Birth Cohort and the Black-White Achievement Gap: The Role of Health Soon After Birth

Kenneth Y. Chay
Jonathan Guryan
Bhashkar Mazumder

July 2008
(Preliminary and Incomplete)

Birth Cohort and the Black-White Achievement Gap: The Role of Health Soon After Birth
Kenneth Y. Chay, Jonathan Guryan, and Bhashkar Mazumder
July 2008

ABSTRACT

A large literature documents the significant gap in average test scores between blacks and whites, while a related literature finds a substantial narrowing of the gap during the 1980's, and a stagnation in convergence during the 1990's. We use two data sources – the Long Term Trends NAEP and AFQT scores for the *universe* of applicants to the U.S. military between 1976 and 1991 – to show that most of the racial convergence in the 1980's is explained by relative improvements across successive cohorts of blacks born between 1963 and the early 1970's and *not* by a secular narrowing in the gap over time. Furthermore, these across-cohort test score gains occurred almost exclusively among blacks in the South.

We then examine the potential *causes* of these large *composition effects* in the test score gap and their significant variation across U.S. states. We demonstrate that the timing of the cohort-based AFQT convergence closely tracks the convergence in measures of black and white infant health for those cohorts. For example, the cohort-specific AFQT gaps (adjusted for age and year effects and selection into test taking) and the racial gaps in post-neonatal mortality rates – deaths between one month and one year of birth – exhibit very similar patterns across states and birth cohorts. We find little evidence that other potential confounders (e.g., schooling desegregation, family background) can explain these patterns. Investments in health at very early ages appear to have large, long-term effects on human capital accumulation. We also discuss preliminary evidence that the staggered timing of hospital integration across the South is consistent with the patterns of gains in black test scores 17 to 20 years after birth – more so than other hypotheses for progress in black infant health (e.g., Food Stamps, AFDC, Medicaid).

Kenneth Y. Chay
Department of Economics
Brown University
Box B
Providence, RI 02912
and NBER
kenneth_chay@brown.edu

Jonathan Guryan
Graduate School of Business
University of Chicago
5807 S. Woodlawn Avenue
Chicago, IL 60637
and NBER
jonathan.guryan@chicagoGSB.edu

Bhashkar Mazumder
Federal Reserve Bank
230 S. Lasalle Street
Chicago, IL 60604
bmazumde@frbchi.org

I. Introduction

By most measures, there is a significant gap in skills between blacks and whites in the U.S. One such measure that has received a great deal of attention by social scientists and the public alike is standardized test scores. Yet, for all of the discussion of the black-white test score gap, little is known for sure about its source or about what policies, if any, could effectively narrow it.

In this paper we argue that the convergence in the measured black-white test score gap that occurred during the 1970's and 1980's is informative about the source of at least part of the gap, and may point to one type of policy that has been successful at narrowing it. We present data from two datasets: one that contains National Assessment of Educational Progress-Long Term Trend (NAEP-LTT) scores from 1971 to 2004, and another with Armed Forces Qualifying Test (AFQT) results from every applicant to the U.S. military between 1976 and 1991. We view the former as a random sample, though one that is too small to make comparisons at detailed geographic levels.[1] The latter is much larger, but only includes those who chose to apply for induction into the military. We correct for this selection using Inverse Probability Weighting (IPW), whereby we weight each observation by an estimate of its likelihood of selection into the sample. Ours is a particularly nice design for IPW because we know the universe of those selected, and therefore only need to estimate the population size for a particular group to have an estimate of that group's likelihood of selection. We use census and natality data to estimate these population sizes.

Using both datasets, we show that the convergence in the black-white test score gap that was observed during the 1980's is better understood as having accrued to successive cohorts of blacks born between 1963 and the early 1970's. For example, in the AFQT data where we have the largest samples, we show that southern black and white test scores tracked each other closely for the cohorts born in the late 1950's and early 1960's. Beginning with the cohort born in 1963, there was a sharp improvement in

---

[1] A random sample of 9-, 13-, and 17-year-old students who are enrolled in school are tested. Selection is therefore induced by high-school dropouts. We include the 17-year-olds so that we can statistically distinguish cohort and year fixed effects, and have checked that trends in high-school completion rates are not different by region in a way that would explain the patterns we see in the data.

black test scores that lasted for about ten cohorts, and which reduced the black-white test score gap by about 40 percent. This cohort-based convergence, which we contrast with a narrowing that affected children of all ages in a given calendar year, explains all of the narrowing of the black-white AFQT gap during the 1980's in the South, and none of that narrowing in the North.

That the convergence in the black-white test score gap during this period lines up by cohort, conditional on time effects, is a remarkable finding on its own. It suggests that the search for causes of the black-white test score gap narrowing should not be restricted to the time at which the convergence became apparent. It opens up the possibility that interventions or environmental changes that affected successive cohorts differently – possibly at earlier ages – were the initial cause of the relative improvement of black test scores.

With this empirical finding as motivation, we then propose and test a specific hypothesis for the cause of the cohort-based convergence in the black-white test score gap — that the cohort-based convergence in the black-white test score gap was caused by relative improvements in black infant health. This *infant health hypothesis* states that some intervention or set of interventions improved the health of black infants born between the mid-1960's and the mid-1970's, and that this health improvement led to long-term improvements in black skills for those cohorts, measured 9 to 20 years later. In the absence of a perfect measure of latent infant health, we take post-neonatal mortality (PNMR) to be a proxy. PNMR is the number of infants that die between one month and one year after birth, divided by the number of live births. We discuss weaknesses of this measure of infant health later in the paper.[2]

We then relate the timing of the black-white test score gap narrowing to the convergence in the black-white PNMR gap in various geographic areas. In short, we find a remarkable correspondence between the two measures. Just as for AFQT scores, black and white PNMR tracked each other closely for the cohorts born between the late 1950's and the early 1960's. Then in the mid 1960's, black PNMR

[2] We discuss, for example, the fact that infant health may improve without affecting the number or fraction of babies that survive, and the fact that mortality rates are inherently linked with selection bias.

began to decline sharply in the South. The size of the AFQT dataset allows us to compare the timing of the convergence of the two series across states within and across regions.

We break states up into groups according to historical racial patterns in PNMR. We then show that the timing of PNMR convergence across states within the South matches remarkably well with the relative in those states in black-white AFQT convergence across birth cohorts. For example, whereas the black-white PNMR gap began to converge two years later in Alabama and Mississippi than in South and North Carolina, the same two year gap is found between the cohorts of blacks that began to experience relative improvements in AFQT scores. We show similar comparisons across other states in the South and across groups of states within the North.

In both traditional and difference-in-difference comparisons, the black-white test score gap begins to converge with the cohort born in one or two years before the black PNMR convergence begins. This pattern suggests that an improvement in infant health in the first 1.5 to 2.5 years for black children might be the cause of the narrowing of the black-white test score gap.[3] In a subsequent section, we investigate one such intervention—the racial integration of Southern hospitals—and ask whether it might explain the narrowing of both racial gaps. Almond, Chay and Greenstone (2008) argue that the forced integration of hospitals led to significant improvements in black infant health, with no corresponding negative effect on whites. We show that the timing of AFQT improvements matches well with the convergence of the black-white gap in the rate of births that took place in a hospital.

We then investigate a number of alternative causes of the racial convergence in both test scores and infant health. Though they are alternatives to hospital integration as a root cause, some of these stories also share the feature that black health improvements are the mechanism for the narrowing of the black-white test score gap. These alternative stories include some of the large social programs of the mid-1960's that were part of the War on Poverty (e.g. AFDC, Medicaid, Food Stamps, Head Start); the relative improvement in black earnings that resulted in part from the Civil Rights Act; and school

[3] PNMR is recorded by the year of death, not the year of birth. If post-neonatal deaths were uniformly distributed across the eleven months, this would mean the dates we report for PNMR are about 5.5 months later on average than the dates of birth.

desegregation. We conclude that some of these stories may have contributed to black infant health, and therefore to the narrowing of the black-white test score gap. Thus far, however, it does not appear that any of these stories match all the patterns in the data as well as the measure of infant health does. Furthermore, the stories that do not appear tightly linked with infant health (e.g. school desegregation) do not appear to match he patterns of cohort-based convergence in test scores. Continued research will compile additional evidence on the roles of each of these stories as root causes for the black-white test score gap convergence.

The paper proceeds as follows. The next section presents some background on trends in infant mortality. In section III, we show results from the NAEP-LTT. In section IV, we describe the military applicant dataset that has the AFQT scores. Section V describes inverse probability weighting which we use to correct for selection in the sample for which we have AFQT scores. In section VI, we present the AFQT results. Section VII states and tests the infant health hypothesis. Section VIII presents evidence on hospital integration. In section IX, we discuss alternative root causes. And, in section X, we conclude.

II. Aggregate trends in infant mortality, 1950-2000

Before turning to the analysis of test scores, we pause here to present some background on trends in infant mortality. Below, we will present evidence that improvements in black AFQT scores accrued to cohorts born in states in which infant health was improving. Though we do not have evidence health at all ages of infancy, natality data provide good data on mortality rates of children in the first year of life. The second half of the 20th century saw a remarkable improvement in these indicators for black children. Figure 1a plots the black-white difference in infant mortality, defined as deaths per 1,000 births within the first year of life, for the years 1950 and 2000. During the 1950's, there was a fairly stable black-white gap in infant mortality of about 20 births per 1,000. Near 1964, this gap began to narrow dramatically. Within about ten years, the black-white infant mortality gap had closed to about 12 births in 1,000. By 1980, the gap was close to half of what it was in the mid 1960's.

Figure 1b separately plots the black-white gap in neonatal (deaths within one month of birth per 1,000) and post-neonatal (deaths between one month and one year following birth per 1,000) mortality for 1950 to 1991. As can be seen in the figure, a majority of the decline in infant mortality from the mid-1960's to the mid 1970's occurred during the post-neonatal period. Below we argue that the dramatic improvement in infant health of which this decline is indicative may have had long-term effects on human capital accumulation for blacks born during this period.

III. NAEP results

We begin by examining data from the National Assessment of Educational Progress Long Term Trends (NAEP-LTT) test. The NAEP-LTT is one of two tests administered by the U.S. Department of Education aimed at documenting national patterns of achievement in the nation's schools. Designed to measure trends over time, the NAEP-LTT maintains a constant testing frame. Since 1971, the NAEP-LTT has been given in various years to a random sample of 9-, 13- and 17-year olds enrolled in U.S. schools.[4] Below we analyze microdata of students' math and reading scores for all available years in which the tests were administered, for both boys and girls.[5] [6]

Figure 2a plots scaled scores separately for black and white students by calendar year, along with the black-white difference.[7] To make use of all available data, we combine math and reading scores and include tests from all three age groups. The figure plots the year effects from a regression that controls for race-specific age and subject effects. Consistent with the previous literature, the figure highlights a

[4] The standard NAEP, sometimes called "The Nation's Report Card," has been given since 1969, and the testing framework changes over time to account for changes in national curricula.

[5] The reading test was given in 1971, 1975 , 1980, 1984, 1988, 1990, 1992, 1994, 1996, 1999, and 2004. The math test was given in 1973, 1978, 1982, 1986, 1990, 1992, 1994, 1996, 1999, and 2004. We were not able to obtain the 1973 math scores. Scores from all other tests listed above are included in the analysis.

[6] Though we restrict our sample to men only in the subsequent analysis of AFQT scores, we include both boys and girls here. We make the restriction to men only in the military sample because we are primarily concerned with non-random selection and we believe the selection process is more constant over our sample period for men than for women. We include girls in the NAEP-LTT analysis because we are primarily concerned with increasing sample size to improve precision, and because boys and girls are selected into the NAEP test-taking sample in the same way.

[7] We show results from scaled scores rather than some other transformation because these are the scores that are reported in public releases of the data, and to match some of the literature (e.g. Dickens & Flynn (2006), Cook & Evans (2000)).

marked convergence in the black-white test score gap during the 1980's. From 1971 to 1978, the black-white gap in NAEP-LTT scores remained fairly constant, declining from 42.7 to 39.3 points (the standard deviation in NAEP-LTT scores is on the order of 45 points). During the decade of the 1980's, black scores increased while white scores remained virtually unchanged. By 1990, the black-white test score gap had declined to 26.4, or about 60 percent of a standard deviation. This convergence halted abruptly in 1990, and for the next 15 years remained virtually constant.

In figure 2b, we plot the NAEP-LTT series separately for each age cohort. When the 9, 13- and 17-year-old series are plotted by year, an interesting pattern emerges. The black-white convergence seen in figure 2a appears to have begun earliest in the 9-year-old test scores, followed by the 13-year-old scores, and lastly by the 17-year old scores. The black-white gap in 9-year-old NAEP-LTT scores begins converging at some point before 1974 (because we have no data prior to 1971 it is impossible to know whether the change from 1971 to 1974 was the continuation of an earlier trend). For 13-year-olds, the convergence appears to begin sometime between 1978 and 1980, while for 17-year-olds the convergence begins sometime between 1980 and 1982. This pattern suggests that a better way to describe the convergence is that successive cohorts of blacks, beginning with the cohort born sometime in the early-to-mid 1960's, experienced relative improvements in NAEP-LTT scores.

In figure 2c and 2d, we examine this possibility more directly. Instead of plotting test scores against the year the test was taken, we plot them against the year the student was born. Figure 2c shows the series separately for blacks and whites, along with the black-white difference. Again, both math and reading scores are included for all three ages, and the year effects shown are from a regression that includes controls for race-specific age and subject effects. As in the figure that shows trends over time, white scores appeared to be fairly constant over successive birth cohorts. And, for blacks born between 1953 and 1964, there was little upward trend in NAEP-LTT scores, though the data do exhibit a good deal of cohort-to-cohort fluctuation among blacks in these early years.

The most striking pattern in the figure is that over 10 birth cohorts—between 1964 and 1973—black NAEP-LTT scores improved by approximately 20 points (about 45 percent of a standard deviation) while white scores remained virtually unchanged. Then around 1974, just as abruptly the convergence stopped and black scores tracked white scores for the next 20 years.

A natural question to ask is whether the convergence in black-white scores accrued to successive cohorts of blacks, or whether the convergence can be fully explained by secular improvements that affected black children of all ages. Furthermore, given the different historical experiences in the U.S. South and North, it is natural to ask whether black-white test score convergence followed different patterns in these two regions.

