-samples to the 1940 census requires that we first define the population of potential matches. In order to make the process more manageable and computationally tractable, we explicitly assume that sex and place of birth is reported accurately as well as consistently over time.[2] As an example, when attempting to find Michael Corcoran, male, born in Massachusetts according to the 1920 Census, we limit the population of potential matches in 1940 to males who were reported as being born in that same state in the 20 years later. Since age (and year of birth) is reported in the censuses rather than date of birth, and because this type of information may have been reported less accurately at the time, we allow for deviations in birth year across data sources. In particular, we stipulate that birth years must be within +/- five years, implying that each unique individual who, according to the 1920 Census, was born in 1919 will be compared and possibly linked to individuals who in the 1940 Census, conditional on sex and place of birth being the same, were recorded as being born between 1914 and 1924. Allowing for a broad a range of potential matches has advantages as well as disadvantages. One advantage is that we are able to successfully link individuals for whom the year of birth in either census was reported or digitized incorrectly by more than a few years. It needs to be emphasized that the census data on which the results of this paper are based contain errors. Considering the data generating process makes it easy to understand why this is the case, between reporting or enumerator error at the time of the census and poor quality digital images of the enumeration form, as well as errors on the part of personnel attempting to correctly

---

[2] The place of birth assumption is probably not entirely accurate, but the implications thereof should not be important.

decipher hand writing of the era. One disadvantage of allowing for a larger window in terms of year of birth is the increased risk of false positives. Continuing the example from above, if the hypothetical Michael Corcoran, born in 1919 and observed in the 1920 Census, dies at the age of two, he will obviously not be enumerated in the 1940 Census. If his parents decide to have another child, born in 1924 and also name him Michael, this identically named but obviously different individual would be one of the candidates for the link to 1940. More generally, the wider the birth year window, the larger the pool of potential matches and, at the same time, the higher the probability of (1) finding the right individual *and* (2) making an incorrect link. This is an issue we will return to later in the paper.

Our linking algorithm, following Feigenbaum (2016), begins by extracting a subsample of the population to be linked manually; this subsample provides "training data." We use the training data to calibrate the algorithm and to evaluate its performance. To start, we randomly selected 2,500 individuals from the 1920 sample (representing all four subsamples); manually linked records to the relevant universe of possible 1940 Census matches; and then restricted the sample to matches where the name similarity scores (using the Jaro-Winkler algorithm) were at least 0.8. This results in a sample of 2,272 individuals, or 91 percent of the 2,500 we began with. Various methods for manually declaring matches have been proposed, i.e., when individual *i* from the 1920 Census is confidently declared to be the same individual as individual *y* in the 1940 Census. We use a "logical" method, where we declare matches when all information on names match, and when there are no other possible matches available. When names are misspelled, we exercise more caution in declaring matches, but nevertheless try to be logical and consistent. In the training data, we manually declare 32.3 percent of the sample as uniquely matched across the 1920 and 1940 censuses. To determine the most appropriate thresholds to use for matches in the full data set, we fit a simple probit model, where the variable containing manually declared matches was regressed on a selection of individual-level predictors. In training the algorithm, we repeatedly split the training data into two, estimating the model on one half and testing how it performs on the other half, where the "truth" is known. The algorithm declares a unique match based on (1) the greatest similarity between any 1-to-1 match within any given 1920 individual (technically the predicted probability based on the probit regression estimates) and (2) the relative difference between the best and the second best possible match within any given 1920 individual. By looping multiple times over a range of realistic values on both parameters, and provided that the training data resembles the population it is meant to represent, we were able to choose values on both that optimized the overall performance of the algorithm. Here, our main objectives were to minimize the presence of false positives (incorrectly matched cases) while at the same time maximizing the amount of true positives (correctly matched cases) and true negatives (correctly unmatched cases).

In selecting thresholds for declaring matches in our data, we use the Matthew's Correlation Coefficient (MCC), which is an especially useful measurement for two-class data where the classes are not very balanced (Chicco 2017). This is definitely the case in our situation, where the 2,272 individuals in the training data are, on average, linked to 40.5 individuals in the 1940 Census and are thus represented by a data set of 92,000 observations. In the data set, 734 observations (less than 1 percent) were declared to be a match. The MCC, in Eq. (1) below, compares the predictions of the algorithm to all possible outcomes (true/false positives/negatives) and provides a single metric (ranging from -1 to +1) to be used to select which thresholds to use. The formula is as follows, where TP represents true positive, TN represents true negative, FP represents false positive, and FN represents false negative:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \tag{1}$$

After running multiple loops over all relevant potential threshold values, the mean optimized MCC value amounts to 76.5. Table 1 below shows the evolution of the sample, from the original 1920 population and through determining unique links in the 1940 Census. Note that the table displays the share of confidently linked individuals across the sample, not conditioning on (for the neighbors, siblings and twins) at least two members within each unit being confidently linked. Naturally, this further diminishes the sample size.

