Selection and Economic Gains in the Great Migration of African Americans: New Evidence from Linked Census Data

William J. Collins and Marianne H. Wanamaker

July 2011: Preliminary and incomplete draft

Abstract: We assemble a new dataset that links census records for more than 5,000 African American males from 1910 to 1930, the first two decades of the "Great Migration" from the South. We use the new dataset to engage major themes in research on the Great Migration. We find that literacy in 1910 is weakly correlated with subsequent inter-regional migration, but distance to major northern cities is a strong predictor even when controlling for personal and local characteristics. New estimates of race- and region-specific earnings indicate that the migrants' nominal and real gains were large on average, even among men from the same county or same household. There is some evidence consistent with positive selection into migration, but this can account for a small portion of the earnings difference between migrants and non-migrants.

Collins is a Professor of Economics at Vanderbilt University and Research Associate of the NBER. Wanamaker is an Assistant Professor of Economics at the University of Tennessee. The authors are grateful for suggestions from Jeremy Atack, Hoyt Bleakley, Leah Boustan, Robert Margo, and seminar participants at the University of Chicago, Northwestern University, University of Tennessee, and Vanderbilt University. Tom Markle and Greg Niemesh have provided excellent research assistance. The Grey Fund and the Dornbush Fund at Vanderbilt University and Office of Research at the University of Tennessee have provided generous research support.

The Great Migration of African Americans from the South began in earnest during World War I as labor demand boomed in northern industrial centers and the era of mass migration from Europe approached its end. Between 1910 and 1970, the share of African Americans living in the South fell from 90 to 53 percent, and the share living in non-southern central cities rose from 5 to approximately 36 percent.[1] The economic, political, and social ramifications of the Great Migration have been profound and long-lasting—promoting racial convergence in labor market earnings (Smith and Welch 1989), fueling the rapid rise of black ghettos (Cutler, Glaeser and Vigdor 1999), and influencing the timing and course of the Civil Rights Movement (Gregory 2005, Sugrue 2008). Moreover, it is an outstanding example of long-distance labor mobility, in which a relatively poor population embarked on and expanded new patterns of migration in response to large inter-regional wage differentials, falling migration costs, and changing labor demand conditions.

Scholarship on the economics of the Great Migration spans nearly a full century and has relied primarily on insights gleaned from cross-sections of federal census data.[2] But the answers to key questions about the migrants and their labor market outcomes are clouded by data limitations and measurement problems that are inherent to these sources. In particular, the quantitative implications of selection into the migrant stream for interpreting differences between migrants' and non-migrants' economic status are unknown and difficult to bound with the information available in public-use samples of the census. These problems are also common in studies of international migration because standard datasets rarely include detailed information on migrants' and nonmigrants' background characteristics (Chiquiar and Hanson 2005, Abramitsky, Boustan, and Eriksson 2010).[3]

[1] The 36 percent figure is calculated using the 1970 F2 State IPUMS sample. The central city status of about 4 percent of non-southern blacks is not reported. If all those are assumed to live in central cities (an upper bound), then 37 percent of all blacks would reside in non-southern central cities; if all those are assumed not to live in central cities (a lower bound), the overall figure is 35 percent. So, 36 percent is between the lower and upper bound.

[2] See, *inter alia*, U.S. Department of Labor 1919, Lewis 1931, Higgs 1976, Vickery 1977, Gill 1979, Wright 1986, Gottlieb 1987, Margo 1988, Collins 1997, Vigdor 2002, Boustan 2009, and Logan 2009.

[3] There are some exceptions to the common reliance on census cross-sections for study of the Great Migration. For the pre-Great Migration period, Logan (2009) examined migration in the Colored Troops Sample of the Civil War Union Army Data. Collins (2000) studied retrospective data for workers in six non-southern cities to characterize

The key to our advance is a newly constructed dataset that provides unprecedented detail about the origins and outcomes of those who participated in the first waves of the Great Migration, as well as those who decided to stay in the South. We have linked African American males, ages 0 to 40, from the public-use microdata sample of the 1910 federal census (a completely transcribed 1% random sample of the 1910 population schedules) to the hand-written manuscripts of the 1930 federal census, and then transcribed key variables from the 1930 manuscripts. We concentrate on the early decades of the Great Migration because the full disclosure of names and locations in the census data is central to our empirical strategy, and this information is revealed only up to the 1930 census. In addition, because early patterns of long-distance migration often set a path for subsequent migration flows, the first decades of the Great Migration are particularly important. The final dataset includes extensive personal and family characteristics *before* and after the start of the Great Migration, exact location information in 1910 and 1930, and labor market outcomes in 1930 (such as employment, occupation, and industry). Because the census did not collect income data in this period, we have pursued new approaches to estimating southern-born, black males' earnings in various occupations, industries, and regions.

Several novel findings emerge from our preliminary investigation of the dataset. First, contrary to conventional wisdom about the Great Migration, we find that census-reported literacy circa 1910 is a weak predictor of inter-regional migration. Distance to major urban centers in the North is a much more robust factor. Although migrants and their fathers were more likely to have selected out of agricultural work prior to 1910 than non-migrants, even these variables are not strong predictors of migration in the regression analysis. On the other hand, veteran status and growing up in owner-occupied housing, a proxy for family wealth, are strongly, positively correlated with migration.

black occupational upgrading during the 1940s and its connection to inter-regional migration. Bodnar *et al.* studied migrants in Pittsburgh prior to 1920, and Maloney (2001) studied migrants in Cincinnati between 1910 and 1920. Black *et al.* (2010) study mortality rates of migrants and non-migrants using Social Security and Medicare data.

Second, we find that the migrants earned far more than non-migrants in nominal and real terms, and it is unlikely that the differences are due to unobserved ability. With and without extensive controls for observable personal and local characteristics, the earnings differential is large. Moreover, men from the same county and brothers from same household had very different earnings in 1930 depending on their migration status. Inter-regional differences in pay within job types and inter-regional differences in the distribution of workers across job types appear to hold roughly equal importance in accounting for the large premium earned by migrants. We stress that these findings are preliminary, as we continue to refine the dataset and analyses.

1. Historical and Theoretical Context

There are strong theoretical connections between the economics of the Great Migration and other migration flows from low wage to high wage regions (Sjaastad 1962, Todaro 1969, Borjas 1987, Carrington et al. 1996, Hatton and Williamson 2005). The absence of formal policy barriers to internal migration allows us to concentrate on the incentives, costs, and constraints facing potential migrants from the South. For simplicity, we assume that the timing of World War I and subsequent restrictions on international migration are exogenous.

A simple starting point posits that worker *i* will move from region 0 to region 1 if the expected benefits of residing in region 1 ($V_{i,1}$), which may include higher lifetime income and consumption and/or better amenities (such as more secure civil rights), exceed the expected costs of relocating, including the value of foregone earnings and amenities in the home location ($V_{i,0}$) and adjustment costs associated with travel and assimilation (C_i). That is, a worker will move if $V_{i,1} - V_{i,0} > C_i$. Note that having an extensive network of friends or family in region 1 could lower C_i and/or raise $V_{i,1}$ (e.g., assistance in finding housing or a job), making migration more likely.

Because different workers perceive different values of $V_{i,1}$, $V_{i,0}$, and C_i, and because these differences may be systematically correlated with workers' characteristics, such as age, skill, or initial location, there may be non-random selection into the migrant stream. Such selection is interesting in its own right because it is essential to the economic character of the migration flow, and also because the non-random selection will influence the interpretation of labor market outcomes for "movers" relative to "stayers." Beyond the scope of this paper, selection also influences the interpretation of migration's effect on competing or complementary workers in sending and receiving economies (Boustan 2009), as well as the likelihood of social assimilation or segregation.

Relatively few southern blacks found that $V_{i,1} - V_{i,0} > C_i$ prior to World War I. Due to harsh restrictions on schooling under slavery (Williams 2005) and the absence of large-scale land or wealth redistribution after the Civil War, newly emancipated African Americans had extremely low levels of human and physical capital. In 1870, only 17 percent of African Americans over age 9 could read and write, and less than 5 percent of men, age 20 to 60, owned real property. Blacks were, on average, a very poor, agricultural population concentrated within a relatively poor region (Easterlin 1971, Ransom and Sutch 1977, Margo 2004). It is common for poor, illiterate, agricultural populations to have low rates of long-distance migration, even when the potential gains to relocation appear to be large at prevailing wage levels (Hatton and Williamson 1998). Furthermore, Wright (1987) and Rosenbloom (2002) argue that post-bellum southern labor markets were poorly integrated with non-southern markets, well into the early twentieth century. Outmigration was also dampened by the widespread unwillingness of northern industrial employers to hire black workers, except as occasional strikebreakers, until the massive influx of European immigration was cut short by World War I and tight immigration restrictions in the 1920s (Myrdal 1944, Thomas 1954, Collins 1997). To the extent there was migration, Margo's evidence from 1900 and 1910 census cross-sections suggests that the migrants were positively selected on literacy, though he noted that omitted variables might obscure the true effect of literacy on migration propensity (1990, pp. 109-114).