Figures 2d provides some evidence on these questions. For the South and North, the figure plots the black-white difference in birth-year effects from regressions run separately for blacks and whites that also includes year, age and subject effects. By including both birth-year and year effects, we are asking the data to tell whether the cohort-based convergence seen in Figure 2c is truly indicative of something that affected successive cohorts between 1963 and 1974 differently, or whether secular improvements in resources afforded blacks during the 1980's happened to affect those cohorts because they were the ones in school at the time.

Notice that Figure 2b illustrates how it is possible to identify unrestricted cohort, age and time effects. For each race, we restrict the pattern of age effects to be constant across time. Year effects allow this age profile to shift up and down in a parallel fashion year to year. Variation in the shape of this age profile in different years is left to identify cohort effects. This cohort-based convergence is exhibited in figure 2c by a convergence that occurs four years later in the 13-year-old series than in the 9-year-old series, and another four years later in the 17-year-old series.

The solid black line in Figure 2d shows that in the South there was indeed a significant improvement in black test scores beginning with the cohort born around 1964 and continuing until the cohort born in 1973. This cohort-based convergence is apparent even after controlling for black×year effects. It implies that the improvement in black 13-year-old NAEP-LTT scores began approximately four years after the improvement in black 9-year-old NAEP-LTT scores began, and approximately four years before the improvement in black 17-year-old scores began.

The dashed line in Figure 2d plots a similar series for the North.[8] In contrast with the pattern for southern blacks, there is not a clear improvement among northern blacks born during the 1960's, once we condition on year effects. Depending on how one chooses beginning and ending years, it may be possible to discern an improvement by blacks in the North, but this improvement is masked by a great deal of cohort-to-cohort variation.

This set of patterns is provocative and suggestive. It raises the possibility that something happened to improve the relative standing of blacks born between 1964 and 1973; that whatever caused this improvement affected successive cohorts more significantly but did not affect children of all ages in a way that would improve current test scores in the same way; did not negatively affect whites; and did not affect blacks in the North in the same way that it did blacks in the South.

Unfortunately, the NAEP-LTT sample is not large enough to explore more detailed comparisons that might allow us to pin down what such a cause might be. Furthermore, the structure of the NAEP-LTT does not allow us to identify cohort effects for each birth year during the critical period. For these reasons, we next turn to a much larger dataset that has test scores for both blacks and whites over a long time period and for the set of cohorts among which we observe this convergence, as well as a set of cohorts born during the decade before.

IV. Dataset with AFQT Scores from applicants to the U.S. Military, 1976-1991

To document whether the cohort-based convergence seen in the NAEP-LTT data is a phenomenon seen in other data and to isolate its source, we would ideally collect test scores from a large random sample of blacks and whites from multiple birth cohorts, measured in different calendar years. To have the best chance of identifying whether the convergence is cohort-based or secular, we would collect the data on an annual basis and make sure to collect data from every birth cohort during the period

[8] To match the subsequent analysis of the AFQT data for the "Rustbelt", we define the North to be the Northeast and Central regions. The vast majority of blacks living outside the South live in the states in these regions.

under study. In addition, the dataset would be large enough and geographically diverse enough to allow comparisons across regions and even states.

In this paper, we present results from a dataset that satisfies many of these criteria, though not all. We describe the data and its strengths and weaknesses here. The data includes the universe of applicants to the U.S. military between 1976 and 1991. The data include the age, completed education, and zip-code of residence, all measured at the time of application. For the analysis, we restrict our attention to males who were ages 17-20 at the time of application. Each applicant takes a battery of tests, called the Armed Services Vocational Aptitude Battery (ASVAB), various components of which are combined to form a summary score used for screening purposes. This summary is called the Armed Forces Qualifying Test (AFQT), and is commonly used by economists as a measure of cognitive ability. The AFQT score is a percentile relative to a nationally representative sample of 18-23 year olds from the Profile of American Youth 1980.[9]

The AFQT data are summarized in Table 1. The AFQT data include 2,649,573 white males and 1,103,748 black males, born between 1957 and 1974. Importantly, this span includes a number of cohorts born in the period before the convergence began in the NAEP-LTT data, and the full decade of cohorts for which we saw the narrowing of the black-white test score gap. Unfortunately, the data does not include cohorts born after the narrowing stopped. The large sample size, however, allows us to compare the timing of black-white test score convergence both across regions (e.g. South v. Rustbelt), across parts of the South (e.g. the Deep South v. the rest of the South) and across states within regions (e.g. Alabama and Mississippi v. South Carolina and North Carolina).

The major weakness of the military applicant data is that it only includes those who chose to apply to the U.S. military. This group of applicants is clearly not a representative sample of all U.S. born

[9] The Profile of American Youth is a sample from the 1979 National Longitudinal Survey of Youth (NLSY79). One reason the AFQT is used so often by economists is that it is included on the NLSY79. The NLSY sample was used to norm the AFQT using the sample of 18-23 year olds tested in 1979. A well-documented misnorming of the AFQT for the period between 1976 and the renorming based on the NLSY79 sample led the military to inadvertently admit many more low-scoring applicants than it intended during this period. All years of our data are normed relative to the same NLSY79 cohort, even those from the misnorming period. The AFQT was subsequently renormed based on the 1997 NLSY, but this occurred after all of the cohorts in our study took the test.

17-20 year olds. Applications to the military are countercyclical, and blacks are more likely to apply than whites. To obtain an unbiased estimate of black and white average test scores for a given cohort in a given year, we must therefore correct for this selection.

V. Inverse Probability Weighting

We correct for selection by Inverse Probability Weighting (IPW) (see e.g. Wooldridge (2002) for a detailed discussion). In this procedure, we estimate the probability that each observation, or group of observations, is selected into the sample, and then weight by the inverse of that probability. Compared with most situations in which researchers correct for selection, our design is well-suited. Researchers typically have to estimate a selection equation using a sample of those selected. The estimated propensity is then either inserted as a control into a second stage estimating equation, or used to construct inverse probability weights. In contrast, because we know the universe of applicants – the selected population – we know the numerator of the fraction used to estimate the true probability of selection. We are left only to estimate the denominators – the size of the population from which applicants were selected. In the following section, we describe the three ways in which we estimate these denominators, using both Natality and Census data.

Estimating the population size

For each observation, or group of observations, we must estimate the probability of selection into the sample. The true probability of selection for any group is the fraction of the population in that group who applied to the U.S. military. Since we have the universe of applicants to the U.S. military for the period under study, we know the numerator of this fraction. We are left only to estimate the denominator, or the population size. We estimate these denominators in three ways.

IPW_st: For the first set of weights, we estimate the population size for each cohort in each state using data from the National Vital Statistics System. We use the birth and death records to count the number of births that survived to age one, by race in each state in each year. We then take the count of applicants in our data and match by race, state of residence and birth year. A strength of this method is that it uses administrative data on the universe, rather than a sample, to calculate both the applicant and population sizes. A weakness is that the natality data counts births by state of birth, while the applicant data can only be linked by state of residence at the time of application.

IPW_Census: A second set of weights is constructed with the goal of estimating both the numerator and denominator by state of residence. To estimate cohort-sizes, we use the 1970, 1980 and 1990 censuses. Each census can be used to compute population counts by race, state of residence and age, as of the census years. In addition, we use a question that asks respondents where they lived five years ago to compute population sizes by race, state of residence and age in 1965, 1975 and 1985. We then use the nearest of these six cross-sections to compute the cohort size by race for those still living in each state at 17, 18, 19, and 20.

A strength of this method is that it calculates population sizes by state of residence at the time of application, which is presumably the time at which the selection process occurs. A weakness is that there may be selective migration between birth and 17 which this weighting does not address (a separate analysis not reported here suggests migration patterns cannot explain the patterns in AFQT scores we report below). Another weakness is that because we can only measure population sizes every five years, we are forced to use nearby cohorts to estimate cohort sizes. So long as cohorts do not change in size quickly, this is unlikely to have a major effect on the estimates.

IPW_Census_Educ: To recover unbiased estimates of population average test scores, selection must be ignorable conditional on the cells within which we calculate selection probabilities. One concern therefore with the first two weights, is that they presume selection is unrelated to education, conditional on race, state of residence, birth year and age. One might argue that this assumption is too strong since the alternative employment options of more highly educated are less cyclical. With that motivation, we allow the selection probabilities, and therefore the weights, to vary by education, in addition to the dimensions described above.

The relevant notion of education is not eventual years of completed education, since the test is taken at the time of application. Instead, what is relevant is years of completed education at the time the test was administered, or equivalently at the time of application to the military. Because we know this for the applicants, once again we can calculate the size of the test-taking population for each group (i.e. by race × state of residence × year × birth year × completed education at time of application).

To estimate the cohort size by completed education, we begin with the cohort size estimates used to estimate the *IPW_Census* weights. We then use the 1980 and 1990 censuses to estimate the fraction of each group that falls into one of three completed education categories: less than 11 years, exactly 11 years, and more than 11 years. With each census, we estimate the fraction by race × age that fall into each of these three categories. For each cohort, we use the probability from the nearest of the two censuses. The cohort size that varies by race × state of residence × year × birth year is then multiplied by this probability to obtain an estimate of cohort size that varies by race × state of residence × year × birth year × completed education at time of application.[10]

Figure 4 plots the black-white difference in selection probability by year, for the South and Rustbelt separately.[11] It shows that blacks are more likely to apply to the military, and black application rates are typically more countercyclical than for whites. In many of the analyses below, we will compare black-white differences in the South and Rustbelt. It is reassuring, therefore, that the black-white difference in selection probabilities in the South and Rustbelt appear to track each other very closely.

VI. AFQT Results (*Cohort-based convergence*)

We now turn to the estimates of black and white average AFQT scores from models similar to those presented above for the NAEP data. We estimate models of the following form

---

[10] Since this final set of weights does not turn out to effect the estimates we do not report results using them. In many of the specifications, we are able to include race-by-education effects, which subsume much of the variation that distinguishes the second and third weights. Controlling for these fixed effects has much the same effect on the estimates as moving from weights based on census counts to weights based on census counts by education.

[11] The Rustbelt is defined as NY, PA, OH, MI and IL. The vast majority of black men in the U.S. live either in the South or in these states. Our South sample includes 1,395,143 white men and 748,592 black men; the Rustbelt sample has 1,254,430 white men and 355,156 black men.

$$T_{icat} = \delta_t^W + \gamma_c^W + \alpha_a^W + X_{ict}\beta^W + \varepsilon_{icat}$$
$$T_{icat} = \delta_t^B + \gamma_c^B + \alpha_a^B + X_{ict}\beta^B + \varepsilon_{icat} \qquad (1.1)$$

where *i* indexes individuals, *c* indexes year of birth, *a* indexes the age at which the test was taken, and *t* indexes the calendar year in which the test was taken. T denotes the test score, X is a vector of controls, β is a vector of parameters to be estimated, and ε is an error term. $\delta_t^r = \delta_1^r + ... + \delta_T^r$ are the race-specific calendar year effects, $\alpha_a^r = \alpha_{18}^r + \alpha_{19}^r + \alpha_{20}^r$ are the race-specific age effects, and $\gamma_c^r = \gamma_1^r + ... + \gamma_C^r$ are the race-specific birth-year dummies, where $r \in (B, W)$ indexes race, *T* is the total number of calendar years represented our sample, and *C* is the total number of birth years represented in our sample.

There is a well-known problem that it is impossible to separately identify linear age, cohort and time effects. An observation's birth year and calendar year fully characterizes its age. We restrict the age profile of test scores to be the same in each year, allowing this profile to shift up and down year-to-year. This restriction allows us to separate the remaining variation in test scores into the amount that accrued to successive cohorts, and the amount that accrued to all individuals of all ages in particular calendar years. It is our intention in this section to partition the convergence in the black-white AFQT gap into these two parts.

We do not attempt in this section to assign a causal mechanism to the estimated cohort effects. The estimates of the $\gamma_c^B - \gamma_c^W$ differences are descriptive, and do not imply that the cause occurred in year *c*. Without some external information, it is impossible to identify whether, for example, something that affected the 1970 cohort was caused by something that affected babies in 1970 or caused by something that affected 10-year-olds in 1980 and/or 11-year-olds in 1981. We will attempt to identify the timing of the cause in subsequent sections.

In Appendix Figures 2a-c, we present estimates of γ_c^B, γ_c^W, and $\gamma_c^B - \gamma_c^W$, the black-white difference in AFQT scores, for birth cohorts 1957 to 1972, separately for the South and the Rustbelt. The cohort effects are estimated conditional on year effects, age effects and completed education effects.

Panels a and b show the levels for whites and the black-white difference, for the South and Rustbelt, respectively, while panel c shows the South-Rustbelt differences in the white level and the black-white difference. All regressions are estimated separately by race. In the Appendix Figure 2a, six series are plotted, three for whites and three for the black-white difference. Each series corresponds to a different selection correction: uncorrected, weighted by *IPW_st*, and weighted by *IPW_census*. The series track each other fairly closely, suggesting that selection correction does not have a large effect on the estimates.

As can be seen in panel a, the estimated black-white test score gap remained fairly constant in the South at about 20 to 21 percentile points for the cohorts born between 1957 and the early 1960's. This gap is slightly smaller than a standard deviation. Then, beginning with the 1962 cohort, black scores began to rise sharply relative to white scores. Black AFQT scores converged towards white scores at a fairly constant rate for the next 10 cohorts. By the 1972 birth cohort, the 21-point gap had shrunk to about 15 percentile points, a 29 percent decline.

This pattern—a stable black-white test score gap for the five or six cohorts beginning in 1957, followed by a sharp upward trend in black AFQT scores beginning with the 1962 or 1963 cohort that lasts through the early 1970's cohorts—is a robust pattern that is seen in many of the figures to follow. The pattern was very different in the Rustbelt, however. There was some convergence among Rustbelt blacks born in the mid-1960's and early 1970's, but this improvement was much smaller in magnitude than we saw among those in the South.