Linking Census Records to Mortality Records

The strategy for linking the 1920-1940 sample to mortality records proceeds in a similar manner, with some slight differences due to differences in data availability. Most importantly, since the Social Security Death Master File (SSDMF) does not contain information on state of birth, nor on sex, the linkage requirements are somewhat less strict. In addition, while deaths are recorded from as early as the 1940s, it is only from well into the 1960s that the annual number of recorded deaths stabilize at a consistent level, at around 2,000,000 deaths annually. We are therefore *de facto* conditioning on survival until a point in time that is defined in a somewhat fuzzy way. The last recorded death in the file at our disposal occurred in December 2013. Another issue is that the file does not contain all deaths of individuals with social security numbers. Manual searches of Social Security claims and death records on Ancestry.com revealed instances where members of our sample had died, but are absent from the SSDMF. Our manual lookups make us confident that this is not a pervasive phenomenon, but it nevertheless deserves to be mentioned.

After observations with unambiguously female names were dropped from the SSDMF, individuals linked across 1920 and 1940 were linked to the SSDMF population on year of birth (+/- 5 years, as before). Since no restriction was possible on place of birth, the universe of possible matches becomes substantially larger.

The training data consists of 1,200 individuals from the linked 1920-1940 census sample who were linked to the universe of potential SSDMF matches. Since the universe of potential matches now includes age-appropriate individuals from the entire United States, the mean number of possible matches is considerably higher, at 423. Despite this, 36 percent of the individuals were uniquely linked to a record in the SSDMF, following the same linking logic as before. When calibrating the algorithm used to assign links in the main data, we fit a similarly specified probit model that also included a variable measuring the euclidian distance between the individual's state of birth and the state where the potential match obtained their Social Security number.

The optimal MCC when linking the two datasets is 81.6, and Table 2, below, displays the final part of the linking procedure, thus the number of successfully linked individuals across all sources.

Study Sample

Our study sample conditions on being linked across all three data sources and, for all groups but the random sample, conditions on at least two individuals belonging to each clustering unit. This restriction entails dropping cases in incomplete twin, sibling, or neighbor pairs. We also drop a small subset of pairs where an individual erroneously has been digitized as two unique individuals (with the same name and birth year). For the twins, displayed in Table 3, this restriction results in the sample size dropping further, to 2,036 unique individuals, in column [2].

In the last step of sample selection, we drop pairs where at least one member of the pair is missing information on education. Furthermore, our methods of analysis (outlined in the next section) for twins, siblings and neighbors rely on within-cluster variation in both the educational attainment and the age at death variable. Consequently, a last sample restriction is that we drop twin/sibling/neighbor units that are not discordant on education, further diminishing the sample size to 1,976 individuals, in column [3].

One concern is that sample selection—occurring either through incomplete record linkage or missing data—results in an analytic sample that differs from the target population in nontrivial ways. Table 3 indeed shows differences by race and geographic region in the likelihood or remaining "in sample" after our various selection filters are in place. Non-white individuals and individuals born in the south are less likely to be in the linked sample than in the original 1920 sample, with the opposite applying to whites and individuals from the northeastern part of the country. Apart from this, there does not seem to be any noteworthy selection on father's occupation, nor on whether the respondent's parents were native born.

Empirical Strategy

To estimate the effects of educational attainment on mortality, we use a standard fixed effects specification for within-pair estimation (we refer to twins here for convenience, but the same models will be fit for non-twin pairs as well, as we note below):

$$M_{ij} = \alpha + \beta S_{ij} + C_j + G_j + \varepsilon_{ij}, \tag{2}$$

where M_{ij} is a continuous measure of age at death for individual i (i = 1, 2) in twinship j (j = 1, 2, …, N); C_j is a measure of unobserved contextual characteristics (e.g., family, peer group, or neighborhood attributes); G_j is a measure of unobserved genetic endowments; and ε_{ij} is a random individual-level error term that is assumed to be uncorrelated with the other explanatory variables in the model.[3] In a within-twin-pair model, the unobserved components in Eq. (2) are "controlled away" by modeling differences within pairs of identical or fraternal twins:

$$M_{1j} - M_{2j} = \Delta M_j = \beta \Delta S_j + \Delta \varepsilon_j, \qquad (3)$$

where the Δ's represent differences between variables for the jth twin pair (i.e., $S_{1j} - S_{2j}$). This approach eliminates the effects of unobserved contextual characteristics (C_j)—since the vast majority of twins experience the same family, school, and neighborhood environments while growing up. It also (at least partially) eliminates the effects of unobserved genetic endowments (G_j)—since MZ twins share 100% of their genes at birth and DZ twins, like non-twin siblings, share 50% of their genes on average.