Several historical trends and events altered the distribution of $V_{i,1}$, $V_{i,0}$, and C_i across potential migrants, eventually propelling the Great Migration. First, to the extent that ignorance and illiteracy constrained inter-regional migration, this constraint loosened with each generation's significant educational advances (Anderson 1988, Collins and Margo 2006). Whereas only 13 percent of southern blacks (age 20-40) were literate in 1870, 83 percent were literate in 1930. Second, after federal troops withdrew from the South in 1877, whites erected more extensive barriers to blacks' social, political, and economic mobility—the pervasive web of constraints known as "Jim Crow."[4] From this perspective, local "amenities" for African Americans deteriorated in many facets between 1880 and World War I—disenfranchisement, mob violence, *de jure* segregation, and so on.[5] C. Vann Woodward wrote, "In the early years of the twentieth century, it was becoming clear that the Negro would be effectively disenfranchised throughout the South, that he would be firmly relegated to the lower rungs of the economic ladder, and that neither equality nor aspirations for equality in any department of life were for him" (1974, pp. 6-7). Third, rising urbanization and improved transportation networks within the South, such as more and better roads and cars, may have facilitated higher migration rates.

World War I decisively raised $V_{i,1}$ by increasing northern demand for black labor, perhaps especially so for less-skilled workers relative to previous decades. Once they had a "foot in the door", some northern firms became accustomed to employing black laborers (Whatley 1990; Foote, Whatley and Wright 2003). Tight immigration restrictions adopted in the 1920s reinforced demand for black workers in northern cities. As the stock of black migrants in the North increased, the dynamics of chain migration unfolded—migration became comparatively easy once friends and family were able to assist with the move. Myrdal surmised that, "Much in the Great Migration after

[4] Woodward notes that the origins of the term "Jim Crow" are "lost in obscurity" (1974, p. 7), but it was in use by the late nineteenth century. The term is sometimes used narrowly with respect to legally enforced segregation in the South, but we will use it more broadly, encompassing the span of racial norms, threats, and laws that proscribed blacks' civil rights.

[5] See Kousser (1974) and Redding and James (2001) on disenfranchisement. See Tolnay and Beck (1995) on mob violence and lynching. See Margo (1990) on widening racial gaps in school quality. Myrdal (1944) emphasized that blacks were shut out of growing employment opportunities in southern manufacturing: "The Industrial Revolution, with all its connotation of modern progress and new opportunity, came to the South later than it did to the North, but it did come. Negroes, however, were not allowed to share in many of its fruits" (p. 188).

1915 is left unexplained if we do not assume that there was before 1915 an existing and widening difference in living conditions between South and North, which did not express itself in a mass migration simply because the latter did not get a start and become a pattern" (2009 [1944], p. 193). Once that pattern of emigration was established, the outflow of black labor from the South was self-reinforcing.

More than 550,000 African Americans left the South (net) between 1910 and 1920, more than the previous four decades combined. Another 900,000 left in the 1920s (Eldridge and Thomas 1964, p. 90). By 1930, approximately 24 percent of 30-to-40 year old black men who had been born in the South lived outside the South, almost all of whom had moved to the East North Central and Middle Atlantic census divisions.[6] The outflow of black emigrants ebbed in the 1930s during the Depression, but resumed at even higher rates in the 1940s and 1950s before slowing dramatically in the 1960s and eventually reversing direction.

Selection on skill has been a central theme in empirical studies of migration (Roy 1951, Borjas 1987, Hanson 2006, Abramitsky, Boustan, and Eriksson 2010), including the Great Migration (Margo 1990, Vigdor 2002). Suppose that region 0 offers relatively high rewards to skill compared to region 1, perhaps because skills are relatively scarce in region 0. All else the same (including C across workers), one would expect relatively *less*-skilled workers to be more likely to move from region 0 to 1, a "negatively selected" migrant flow. Highly skilled workers, perhaps black professionals providing services to black clients in the South, would be more likely to find region 0 appealing in this case. However, if migration costs or the valuation of region-specific amenities vary with skill level, or if there are credit or other constraints on the poorest workers, then the predicted patterns of selection may be more complex (Margo 1990, Borjas 1995, p. 1690; Hatton and

[6] This compares to 11 percent of 30-40 year old southern-born black men in 1910 and 7 percent in 1880. These calculations are all based on the census region definition in the IPUMS data, which includes some states that we do not classify as southern in our linked dataset, specifically Delaware, Maryland, and Washington, DC.

Williamson 1998).[7] For instance, Chiquiar and Hanson (2005) show that "intermediate selection" might result if migration costs are high at low levels of skill but fall as skill rises. "Polarized selection" could result if returns to skill are low in region 1 (inducing migration from the low end of the skill spectrum) *but* high-skill workers are especially attracted to the amenities in region 1 (leading to a rise in $V_{i,1}$ at high levels of skill), such that $V_1 - C > V_0$ at both low and high levels of skill but not in between.[8] The linked dataset created and examined in this paper allows us to address selection issues more thoroughly than is possible with cross-sectional data.

2. Data and Descriptive Statistics

The great advantage of the linked dataset is that it allows us to observe the personal attributes of migrants and non-migrants before and after the onset of the Great Migration. Although we cannot literally follow a person at each point in time between 1910 and 1930, we can observe the younger men in our sample while they still lived with their parents and siblings in 1910, providing detailed knowledge of their background.[9] We can observe the older men in our sample while they still lived and worked in the South in 1910, providing information on their early labor market status and occupations. The data also reveal, up to the local census tract, exactly where the person was located in 1910, meaning that one can evaluate or control for the influence of local characteristics at a much finer level of geographic specificity than with other datasets.[10]

To construct the new longitudinal sample, we started with the IPUMS one-percent cross-section of the 1910 Census of Population (Ruggles *et al.* 2010), limiting it to black male residents of

[7] Since borrowing against future earnings to finance migration costs is difficult for poor workers, especially those without a network of higher-income relatives, it is possible that some workers who would like to move are unable to do so. In this setting, a rising level of income in the home region may lead to a counter-intuitive increase in the propensity to emigrate, as the credit constraint is relaxed.

[8] Along these lines, Margo (1990, p. 121) suggests that, "For younger blacks, an unwillingness to acquiesce to Jim Crow seemed to be a consequence of being better educated."

[9] We are attempting to fill in 1920 data for as many of our observations as possible.

[10] In other studies of migration over this time period, pre-migration residence is defined as state-of-birth, a limitation of cross-section census data.

southern states between the ages of 0 and 40.[11] This generates an initial sample of 28,215 individuals, some of whom reside within the same 1910 household (e.g., brothers).[12] Images of the hand-written manuscripts of the 1930 Census of Population are digitized and have some searchable information on Ancestry.com. We used each individual's name and place of birth information from the 1910 IPUMS sample as search criteria in the 1930 manuscript database.[13] A successful "match" was generated by locating exactly one person with these characteristics in the 1930 census manuscripts. This matching process generated 5,929 successful matches, a 21 percent match rate. Deleting duplicate matches (different individuals in 1910 matched to the same individual in 1930) and other discrepancies leaves a sample size of 5,465 individuals.[14]

Table 1 compares the 1910 characteristics of the matched sample and the full 1910 IPUMS sample of southern black males (age 0 to 40). Reassuringly, the matched sample's properties are very similar to those of the full sample in terms of state-of-residence, literacy and school attendance, likelihood of residing in owner-occupied housing, urban residence, and age distribution. For example, among those under age 21 in 1910 whose father was head-of-household, 64 percent of the fathers in the matched sample and 63 percent of the fathers in the base IPUMS sample were farmers. In short, there is no strong evidence of biased selection into the matched sample relative to the base 1910 IPUMS cross-sectional sample. A separate check with the 1930 IPUMS cross-sectional sample of southern-born black men, age 20 to 60, reveals that 22.0 percent resided outside the South at the

[11] Southern states for our purposes are Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia. Despite the official census classification, we exclude Delaware, Maryland, and the District of Columbia from the list of Southern states.
[12] Expanding our data to use the 1910 oversample would increase the sample size by approximately 80%.
[13] Our search criteria include a SOUNDEX version of the individual's last name, the first three letters of the individual's first name, the individual's state of birth and their birth year within two years. SOUNDEX is a common algorithm used to generate alternative spellings of a surname. SOUNDEX matches include the exact last name and any reasonably close approximation to that last name.
[14] Of those, 2,233 are also matched using an exact match rule on the last name (i.e., no SOUNDEX matches).

time of the 1930 census. This is close to the 20.2 percent of our matched sample who resided in the South in 1910 but not in 1930.[15]

In addition to the individual and household data from the census manuscripts, we have appended data describing the 1910 county-of-residence from the National Historical Geographical Information System (www.nhgis.org) and Haines (2010). These data include variables from the population, manufacturing, and agricultural censuses which allow us to characterize each person's economic environment circa 1910. Geographic coordinates for each county are used to create the distance-to-North variables described below. We have also incorporated information on the number of recorded lynchings in each Southern county, and future versions of the dataset may include estimates of black voting rates and the extent of flooding from the Mississippi River in 1927.[16]

Table 2A splits the sample into two groups: those who left the South after 1910 and resided in the North in 1930 ("migrants") and those who resided in the South in both 1910 and 1930 ("non-migrants"). At present, we cannot tell whether someone left the South after 1910 and then returned before 1930, but we will match the men to the 1920 hand-written manuscripts to provide a sense of how much return migration occurred.[17,18] For some variables, it makes sense to tabulate the data for specific age ranges (e.g., 0 to 20, or 21 and over).