Appendix Figure 2c plots the South-Rustbelt differences. Among whites, trends over time were quite similar in the South and Rustbelt. There was some variation year to year, but the difference remained between 0.5 and 1.5 percentile points. In contrast, the South-Rustbelt difference in black-white differences follows the pattern described above. Black-white gaps tracked each other closely in the two regions among those born between 1957 and 1962. Then in 1963, AFQT scores of blacks in the South begain to improve sharply. This convergence continued for the next 10 cohorts until for those born in 1972, there was close to no difference in the black-white test score gap across regions.

VII. A Proposed Explanation: Relative Improvements in Black Infant Health

We propose a possible explanation for the sharp decrease in the black-white test score gap that accrued to cohorts born between 1963 and the mid-1970's, particularly in the South. The period during which these cohorts were born was a time of significant improvements in black infant health. Black infant mortality rates, nearly twice that of whites nationally in 1960, fell sharply between 1964 and 1974, closing about half of the gap with white infant mortality rates over that ten-year period. This racial convergence was greatest in the Deep South (AL, MS, GA, LA and SC), the region where the black-white gap in infant health was the greatest in 1960. Absolute and relative improvements in black infant mortality rates were slightly smaller in the rest of the South, and least significant in the North.

We hypothesize that this improvement in black infant health, and the corresponding convergence in the black-white infant health gap, explains a significant portion of the cohort-based convergence in the black-white test score gap that we have shown in both NAEP and AFQT scores. This hypothesis implies that investments in early-life health have long-term effects on human capital accumulation, possibly because the cost of future human capital accumulation is affected.

Consider for example an intervention in 1964 that improves the health of blacks between the ages of zero and 3 years. If the improvement in infant health has long-term effects on measured cognitive skills, we should expect to see improvements in AFQT scores for blacks born in 1962, 1963 and 1964 relative to those born in 1960 and 1961. If a similar intervention affected the health of children up to 4 years of age, we would then expect to see improvements in AFQT scores for those born after 1960.

A. Regional comparisons of black and white test scores and infant health convergence

We begin to investigate the infant health hypothesis with the evidence shown in Figure 3. Figure 3a shows four series, in blue the post neonatal mortality rates (PNMR) for white and in black the black-white difference in PNMR, for both the South and Rustbelt.

The PNMR is the number of deaths between ages one month and one year divided by the number of births. Ideally, we would have a measure of the health status in each year of all infants, both surviving and otherwise. Unfortunately, we do not have a measure of this latent variable. Instead, we take PNMR to be a proxy for latent infant health. We focus on post-neonatal mortality as opposed to neo-natal mortality for two reasons. First, the timing in PNMR matches the timing in AFQT convergence, while the same is not true for neo-natal mortality rates. And second, we believe PNMR is a better measure of the conditions that affect infant health specifically, rather than those that affect health *in utero*. It is not that these other conditions are unimportant; rather changes in those factors do not appear to have caused the specific convergence in black-white skills that we documented in the previous section.

PNMR has two main weaknesses as a measure of infant health. First, there may be improvements in infant health that do not affect mortality rates. In fact, Almond, Chay and Greenstone (2008) find that the integration of hospitals in the South caused increases in black hospital delivery rates about one year before it caused increases in PNMR, suggesting that access to healthcare might have positive effects on other health outcomes before the improvements in health are great enough to affect mortality.

Second, decreases in PNMR are mechanically associated with selection. To the extent that the marginal surviving infant is negatively selected, decreases in PNMR would be expected to be associated with decreases in average human capital of the survivors. To the extent that we find positive effects, our estimates can therefore be viewed as lower bounds on the effect of improvements in infant health on long run increases in test scores.

As can be seen in the figure, white PNMR was declining fairly gradually during this period in both the South and Rustbelt. The black PNMR tracked the white PNMR in the Rustbelt, the black-white gap declining slightly between 1966 and 1974. The black PNMR in the South exhibited a very different pattern. From the late 1950's through the first few years of the 1960's the black-white PNMR gap in the South remained constant at about 14 deaths per 1,000. Then around 1964, black PNMR began to decline sharply in the South. The resulting convergence in black PNMR continued at a constant rate at least through 1972. What began as a difference of 8 post-neonatal deaths per thousand live births, between 1964 and 1972, the difference in black PNMR between the South and Rustbelt virtually disappeared.

The analogous series for AFQT scores are plotted in Figure 3b and the regression coefficients underlying this table can be found in Table 2. Remarkably, the pattern in the convergence in black PNMR mirrors that of the cohort effects in black AFQT scores. AFQT scores of Southern blacks born in the late 1950's into the early 1960's tracked those of Southern whites. Then beginning with the cohort born in 1962, Southern blacks began to improve relative to Southern whites. This relative improvement continued for the next ten birth cohorts. The South-Rustbelt gap in both PNMR and AFQT were fairly constant in the years leading up to the early 1960's. Then, each began a sharp convergence, PNMR in 1964 and AFQT in 1962. By 1972, the South-Rustbelt gap in each was virtually erased. This close relationship in timing, estimated non-parametrically, is consistent with the infant health hypothesis.

Why however does the convergence in the black PNMR series begin two years before the convergence in black AFQT scores? PNMR is measured by year of death not birth, so some of the deaths included in year *t* are of babies born in year *t-1*. More significantly, recall that we take PNMR to be an indicator of infant health. If the intervention that led to the improvement in PNMR positively affected the health of infants between zero and 18-24 months, we would expect to see improvements in health in cohorts born a year before the improvement in PNMR. We investigate one possible such intervention, the integration of Southern hospitals, in section VII.

Figure 3c shows a complementary interesting fact. Here we plot the year effects from the same regressions estimated to produce Figures 3a and 3b, along with the unconditional trend in black and white AFQT scores by year in the South. The dashed lines show the unconditional trend, an increase in black relative to white scores that was observed in the 1980's. The solid black line shows that after conditioning on birth-year effects, this black-white test score gap narrowing during the 1980's is significantly less pronounced. Most of the relative improvement in black AFQT scores appears to have accrued to successive birth cohorts, rather than to all test-takers in successive test years.

One might be concerned that the contemporaneous convergence in black PNMR and AFQT scores between the South and Rustbelt simply captures general convergence between the two regions that occurred in the mid- to late-1960's. On this point, recall two features of the estimates shown in figure 3.

First, the plotted AFQT differences are residual differences after controlling for year effects, year-specific education and age effects. And second, the AFQT is measured 17-20 years later. So, if it is the case that changes in the South relative to the Rustbelt that occurred in the 1960's caused improvements in black AFQT scores almost two decades later, this is strong evidence that investments early in life have important long run effects. Though we think the data we will present in this paper argues strongly that the intervention that caused the convergence in the black-white test score gap was something that affected infants, even this more general statement—that early life investments in health caused the convergence in the black-white test score gap—is both important and striking.

In spite of these arguments, one might still be concerned that regional trends in these two series just happen to line up well. Figure 4 presents results that argue strongly against this possibility. Whereas figure 3 plotted the black-white difference in Southern PNMR and AFQT cohort effects, figure 4 introduces another level of differencing. Now, the South-Rustbelt difference in black-white differences in the two series are shown. The results are almost identical, implying there was no differential cohort-based convergence among whites like what is apparent for blacks, nor did the same black-white convergence occur in the Rustbelt as in the South.

The pre-1963 trend in AFQT cohort effects remains fairly flat. Then, beginning with the 1963 birth cohort, black AFQT scores began a sharp increase relative to whites, and this increase occurred only among Southern blacks. The black-white PNMR gap in the South relative to the Rustbelt began to converge sharply beginning in 1964. By the 1972 birth cohort, both PNMR and AFQT scores were relatively better for blacks relative to whites in the South as compared with the Rustbelt. The fact that the black-white South-Rustbelt difference-in-difference for both series is so similar to the black South-Rustbelt difference implies that all of the convergence in both PNMR and AFQT was driven by improvements among blacks. In fact, this can be seen directly by the fact that the South-Rustbelt difference in PNMR and AFQT is remarkably flat during the entire period.

Figure 4b has another interesting comparison. The black-white AFQT gap is plotted by birth year with two different sets of controls. The dashed line comes from a regression that allows the education fixed effects to vary by race. This is the specification that is used in all subsequent figures. In contrast, the solid black line comes from a regression that restricts the education effects to be the same for blacks and whites. This specification would be appropriate if one wanted to take racial differences in the returns to education to be an outcome of the early life health improvements we document with the PNMR series. Under this assumption, the gap between the two series can be interpreted as the amount of black-white AFQT convergence that is attributable to improvements in human capital accumulation rates as a result of early life health improvements for blacks.

B. State-by-state comparisons of black and white infant health and test score convergence

The South-Rustbelt comparisons exploit regional differences in the rate and timing of convergence in PNMR by race between these regions. There was also variation within region, across Southern states in the speed and timing of this convergence in racial gaps in infant health. As shown in Almond, Chay, and Greenstone (2008), the improvement in black PNMR was greatest in the Deep South, where the initial gap was largest, smaller but significant in the rest of the South, and least significant in the North where the initial gap was smallest.

The large sample size in the military applicant data allows for statistically meaningful comparisons across smaller geographic areas, such as states. Was the convergence in the black-white AFQT gap also largest in the Deep South? Did the variation in the timing of black-white AFQT gap convergence across states match the variation in the timing of black-white infant health convergence?

To investigate, we divide the Southern states into four groups based on the patterns of black-white PNMR convergence: Alabama and Mississippi, South Carolina and North Carolina, Tennessee and Virginia, and Florida and Louisiana.[12] The patterns in black-white PNMR convergence can be seen for each of these four pairs of states in Figure 5A. The dashed line with triangles shows the series for North and South Carolina. The experience in these two states was fairly similar to the overall pattern in the South. The black-white PNMR gap remained near 17 between 1957 and 1963. In 1964, black PNMR

[12] We leave out some states to keep the number of series plotted on the figure a readable amount.

began to decline sharply and by 1974 the gap narrowed to about 5. Compare this pattern with that of Mississippi and Alabama. The racial PNMR gap tracked that of the Carolinas, though at a slightly lower rate until 1963. However, the gap does not converge in AL-MS in 1964. Instead it remains constant for another two years, until it begins to converge sharply in 1966. The TN-VA gap started at a lower level and remained flat until it also began to converge in 1964. The most dissimilar series is for FL-LA. Convergence in these states was gradual during the entire period under study.

In figure 5B, we compare these cross-state patterns in racial PNMR gaps to those for AFQT scores. The correspondence is striking. To begin, compare AL-MS to NC-SC and TN-VA. Just as for PNMR, the black-white gap in all three pairs of states was flat for the cohorts born between 1957 and 1961, the smaller PNMR gap in TN-VA corresponding to a smaller AFQT gap during these years. Then in 1962, black AFQT scores in NC-SC and TN-VA began to converge to whites. Remarkably, just as for PNMR, the racial convergence in AL-MS began two years after that in NC-SC and TN-VA. Lastly in FL-LA, the pair of states in which the black-white PNMR gap narrowed continuously over the period, the racial AFQT gap also narrowed continuously.

This cross-state comparison is most striking within the South, where the convergence in PNMR and AFQT were the most dramatic, but it is possible to make similar comparisons within the Rustbelt. Figure 5C plots the black-white PNMR gap for IL-NY and MI-OH. The figure also plots the series for TN-VA for comparison purposes. Though the relative improvement in black PNMR was smaller in these Northern states—the gaps were smaller to begin with—there was a difference in the patterns between these two pairs of states. In MI-OH, there was virtually no improvement in black relative to white PNMR. However, between 1957 and 1974 black PNMR gradually declined relative to whites in IL-NY.

Figure 5D plots the AFQT cohort effects for these pairs of states. Just as for PNMR, there is no improvement in black AFQT relative to whites in MI-OH. There is, however, a gradual decline in the black-white AFQT gap in IL-NY over the cohorts born between 1957 and 1972. Furthermore, the racial convergence in both PNMR and AFQT in IL-NY was about a third as large as in TN-VA.

C. Magnitudes of changes in PNMR and AFQT

Thus far, the analysis has focused on the year to year timing of when racial convergence in PNMR and AFQT began in different areas. We have not focused on the other important feature of these series, the large change in black-white gaps over a ten-year period. In Tables 3-6 we present regression estimates of these magnitudes. Table 3 begins with the South-Rustbelt comparison. Column 1a shows the average white AFQT score for the South, Rustbelt and their difference for the 1960-62 birth cohorts, after controlling for race*age, race*time and education fixed effects. All regressions are estimated separately by region. In parentheses are standard errors, which allow for unrestricted clustering at the state level, while in curved brackets are PNMR for the relevant group. Column 1b, then shows the change in AFQT scores in each region, along with the double-difference, between the 1960-62 cohorts and the 1970-72 cohorts. Columns 1c and 1d then show the same set of outcomes, but for black-white differences. The South-Rustbelt, black-white difference in difference is thus found in row C of column 1d. It shows that relative to the Rustbelt, the black-white AFQT gap in the South rose by 4.94 percentile points, while the relative black-white PNMR gap fell by 6.92 births per 1,000. The right-hand panel (columns 2a-d) replicates this analysis while allowing education effects to vary by race. As discussed above, the relative AFQT improvement is slightly smaller when racial differences in test-score-returns-to education are assumed to be controls rather than outcomes.

Table 4 repeats the analysis adding various combinations of controls. Column 1 corresponds to the left-hand panel of Table 3, while column 3 corresponds to the right-hand panel of Table 3. The inclusion of further controls, especially those that allow returns to education to vary, generates slightly smaller estimates, though all are significant at the 1-percent level. The model in column 9 includes a comprehensive set of interactions: region-race-cohort, region-race-age, region-race-year, region-education, race-education, region-race-education, age-time[13], region-age-time, race-age-time, education-time, region-education-time, race-education-time. This model yields an estimated improvement of 2.59 percentile points, which is significant at the 1-percent level.

[13] The age-time interaction is a dummy for being or 20 interacted with year effects.

In Table 5, we present regression estimates of between state comparisons of cohorts born in 1960-62 and in 1970-72. The top panel shows the black-white difference in AFQT and PNMR for the 1960-62 cohorts for SC-NC, TN-VA, IL-NY and MI-OH, along with the change between 1970-72. As seen earlier in the figures, there is a monotonic negative relationship between the change in AFQT scores and the change in PNMR. Furthermore, the change in both variables is largest in the states where the gap was the biggest to begin with; and in the states where there is no change in PNMR, there is also no change in AFQT.