The modeling strategy described above can be modified in three respects to assess the degree (and nature) of omitted variable biases in prior U.S.-based work that uses data on unrelated individuals and more standard estimation procedures. First, we can estimate OLS models of age at death for our sample of unrelated males who live in different neighborhoods. In these analyses, we include covariates for family socioeconomic origins, family structure and composition, and geography; we expect the results to reproduce findings of prior work. Second,

---

[3] Supplementary analyses using survival analysis (i.e., stratified partial likelihood models) in place of the within-pair fixed effects estimator described above produced substantively similar results.

we can estimate within-neighborhood fixed effects models for our sample of unrelated pairs of males who live in the same neighborhood. These models allow us to consider (at least tentatively, for reasons we discuss later) the degree to which the association between education and mortality is confounded by geographic and neighborhood factors. Third, we can estimate within-family fixed effects models for our sample of pairs of male-male non-twin siblings. These models assess the degree to which the association between education and mortality is confounded by shared environmental and genetic conditions, but it is a less stringent test than the within twin pair analyses because non-twin siblings generally share fewer environmental and genetic endowments. Together with the estimates obtained using our sample of twins, these analyses will provide useful information about the magnitude of education effects on mortality and the role played by different confounds.

RESULTS

In Table 4 we present key descriptive statistics for education and mortality variables for each of the four analytic subsamples: Unrelated individuals ("random"), unrelated neighbors ("neighbors"), non-twin siblings ("siblings"), and twins. The mean years of education is between 10.27 and 10.59 across analytic samples. Within pairs, the rate of educational discordance—the percentage of pairs in which the two differed in their years of schooling completed—was high (77%) among neighbors, lower among siblings (63%), and lowest among twins (38%). Likewise, the within-pair absolute difference in years of schooling was highest among neighbors (2.4 years), lower among siblings (1.5 years), and lowest among twins (1.0 years). The mean age at death is between 73.01 and 73.92 across the four groups.

In Table 5 we present estimates of the effect of education—expressed as years of schooling completed—on age at death. In the four leftmost columns of results we present unpaired models for unrelated people, neighbors, siblings, and twins. That is, we treat people in these groups as individuals, and ignore pair structures; these models adjust for the race, family socioeconomic, parental nativity, region, and age variables listed in Table 2. As expected, across all four groups there is a positive and significant effect of education on age at death. Each additional year of schooling leads to between 0.30 and 0.52 more years of life. Importantly, this means that—in an unpaired model—twins who complete more schooling live longer. The estimated effect is actually larger among neighbors, siblings, and twins than among unrelated people.

In the right three columns of results in Table 5 we present paired models—corresponding to Eq. 3 above—for pairs of unrelated neighbors, non-twin siblings, and twins. The model for neighbors adjusts for the same covariates as in the unpaired models (with the exception of state of residence); those for siblings and twins do not (because those things fall out of the model since siblings and twins are equal on all of those covariates). For non-twin siblings, the within-pair estimate of the effect of years of schooling on age at death (0.55) is about the same as the unpaired estimate (0.52). However, among pairs of twins, there is no statistically significant effect of years of schooling on age at death. Twins who complete additional years of schooling do not live longer (or shorter) than their twin pair with fewer years of schooling.

In Table 6 we repeat the analyses in Table 5 using a categorical operationalization of education. Here, we classify people as having completed fewer than 12 years schooling, exactly 12 years of schooling, or more than 12 years of schooling. That is, we classify people as either not completing secondary school, completing it but going no further, or completing education beyond secondary school. In the unpaired models, there are sizable and significant effects of education on age at death. Compared to those who did not complete secondary school, those who only completed it lived between 1.64 years longer (for unrelated people) and 2.53 years longer (for non-twin siblings).

In the right three columns of Table 6 we present pairs models for neighbors, siblings, and twins using this categorical operationalization of education. Among siblings and twins, the effects of completing more than 12 years of school remain significant but are modestly smaller in magnitude; the effects of completing only secondary school are significant for siblings but not for twins. In other words, even among pairs of twins, individuals who went beyond secondary school lived longer (almost 4 years longer) than those who did not complete secondary school. However, note that only 36 pairs of twins are discordant in this way; the vast majority of discordant pairs have one twin with fewer than 12 years of schooling and one with exactly 12 years of schooling. For that contrast, the coefficient is not significant (not is a joint test of the two dummies).

Limitations and Robustness Checks

A major concern about research based on linked historical records has to do with the quality of the links. This limitation is inherent in the process—we ultimate have no way of knowing with complete certainty that we have accurately linked individuals. However, to assess the robustness of our twin pair results, we re-estimated our models in two ways designed to increase our confidence in the quality of our links.

First, we estimated twin pair models using data only on the 680 twin pairs who were confirmed to have exactly the same birthdate in Social Security records. Those with different dates are probably still the correct individuals—dates in both the 1920 Census and Social Security records are subject to reporting errors—but we are less certain the quality of our links in those cases. As shown in Table 7 (which uses a continuous measure of education) and Table 8 (which uses a categorical measure) our results for twins are essentially unchanged when we limit the sample in this way.