There are notable differences between the migrant and non-migrant groups, but they are not as stark as some of the historical literature would suggest. Migrants had a slightly higher rate of literacy than non-migrants (68 compared to 65 percent), were more likely to be attending school if age 5-20 (51 compared to 48 percent), and were more likely to reside in owner occupied housing (25 compared to 22 percent) in 1910.

[15] We do not expect these numbers to be exactly the same because of interregional mobility prior to 1910 (and sample variability).

[16] See http://people.uncw.edu/hinese/HAL/HAL%20Web%20Page.htm. The Project HAL data are based on a dataset compiled by Stewart Tolnay and E.M. Beck (1995) who examined the NAACP Lynching Records at Tuskegee Institute. Lynching data for Virginia are from Brundage (1993).

[17] A preliminary analysis of the 1920 data indicates that less than 5% of our sample are return migrants.

[18] Place of birth for an observation's children might also shed light on the likelihood of return migration.

The more stark differences pertain to pre-migration occupation and geographic location. Migrants had disproportionately sorted out of agricultural occupations *before* leaving the South—only 43 percent of migrants worked as farmers or farm laborers in 1910 compared to 57 of those who chose to stay in the South. The migrants were also closer on average to major destinations in the urban north in 1910, by about 70 miles (13 percent).

These differences in characteristics extend backward at least one generation, to the cohort of parents who would have been born soon after Emancipation.[19] The fathers of migrants were 7 percentage points less likely to be farmers than the fathers of non-migrants (58 compared to 65 percent), and migrants were more likely to reside in a household with at least one literate parent. There is no measurable difference in terms of average family size or birth order between migrants and non-migrants.

The county-level characteristics are about what one would expect given the slightly north-eastern average location of migrants relative to non-migrants. The average black proportion of the county population is lower, and the proportion of tenants among black farmers and the proportion of cotton among farm acreage are lower. In terms of literacy and children's schooling, the average characteristics of the local *white* population are similar across the two columns. The black population tends to have slightly higher literacy and school attendance rates in the migrants' initial counties. On average, lynching between 1900 and 1930 appears to have been slightly less common in the migrants' counties (2.1 compared to 2.5 among non-migrants), though there is no difference in the median (1.0).[20]

Taken together, the evidence in table 2A, none of which is discernable in a census cross-section, is consistent with weakly positive selection into the migrant stream. Migrants were better

[19] For instance, a 20-year old in 1910 was born in 1890. If the mother was 25 years old at the time of birth, she would have been born in 1865.

[20] Lynching data are not available for Texas, Oklahoma, and West Virginia. Expressing lynching relative to black population size in the county does not reverse the finding that migrants were less exposed to local lynching than non-migrants, nor does using lynching counted over the full 1870 to 1930 period.

educated and had better educated parents, were more likely to have been exposed to urban environments and non-agricultural work, and were more likely to reside in owner-occupied housing (an indicator of wealth) before moving. However, the differences across the groups are less sharp than one might expect on the basis of a cross-section of 1930 micro data, which would suggest a far stronger degree of migrant selection. In our linked dataset, fully 95 percent of the migrants are coded as literate *in 1930*.[21] Only 68 percent of these same individuals were literate in 1910, before leaving the South. [22] It is possible that some were still pursuing elementary-level schooling in 1910 (despite being over age 10) and that they acquired literacy before leaving the South. But we see the same pattern for those who were age 16 and over in 1910, which is less plausibly explained by pre-migration schooling. Three other possibilities seem more likely: a) some migrants may have acquired literacy after moving to the North (e.g., at evening schools); b) migrants may have systematically misrepresented their literacy status to enumerators in either the South or the North; c) census enumerators in the North may have used different standards or assumptions than those in the South when assessing literacy. More research will be necessary to understand these differences, but this finding raises an important caveat regarding the usefulness of census literacy variables in making comparisons across groups, regions, and time. In this case, the 1930 cross section appears to give a misleading characterization of the difference in *pre-existing* educational attainment of migrants compared to non-migrants.

Table 2B reports the 1930 literacy rates, as well as several other characteristics of migrants and non-migrants. Although the two groups were fairly similar in terms of their 1910 observables, by 1930 their lives had taken divergent paths. More than half of the non-migrants were farmers or

[21] This is comparable to the literacy rate among black males born in the South but residing elsewhere in the full 1930 IPUMS cross section (limited to ages 20 to 60 for comparability), and so it is not peculiar to our sample.

[22] There is some evidence of a rise in literacy rates among stayers in the South between 1910 and 1930 in the linked data, but the change is smaller than for the migrants. Among non-migrants, 65 percent were coded as literate in 1910 and 75 percent were coded as literate in 1930. Among migrants, the comparable figures are 68 and 95 percent. So, a difference of 3 percentage points in 1910 widens to 20 percentage point in 1930.

farm laborers, compared to just 2 percent of the migrants. The migrants were disproportionately employed as operatives and unskilled laborers (in sum, nearly 60 percent), and nearly all worked for wages. Given their overwhelmingly urban destinations, it is not surprising that migrants had higher unemployment rates and lower rates of owner-occupancy (Collins and Margo 2011). Between 1910 and 1930, the average location of non-migrants, measured by average latitude and longitude, barely changed (though there is considerable evidence of within-South movement), whereas the average migrant was drawn far to the north and somewhat to the east of the average starting point.

Table 3 reports simple transition matrices for broad job categories, tracing out the cell-to-cell links from 1910 to 1930 for those aged 21 to 40 in 1910. This age group, which had reached maturity and entered the labor market before the start of World War I, was somewhat less likely to move than the younger males in our sample (15.4 percent compared to 22.5 percent), but their transitions are still interesting. The table is broken into three panels: Panel A includes everyone; B includes stayers only; and C includes movers only. The majority of those who worked in farming in 1910, whether as a farmer or farm laborer, still worked in farming in 1930 (panel A, 58 percent = 33.1/56.8), especially if they stayed in the South (panel B, 66 percent = 38.8/59.1). The single largest cell among the regional migrants in panel C is the group that shifted from farm to non-farm laborer (33.5 percent), but this group is nearly matched in size by the group that worked as non-farm, unskilled laborers before leaving the South. Among non-farm, unskilled laborers in 1910 who stayed in the South (panel B), many ended up in farming by 1930, suggesting that movement between farm and non-farm occupations was fairly fluid within the South. Although the group is small in number in our sample, it is interesting that nearly all those in the professional/clerical category in 1910 stayed in the South.

Table 4 presents descriptive linear-probability regressions of inter-regional migration on personal, household and local characteristics in 1910. The available and relevant variables differ across age groups, and so the specifications and samples across columns in Table 4 vary accordingly.[23] It is striking that distance is a powerful predictor of regional migration status even when controlling for an extensive set of personal background and local characteristics, including cotton acreage. Across subsamples and specifications, distance is the most consistent correlate of inter-regional migration, suggesting that 200 extra miles to major northern cities (slightly more than one standard deviation) lowered the likelihood of migration by more than 6 percentage points. Excluding Kentucky, Virginia, and West Virginia (the so-called "border states") weakens the coefficient on distance but does not undermine the basic result. In contrast, own literacy in 1910 is a surprisingly weak predictor of inter-regional migration in columns 2 and 5.[24] Being resident in a small or large city in 1910 is correlated with higher inter-regional migration propensities, but the coefficients' magnitude and statistical significance varies across columns.