The bottom panel shows various comparisons across these state pairs. In each cell, the top number is the black-white AFQT gap. In square brackets are the absolute values of the *t*-statistics. The subsequent number is the change in the black-white AFQT gap between the 1960-62 cohorts and the 1970-72 cohorts. Moving across the columns, more and more controls are added to the regressions. The results show that even between groups of states within the South and within the North, the relative changes in black-white AFQT gaps are statistically significant.

Table 6 is similar in set up to Table 5, except that all comparisons are between AL-MS and another pair of states. Because the drop in black PNMR was so sharp in Alabama and Mississippi, we narrow the window of cohorts over which we difference. In this table, comparisons are made between the 1961-63 cohorts and the 1966-68 cohorts. Even over this short time period, relative changes between AL-MS and TN-VA are statistically distinguishable and significant in magnitude. Depending on the set of controls, estimates range from 1.84 percentile points to 2.54 percentile points. This change corresponded to a relative change of 2.59 in black-white PNMR. Comparisons of AL-MS to IL-NY and to MI-OH yield estimated improvements of 3.57-5.69 and 5.95 to 9.82 percentile points. Notably, the size of the relative racial PNMR improvement (shown in curved brackets) correlates strongly with the size of the estimated racial AFQT improvements.

VIII. A Possible Cause of Black Infant Health Improvement in the 1960's: Hospital Integration

Having established the strong relationship between black infant health in a cohort and the AFQT scores measured 17-20 years later for blacks in that cohort relative to whites, we now turn to the natural question of what caused the rapid improvement in black infant health documented in the previous section. In an earlier paper, Almond, Chay and Greenstone (2008) argue that the forced integration of previously segregated in the hospitals in the South caused a decrease in black PNMR. In segregated hospitals, there were separate waiting rooms for black and white patients, and in many cases black patients who came to the emergency room for treatment were forced to wait until all white patients were treated, regardless of the severity or urgency of the patients' conditions. In other cases, care was refused to black patients outright. Almond, Chay and Greenstone (2008) argue that after the passage of the Civil Rights Act of 1964, previously segregated hospitals integrated, whether as a result of fear a of legal punishment or to avoid losing federal Medicare funding. After integration, they show that access to hospital care, measured for example by the fraction of births that took place in hospitals, increased significantly for blacks. No corresponding decrease in such measures was apparent for whites.

Almond, Chay and Greenstone (2008) argue that this increase in access to hospital care had a particularly important effect on post-neonatal infant health. They find no effect on neonatal mortality rates (deaths in the first month after birth), but large effects on PNMR. They hypothesize that access to hospital care has the largest effect on health outcomes that depend on treatable conditions. The most treatable conditions, they argue, that affect a significant portion of infants, are dehydration caused by diarrhea, and pneumonia. These conditions are serious if left untreated for all infants. They are easily treatable for babies older than one month. Consistent with this hypothesis, Almond, Chay and Greenstone (2008) find that hospital integration lead to improvements in black PNMR due to these specific causes and not to other causes that are harder to treat.

The relationship between access to hospital care and PNMR is summarized nicely by a set of figures from their work. We present these borrowed figures and label them Figures 6A-D. In panel A, the black-white difference in hospital birth rate is plotted against each state's percent black for the cohorts born between 1958 and 1960. States are only included if at least 95 percent of non-white births are black. The figure shows a strong negative relationship. The highest-fraction-black states, predominantly those in the Deep South, have lower relative hospital birth rates than the lowest-fraction-black states, predominantly those in the North. The rest of the southern states fall in between both in terms of fraction black and black-white hospital birth rate gaps. Panel B shows the same relationship for the cohorts born between 1972 and 1974. The relationship is severely flattened. The black-white hospital birth rate gap narrowed in virtually all states, getting very close to zero in the majority of states. The convergence was greatest in the Deep South, smaller in the rest of the South, and smallest in the North.

Panels C and D show a similar relationship, except that the black-white hospital birth rate gap is replaced with the black-white PNMR gap. The figure shows a strong positive relationship. The highest-fraction-black states, predominantly those in the Deep South, have higher relative black PNMR than the lowest-fraction-black states, predominantly those in the North. The rest of the southern states fall in between both in terms of fraction black and black-white PNMR gaps. Panel D shows the same relationship for the cohorts born between 1972 and 1974. The relationship is severely flattened. The black-white PNMR declined in virtually all states, and the reduction was greatest in the Deep South, smaller in the rest of the South, and smallest in the North.

This pattern, along with the other evidence presented in Almond, Chay and Greenstone (2008), suggests that the improvement in black PNMR, particularly in Deep South, was driven at least in part by improved access to medical care—due in significant part to the forced integration of Southern hospitals.

We present here further evidence that this improved access to hospital care may have had long-term effects on human capital accumulation, and may explain some of the cohort-based convergence in the black-white test score gap that is documented above. Motivated by the results in Figure 6, Figure 7 shows a comparison of AL-MS to TN-VA and to IL-NY. Panel A shows comparisons of AL-MS and TN-VA (or IL-NY), for both PNMR and hospital birth rates. The double differences show that black-white gaps in both series followed similar patterns in AL-MS as in the other two pairs of states from 1957 until the mid-1960's. Then, the relative racial gaps in both PNMR and hospital birth rates began to converge. As can be seen in panel B, the timing in the cross-state convergence in both of these series line up fairly well in general with the timing of the relative improvement in the black-white AFQT gaps.

The correspondence of these three series, measured 17 to 20 years apart is notable, and suggestive that access to health care during the early years of life may have played an important role in shaping the skill gap between blacks and whites born in the South during this period.

IX. Alternative causes for the improvement in black infant health in the South during the 1960's

Thus far we have presented evidence in support of three related hypotheses. First, that much of the convergence in the black-white test score gap measured during the 1980's accrued to successive cohorts, particularly those born in the South, rather than to blacks of all ages in particular years. Second, that this cohort-based convergence was caused at least in part by improvements in black infant health among the cohorts that experienced the relative improvements in test scores measured anywhere from 9 to 20 years later. And third, that the improvement in infant health, and therefore the resulting improvement in test scores, was originally caused by increased access to healthcare at early ages—a change that was driven by forced integration of hospitals in the South during the mid- to late-1960's.

Though we have presented evidence that strongly support the second and third hypotheses above, there are a number of alternative causes. The cohorts that experienced the convergence in the black-white test score gap lived the early part of their life at a time when a number of public policies and general changes in the economic environment benefited blacks relative to whites. In this section, we discuss a number of these alternative explanations for the cohort-based convergence in the black-white skill gap.

The War on Poverty: Food Stamps, AFDC, Medicaid and other social programs

As a part of President Lyndon Johnson's War on Poverty, a number of large social programs were initiated in the mid-1960's. Since the programs were aimed at helping the poor, many of them arguably benefited blacks relative to whites. And, the general timing matches the childhood of the black cohorts that experienced the sharp gains. If any of these particular social programs – Food Stamps, AFDC, or

Medicaid – were the cause of the narrowing of the black-white test score gap, it is likely that an improvement in health was a part of the mechanism. Each of the three was aimed at either helping poor families to buy sufficient supplies of food or subsidizing their medical care. Each also had an income effect, however, so it would be difficult to rule out direct effects of other expenditures.

It is worth recalling, however, the patterns in the test score convergence documented above. For it to be the case that any or all of these social programs caused the narrowing of the black-white skill gap, a few empirical facts would have to be reconciled. First, the social program(s) would have to cause improvements in the South and not the North. Second, the programs would have to affect the test scores of successive cohorts. Third, the particular cohorts that experienced the improvements would have to be the same ones for whom we observe the relative test score improvements. And fourth, the cross-state variation in the timing of the improvement would have to match what we see for test scores.

Did the timing of these social programs match the timing of the black-white test score convergence? Without some restrictions on the mechanisms (i.e. the ages at which the programs had effects on human capital accumulation), this is a very difficult question to answer. The cohort-based convergence that we document did not have to be caused by something that happened at birth. It is possible that later interventions, by affecting the correct ages, could have affected the right cohorts. It is also possible, though less likely, that variation across states in the timing of these programs' implementation could be matched with variation in the ages they affected to match the timing of the cohort-based convergence in test scores. However, many of these programs effectively began a number of years after the births of the black children for whom we observe AFQT score improvements. Medicaid, for example, did not begin in Alabama and Mississippi until 1970, and was adopted several years earlier by nearly all of the states outside of the South. For Medicaid to explain the improvements in black test scores it would have to be that this program's long term effects were due to its effects on health of four and five year olds, but due to its effects on one, two or three year olds.

Regarding Food Stamps, Hoynes and Schanzenbach (2008) show: i) Alabama, Mississippi and North Carolina were particularly slow to roll out the program across its counties relative to Illinois, Ohio and Michigan; ii) much of the rollout in these Southern states occurred after 1967; and iii) the earlier rollout in the South may have targeted predominantly white rural counties over those with majority black populations. Further, AFDC (Aid to Families with Dependent Children) caseloads in Alabama and Mississippi grew at less than half the national rate between 1965 and 1970 (Department of Health and Human Services 1998). Finally, a preliminary examination of the data on Head Start programs provided by Ludwig and Miller (2007) show lower participation rates in Alabama and Mississippi relative to Illinois and Michigan as of 1968 and less growth in Head Start funding between 1968 and 1972.

More research is warranted testing the long-term effects of these important social programs on black and white skill accumulation. Though the evidence presented above suggests that at least some part of the convergence in test scores was due to improvements in infant health, we would regard it just as important a finding if the major social programs of the War on Poverty also contributed.

Decrease in the black-white wage gap

Another alternative hypothesis is that the parents of black children born during the 1960's, particularly in the South, had more wealth and earnings than their predecessors. As a result, their children had access to better nutrition and healthcare, and in general grew up in an environment that was more supportive of human capital accumulation. Some of the causes of improving black economic conditions were too gradual to have caused the sharp convergence we see for the test scores of the mid-1960's cohorts. Black educational attainment, for example, had been improving for decades prior to the 1960's. And, the quality of black schools, on observable dimensions, had been gradually improving since the early part of the century (Margo (1990), Card and Krueger (1992)). There is evidence, though, that the Civil Rights Act of 1964 did have a more immediate effect (Chay (1995), Donohue and Heckman (1991), Heckman and Payner (1989)). It is worth investigating whether the changes in black earnings during this period accrued contemporaneously to parents of black children from the relevant cohorts, and whether the variation across states matches the patterns in test score convergence. As a first pass, Panel C of Figure 8 – based on the merged data of Social Security earnings records to the 1978 Current Population Survey used in Chay (1995) – shows differences between states in the log-earnings of (19-51 year-old) black men that do not seem to match the patterns in AFQT scores.

We emphasize, however, that the timing of the Civil Rights Act and the resulting increase in black earnings imply that this explanation would have to involve a mechanism in which parental earnings has a particularly large effect at early ages. If parental earnings affects the human capital accumulation of older children (e.g. 5 year olds), then improvements in AFQT scores should have been apparent for blacks born before 1964 (e.g. 1959).

Increases in black parental earnings

The improvements in black wages relative to whites may also have signaled to black parents that the return to investing in their children's human capital had increased (Neal, 2006). While there is no direct evidence on changes in direct investments by black parents during the 1960's, there is evidence that racial differences in pre-labor-market conditions have important effects on earnings gaps (Neal and Johnson (1996)). Such a story is hard to reconcile with the facts, though. The sharp changes in the black-white test score gap from year to year suggest that parents would have had to react quickly to expected increasing returns to investment. Furthermore, it is not clear what would explain the difference in timing between two southern states like Alabama and South Carolina.

School desegregation

School desegregation has been posited as an alternative explanation for black-white test score convergence during the period we study (see e.g. Grissmer, 1998). A cursory examination of the timing of school desegregation in large urban school districts, particularly in the South, appears to match the timing of the test score convergence documented above. The vast majority of integration court orders for Southern school districts took effect between 1968 and 1972. As these were the years when those born between 1963 and 1967 entered school, one might argue that court ordered integration is a proximate cause of the black-white test score convergence. This argument overlooks a number of important considerations. First, the school districts affected by court ordered integration accounted for less than half of black enrollment, and far less for whites. Court orders typically applied to very large urban school districts. Cascio, Gordon, Lewis and Reber (2007) show that many southern districts that were not integrated by court order integrated earlier, possibly in response to fears of losing federal funding and to avoid court orders. Second, the court orders varied as to whether they were implemented to affect all grades at once or to gradually integrate grades as cohorts moved through the system. The former design would be more likely to present as a convergent pattern of year effects in test scores, while the latter would be more likely to appear as cohort effects. Third, it is not clear whether integration at earlier grades has a larger effect on test scores than integration at later grades. This is relevant since integration first affected different students in different grades.

To assess whether school desegregation can explain the patterns of black-white test score convergence documented above, ideally we would examine grade-level enrollment data by race. We would compute indices of racial integration (e.g. the exposure index and dissimilarity index) at the district-grade-year level. We would then estimate whether changes in integration line up with year effects or cohort effects, and examine how these patterns differed in the North and South. Unfortunately grade level racial enrollment data by school do not exist on a consistent basis for a large sample of districts.

As a substitute, we examine annual school-level racial enrollment data for the set of large urban school districts that were most likely to be integrated by court order between 1961 and the early 1980's. The data cover the years 1968 to 1984, and were complied by Welch and Light from data collected by the Office of Civil Rights of the U.S. Department of Education and from individual school districts. These are the data that were documented by Welch and Light (1987) and were further analyzed in Guryan (2004) and Reber (2005). In addition to school-level enrollment counts by race for each school in surveyed districts, the data includes indicators for the grades served by each school. These indicators allow us to compute integration indices that vary by grade within districts, but they force us to assume that all students within a school were exposed to the same level of integration. Specifically, we use the grade-served information to create district-by-grade measures of racial integration. We compute a different black-exposure-to-whites index for the set of schools in the district that serve each grade (from K through 12).[14]

We then take these data and estimate specifications analogous to those reported above. The dependent variable is the measure of racial integration, and we assume that year of birth is year-grade-6. We estimate three basic specifications: (1) with year effects only, (2) with birth-year effects only, and (3) with both year and birth-year effects. We present the result for the South in figures 13a-d. Figure 13a shows how black exposure to whites, a commonly used measure of racial integration that measures the average fraction white weighted by black enrollment, varied over time. From 1968 to 1973, black exposure to whites in school increased fairly constantly. From 1974 through the mid 1980's the exposure index declined as whites moved out of large urban school districts. In figure 13b, we explore trends in integration by birth year rather than by calendar year. The data allow us to estimate exposure rates for cohorts born between 1950 and 1978.[15] Most interestingly for the purposes of this paper is the fact that black exposure to whites began to increase among cohorts born in 1958 and continued increasing until the mid-1960's. By the 1966 or 1967 cohort, integration began to stabilize. This pattern of average exposure to whites does not match the pattern of test score convergence. Integration began about five years earlier, and stopped before test score convergence.