Second, we re-estimated all of the models in Table 5 after dropping records for which we were less certain of the quality of the link. Recall that we used a probabilistic record linkage algorithm. A product of this algorithm is a score assigned to each possible link that described the probability of a true match. For this analysis, we dropped cases in which that probability was relatively low; for neighbors, siblings, and twins we dropped entire pairs in which either member's probability was relatively low. As shown in Table 9, however, the results are basically unchanged (as compared to Table 3). In unpaired models, years of schooling has a sizable and significant effect on age at death. In paired models, that effect persists for siblings but is greatly reduced in magnitude and no longer statistically significant for twins.

QUALIFICATIONS AND NEXT STEPS

Our research is potentially limited in some other respects, beyond the quality of our record links. Most obviously, our results pertain only to individuals born between 1915 and 1920. Whether our findings extend to more recent cohorts of adults is not knowable given the inherent lag time involved. We should note that this is an issue that is generic to *all* studies of educational gradients in mortality.

Beyond this basic limitation, the within-pair comparison strategy described above for identifying causal effects of education on mortality has been subject to a number of important critiques. Questions about measurement error, residual within-pair variation, and external validity have all been raised in response to prior studies, especially those that involve twins (Boardman and Fletcher 2015; Bound and Solon 1999; Gilman and Loucks 2014; Kaufman and Glymour 2011). To assess the sensitivity of our estimates to these concerns, we plan to carry out a series of additional robustness checks (the results of which are not yet available). We describe these checks below.

Measurement Error. It is well known that attenuation bias is more pronounced in fixed-effects models due to the less favorable signal-to-noise ratio (Ashenfelter and Krueger 1994; Griliches 1979). This leads to an increased chance of making a Type II error. Although our measure of education is presumably subject to less measurement error than typical measures (because of the relatively short recall period involved for most members of our sample), we will nevertheless mitigate against this problem by (1) identifying people who were double-enumerated in 1940 and (2) linking members of our samples observed in the 1920 Census to the World War II Army Enlistment file. Both steps will provide second and independent measures of education for linked cases; we can then use that second measure to obtain reliability ratios and measurement-error adjusted estimates of β (the coefficient relating education to mortality). Similar approaches have been used in past work (Lundborg 2012).

Residual Variation. The model specified in Eq. (3) differences out all characteristics that are shared by both members of a twin (or non-twin) pair, but it does not account for characteristics that vary between individuals within pairs. This could bias our estimates if there are unobserved individual-specific factors, X_{ij}, that are correlated with both education and mortality (e.g., childhood health, intelligence, or genetic differences due to the presence of DZ twins in our twin sample). The direction of the bias depends on the nature of the relationship: if X_{ij} correlates with schooling and longevity in the same (opposite) way, our estimate of β will be an overestimate (underestimate) of the true effect (Kohler, Behrman and Schnittker 2011).[4] Although we cannot resolve this issue directly, we can (and will) determine how strong these associations would have to be in order to compromise our results using a modified version of the bounding approach described by Altonji et al. (Altonji, Elder and Taber 2005) and others.

External Validity. A third concern raised—particularly about discordant twin studies—is that the results may not generalize to the broader adult population (Boardman and Fletcher 2015; Conley, Strully and Bennett 2006; Kaufman and Glymour 2011). Twins—and in our case, twins that are discordant with respect to education—may be a unique and non-representative group. The design of our study allows us to examine this issue in two ways. First, as described above, we fit "unpaired" regression models that treat twins as if they are non-twin singletons (with standard errors clustered at the twin-level). As shown in the tables, the results obtained from these unpaired models resemble those obtained using the same techniques on our subsample of unrelated individuals. This implies that twins are unremarkable with respect to the processes in question.

[4] We suspect that over-estimates are more likely than under-estimates, given that the most likely sources of within-pair variation (i.e., early-life health shocks) negatively correlate with educational attainment and longevity. If this speculation is correct, it would imply that the estimates presented in this paper are too large rather than too small, conditional on measurement error being small in magnitude.

As a second check, we plan to compare the marginal distributions of variables describing males in the twin pair sample to the marginal distributions of variables describing males in our non-twin pair samples. If the distributions look similar—or if we can cause them to look similar by utilizing post-stratification weighting techniques—we will be able to make stronger claims with respect to external validity.

REFERENCES

Altonji, Joseph G., Todd E. Elder, and Christopher R. Taber. 2005. "Selection on Observed and Unobserved Variables: Assessing the Effectiveness of Catholic Schools." *Journal of Political Economy* 113(1):151-84.

Amin, Vikesh, Jere R. Behrman, and Hans-Peter Kohler. 2015. "Schooling has smaller or insignificant effects on adult health in the US than suggested by cross-sectional associations: New estimates using relatively large samples of identical twins." *Social Science & Medicine* 127:181-89.

Ashenfelter, Orley, and Alan Krueger. 1994. "Estiamtes of the Economic Return to Schooling from a New Sample of Twins." *The American Economic Review* 84(5):1157-73.