Among children in 1910 (columns 3 and 4), residing in owner-occupied housing is associated with a 4 to 5 percentage point increase in migration propensity, perhaps because their parents had more resources to facilitate migration. Among the older males in 1910 (column 5), it is interesting that while one's own literacy is a weak predictor of migration, the local literacy rate is positively correlated with migration: a 10 percentage point increase in local literacy is associated with a 3 percentage point increase in migration propensity, *ceteris paribus*. This is consistent with having more access to information about migration opportunities in places where the black community had higher literacy. It is also notable that after conditioning on other observables, local school enrollment rates are negatively correlated with outmigration among black adults, which is consistent with black families "voting with their feet" in response to poor educational opportunities for their children (Margo 1990).

[23] We have run regressions that use the same parsimonious base specification for all of the age group samples used in table 4. The qualitative results discussed in the text are not sensitive to this. A future version of the paper will include an appendix table for comparison.

[24] Literacy is only recorded for those age 10 and above, so the underlying sample must be restricted in specifications that include this variable.

3. Selection and the Returns to Migration

It is well understood that simple comparisons of the earnings of migrants and non-migrants could greatly mistake the earnings gains associated with migration—problems of selection on observables and unobservables loom over any such comparison. This section attempts to measure the returns to migration absent the effects of selection on observables and unobservables, to the extent that historical data allow. The linked dataset allows us to go much further toward resolving these problems than is possible with a single census-cross section, in large part because so many things that are unobservable in a cross-section are observable in the linked data.

Earnings for Southern-born African American Men circa 1930

Estimating the economic returns to migration for the typical migrant is complicated by the scarcity of race-, place-, and occupation-specific income in the early twentieth-century. The census first inquired about wage and salary earnings in 1940, which provides one potentially useful foothold for estimation, but which fails to cover self-employed workers such as farmers. We report results for two independent approaches to imputing earnings for the observations in our sample. The approaches are described briefly here and discussed in more detail in the data appendix.

To start, we used the individual-level data from 1960 IPUMS sample to calculate median annual earnings (wage, salary, and self-employment income) for southern-born black men in each three-digit occupation category within each census region (South, Northeast, Midwest, and West).[25] The advantage of this approach is that it provides consistent and detailed coverage across hundreds of race/gender/occupation/region-specific cells. Of course, 1960 is far from 1930, and this method could easily understate earnings differences to the extent that regional convergence in wages

[25] This is similar in spirit to the IPUMS “occscore” variable, which is based on median income in occupations in 1950, but it improves on the occscore variable by focusing on the within-occupation and within-region earnings of southern-born black men.

occurred between 1930 and 1960 (among black workers) or the cross-occupational wage structure compressed.[26]

For an approach that brings us closer in timing to 1930, we have taken *industry*-specific average annual earnings data for 1928 (Margo 1996 based on Lebergott 1964), and then used the micro-level data from 1940's census to provide adjustment factors that are based on the ratio of black men's earnings to all workers' earnings in each industry within each region. For example, if the average construction worker earned $X in 1928, and the median southern black male construction worker earned 60 percent of the median for all construction workers in the 1940 microdata, then the imputed earnings for southern black construction workers is 0.6X.

We present results using both approaches to estimating earnings. The key results are similar despite the completely different bases of imputation, which gives us some confidence that the estimates are reasonable, but we plan to extend the analysis to include more sensitivity checks.

In addition to issues of selection on observables and unobservables, nominal income differences between migrants and non-migrants will overstate the real income gains associated with migration because price levels were, on average, higher outside the South and in cities.[27] We have taken two approaches to adjusting the nominal earnings measures. One is based on estimates of cost-of-living (COL) differences across states from Williamson and Lindert (1980), which are based on work by Stecker (1937) and Koffsky (1949).[28] In essence, Williamson and Lindert use Koffsky's

[26] By using the 1940 micro-level census data rather than the 1960 data, we can take a similar approach and get closer to the time period of interest. However, because the 1940 census does not report the sum of self-employment earnings, we must either proceed without a large segment of the southern black labor force (farmers) or make an ad hoc assignment of income to farmers. We have experimented with this approach, but for brevity we do not report the results here.

[27] Nonetheless, the nominal earnings differences are interesting in the context of models of spatial equilibrium, in which wage differences are indicative of local productivity differences (Roback 1982). Applying the spatial equilibrium framework in this paper's context is questionable given Wright's interpretation of regionally segmented labor markets.

[28] Stecker (1937) estimates cost of living levels in 59 cities based on consumer expenditure surveys from the mid-1930s. Koffsky (1949) estimates the cost-of-living difference between farm and non-farm locations in the 1940s. Williamson and Lindert (1980) use the Koffsky differential and the agricultural share of the workforce in each state to adjust the Stecker-based cost of living figures to state averages. Another approach is based on state-level

estimate of farm/non-farm cost-of-living differences to scale down Stecker's urban-based figures, depending on the proportion of each state's population in agricultural employment. We have pursued an alternative cost-of-living adjustment that is more sensitive to the distribution of the black population across places within states. This hews closer to Stecker's original city-based data and is described in the data appendix. Further work remains to be done on these adjustments, but the results here can provide a sense of the magnitude such price adjustments entail.

Results

We estimate earnings differences between inter-regional migrants and those who stayed in the South in table 5. Differences in average earnings could reflect several different factors that are not mutually exclusive: migrants could be selected on ability; given ability, migrants could move into higher paying occupations ("upgrading"); and within occupations, places may differ in levels of productivity and pay. A key advantage of the linked dataset is that we can narrow the comparisons in a manner that limits the scope for omitted variable bias and that can reveal patterns of selection.

Table 5, column 1 reports coefficients on the migrant variable from log annual earnings regressions *without* control variables—these are the unadjusted migrant-group differences in log income. The coefficients on the migration variable are our focus, and each is expressed relative to the omitted category (non-migrants). The first row uses the 1928 industry-based earnings estimates, unadjusted for cost of living differences, and the second row uses the 1960 microdata-based earnings estimates. The group differences are large—on average migrants earned between 110 and 115 log points more than non-migrants. Comparing down the rows shows how cost-of-living adjustments affect the magnitude of the earnings differences. These adjustments do significantly scale down the

estimates of cost-of-living differences in 1960 by Berry et al. (2000), which may be more appropriate for the income proxies that are based on the 1960 IPUMS data. The results are not sensitive to the choice of COL deflator.

magnitude of the migrants' gains, but the remaining earnings differential is still large at 90 to 100 log points.

Column 2 includes a full set of background control variables. To keep the sample as large as possible, we include control variables that are available for the vast majority of our sample: age fixed effects, veteran status, small and large city status, distance from Chicago or Philadelphia, owner-occupied housing interacted with headship status, state-level log income per capita, black percentage of county population, black adult literacy rate, black children's school attendance rate, and the percent of farm acres in cotton.[29] In each row, the coefficient on migration is diminished relative to the first column, but only slightly, implying that selection on observables can account for a modest share of the raw difference in earnings between migrants and non-migrants.[30]

Column 3 adds county fixed effects to the specification and therefore identifies the migrant coefficient by comparing across men who lived in the same county in 1910. Unobservable local fixed effects are absorbed. The regression still controls for individual-level differences in age, veteran status, city status, and homeownership interacted with headship. The coefficients are slightly diminished in magnitude in comparison with the previous specification, which is consistent with some positive selection into the migrant flow based on unobservable local characteristics. However, the change in coefficients from column 2 to column 3 is very small, typically just 1 or 2 log points.

Going further and following Abramitzky, Boustan, and Eriksson (2010), we can base estimates of the gains from migration on cross-brother comparisons. This eliminates the influence of unobserved household-level effects on the children's labor market outcomes, thereby further narrowing the scope for omitted variable bias. Of course, it comes at the price of using a much smaller sample. In Column 4, we repeat the specification from Column 3 but restrict the sample to

[29] All of these variables, except veteran status, pertain to the observation and his county in 1910.

[30] To maintain a large and consistent sample, we do not control for own literacy in table 5's regressions. Adding 1910's literacy status to the regressions drops everyone under age 10, but has almost no effect on the magnitude of the migrant coefficient.

observations from households with more than one linked record. The sample is further restricted to those whose relation to the household head was “child” in 1910 (i.e., brothers). This helps bridge the results from column 3 with those of column 5, where we include household fixed effects. The coefficient on migrant status is smaller in column 4 than in column 3, but this is entirely due to the change in sample composition. In column 5, the migrant coefficient is identified from comparisons of earnings for brothers with different migrant status (approximately 20 percent of the brothers migrated). In general, the coefficients are similar in columns 4 and 5; if anything, they are somewhat larger in column 5. This would be consistent with negative selection across households within counties, but the relatively large standard errors preclude any strong inference along these lines. There remains scope for selection on ability within households, but we have not found an effective way to isolate it.[31]

For a sample restricted to men over age 20 in 1910, we can estimate specifications that control for occupational status in 1910, measured as log median earnings by southern-resident black men in detailed occupations from the 1960 microdata. This controls for person-specific differences in early labor market outcomes. The results are omitted from table 5 for brevity (and because they are very preliminary). As one would expect, the 1910 occupational status measure is positively correlated with 1930’s outcome, but the migration coefficients are comparable in size to those reported in table 5, even controlling for both 1910 occupational status and county fixed effects.