As described above, the ability to compute integration indices that vary by grade and year within district allows us to separately identify cohort and time effects. The year effects conditional on birth year dummies, shown in figure 13c, exhibit patterns in integration very similar to the unconditional ones shown in figure 13a. This suggests that integration was more likely to happen to all grades in a school at once than to be phased in over time as cohorts moved through the school system.

[14] For example, consider a district has three schools, one that serves grades K-5, one that serves grades 6-12 and one that serves grades K-12. The exposure index for grades K-5 use the enrollment data from the first and third schools, and would yield the same measure for grades K-5, while the exposure index for grades 6-12 would use the enrollment data for the second and third schools, and yield the same measure for grades 6-12.

[15] Exposure to whites appears to have increased among cohorts born in the early 1950's, and then decreased among cohorts born between 1953 and 1958. Part of this pattern is explained by the fact that the earliest cohorts in the data are overrepresented in high schools where exposure rates are higher during this time period, partially due to the way desegregation was implemented.

In contrast, the conditional birth year effects appear to trend downward instead. The downward trend implies that when districts chose to integrate some grades earlier than others they integrated higher grades first. Importantly, there appears to be no break in trend near 1963. In short, this pattern is remarkably different from the one documented earlier for the racial gap in test scores.[16]

X. Conclusion

The black-white test score gap has rightfully captured the attention of economists, policymakers and the public. Yet, for all of the attention and discussion, little is understood about its source or about policies that could reduce the gap. Here, we have documented an important set of facts that should guide researchers in the search for root causes. Using two data sources, we have shown that the narrowing of the black-white test score gap that occurred in the 1980's can be better understood as an improvement by successive birth cohorts of blacks, rather than an improvement that affected blacks of all ages during that decade. The cohort-based convergence began fairly suddenly with the cohort born in 1963, and is only apparent in the South. Though the usual problem of separately identifying cohort, age and time effects is not solved here, we argue that the cohort-based convergence that we document opens the set of potential explanations to those that occurred well before the convergence in test scores was observed.

We test one such explanation, which we call the infant health hypothesis. This hypothesis states that an improvement in black infant health during the mid- to late-1960's had long-term effects on human capital accumulation for the cohorts that experienced these improvements. In the absence of a perfect measure of latent infant health, we take post-neonatal mortality to be a proxy. We show that the timing of black PNMR improvements matches the timing of black AFQT improvements, measured 17 to 20 years

[16] The analysis in this section has implicitly assumed that it is a student's average level of integration, or average school quality, over the 13 years of school that determines his skill accumulation. An alternative model is that exposure to students of a different race or to better schools is more important at earlier grades. To address whether changing this assumption can help the integration data to explain the test score cohort patterns, we estimate similar models but weight early grades more than later grades. Specifically, we weight each district-grade-year observation by 13-grade, so that integration in 1st grade is weighted twelve times as much as integration in 12th grade. When we weight the data this way, we find that most of the patterns shown in the unweighted figures are unchanged. Most importantly, the birth year effects conditional on year effects continue to be declining through the 1960's further suggesting that it would be hard to reconcile the timing of desegregation with the timing of cohort-based convergence in the black-white test score gap.

later, remarkably well. Comparisons across region (South versus North), and within region (AL-MS v. NC-SC, AL-MS v. TN-VA, IL, NY v. MI-OH) show mirroring relationships between the black-white PNMR gap and the black-white AFQT gap. The specific timing across states suggests that improvements in infant health in the first 1.5 to 2.5 years of life had long-term effects on human capital accumulation and explain a significant portion of the narrowing of the black-white test score gap during the 1980's.

We then turn to possible explanations for the improvement in black infant health, focusing on one in particular: the racial integration of southern hospitals during the mid- to late-1960's. Following work by Almond, Chay and Greenstone (2008), we show that there is a strong relationship between black access to hospitals and the black-white PNMR gap. We also show preliminary evidence, that the narrowing of the racial hospital birth rate gap (a measure of access to hospitals) correlates with the timing of the narrowing of the black-white AFQT gap.

Hospital integration was of course not the only policy change during the second half of the 1960's that disproportionately benefited southern blacks (e.g. the Civil Rights Act, Medicaid, Food Stamps, AFDC, and school desegregation). As we discuss, some of the narrowing of the black-white test score gap may have been accounted for by some of these policy changes, though further work is warranted to investigate whether any of them can match the specific patterns we document across states and across cohorts conditional on time effects. Importantly, the cohort-based convergence that we document suggests that to the extent that any of these other policies caused a narrowing of the black-white test score gap, they likely worked through their effects on health and human capital accumulation at young ages.

Finally, our results imply that a portion of the black-white skills gap has at its root differences in investment in children of very young ages. Since current black-white PNMR gaps are much smaller than they were in 1960, the potential of a policy that aims at narrowing this particular gap is not as great as it once was. However, the results suggest more generally that investments at very early ages in health and human capital defined broadly can have important and lasting long-term effects on human capital accumulation. This conclusion is consistent with the findings of Heckman and Carneiro (2003), Bleakley

(2007) and others, and suggests efforts by policymakers that focus exclusively on school-aged children and adults may overlook some of the most effective ways to narrow the black-white skills gap.

References

Almond, Douglas and Kenneth Y. Chay. "The Long-Run and Intergenerational Impact of Poor Infant Health: Evidence from Cohorts Born during the Civil Rights Era." University of California-Berkeley, mimeograph, 2006.

Almond, Douglas V., Kenneth Y. Chay and Michael Greenstone (2008). "The Civil Rights Act of 1964, Hospital Desegregation and Black Infant Mortality in Mississippi," mimeo. February.

Card, David, and Alan B. Krueger (1992). "School Quality and Black-White Relative Earnings: A Direct Assessment," *Quarterly Journal of Economics*, 107(1) 151-200.

Cascio, Elizabeth, Nora Gordon, Ethan Lewis and Sarah Reber (2007). "From Brown to Busing," *NBER Working Paper No. 13279*. July.

Chay, Kenneth Y. (1995). "Evaluating the Impact of the 1964 Civil Rights Act on the Economic Status of Black Men Using Censored Longitudinal Earnings Data," mimeo. October.

Cook, Michael D., and William N. Evans (2000). "Families or Schools? Explaining the Convergence in White and Black Academic Performance," *Journal of Labor Economics*, 18(4) 729-754.

Dickens, William T., and James R. Flynn (2006). "Black Americans Reduce the Racial IQ Gap," *Psychological Science*, 17(10) 913-290.

Donohue, John J. III, and James Heckman (1991). "Continuous Versus Episodic Change: The Impact of Civil Rights Policy on the Economic Status of Blacks," *Journal of Economic Literature*, 29(4) 1603-1643.

Grissmer, David, Ann Flannagan and Stephanie Williamson (1998). "Why Did the Black-White Test Score Gap Narrow in the 1970's and 1980's?" in *The Black-White Test Score Gap,* Christopher Jencks and Meredith Phillips, eds.

Guryan, Jonathan (2004). "Desegregation and Black Dropout Rates," *American Economic Review*, 94(4) 919-943.

Heckman, James J. and Brook S. Payner (1989). "Determining the Impact of Federal Antidiscrimination Policy on the Economic Status of Blacks: A Study of South Carolina," *American Economic Review*, 79(1) 138-177.

Heckman, James J. and Pedro Carneiro (2003). "Human Capital Policy," in *Inequality in America*, Benjamin M. Friedman, ed. Cambridge, MA: MIT Press.

Hoynes, Hilary W. and Diane Whitmore Schanzenbach. "Consumption Responses to In-Kind Transfers: Evidence from the Introduction of the Food Stamp Program." University of California-Davis, mimeograph, 2008.

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

Neal, Derek (2006). "Why Has Black-White Skill Convergence Stopped?" *Handbook of the Economics of Education: Volume 1*, Eric A. Hanushek and Finis Welch, eds.

Reber, Sarah (2005). "Court-Ordered Desegregation: Successes and Failures in Integration Since *Brown v. Board of Education*," *Journal of Human Resources*, 40(3) 559-590.

Welch, Finis and Audrey Light (1987). "New Evidence on School Desegregation," *United States Commission on Civil Rights Clearing House Publication 92*. June.

Wooldridge, Jeffrey M. (2002). "Inverse Probability Weighted M-Estimators for Sample Selection, Attrition, and Stratification," *Portuguese Economic Journal*, 1(2) 117-139.

Table 1: Summary statistics for AFQT samples

	Men born between 1957 and 1972 who took AFQT between 1976 and 1991					
	Ages 16 to 28			Ages 17 to 20		
	Total	Black	White	Total	Black	White
	(1a)	(1b)	(1c)	(1a)	(1b)	(1c)
Average percentile score	46.44	31.03	53.13	45.29	30.31	51.95
[standard deviation]	[24.87]	[19.61]	[23.92]	[24.34]	[19.06]	[23.46]
South	46.19	29.93	54.17	42.58	29.11	51.83
	[25.14]	[19.40]	[23.75]	[24.13]	[18.50]	[23.17]
Rustbelt	48.31	32.78	52.96	47.68	32.26	52.04
	[24.72]	[20.18]	[24.05]	[24.28]	[19.77]	[23.66]
Age distribution (percent)						
16 years old	0.7	0.5	0.9			
17 years old	26.0	21.6	27.9	32.1	27.2	34.2
18 years old	27.6	28.2	27.4	34.0	35.3	33.5
19 years old	17.4	19.1	16.7	21.4	23.8	20.4
20 years old	10.1	11.1	9.7	12.5	13.7	11.9
21 years old	6.2	6.8	6.0			
22 to 28 years old	11.8	12.8	11.4			
Education distribution (percent)						
Less than 10 years	3.4	1.9	4.0	3.6	2.1	4.3
10 years	7.9	7.0	8.2	8.5	7.6	8.9
11-12 years	35.8	35.9	35.7	42.1	42.6	41.8
HS GED	3.7	3.0	4.0	3.1	2.4	3.4
HS graduate	44.8	48.2	43.3	40.8	43.6	39.5
1 year college	1.6	1.5	1.7	1.3	1.2	1.4
2 years college	1.4	1.3	1.5	0.6	0.5	0.6
3+ years college	1.5	1.3	1.6	0.1	0.1	0.1
Percent who take AFQT				6.10	9.98	5.77
South				6.49	10.62	5.62
Rustbelt				5.98	9.21	5.81
Number of observations	4,123,138	1,249,247	2,873,891	3,020,558	929,572	2,090,986

Notes: Data consist of the universe of men who were born between 1957 and 1972 and took the AFQT between 1976 and 1991 in the South and Rustbelt. The South consists of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia. The Rustbelt consists of Illinois, Michigan, New York, Ohio and Pennsylvania. The percent who take the AFQT is calculated from the ratio of the number of men in a state-race-age-year cell who take the AFQT to the total population of men in that cell taken from Decennial Census counts.

Table 2: Estimates of AFQT effects from inverse probability weighted regressions

	South		Rustbelt	
	White	Black-white	White	Black-white
	(1a)	(1b)	(2a)	(2b)
Education effects				
1 year of HS	-15.56***	8.41***	-15.70***	9.72***
	(0.12)	(0.20)	(0.13)	(0.26)
2 years of HS	-10.13***	5.43***	-11.09***	6.78***
	(0.10)	(0.13)	(0.08)	(0.14)
3-4 years of HS	---	---	---	---
HS GED	2.61***	1.85***	4.98***	2.92***
	(0.12)	(0.23)	(0.15)	(0.25)
HS graduate	7.95***	-5.45***	9.50***	-4.94***
	(0.07)	(0.09)	(0.06)	(0.10)
1 year of college	25.97***	-10.96***	29.21***	-9.39***
	(0.21)	(0.35)	(0.19)	(0.45)
2 years of college	29.47***	-9.78***	33.00***	-9.20***
	(0.28)	(0.49)	(0.23)	(0.70)
Age effects				
17 years old	4.42***	-0.58***	6.22***	-2.01***
	(0.08)	(0.11)	(0.07)	(0.13)
18 years old	---	---	---	---
19 years old	-1.06***	-0.48***	-2.43***	0.77***
	(0.09)	(0.12)	(0.08)	(0.14)
20 years old	0.15	-1.40***	-1.25***	0.54**
	(0.14)	(0.18)	(0.12)	(0.22)
Year effects				
1976	-0.81	-1.38**	2.61***	- 4.57***
	(0.51)	(0.68)	(0.44)	(0.77)
1977	-5.60***	-1.42**	-3.13***	-2.70***
	(0.44)	(0.60)	(0.39)	(0.68)
1978	-3.83***	-2.18***	-1.75***	-2.56***
	(0.39)	(0.53)	(0.34)	(0.60)
1979	-3.36***	-2.98***	-2.02***	-3.08***
	(0.34)	(0.46)	(0.30)	(0.52)
1980	2.52***	-1.66***	3.67***	-0.86*
	(0.28)	(0.39)	(0.24)	(0.44)
1981	-3.30***	-2.46***	-1.45***	-2.69***
	(0.23)	(0.33)	(0.21)	(0.37)
1982	-0.92***	-0.72***	0.27	-1.21***
	(0.19)	(0.27)	(0.17)	(0.31)
1983	0.73***	0.53**	0.98***	-0.16
	(0.16)	(0.23)	(0.14)	(0.25)
1984	---	---	---	---
1985	0.19	0.15	-0.25*	0.60**
	(0.16)	(0.23)	(0.14)	(0.25)
1986	-0.04	1.07***	-0.36**	1.07***
	(0.19)	(0.28)	(0.17)	(0.31)
1987	-0.07	0.44	-0.20	0.54
	(0.24)	(0.34)	(0.21)	(0.39)
1988	-1.36***	-0.46	-1.68***	0.43
	(0.29)	(0.41)	(0.26)	(0.48)
1989	-3.30***	-1.46***	-3.99***	0.73
	(0.34)	(0.47)	(0.31)	(0.55)
1990	-3.16***	-1.28**	-3.69***	0.90
	(0.38)	(0.53)	(0.35)	(0.62)
1991	-4.84***	-0.72	-5.17***	1.61**
	(0.43)	(0.61)	(0.39)	(0.73)