Behrman, Jere R. , Hans-Peter Kohler, Vibeke Myrup Jensen, Dorthe Pedersen, Inge Petersen, Paul Bingley, and Kaare Christensen. 2011. "Does More Schooling Reduce Hospitalization and Delay Mortality? New Evidence Based on Danish Twins." *Demography* 48(4):1347-75.

Benjamini, Yoav, Abba M. Krieger, and Daniel Yekutieli. 2006. "Adaptive Linear Step-Up Procedures that Control the False Disovery Rate." *Biometrika* 93:491-507.

Boardman, Jason D., and Jason M. Fletcher. 2015. "To cause or not to cause? That is the question, but identical twins might not have all of the answers." *Social Science & Medicine* 127:198-200.

Bound, John, and Gary Solon. 1999. "Double trouble: on the value of twins-based estimation of the return to schooling." *Economics of Education Review* 18(2):169-82.

Conley, Dalton, Kate W. Strully, and Neil G. Bennett. 2006. "Twin differences in birth weight: The effects of genotype and prenatal environment on neonatal and post-neonatal mortality." *Economics & Human Biology* 4(2):151-83.

Elo, Irma T., and Samuel H. Preston. 1996. "Educational differentials in mortality: United States, 1979–1985." *Social Science & Medicine* 42(1):47-57.

Fu, Zhichun, H.M. Boot, Peter Christen, and Jun Zhou. 2014. "Automatic Record Linkage of Individuals and Households in Historical Census Data." *International Journal of Humanities and Arts Computing* 8(2):204-25.

Gelman, Andrew, and John B Carlin. 2000. "Poststratification and weighting adjustments." in *Survey Nonresponse*, edited by R Groves, D Dillman, J Eltinge, and R Little. New York: Wiley.

Gilman, Stephen E., and Eric B. Loucks. 2014. "Another casualty of sibling fixed-effects analysis of education and health: An informative null, or null information?" *Social Science & Medicine* 118(0):191-93.

Goeken, Ron, Lap Huynh, T.A. Lynch, and Rebecca Vick. 2011. "New Methods of Census Record Linking." *Historical Methods* 44(1):7-14.

Griliches, Zvi. 1979. "Sibling Models and Data in Economics: Beginnings of a Survey." *Journal of Political Economy* 87(5):S37-S64.

Hayward, Mark D., Robert A. Hummer, and Isaac Sasson. 2015. "Trends and group differences in the association between educational attainment and U.S. adult mortality: Implications for understanding education's causal influence." *Social Science & Medicine* 127:8-18.

Hummer, Robert A., and Elaine M. Hernandez. 2013. "The Effect of Educational Attainment on Adult Mortality in the United States." *Population Bulletin* 68(1):1-16.

Hummer, Robert A., and Joseph T. Lariscy. 2011. "Educational Attainment and Adult Mortality." Pp. 241-61 in *International Handbook of Adult Mortality*, edited by Richard G. Rogers and Eileen M. Crimmins: Springer Netherlands.

Jaro, Matthew A. 1989. "Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida." *Journal of the American Statistical Association* 84(406):414-20.

Kaufman, Jay S., and M. Maria Glymour. 2011. "Splitting the Differences: Problems in Using Twin Controls to Study the Effects of BMI on Mortality." *Epidemiology* 22(1):104-06.

Kitagawa, Evelyn M., and Philip M. Hauser. 1968. "Education Differentials in Mortality by Cause of Death: United States, 1960." *Demography* 5(1):318-53.

—. 1973. *Differential Mortality in the United STates: A Study in Socioeconomic Epidemiology*. Cambridge: Harvard University Press.

Kohler, Hans-Peter, Jere Behrman, and Jason Schnittker. 2011. "Social Science Methods for Twins Data: Integrating Causality, Endowments, and Heritability." *Biodemography and Social Biology* 57(1):88-141.

Link, Bruce G., Jo C. Phelan, Richard Miech, and Emily Leckman Westin. 2008. "The Resources That Matter: Fundamental Social Causes of Health Disparities and the Challenge of Intelligence." *Journal of Health and Social Behavior* 49(1):72-91.

Lleras-Muney, Adriana. 2005. "The Relationship Between Education and Adult Mortality in the United States." *The Review of Economic Studies* 72(1):189-221.

Lundborg, Petter. 2012. "The Health Returns to Schooling—What Can We Learn from Twins?" *Journal of Population Economics* 26(2):673-701.

Lundborg, Petter, Carl Hampus Lyttkens, and Paul Nystedt. 2016. "The Effect of Schooling on Mortality: New Evidence From 50,000 Swedish Twins." *Demography* 53(4):1135-68.

Madsen, Mia, Anne-Marie Nybo Andersen, Kaare Christensen, Per Kragh Andersen, and Merete Osler. 2010. "Does Educational Status Impact Adult Mortality in Denmark? A Twin Approach." *American Journal of Epidemiology* 172(2):225-34.

Montez, Jennifer Karas, Robert A. Hummer, Mark D. Hayward, Hyeyoung Woo, and Richard G. Rogers. 2011. "Trends in the Educational Gradient of U.S. Adult Mortality From 1986 Through 2006 by Race, Gender, and Age Group." *Research on Aging* 33(2):145-71.