Overall, comparing the coefficients in column 1 (no controls) to those in column 3 (observable controls and county fixed effects) suggests that only a small portion of the migrants’ earnings advantage may be attributable to selection on observable personal characteristics or local characteristics. Comparing columns 4 and 5, where household fixed effects are added, suggests that

[31] One candidate for a within-household IV is the timing of World War I induction which strongly influenced the likelihood of migration for specific birth cohorts. However, separating the effect of military service from the effect of migration on earnings may prove impossible. Another candidate is birthorder (see Abramitsky, Boustan, and Erikkson 2010), but that seems to have little influence on migration propensity in our preliminary results.

within-county selection on cross-household heterogeneity is weak. Finally, in a sample of older men, controlling for 1910 occupational status has little effect on the migration coefficients. Taken together, the magnitude of the estimates leaves little doubt that the migrants' average gains were large, consistent with the sheer volume of migration from the South both before and after 1930.

Distance as an Instrumental Variable

In the OLS regressions discussed above, adding an extensive range of observable characteristics, local fixed effects, and even household fixed effects has a relatively small effect on the magnitude of the coefficient on migrant status. Nonetheless, there remains some scope for selection on unobserved ability, health, motivation, or other factors that could lead to an overstatement or understatement of migrant gains in the OLS regressions. For another empirical perspective, we rely on the single most robust predictor of migration to emerge from the regressions in table 3—the distance to Chicago or Philadelphia, whichever is closer—to serve as an instrumental variable for migration.[32,33]

Distance exogenously affects the cost of moving, and table 3's results indicate that it has a strong first-stage relationship with the likelihood of inter-regional migration. The F-statistics on the excluded instrument in the IV regressions described below are above 40, and so we are not concerned about weak-instrument bias. The excludability of distance from the second-stage is more difficult to establish. While distance from northern cities in 1910 *per se* may have no direct effect on labor market outcomes (aside from its influence via migration costs), it is possible that distance is correlated with local economic development and therefore 1930 outcomes. More work remains to be

[32] If *within households* migrants were strongly selected on unobservable ability (either positively or negatively), then the fixed effect estimates above would be biased. This type of unobserved selection cannot be treated with the distance to Philadelphia/Chicago instrument.

[33] In practice, the choice of these two cities does not strongly influence the results. We selected them because they had large black populations before the Great Migration and were prime destinations during the Great Migration. It so happens that almost exactly half of our observations were closer to Chicago in 1910 than to Philadelphia.

done here. For now, we proceed under the assumption that distance is a valid instrument when the second-stage regression controls for a long list of observable personal and local characteristics circa 1910, including the importance of cotton agriculture, the local literacy and school attendance rate among blacks, and state income per capita.

Regression specifications that are similar to the first two columns in table 5, but with distance instrumenting for migration status, yield results that reinforce the finding that the migrants' gains were large.[34] If anything, the IV coefficients on migrant status are larger than the OLS results, suggesting that there may have been negative selection on unobservables during the 1910s and 1920s. The results are not undermined even when we include latitude as a control variable, nor when we add a control variable for 1910 occupational status. We intend to investigate further the IV results and their validity in future research.

Decomposing the migrants' gains

As noted above, the differences in average earnings between migrants and non-migrants could reflect several different factors: migrants could be selected on ability; given ability, migrants could move into higher paying jobs; and within jobs, places may differ in levels of productivity and pay. In this subsection we evaluate the extent to which regional differences in pay within job-types account for the migrants' earnings differential.

We assign all of the northern-resident men in our sample the earnings level of southern-resident men who were employed in the same industry category, based on the 1928 industry-based earnings estimates. In other words, we create a counterfactual distribution of earnings for the northern migrants, based on the prevailing earnings of southern black men who worked in the same industry.[35] Of course, this procedure ignores the general equilibrium effects of migration on regional

[34] It is infeasible to have county and household fixed effects while identifying off variation in distance.
[35] Since everyone lives in the South in the counterfactual, we ignore cost of living issues.

earnings levels. If the entire difference between migrants' and non-migrants' earnings were due to regional differences in pay *within*-industries (or selection on ability within job-types), then the regression coefficient on migrant status would be zero when migrants are assigned the counterfactual (southern) earnings level. If the entire difference in earnings between migrants and non-migrants were due to differences in their distributions across job-types, then the regression coefficient on migrant status would be the same using the counterfactual earnings as when using the region-specific earnings (i.e., same as in previous section).

Table 6 shows results for specifications that are similar to those in the first row of table 5, which is replicated for easier comparison. In each column of row 2, where the migrant coefficient is estimated with the counterfactual earnings data, the coefficient is significantly lower than with the original earnings data (row 1), implying that a sizable share of the earnings gap is accounted for by regional differences in pay within job-types. Nonetheless, because the migrant coefficient remains large in magnitude, it is clear that inter-industry mobility was also a critical factor in determining the size of migrants' gains. In practice, movement out of agricultural jobs and into other lines of work accounted for approximately half of the migrants' earnings gains.

4. Conclusions

In any study of migration, it is helpful to have information on people at more than one point in time. The pitfalls of answering questions about migration from cross-sectional datasets have been pointed out many times, but large, representative, longitudinal datasets simply did not exist for the study of the Great Migration until now. For this paper, we have assembled a dataset that links African American males from the 1910 to the 1930 census manuscripts, covering the first two decades of the Great Migration.

We focus on two questions. First, what personal and local characteristics influenced the likelihood of participating in the Great Migration? Our dataset includes many observations while they still resided with their parents, and it includes many others who had already entered the labor force by 1910. Both aspects allow us to characterize the migrants' origins (and non-migrants) in unprecedented detail. Literacy seems to have mattered little. Family wealth and proximity to major northern cities seem to have mattered more.

Second, how large were the migrants' economic gains? We approached this measurement problem from several perspectives—including novel efforts to impute earnings, estimate cost-of-living differences, and to measure econometrically the migrants' unconditional and conditional earnings premium relative to non-migrants. Every empirical perspective indicates that the gains in earnings were large on average. A simple counterfactual suggests that roughly half of the earnings gains were associated with regional differences in pay within job categories, and about half were associated with regional differences in workers' distribution across job categories. Both differences were largely a reflection of the South's economic underdevelopment in comparison with the rest of the United States.

Although the volume of migration is consistent with the existence of large regional gaps in earnings, the results are not a foregone conclusion. First, if non-southern amenities were sufficiently attractive (e.g., more secure civil rights or better education for one's children), then a theory of compensating differentials suggests that African Americans would accept lower real pay to reside and work outside the South. Second, new evidence indicates that migrants did not experience gains in longevity relative to non-migrants, implying that either income and health were weakly correlated in this context or that the income gains were small (Black *et al.* 2010). Third, general equilibrium effects could narrow gaps by lowering blacks' earnings in the North and raising them in the South. All three of these potentially offsetting factors merit more attention, though the first-order fact of large earnings gains seems robust.

Data Appendix

Assigning Occupation Codes, Industry Codes, and Relative Annual Earnings Estimates

The data for 1930 are transcribed from the hand-written manuscripts of the Census, including string variables for occupation and industry and a four-digit occupation/industry code that is unique to the 1930 Census. There is no precise, 1-to-1 crosswalk between the 1930 occupation/industry codes and the 1950 occupation and industry codes that are fundamental to the IPUMS microdata and embedded in our data for 1910. Therefore, we assigned each individual observed in 1930 with an occupation and industry coding based on the 1950 classification system following the steps described here.

The 1930 IPUMS microdata include both the 1930 and the 1950 coding schemes. Our 1930 data from the census manuscripts include only the 1930 coding, so we constructed an algorithm to assign observations a 1950-based occupation code. The algorithm consisted of a number of passes through the data using the 1930 code and text strings of occupation and industry. First, for each 1930 occupation code, we tabulated the 1930 occupation and industry text strings. When the overwhelming majority of occupation strings fell into a single 1950 occupation code, that code became the initial assignment for the manuscript dataset. For occupation strings that were not prevalent, a second pass assigned a 1950 code based on occupation and industry strings alone. We then tabulated occupation strings within the assigned 1950 codes to check for any anomalies. A third pass through the data corrected those anomalies by individually assigning a 1950 code using the occupation and industry text strings.