(Table 2 continued)

	South		Rustbelt	
	White	Black-white	White	Black-white
	(1a)	(1b)	(2a)	(2b)
Birth cohort effects				
1957	47.3***	-21.4***	46.1***	-17.0***
	(0.52)	(0.69)	(0.45)	(0.79)
1958	46.8***	-21.1***	45.5***	-17.2***
	(0.46)	(0.61)	(0.40)	(0.70)
1959	46.3***	-21.2***	44.8***	-17.1***
	(0.40)	(0.54)	(0.35)	(0.62)
1960	45.9***	-21.0***	45.2***	-17.2***
	(0.35)	(0.48)	(0.31)	(0.55)
1961	46.5***	-21.4***	45.7***	-17.5***
	(0.30)	(0.42)	(0.27)	(0.47)
1962	46.2***	-20.7***	45.2***	-16.9***
	(0.25)	(0.36)	(0.22)	(0.40)
1963	46.6***	-20.1***	45.7***	-16.9***
	(0.21)	(0.30)	(0.19)	(0.34)
1964	46.7***	-19.1***	46.1***	-16.2***
	(0.17)	(0.23)	(0.15)	(0.27)
1965	47.6***	-17.7***	47.0***	-16.0***
	(0.14)	(0.19)	(0.12)	(0.22)
1966	49.0***	-17.5***	48.4***	-15.9***
	(0.14)	(0.20)	(0.12)	(0.23)
1967	51.1***	-17.2***	50.3***	-16.2***
	(0.19)	(0.27)	(0.16)	(0.30)
1968	53.2***	-17.4***	52.0***	-16.7***
	(0.23)	(0.33)	(0.21)	(0.38)
1969	55.1***	-17.1***	53.5***	-16.1***
	(0.28)	(0.39)	(0.25)	(0.45)
1970	56.4***	-16.3***	55.2***	-15.9***
	(0.32)	(0.45)	(0.29)	(0.53)
1971	57.2***	-15.8***	56.2***	-15.4***
	(0.37)	(0.52)	(0.33)	(0.61)
1972	58.0***	-15.1***	57.1***	-14.8***
	(0.42)	(0.58)	(0.38)	(0.70)
Sample size	1,410,972		1,609,586	

Notes: Sample contains all takers of the AFQT in the South and Rustbelt who took the test between 1976 and 1991, between the ages of 17 and 20, and were born between 1957 and 1972. Separate regressions are estimated by region and include unrestricted race-by-birth cohort, race-by-time, race-by-age, and race-by-education fixed effects. The regressions are weighted by the inverse probability of individuals in a state-race-birth cohort-age cell taking the test (based on birth counts). The estimated standard errors are in (parentheses) and corrected for heteroskedasticity.

*** significant at 1-percent level, ** significant at 5-percent level, * significant at 10-percent level

Table 3: Changes in AFQT scores between 1960-1962 and 1970-1972, South versus Rustbelt

	Education fixed effects (by region)				Race by education fixed effects (by region)			
	White birth cohorts		Black-white		White birth cohorts		Black-white	
	Average in 1960-1962	Change by 1970-1972	Average in 1960-1962	Change by 1970-1972	Average in 1960-1962	Change by 1970-1972	Average in 1960-1962	Change by 1970-1972
	(1a)	(1b)	(1c)	(1d)	(2a)	(2b)	(2c)	(2d)
A. South	48.22***	7.93***	-24.22***	6.87***	46.97***	8.94***	-20.50***	3.93***
	(0.71)	(0.72)	(0.60)	(0.50)	(0.74)	(0.75)	(0.62)	(0.59)
{PNMR, birth year}	{5.57}	{-1.50}	{14.05}	{-8.27}				
B. Rustbelt	46.47***	8.47***	-19.41***	1.94***	46.01***	8.86***	-16.43***	-0.04
	(0.56)	(0.66)	(0.51)	(0.56)	(0.57)	(0.67)	(0.36)	(0.36)
{PNMR, birth year}	{5.03}	{-1.22}	{5.85}	{-1.35}				
C. South – Rustbelt	1.75*	-0.54	-4.82***	4.94***	0.96	0.08	-4.07***	3.97***
	(0.87)	(0.93)	(0.76)	(0.71)	(0.90)	(0.96)	(0.69)	(0.67)
{PNMR, birth year}	{0.53}	{-0.28}	{8.20}	{-6.92}				

Notes: Sample contains all takers of the AFQT in the South and Rustbelt who took the test between 1976 and 1991, between the ages of 17 and 20, and were born between 1957 and 1972. The sample sizes are 1,410,972 in the South; 1,609,586 in the Rustbelt; and 3,020,558 in the pooled regression. All analyses include unrestricted race-by-birth cohort, race-by-time, and race-by-age fixed effects – interacted with region. Columns (1a) to (1d) include unrestricted education-by-region fixed effects; columns (2a) to (2d) include interactions of the education-by-region effects with race. The regressions are weighted by the inverse probability of individuals in a state-race-birth cohort-age cell taking the test (based on birth counts). The estimated standard errors are in (parentheses) and corrected for heteroskedasticity and unrestricted clustering at the state-level. Post-neonatal mortality rates in the corresponding birth year are in {} and are for 1961-1963 and 1971-1973, respectively.

*** significant at 1-percent level, ** significant at 5-percent level, * significant at 10-percent level

Table 4: Changes in black-white AFQT differences, South versus Rustbelt

	South-Rustbelt difference in black-white AFQT differences								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
1960 to 1962 average	-4.82***	-4.41***	-4.07***	-4.23***	-4.31***	-4.25***	-3.88***	-3.80***	-3.42***
	(0.76)	(0.71)	(0.69)	(0.74)	(0.73)	(0.74)	(0.71)	(0.72)	(0.71)
1960-1962 to 1970-1972 difference	4.94***	4.14***	3.97***	3.68***	3.68***	3.51***	3.02***	2.77***	2.59***
	(0.71)	(0.62)	(0.67)	(0.70)	(0.70)	(0.71)	(0.65)	(0.66)	(0.71)
Region-race-cohort	Y	Y	Y	Y	Y	Y	Y	Y	Y
Region-race-time	Y	Y	Y	Y	Y	Y	Y	Y	Y
Region-race-age	Y	Y	Y	Y	Y	Y	Y	Y	Y
Region-education	Y	Y	Y	Y	Y	Y	Y		Y
Race-education		Y	Y				Y	Y	Y
Region-race-education			Y						Y
Age-time				Y	Y	Y	Y	Y	Y
Region-age-time					Y	Y	Y		Y
Race-age-time						Y	Y	Y	Y
Education-time				Y	Y	Y	Y	Y	Y
Region-education-time					Y	Y	Y		Y
Race-education-time								Y	Y

Notes: See notes to Table 3.
*** significant at 1-percent level, ** significant at 5-percent level, * significant at 10-percent level

Table 5: Between state-group comparisons of changes in black-white AFQT differences, 1960-1962 to 1970-1972

	Black-white Average in 1960-1962	Black-white Change by 1970-1972	Between state group difference in black-white AFQT differences			
	(1a)	(1b)	(2)	(3)	(4)	(5)
A. State group averages						
SC, NC	-24.13	7.42				
	[17.43]	[8.70]				
{PNMR, birth year}	{17.41}	{-10.53}				
TN, VA	-21.96	5.57				
	[16.34]	[10.09]				
{PNMR, birth year}	{10.29}	{-6.67}				
IL, NY	-19.44	2.57				
	[22.18]	[14.16]				
{PNMR, birth year}	{6.56}	{-2.08}				
MI, OH	-19.49	0.25				
	[20.17]	[1.22]				
{PNMR, birth year}	{4.22}	{0.29}				
B. Between state group differences						
1. *SC, NC vs. TN, VA*						
1960 to 1962 average			-2.18	-2.64	-2.12	-2.57
{7.12}			[1.87]	[3.36]	[2.02]	[3.39]
1960-62 to 1970-72			1.86	2.00	1.65	1.86
difference			[2.65]	[5.49]	[2.58]	[4.75]
{-3.85}						
2. *SC, NC vs. MI, OH*						
1960 to 1962 average			-4.65	-4.28	-4.08	-3.54
{13.19}			[3.37]	[4.12]	[2.83]	[3.00]
1960-62 to 1970-72			7.17	6.45	5.95	5.00
difference			[10.29]	[14.77]	[6.90]	[13.87]
{-10.82}						
3. *TN,VA vs. MI, OH*						
1960 to 1962 average			-2.47	-1.64	-1.93	-1.06
{6.07}			[2.96]	[2.33]	[2.02]	[1.20]
1960-62 to 1970-72			5.32	4.45	4.20	3.32
difference			[10.06]	[8.33]	[8.67]	[5.61]
{-6.97}						
4. *IL, NY vs. MI, OH*						
1960 to 1962 average			0.04	0.47	0.07	0.61
{2.35}			[0.04]	[0.56]	[0.07]	[0.67]
1960-62 to 1970-72			2.31	1.35	2.28	1.43
difference			[7.83]	[3.20]	[4.97]	[5.93]
{-2.37}						
State-education	Y	Y	Y	Y	Y	Y
Race-education				Y		Y
State-race-education				Y		Y
Age-time					Y	Y
Race-age-time					Y	Y
Education-time					Y	Y
Race-education-time						Y

Notes: Absolute values of t-ratios in [square brackets]. Post-neonatal mortality rates in the corresponding birth year are in {} and are for 1961-1963 and 1971-1973, respectively.

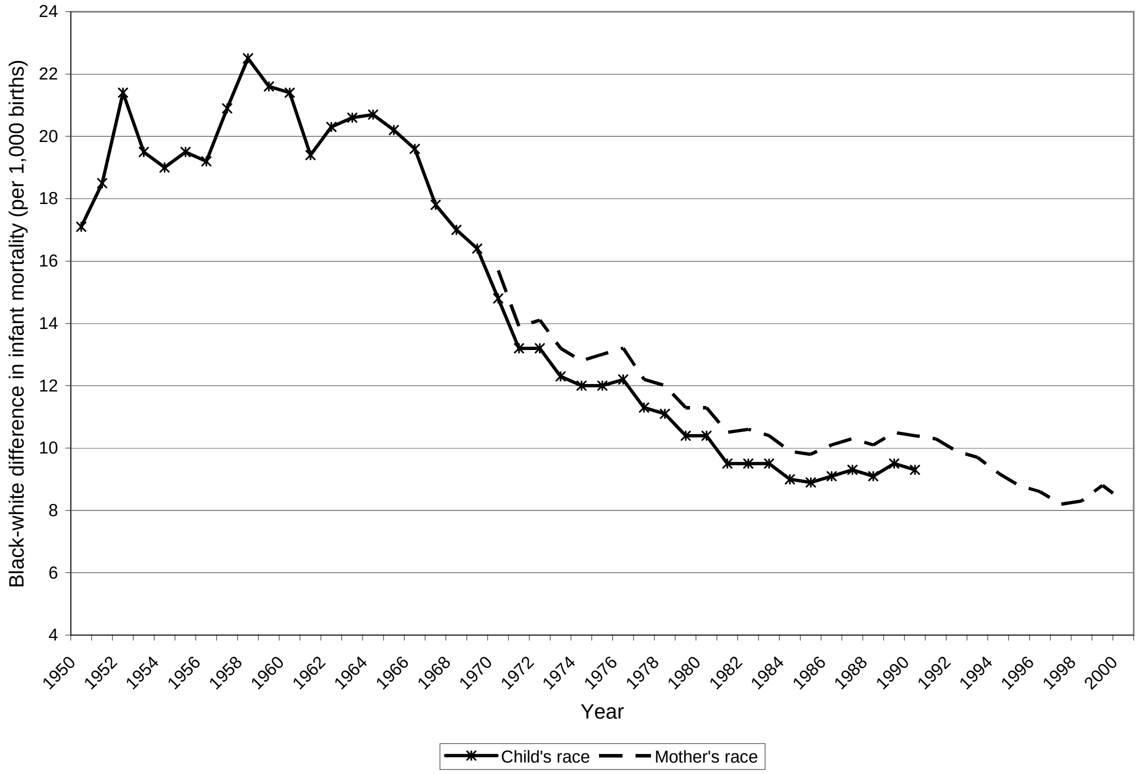
Table 6: Comparisons of changes in black-white AFQT differences between Alabama-Mississippi and other state groups, 1961-1963 to 1966-1968

	Black-white		Between state group difference in black-white AFQT differences			
	Average in 1961-1963	Change by 1966-1968				
	(1a)	(1b)	(2)	(3)	(4)	(5)
A. State group averages						
AL, MS	-24.48	5.13				
	[20.95]	[9.59]				
{PNMR, birth year}	{16.41}	{-5.32}				
TN, VA	-21.14	2.59				
	[18.33]	[7.70]				
{PNMR, birth year}	{9.70}	{-2.73}				
IL, NY	-19.74	2.25				
	[16.11]	[7.26]				
{PNMR, birth year}	{6.90}	{-1.48}				
MI, OH	-19.61	0.83				
	[15.99]	[2.45]				
{PNMR, birth year}	{4.60}	{-0.31}				
B. Between state group differences						
1. *AL, MS vs. TN, VA*						
1961 to 1963 average			-3.34	-3.03	-3.18	-2.91
{6.71}			[12.74]	[4.86]	[13.38]	[4.83]
1961-63 to 1966-68			2.54	2.02	2.21	1.84
difference			[4.92]	[3.35]	[5.45]	[4.02]
{-2.59}						
2. *AL, MS vs. IL, NY*						
1961 to 1963 average			-4.75	-3.92	-4.49	-3.58
{9.52}			[4.65]	[5.41]	[4.65]	[4.51]
1961-63 to 1966-68			2.88	2.33	2.46	1.95
difference			[5.69]	[3.83]	[5.10]	[3.57]
{-3.84}						
3. *AL, MS vs. MI, OH*						
1961 to 1963 average			-4.87	-3.63	-4.58	-3.26
{11.81}			[4.76]	[3.68]	[4.19]	[2.98]
1961-63 to 1966-68			4.30	3.46	3.83	3.08
difference			[9.82]	[5.63]	[8.88]	[5.95]
{-5.01}						
State-education	Y	Y	Y	Y	Y	Y
Race-education				Y		Y
State-race-education				Y		Y
Age-time					Y	Y
Race-age-time					Y	Y
Education-time					Y	Y
Race-education-time						Y

Notes: Absolute values of t-ratios in [square brackets]. Post-neonatal mortality rates in the corresponding birth year are in {} and are for 1962-1965 and 1967-1970, respectively.