Montez, JenniferKaras, and MarkD Hayward. 2011. "Early Life Conditions and Later Life Mortality." Pp. 187-206 in *International Handbook of Adult Mortality*, edited by Richard G. Rogers and Eileen M. Crimmins: Springer Netherlands.

Osler, M., M. McGue, and K. Christensen. 2007. "Socioeconomic position and twins' health: a life-course analysis of 1266 pairs of middle-aged Danish twins." *Int J Epidemiol* 36(1):77-83.

Preston, Samuel H., and Paul Taubman. 1994. "Socioeconomic Differences in Adult Mortality and Health Status." Pp. 279-318 in *Demography of Aging*, edited by Linda G. Martin and S.amuel H. Preston. Washington, D.C.: National Academy Press.

Ruggles, Steven. 2014. "Big microdata for population research." *Demography* 51(1):287-97.

Steinwart, Ingo, and Andreas Christmann. 2008. *Support Vector Machines*. New York: Springer.

Stone, Mervyn. 1974. "Cross-validatory choice and assessment of statistical predictions." *Journal of the Royal Statistical Society. Series B (Methodological)*:111-47.

Vapnik, Vladimir N. 1995. *Statistical Learning Theory*. New York: Wiley-Interscience.

Winkler, William E. 2004. "Methods for evaluating and creating data quality." *Information Systems* 29(7):531-50.

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of Econometrics* 141(2):1281-301.

Table 1: 1920-1940 census linkage

	(A) 1920 sample	(B) 1920-1940 potential sample	(C)=(B)/(A) %	(D) 1920-1940 linked sample	(E)=(D)/(B) %
Twins	81,000	75,276	92.9	22,896	30.4
Siblings	100,000	92,503	92.5	28,427	30.7
Unrelated neighbc	99,545	92,252	92.7	27,356	29.7
Random sample	100,000	92,510	92.5	27,583	29.8

Table 2: 1920-1940 to SSDMF linkage

	Linked individuals 1920-1940-SSDMF
Twins	4564*
Siblings	4938*
Unrelated neighbors	8244
Random sample	8180

Note: * The number of linked individuals is out of the confidently linked pairs between 1920-1940

Table 3. Sample selection for twin sub-sample

	Step of sample selection			
	Full sample	[1]	[2]	[3]
N	**81,193**	**4,564**	**2,036**	**1,976**
White	89.09	96.23	97.74	97.77
Region				
New England	6.02	5.61	5.89	5.67
Mid-Atlantic	17.84	14.59	17.09	17.21
East North Central	18.46	24.56	28.00	28.34
West North Central	13.54	20.77	18.66	18.12
South Atlantic	15.20	9.36	8.94	9.11
East South Central	10.14	7.34	6.29	6.38
West South Central	11.82	9.38	7.56	7.69
Mountain	3.22	3.07	2.85	2.83
Pacific	3.76	5.32	4.72	4.66
Householder's occupation				
Blue collar	44.82	40.45	42.68	42.76
White collar	7.36	8.33	9.04	9.01
Farm (owner/tenant/manager	34.76	38.69	37.18	37.30
Not classified	13.06	12.53	11.10	10.93
Family size (mean)	7.01	6.99	6.95	6.96
US born father	75.42	77.96	77.65	77.78
US born mother	77.62	81.75	82.51	82.54
Age in 1920 (mean)	2.38	2.47	2.54	2.55

Note: The steps of sample selection are [1] drop cases that cannot be uniquely linked to the 1940 census and/or the SSDI; [2] drop incomplete pairs (i.e., pairs where one member was linked but not the other) and pairs made up of duplicate records; and [3] drop pairs where one or both members of the twinship is missing information on educational attainment.

Table 4. Descriptive statistics, by subsample

	Random	Neighbors	Siblings	Twins
Education				
Mean	10.27	10.59	10.28	10.35
SD	2.95	2.81	2.70	2.71
Percent discordant		77.83	63.28	37.85
Correlation within pair		0.31	0.65	0.74
Absolute mean difference within pair		2.41	1.51	0.95
SD of absolute difference		2.25	1.68	1.71
Age at death				
Mean	73.01	73.60	73.44	73.92
SD	11.87	11.83	11.84	1.71
Percent discordant		96.31	96.46	97.37
Correlation within pair		0.08	0.13	0.18
Absolute mean difference within pair		12.46	12.17	11.55
SD of absolute difference		10.16	9.83	9.17
n	7,996	812	2,146	1,976

Note: The subsamples are restricted to cases (in the random subsample) or pairs (in the neighbor, sibling, and twin subsamples) where there is full information on education and age at death.