Once the “occ1950” codes are in place, it is possible to assign occupation-based income levels to each observation. The IPUMS includes a variable called “occscore” which assigns annual earnings levels to occupations based on the 1950 median earnings of all workers in that specific occupation category. Our approach is similar in spirit, but attempts to assign income levels that are specific to southern-born black men residing in each of four census regions (South, Northeast, Midwest, and West). In other words, the annual earnings assignments are specific to race/gender/region-of-birth/region-of-residence/ occupation categories. To construct the estimates, we started with the comparatively large 1960 IPUMS dataset. The 1960 dataset has the advantage of reporting both wage and self-employment income in the previous year for a large number of southern-born black men, which allows us to include self-employed farmers. After sorting men (age 18 to 65, in the labor force, worked at least 1 week in the previous year) into detailed cells, we collapsed the data and retrieved the median value of earnings within each cell. These became the basis of our program that automatically assigns earnings to each observation in our main dataset. In cases where there were less than 10 observations in the cell, we moved to broader region-of-residence groupings (South and Non-South) to estimate median earnings. If there were still less than 10 observations, we moved to broader occupational groupings (basically one-digit of detail), while returning to the original four region-of-residence distinctions.

We followed a similar procedure using the IPUMS microdata from 1940, which is obviously closer to the time period of interest, but which reports only wage and salary income. We include only wage and salary workers in the calculation of median earnings. This means that assignments of earnings to farmers (or other self-employed workers) must be forgone or rely on ad hoc assignments. We have omitted these results from the paper for brevity.

The final set of annual earnings assignments are based on broad industry-level data from 1928. These data are reported in *Historical Statistics of the United States* (Margo 1996, 2-273) and were taken from Lebergott (1964). This provides a completely different basis for the assignment of income compared to that take above, and it brings us much closer to our main dataset's year of observation, 1930. Because there is considerable scope for differences between the average earnings of black men within any industry and the average earnings of all workers, we have adjusted the Lebergott data as follows. Using the 1940 IPUMS microdata, we calculated the median earnings of southern-born black men in each broad industry category and the median earnings of all workers in each industry category. We did this for the South and Non-South separately, with a sample that included workers who were employed for at least 40 weeks in the previous year and were in the labor force at the time of the census. We then scaled the Lebergott data by the ratio of (southern-born black men)/(all workers) median earnings in each region.

Alternative cost of living adjustments

Williamson and Lindert (1980) provide one basis for adjusting nominal earnings for geographic differences in the cost of living circa 1930. Their data are reported at the state level and are built up from city-level information located in Stecker (1937). Williamson and Lindert essentially create a weighted average of cost-of-living at the state level by adjusting the Stecker data according to the share of the labor force in agriculture in each state. The adjustment reflects Koffsky's (1949) estimate of the difference between farm and city price levels in 1941. Our alternative approach works with the same underlying data, but it stays closer to Stecker's city-specific data when possible. First, we assign Stecker's city-specific values to those living in the cities she covered (e.g., this fixes Chicago relative to Birmingham). Then, for residents of cities not covered by Stecker but with at least 25,000 residents, we assign values that are equal to the black-population-weighted average for Stecker-covered cities in the same state (e.g., this assigns Montgomery the average cost-of-living of Birmingham and Mobile). In a few states, Stecker covers no cities, the most important of which for our purposes is Mississippi. We use Alabama's data as a substitute. Then, we assign values to non-city residents in each state by applying the same "Koffsky adjustment" factor as Williamson and Lindert—this scales down Stecker's city-based values for application to non-city residents within each state. In Illinois, for example, where blacks were heavily concentrated in cities, the second approach yields a higher COL index value than in Williamson and Lindert.

References

Abramitzky, Ran, Leah Platt Boustan, and Katharine Eriksson. 2010. “Europe’s Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration.” NBER Working Paper 15684.

Anderson, James D. 1988. *The Education of Blacks in the South, 1860-1935*. Chapel Hill: University of North Carolina Press.

Berry, William D., Richard C. Fording, and Russell L. Hanson. 2000. “An Annual Cost of Living Index for the American States, 1960-1995.” *Journal of Politics* 62, 2: 550-567.

Black, Dan, Magdalena Muszynska, Seth Sanders, and Lowell Taylor. 2010. “The Great Migration and African-American Mortality: Evidence from Mississippi.” Working paper.

Bodnar, John, Roger Simon, and Michael P. Weber. 1982. *Lives of Their Own: Blacks, Italians, and Poles in Pittsburgh, 1900-1960*. Champaign, IL: University of Illinois Press.

Borjas, George. 1987. “Self-Selection and the Earnings of Immigrants.” American Economic Review 77: 531-553.

Borjas, George. 1994. “The Economics of Immigration.” *Journal of Economic Literature* 32, 4: 1667-1717.

Boustan, Leah Platt. 2009. “Competition in the Promised Land: Black Migration and Northern Labor Markets, 1940-1970.” *Journal of Economic History* 69,3: 756-783.

Brundage, W. Fitzhugh. 1993. *Lynching in the New South: Georgia and Virginia, 1880-1930*. Champaign, IL: University of Illinois Press.

Carrington, William J., Enrica Detragiache, and Tara Vishwanath. 1996. “Migration with Endogenous Moving Costs.” *American Economic Review* 86, 4: 909-930.

Chiquiar, Daniel, and Gordon H. Hanson “International Migration, Self-Selection, and the Distribution of Wages: Evidence from Mexico and the United States.” *Journal of Political Economy* 113, 2: 239-281.

Collins, William J. 1997. “When the Tide Turned: Immigration and the Delay of the Great Black Migration.” *Journal of Economic History* 57, 3: 607-632.

Collins, William J. 2000. “African-American Economic Mobility in the 1940s: A Portrait from the Palmer Survey.” *Journal of Economic History* 60, 3: 756-781.

Collins, William J. and Robert A. Margo. 2006. “Historical Perspectives on Racial Differences in Schooling in the United States.” In *Handbook of the Economics of Education: Volume 1*, edited by E. Hanushek and F. Welch. New York: North-Holland: 107-154.

Collins, William J. and Robert A. Margo. 2011. “Race and Home Ownership from the End of the Civil War to the Present.” *American Economic Review: Papers and Proceedings* 101, 3: 355-359.

Cutler, Glaeser, Vigdor. 1999. “The Rise and Decline of the American Ghetto.” *Journal of Political Economy* 107, 3: 455-506.

Easterlin, Richard A. 1971. “Regional Income Trends, 1840-1950.” In *The Reinterpretation of American Economic History,* edited by Robert W. Fogel and Stanley L. Engerman. New York: Harper & Row: 38-49.

Eichenlaub, Suzanne C., Stewart E. Tolnay, and J. Trent Alexander. 2010. “Moving Out but Not Up: Economic Outcomes in the Great Migration.” *American Sociological Review* 75(1):101-125.

Eldridge, Hope T., and Dorothy Swaine Thomas. 1964. *Population Redistribution and Economic Growth, United States, 1870-1950, Vol. 3: Demographic Analyses and Interrelations*. Philadelphia: American Philosophical Society.

Foote, Christopher L., Warren C. Whatley, and Gavin Wright. 2003. “Arbitraging and Discriminatory Labor Market: Black Workers at the Ford Motor Company, 1918-1947.” *Journal of Labor Economics* 21, 3: 493-532.

Gill, Flora. 1979. *Economics and the Black Exodus*. New York: Garland Publishing.

Gottlieb, Peter. 1987. *Making Their Own Way: Southern Blacks' Migration to Pittsburgh, 1916-1930*. Chicago: University of Illinois Press.

Haines, Michael R. and Inter-university Consortium for Political and Social Research. 2010. *Historical, Demographic, Economic, and Social Data: The United States, 1790-2002* [computer file]. ICPSR02896-v3. Ann Arbor, MI: ICPSR [distributor].

Hanson, Gordon H. 2006. “Illegal Immigration from Mexico to the United States.” *Journal of Economic Literature* 44: 869-924.

Hatton, Timothy J. and Jeffrey G. Williamson. 1998. *The Age of Mass Migration: Causes and Economic Impact*. New York: Oxford University Press.

Hatton, Timothy J. and Jeffrey G. Williamson. 2005. *Global Migration and the World Economy: Two Centuries of Policy and Performance*. Cambridge, MA: MIT Press.

Higgs, Robert. 1976. “The Boll Weevil, the Cotton Economy, and Black Migration 1910-1930.” *Agricultural History* 50, 2: 335-50.

Hines, Elizabeth and Eliza Steelwater. 2011. Project HAL: Historical American Lynching Data Collection Project. Electronic Database: http://people.uncw.edu/hinese/HAL/HAL%20Web%20Page.htm.