Figure 1: Black-white difference in infant mortality rates in the United States, 1950 to 2000

A. Infant mortality over time by child and mother's race



B. Infant, post-neonatal and neonatal mortality rates (child's race)

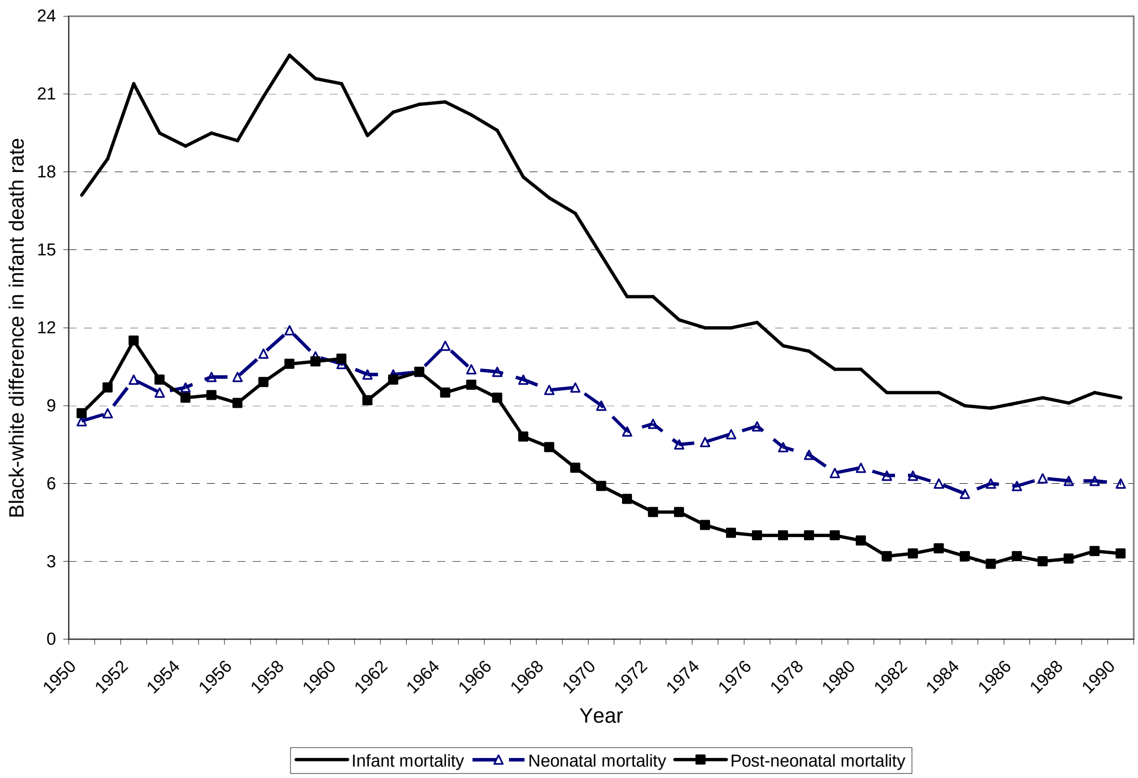
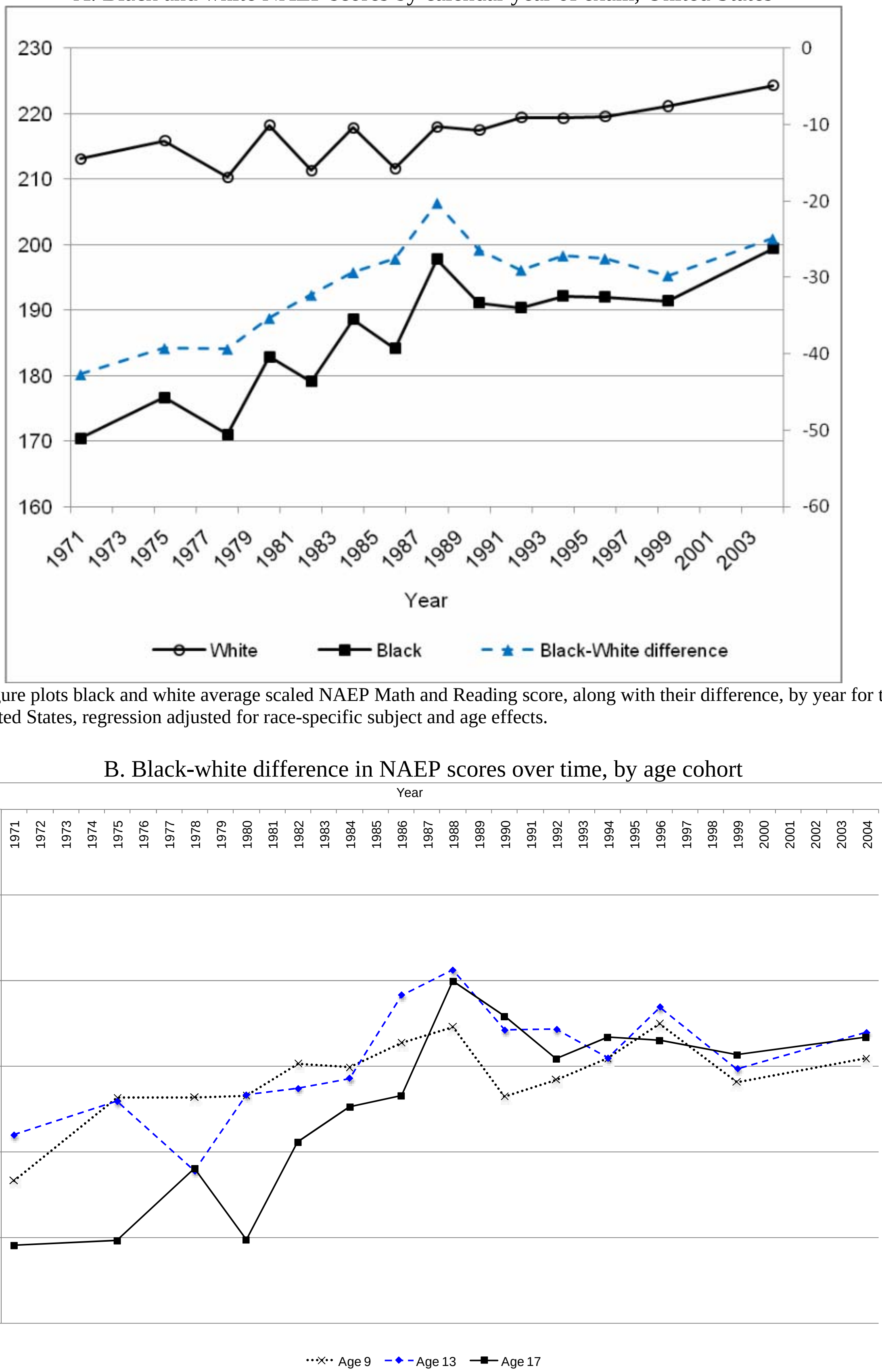


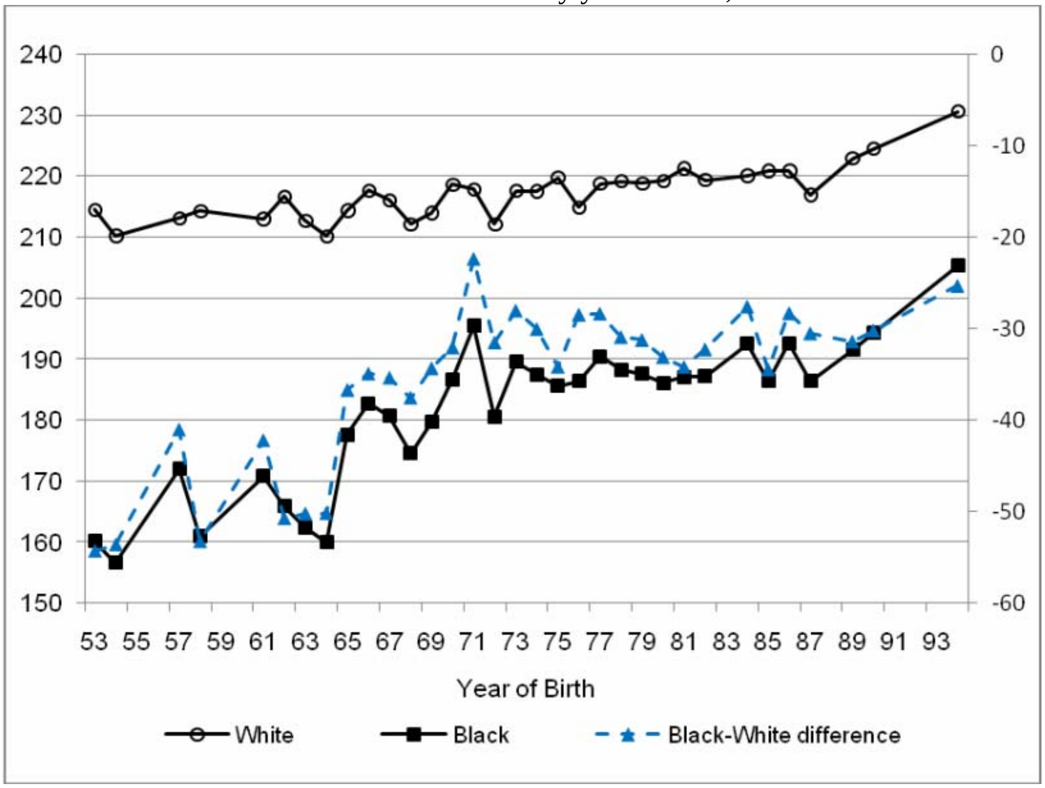
Figure 2: National Assessment of Educational Progress Scores



Notes: Figure plots black and white average scaled NAEP Math and Reading score, along with their difference, by year for the entire United States, regression adjusted for race-specific subject and age effects.

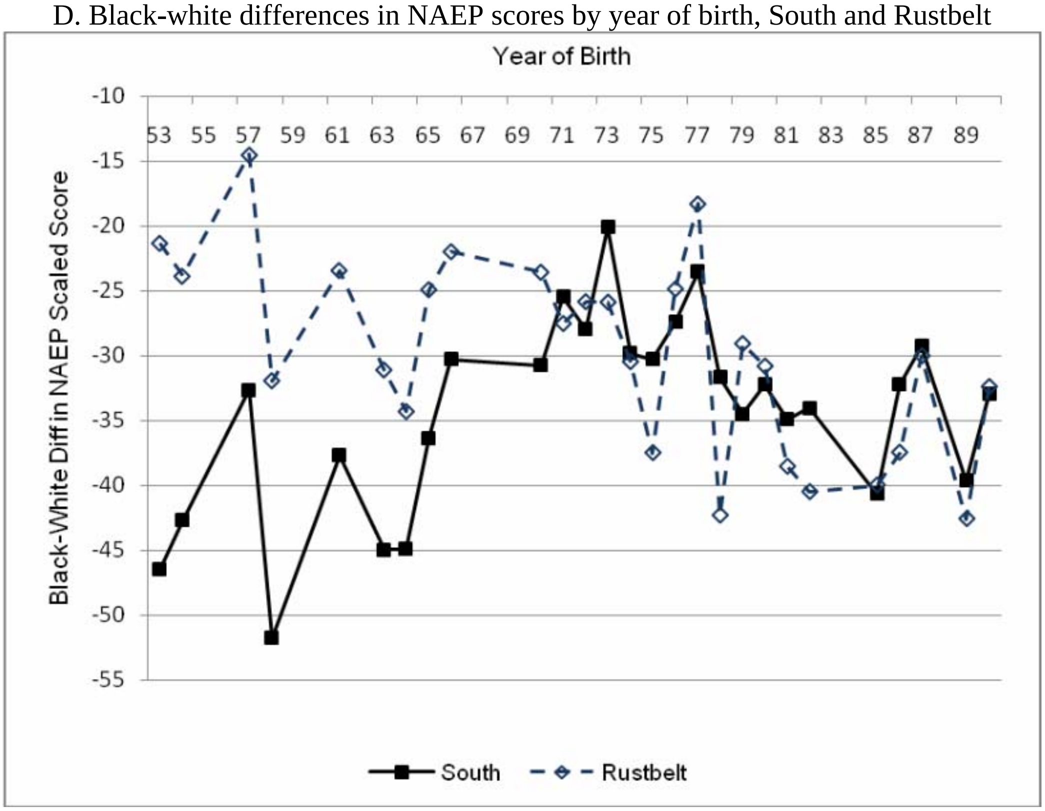
Notes: Figure plots the black-white difference in scaled NAEP score separately for 9-, 13- and 17-year-olds, after conditioning on race-specific subject effects.

C. Black and white NAEP scores by year of birth, United States



Notes: Figure plots black and white average scaled NAEP Math and Reading score, along with their difference, by year of birth for the entire US, regression adjusted for race-specific subject and age effects.