Table 5.. Unpaired (OLS) and paired (fixed effects) models predicting age at death, by subsample

	Unpaired				Paired		
	Random	Neighbors	Siblings	Twins	Neighbors	Siblings	Twins
Years of schooling	0.30 ***	0.36 **	0.52 ***	0.42 ***	0.02	0.55 ***	0.19
	(0.05)	(0.17)	(0.11)	(0.10)	(0.25)	(0.21)	(0.24)
White	-0.81	-0.73	-1.87	-0.07	-2.97		
	(0.69)	(3.20)	(2.26)	(1.85)	(6.71)		
Householder's occupation							
White collar	0.53	0.70	0.88	1.00	-1.11		
	(0.49)	(1.36)	(1.08)	(1.01)	(2.10)		
Farmers (owner/tenant/manager)	1.21 ***	1.33	2.15 ***	2.21 ***	0.97		
	(0.34)	(1.11)	(0.71)	(0.72)	(2.24)		
Not classified	-0.30	-1.03	1.05	0.34	0.03		
	(0.41)	(1.55)	(0.80)	(0.90)	(1.89)		
Family size	-0.06	0.04	0.01	0.11	-0.13		
	(0.06)	(0.21)	(0.14)	(0.12)	(0.30)		
US born father	0.11	-1.81	1.10	-1.77 **	-1.90		
	(0.46)	(1.38)	(0.73)	(0.90)	(1.97)		
US born mother	-0.65	1.28	-0.33	0.97	0.34		
	(0.49)	(1.41)	(0.79)	(0.98)	(2.23)		
Age in 1920	0.37 ***	0.27	0.25 *	0.36 **	0.44	0.33 *	
	(0.08)	(0.26)	(0.15)	(0.16)	(0.35)	(0.18)	
Constant	70.32 ***	67.06 ***	65.50 ***	64.95 ***	76.86 ***	66.88 ***	71.93 ***
	(1.40)	(4.54)	(4.21)	(3.35)	(7.84)	(2.23)	(2.50)

Note: Unpaired models using the unrelated neighbor, sibling, and twin subsamples cluster standard errors at the pair level. All unpaired models also included state dummies (not shown). The reference category for white is non-white. The reference category for householder's occupation is blue collar. Sample sizes are as follows: random (*n* = 7,996), unrelated neighbors (*n* = 812), non-twin siblings (*n* = 2,146), twins (*n* = 1,976).

*** $p < .01$; ** $p < .05$; * $p < .10$

Table 6. Unpaired (OLS) and paired (fixed effects) models using a categorical measure of education

	Unpaired				Paired		
	Random	Neighbors	Siblings	Twins	Neighbors	Siblings	Twins
Educational attainment							
12 years	1.62 ***	1.84 *	2.53 ***	1.22 **	1.06	3.11 ***	1.64
	(0.31)	(1.02)	(0.61)	(0.60)	(1.42)	(1.01)	(1.30)
More than 12 years	2.13 ***	1.76	3.87 ***	4.39 ***	-0.76	2.90 *	3.93 **
	(0.42)	(1.39)	(0.86)	(0.86)	(1.92)	(1.55)	(1.92)
F -test of joint significance	20.35	1.87	14.35	12.99	0.58	4.96	2.25
p -value	<0.01	0.16	<0.01	<0.01	0.56	<0.01	0.11
White	-0.50	0.20	-1.22	0.37	-3.47		
	(0.69)	(3.44)	(2.34)	(1.94)	(6.72)		
Householder's occupation							
White collar	0.55	1.02	0.90	0.76	-0.79		
	(0.49)	(1.43)	(1.07)	(1.00)	(2.14)		
Farmers (owner/tenant/manager)	1.19 ***	1.28	2.06 ***	2.10 ***	0.97		
	(0.34)	(1.10)	(0.70)	(0.72)	(2.24)		
Not classified	-0.31	-0.93	1.01	0.06	0.10		
	(0.41)	(1.56)	(0.79)	(0.90)	(1.89)		
Family size	-0.06	0.02	0.01	0.10	-0.16		
	(0.06)	(0.21)	(0.14)	(0.11)	(0.30)		
US born father	0.11	-1.90	1.05	-1.73 *	-1.96		
	(0.46)	(1.38)	(0.73)	(0.89)	(1.97)		
US born mother	-0.63	1.30	-0.32	0.97	0.42		
	(0.49)	(1.41)	(0.79)	(0.96)	(2.23)		
Age in 1920	0.38 ***	0.29	0.26 *	0.37 **	0.45	0.38 **	
	(0.08)	(0.26)	(0.15)	(0.16)	(0.35)	(0.18)	
Constant	72.16 ***	68.90 ***	68.49 ***	67.64 ***	77.39 ***	71.14 ***	72.86 ***
	(1.34)	(4.51)	(3.98)	(3.28)	(7.46)	(0.69)	(0.63)

Note: Unpaired models using the unrelated neighbor, sibling, and twin subsamples cluster standard errors at the pair level. All unpaired models also included state dummies (not shown). The reference categories are less than 12 years of education (educational attainment), non-white (white), and blue collar (householder's occupation). Sample sizes are as follows: random (*n* = 7,996), unrelated neighbors (*n* = 812), non-twin siblings (*n* = 2,146), twins (*n* = 1,976).