Koffsky, Nathan. 1949. “Part II: Farm and Urban Purchasing Power.” In *Studies in Income and Wealth, Volume 11,* pp. 151-220. New York, NY: National Bureau of Economic Research.

Kousser, J. Morgan. 1974. *The Shaping of Southern Politics: Suffrage Restriction and the Establishment of the One-party South, 1880-1910*. New Haven: Yale University Press.

Lebergott, Stanley. 1964. *Manpower in Economic Growth*. New York, NY: McGraw-Hill.

Lewis, Edward E. 1931. *The Mobility of the Negro: A Study in the American Labor Supply*. New York: Columbia University Press.

Logan, Trevon D. 2009. “Health, Human Capital, and African-American Migration before 1910.” *Explorations in Economic History* 46: 169-185.

Maloney, Thomas N. 2001. “Migration and Economic Opportunity in the 1910s: New Evidence on African-American Occupational Mobility in the North.” *Explorations in Economic History* 38: 147-165.

Margo, Robert A. 1990. *Race and Schooling in the South 1880-1950*. Chicago: University of Chicago Press.

Margo, Robert A. 1996. “Wages.” In *Historical Statistics of the United States, Millennial Edition*, edited by Susan B. Carter et al., pp. 2-254-2-300. New York: Cambridge University Press.

Margo, Robert A. 2004. “The North-South Wage Gap, Before and After the Civil War,” in D. Eltis, F. Lewis, and K. Sokoloff, eds., *Slavery in the Development of the Americas*, pp. 324-351. New York: Cambridge University Press.

Minnesota Population Center. 2004. National Historical Geographic Information System: Pre-release Version 0.1. Minneapolis, MN: University of Minnesota.

Myrdal, Gunnar. 1944. *An American Dilemma: The Negro Problem and Modern Democracy*. New York: Harper & Brothers Publishers.

Ransom, Roger L. and Richard Sutch. 2001 [1977]. *One Kind of Freedom: The Economic Consequences of Emancipation, Second Edition*. New York: Cambridge University Press.

Roback, Jennifer. 1982. “Wages, Rents, and the Quality of Life.” *Journal of Political Economy* 90, 6: 1257-1278.

Rosenbloom, Joshua L. 2002. *Looking for Work, Searching for Workers: American Labor Markets during Industrialization*. New York: Cambridge University Press.

Roy, A.D. 1951. “Some Thoughts on the Distribution of Earnings.” *Oxford Economic Papers* 3: 135-146.

Ruggles, Steven, J. Trent Alexander, Katie Genadek, Ronald Goeken, Matthew B. Schroeder, and Matthew Sobek. 2010. Integrated Public Use Microdata Series: Version 5.0 [Machine-readable database]. Minneapolis: University of Minnesota.

Sjaastad, Larry. 1962. “The Cost and Returns of Human Migration.” *Journal of Political Economy* 52 (supp.): 80-93.

Stecker, Margaret Loomis. 1937. *Intercity Differences in Costs of Living in March 1935, 59 Cities*. Works Progress Administration, Division of Social Research, Research Monograph XII. Washington, DC: GPO.

Sugrue, Thomas J. 2008. *Sweet Land of Liberty: The Forgotten Struggle for Civil Rights in the North*. New York: Random House.

Thomas, Brinley. 1954. *Migration and Economic Growth*. New York: Cambridge University Press.

Todaro, Michael P. 1969. “A Model of Labor Migration and Urban Unemployment in Less Developed Countries.” *American Economic Review* 59: 138-148.

Tolnay, Stewart E. and E. M. Beck. 1995. *A Festival of Violence: An Analysis of Southern Lynchings, 1882-1930*. Champaign: University of Illinois Press.

U.S. Department of Labor. *Negro Migration in 1916-17*. Washington, DC: GPO, 1919.

Vickery, William E. 1977. *The Economics of Negro Migration, 1900-1960*. New York: Arno Press.

Vigdor, Jacob L. 2002. “The Pursuit of Opportunity: Explaining Selective Black Migration.” *Journal of Urban Economics* 51: 391-417.

Whatley, Warren. 1990. “Getting a Foot in the Door: Learning, State Dependence, and the Racial Integration of Firms.” *Journal of Economic History* 50, 1: 43-67.

Wilkerson, Isabel. 2010. *The Warmth of Other Suns: The Epic Story of America’s Great Migration*. New York, NY: Random House.

Williams, Heather. 2005. *Self-Taught: African American Education in Slavery and Freedom*. Chapel Hill: University of North Carolina Press.

Williamson, Jeffrey G. and Peter H. Lindert. 1980. *American Inequality: A Macroeconomic History*. New York, NY: Academic Press.

Woodward, C. Vann. 1974. *The Strange Career of Jim Crow, Third Revised Edition*. New York: Oxford University Press.

Wright, Gavin. 1986. *Old South, New South: Revolutions in the Southern Economy since the Civil War*. New York, NY: Basic Books.

Wright, Gavin. 1987. “Postbellum Southern Labor Markets.” In *Quantity and Quiddity: Essays in US. Economic Histor*y, edited by Peter Kilby, 98-134. Middletown, CT: Wesleyan University Press.

Table 1: Comparison of Matched and Full Sample Characteristics, Southern Black Males, 1910

	Matched Sample	Full IPUMS Sample
Panel A: Distribution of state of residence		
Alabama	9.8	10.3
Arkansas	5.0	4.9
Florida	4.0	3.8
Georgia	13.5	13.9
Kentucky	3.2	3.0
Louisiana	8.2	8.7
Mississippi	13.0	12.2
North Carolina	8.9	8.0
Oklahoma	1.5	1.7
South Carolina	10.9	10.2
Tennessee	5.3	5.5
Texas	9.2	9.0
Virginia	6.8	7.9
West Virginia	0.7	1.0
Panel B: Personal characteristics		
Attending school (age 0-20)	36.8	36.3
In owner-occupied housing	22.4	23.9
Literate (age 10-20)	62.4	62.6
Literate (age 10-40)	65.7	65.6
Father is farmer (age 0-20)	64.4	63.3
1910 city population		
Not in city	74.5	73.2
City pop. <=25,000	16.3	17.3
City pop. >25,000	9.2	9.5
Panel C: Age Distribution		
Min Age	0	0
Max Age	40	40
Median Age	16	15
Mean Age	17.0	16.7
Std. Dev.	11.2	11.3

Notes and sources: See the text for a description of how the matched dataset was created. The IPUMS data are from Ruggles et al. (2010).

Table 2A: 1910 Summary Statistics of Males in Linked Dataset,
by Subsequent Inter-regional Migration Status

	Non-Migrants (N=4,361)	Migrants (N=1,104)
Personal characteristics		
Attending school (age 5-20)	47.6	51.2
Literate (age 10-40)	65.1	68.4
Owner-occupied housing	21.7	25.1
Occupation is farmer (age 21-40)	38.9	26.3
Occupation is farm laborer (age 21-40)	18.2	16.7
Mean age in 1910	17.3	15.7
1910 city population		
Not in city	75.8	69.4
City pop. <=25,000	15.4	19.8
City pop. > 25,000	8.9	10.8
Latitude (county)	33.4	34.1
Longitude (county)	86.6	84.9
Distance to Chicago or Philadelphia (min.)	578.2	510.3
Veteran status (observed in 1930)	7.3	13.1
Household characteristics		
Parent present (age 0-20)	87.4	86.5
Parent literate (age 0-20)	66.9	69.6
Father is farmer (age 0-20)	65.0	58.2
Father is farm laborer (age 0-20)	11.8	12.0
Number of siblings in household (0-20)	4.7	4.7
Place in birthorder (among those in hh)	2.9	2.9
Local characteristics		
Black percent of population	49.0	47.3
Black percent of farmers	45.9	43.5
Percent of black farmers who were tenants	69.3	65.7
Percent of white farmers who were tenants	40.8	39.6
Percent of farm acres in cotton	16.0	14.5
Percent of crop value in cotton	40.4	36.4
Adult black literacy	60.8	61.9
Adult white literacy	91.8	92.1
Black school attendance (6-14)	57.9	59.2
White school attendance (6-14)	75.5	75.5
Number of recorded lynchings, 1900-30	2.5	2.1

Notes: The farm laborer category includes unpaid family worker.

Sources: Personal and household characteristics are based on the linked census data, which is described in the text. Most county-level characteristics are from Minnesota Pop Center, NHGIS (2004). Data on cotton acreage and value are from Haines (2010). Lynching counts are from Brundage (1993) and Hines and Steelwater (2011), which is largely based on data compiled by Tolnay and Beck (1995)

Table 2B: 1930 Summary Statistics of Men in Linked Dataset, by Inter-regional Migration Status

	Non-Migrants	Migrants
Personal characteristics		
Literate	76.0	95.2
Owner-occupied housing	23.3	18.5
Farmer	40.2	1.0
Farm laborer	11.8	1.2
Operative	7.8	14.0
Non-agricultural laborer	25.8	44.5
Mean age	37.3	35.7
Employed (define)	94.1	84.0
Class of worker, “own account”	39.4	4.0
Class of worker, wage or salary employee	57.3	94.4
Marital status	81.6	73.2
Veteran status	7.3	13.1
Latitude	33.5	40.3
Longitude	86.6	83.4

Notes: All men in the underlying data resided in the South in 1910. Inter-regional migration status pertains to region of residence in 1930. The farm laborer category includes a small number of unpaid family workers. Data for 1930 are transcribed from the hand-written census manuscripts.
Sources: Linked census data.

Table 3: Occupational Transition Matrices, 1910 to 1930

	Distribution in 1910	Professional/ Clerical in 1930	Farm in 1930	Crafts/Semi-Skill in 1930	Non-Ag Laborer in 1930
Panel A: Full Sample					
(N=1,829)					
Professional/Clerical	1.5	0.4	0.7	0.2	0.3
Farm	56.8	1.8	33.1	4.7	17.2
Crafts/Semi-Skill	8.0	0.9	2.5	1.1	3.5
Non-Ag Laborer	33.8	1.6	13.8	4.3	14.1
Panel B: Non-Migrants					
(N=1,548)					
Professional/Clerical	1.6	0.5	0.8	0.1	0.3
Farm	59.1	1.7	38.8	4.4	14.3
Crafts/Semi-Skill	7.6	0.8	3.0	1.0	2.8
Non-Ag Laborer	31.7	1.3	15.9	3.0	11.6
Panel C: Migrants					
(N=281)					
Professional/Clerical	0.7	0.0	0.0	0.4	0.4
Farm	43.8	2.5	1.8	6.1	33.5
Crafts/Semi-Skill	10.3	1.4	0.0	1.4	7.5
Non-Ag Laborer	45.2	3.2	2.5	11.4	28.1

Notes: The base sample for this table includes men who were age 21 to 40 in 1910 and had occupation reported in both 1910 and 1930. Each cell reports the percentage of the panel’s sample that transitioned from one category to another between 1910 and 1930 (e.g., 17.2 percent of all farm workers in 1910 transitioned to non-farm, unskilled labor by 1930). Within each panel, the 1930 percentages sum to 100.
Sources: Linked census data.

Table 4: Inter-regional Migration Propensities, Descriptive Regressions

	Age 0-40	Age 10-40	Age 0-16	Age 5-16	Age 17-40
Personal characteristics					
Literacy	---	0.0135	---	---	0.00254
		(0.0160)			(0.0200)
Min. distance to Chicago or Philadelphia, 100 mile units	-0.0353**	-0.0317**	-0.0359**	-0.0359**	-0.0341**
	(0.00741)	(0.00790)	(0.00763)	(0.00934)	(0.00709)
Small city resident	0.0557**	0.0756**	0.0200	0.0625	0.0451
	(0.0230)	(0.0246)	(0.0445)	(0.0551)	(0.0450)
Large city resident	0.0371	0.0289	0.0673	0.0982	-0.0217
	(0.0371)	(0.0411)	(0.0480)	(0.0722)	(0.0590)
School attendance	---	---	---	0.0297	---
				(0.0176)	
Farm occupation	---	---	---	---	-0.0453
					(0.0360)
Household characteristics					
Parent literate	---	---	0.0290	0.00801	---
			(0.0178)	(0.0211)	
Parent in farm occupation	---	---	-0.0182	0.00163	---
			(0.0323)	(0.0374)	
Owner-occupied housing	0.0106	-0.00336	0.0540**	0.0416	-0.0233
	(0.0141)	(0.0163)	(0.0210)	(0.0294)	(0.0203)
Number of siblings in hh	---	---	-0.00198	-0.00342	---
			(0.00377)	(0.00536)	
Local characteristics					
Log state income per capita	0.0437	0.0356	0.0414	0.0459	0.0454
	(0.0711)	(0.0812)	(0.0654)	(0.0771)	(0.0700)
Black percent of population	0.0255	0.0117	0.0490	-0.0110	0.0202
	(0.0308)	(0.0401)	(0.0577)	(0.0641)	(0.0452)
Black adult literacy	0.157*	0.201*	-0.0350	-0.0259	0.325**
	(0.0840)	(0.0939)	(0.106)	(0.133)	(0.100)
Black school attendance	-0.0293	-0.0914	0.0354	-0.0438	-0.154**
	(0.0523)	(0.0647)	(0.0613)	(0.0731)	(0.0708)
Percent of farm acres in cotton, 1909	-0.0217	-0.0180	0.0394	0.0658	-0.0586
	(0.0836)	(0.103)	(0.110)	(0.112)	(0.108)
N	5,421	3,671	2,444	1,691	1,987

Note: Regressions are linear probability models. All specifications include age fixed effects. Standard errors are clustered by state of residence in 1910. Local characteristics are county-level variables in 1910 unless otherwise specified. ** denotes p-value <= 0.05; * denotes 0.5 < p-value <=0.10.

Table 5: Log Earnings Differentials by Migrant Status, 1930

	1	2	3	4	5
Nominal, 1928-based	1.112	1.064	1.052	0.990	1.017
	(0.0157)	(0.0169)	(0.0215)	(0.123)	(0.193)
Nominal, 1960-based	1.173	1.108	1.091	0.987	1.000
	(0.0155)	(0.0171)	(0.0217)	(0.118)	(0.174)
Real, 1928-based, COL1	0.968	0.916	0.903	0.840	0.875
	(0.0158)	(0.0169)	(0.0214)	(0.123)	(0.193)
Real, 1960-based, COL1	1.018	0.960	0.942	0.837	0.858
	(0.0157)	(0.0172)	(0.0216)	(0.117)	(0.173)
Real, 1928-based, COL2	0.917	0.867	0.856	0.806	0.831
	(0.0150)	(0.0161)	(0.0204)	(0.119)	(0.187)
Real, 1960-based, COL2	0.967	0.911	0.895	0.803	0.814
	(0.0151)	(0.0165)	(0.0207)	(0.114)	(0.170)
Controls for personal and county characteristics in 1910	No	Yes	Yes	Yes	Yes
1910 County fixed effects	No	No	Yes	Yes	---
1910 Household fixed effects	No	No	No	No	Yes
N	4259	4259	4259	448	448

Notes and sources: Each coefficient is from a separate regression of log earnings on migrant status (=1 if inter-regional migrant). All are statistically significant at the 5 percent level. The list of control variables differs across the columns. Standard errors are adjusted for clustering at the household level. Column 1 has no control variables. Column 2 controls for age fixed effects, veteran status, city status, distance to Chicago or Philadelphia, owner-occupied housing interacted with headship status, state-level log income per capita, black percent of county population, black adult literacy rate in the county, black children's school attendance in the county, and percent of farm acres in cotton. All controls pertain to 1910 except veteran status. Column 3 adds county fixed effects. Column 4 is the same specification as column 3 but includes only households with multiple observations. Column 5 adds household fixed effects. County fixed effects are irrelevant when household fixed effects are included, but coefficients on personal characteristics (age, veteran status) are still identified. COL1 is the state-level cost of living index from Williamson and Lindert (1980), which is based on Stecker (1937) and Koffsky (1949). It makes an adjustment to the urban COL index from Stecker using the proportion of the labor force in agriculture (presumably in 1930) and the gap between farm and non-farm costs of living from Koffsky (1949). We created COL2 by assigning indices to residents in cities specifically covered by Stecker, then assigning indices to residents in cities with at least 25,000 residents but not covered specifically in Stecker (these get the weighted average of indices for cities in the same state that are covered by Stecker), then finally assigning indices to other residents in each state by making a downward "Koffsky adjustment" to Stecker's figures for cities in that state. See the text and data appendix for more discussion.

Table 6: Log Earnings Differentials by Migrant Status, 1930,
with Actual and Counterfactual Earnings

	1	2	3	4	5
Actual, 1928-based earnings	1.112 (0.0157)	1.064 (0.0169)	1.052 (0.0215)	0.990 (0.123)	1.017 (0.193)
Counterfactual, 1928-based earnings	0.664 (0.0172)	0.603 (0.0183)	0.588 (0.0227)	0.532 (0.126)	0.577 (0.196)
N	4259	4259	4259	448	448

Notes and sources: See table 5. The counterfactual earnings level assigns southern-specific earnings to northern migrants in the same job type (defined on basis of industry). The difference in coefficients between row 1 and row 2 is due to regional differences in pay within industries. The remaining difference is due to differences in the distributions of migrants and non-migrants over job types.