*** $p < .01$; ** $p < .05$; * $p < .10$

Table 7. Twin pairs with matching day and month of birth (continuous measure of education)

	Unpaired	Paired
Years of schooling	0.54 ***	0.26
	(0.14)	(0.31)
White	(1.24)	
	(2.46)	
Householder's occupation		
White collar	0.18	
	(1.18)	
Farmers (owner/tenant/manager)	2.73 ***	
	(0.87)	
Not classified	0.00	
	(1.08)	
Family size	0.21	
	(0.13)	
US born father	1.61	
	(1.07)	
US born mother	0.90	
	(1.16)	
Age in 1920	0.33 *	
	(0.19)	
Constant	59.73 ***	71.59 ***
	(3.76)	(3.32)

Note: In these models, the sample is restricted to pairs with matching day and month of birth, as recorded in the SSDI (n = 1,360 cases). The unpaired model clusters standard errors at the pair level and includes state dummies (not shown). The reference category for white is non-white. The reference category for householder's occupation is blue collar.

*** $p < .01$; ** $p < .05$; * $p < .10$

Table 8. Twin pairs with matching day and month of birth (categorical measure of education)

	Unpaired	Paired
Educational attainment		
12 years	1.93 ***	0.76
	(0.73)	(1.62)
More than 12 years	5.04 ***	4.95 **
	(1.04)	(2.36)
F -test of joint significance	12.28	2.23
p -value	<0.01	0.11
White	1.55	
	(2.57)	
Householder's occupation		
White collar	0.09	
	(1.17)	
Farmers (owner/tenant/manager)	2.71 ***	
	(0.86)	
Not classified	0.34	
	(1.08)	
Family size	0.19	
	(0.13)	
US born father	-1.69	
	(1.07)	
US born mother	0.97	
	(1.15)	
Age in 1920	0.36 *	
	(0.19)	
Constant	63.37 ***	73.38 ***
	(3.43)	(0.80)

Note: In these models, the sample is restricted to pairs with matching day and month of birth, as recorded in the SSDI (n = 1,360 cases). The unpaired model clusters standard errors at the pair level and includes state dummies (not shown). The reference categories are less than 12 years of education (educational attainment), non-white (white), and blue collar (householder's occupation).

*** $p < .01$; ** $p < .05$; * $p < .10$

Table 9. Results after dropping low-confidence links

	Unpaired				Paired		
	Random	Neighbors	Siblings	Twins	Neighbors	Siblings	Twins
Years of schooling	0.33 ***	0.51 **	0.51 ***	0.38 ***	0.18	0.55 **	0.10
	(0.05)	(0.20)	(0.12)	(0.11)	(0.28)	(0.22)	(0.26)
White	-0.60	-0.20	-3.39	0.55	2.06		
	(0.75)	(3.65)	(2.26)	(2.17)	(7.87)		
Householder's occupation							
White collar	0.41	-0.92	0.88	0.70	-2.91		
	(0.51)	(1.45)	(1.17)	(1.08)	(2.27)		
Farmers (owner/tenant/manager)	1.17 ***	1.63	2.31 ***	2.04 ***	-0.28		
	(0.36)	(1.27)	(0.75)	(0.78)	(2.30)		
Not classified	-0.41	-2.49	0.82	-0.18	2.82		
	(0.44)	(1.67)	(0.88)	(0.93)	(2.02)		
Family size	-0.06	0.05	0.00	0.03	-0.13		
	(0.07)	(0.24)	(0.15)	(0.12)	(0.33)		
US born father	0.29	-1.54	0.80	-1.35	2.22		
	(0.48)	(1.63)	(0.79)	(1.02)	(2.12)		
US born mother	-0.64	1.97	-0.39	1.02	1.07		
	(0.51)	(1.59)	(0.85)	(1.08)	(2.38)		
Age in 1920	0.36 ***	0.23	0.20	0.38 **	0.27	0.30	
	(0.08)	(0.29)	(0.16)	(0.17)	(0.36)	(0.19)	
Constant	69.55 ***	64.73 ***	68.95 ***	64.12 ***	71.26 ***	67.09 ***	73.06 ***
	(1.51)	(4.74)	(6.19)	(3.58)	(8.99)	(2.36)	(2.75)

Note: Pairs where one or both members had low true match probabilities, i.e., Pr(True match = 1) < 0.2 on either the 1920-to-1940 link or the 1940-SSDI link were deleted, as were individual cases in the random subsample. Unpaired models using the unrelated neighbor, sibling, and twin subsamples cluster standard errors at the pair level. All unpaired models also included state dummies (not shown). The reference category for white is non-white. The reference category for householder's occupation is blue collar. Sample sizes are as follows: random (*n* = 7,181), unrelated neighbors (*n* = 614), non-twin siblings (*n* = 1,860), twins (*n* = 1,688).

*** $p < .01$; ** $p < .05$; * $p < .10$