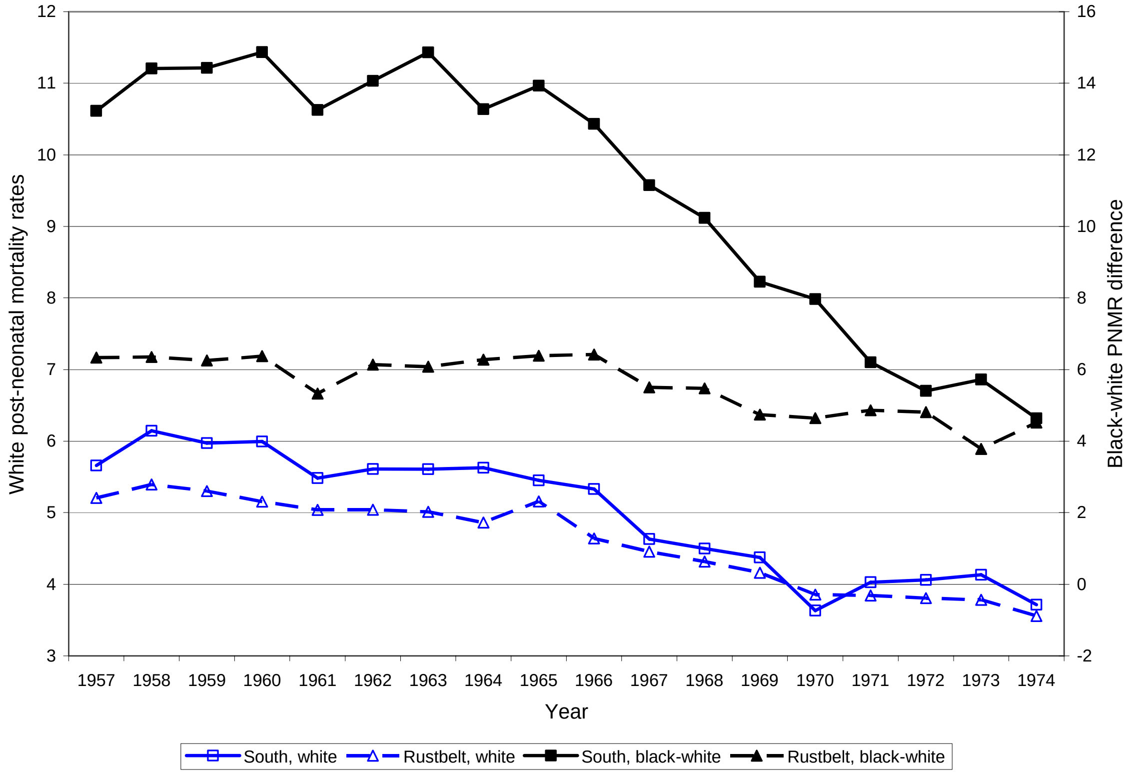
D. Black-white differences in NAEP scores by year of birth, South and Rustbelt



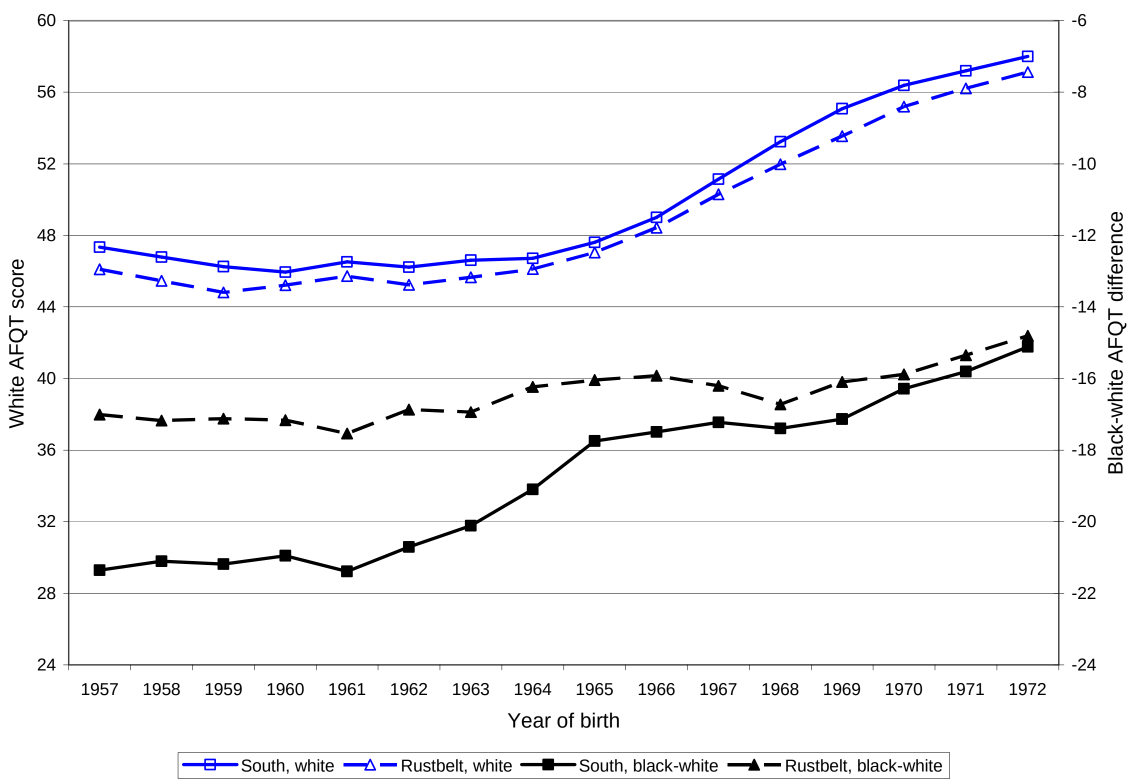
Notes: Figure plots the black minus white difference in average scaled NAEP Math and Reading score by year for the South and Rustbelt, regression adjusted for race-specific year, subject and age effects. The Rustbelt is defined here as the Northeast and Midwest Census regions.

Figure 3: White and black-white differences in the South and Rustbelt

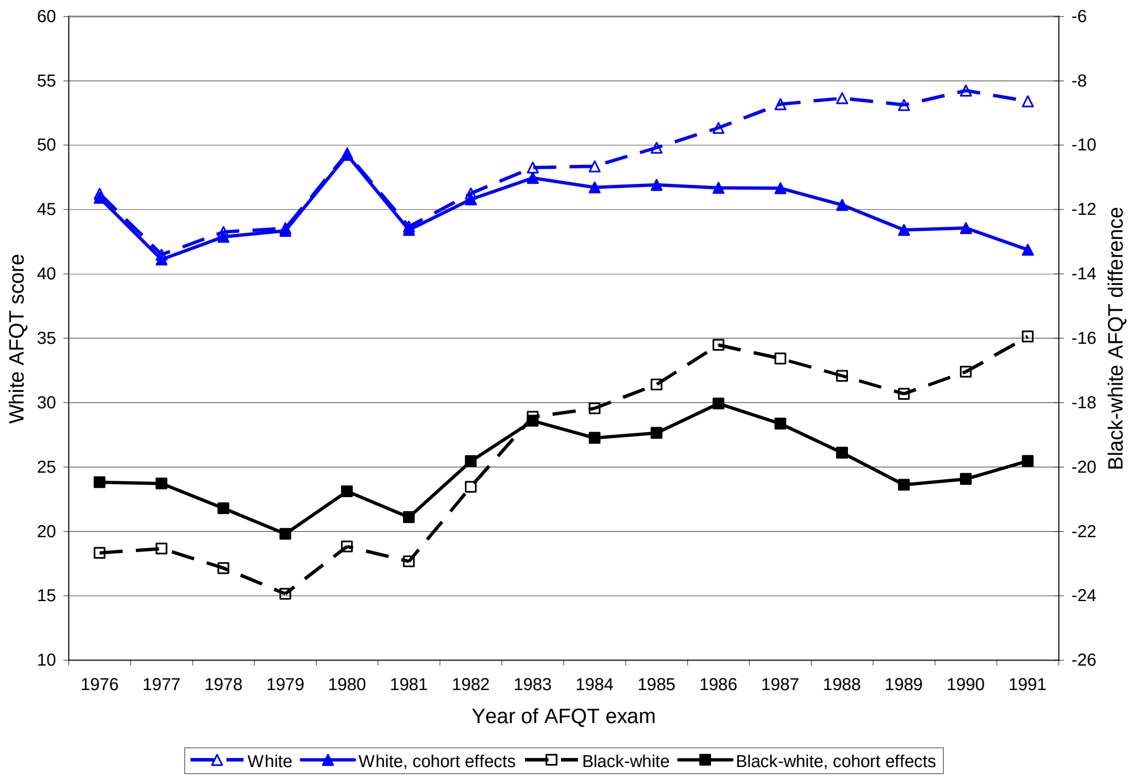
A. Post-neonatal mortality rates by year



B. AFQT scores by year of birth



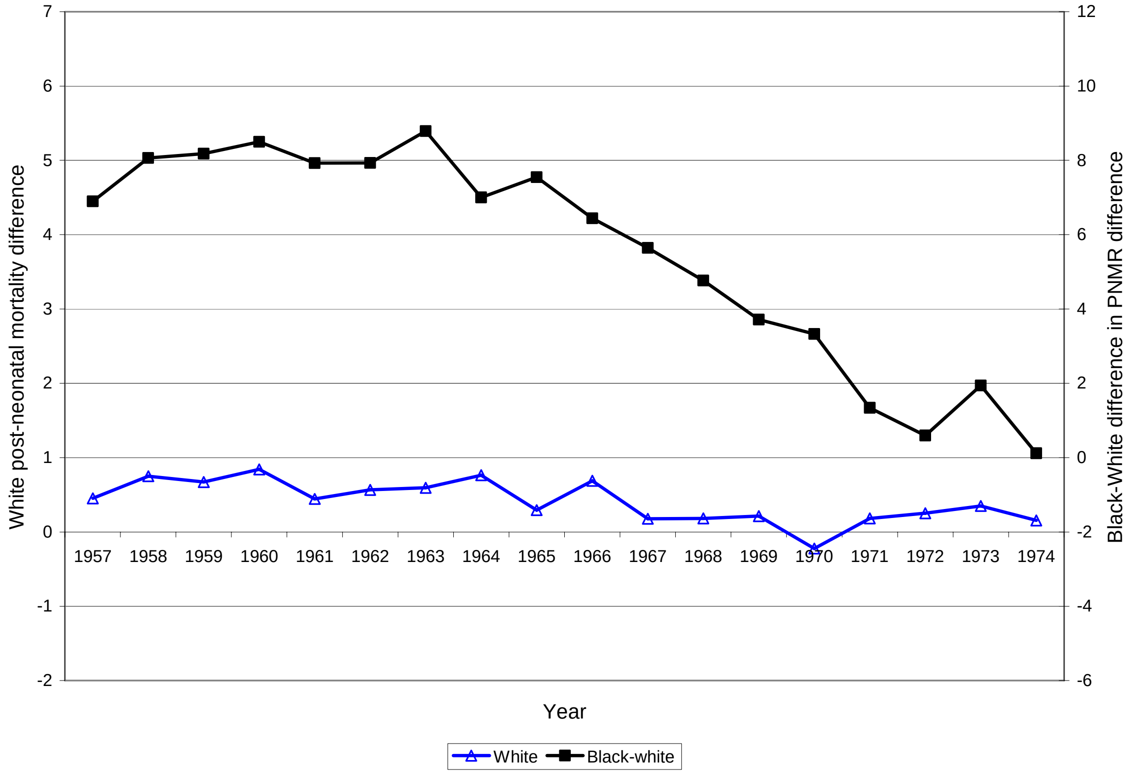
C. AFQT scores by year of exam, South



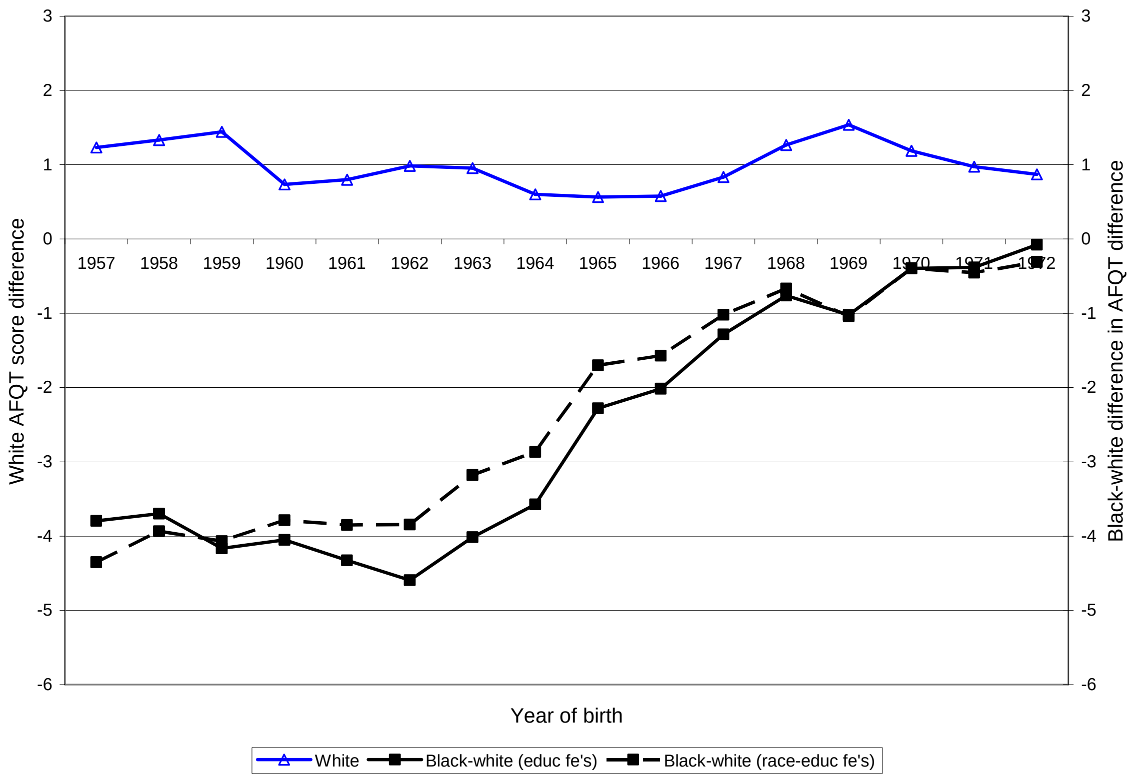
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age and year effects interacted with race, and unrestricted education effects interacted with race. The baseline group is men who were 18 years-old and had 3 to 4 years of high school education when they took the exam.

Figure 4: South-Rustbelt differences in white levels and black-white differences

A. Post-neonatal mortality rates



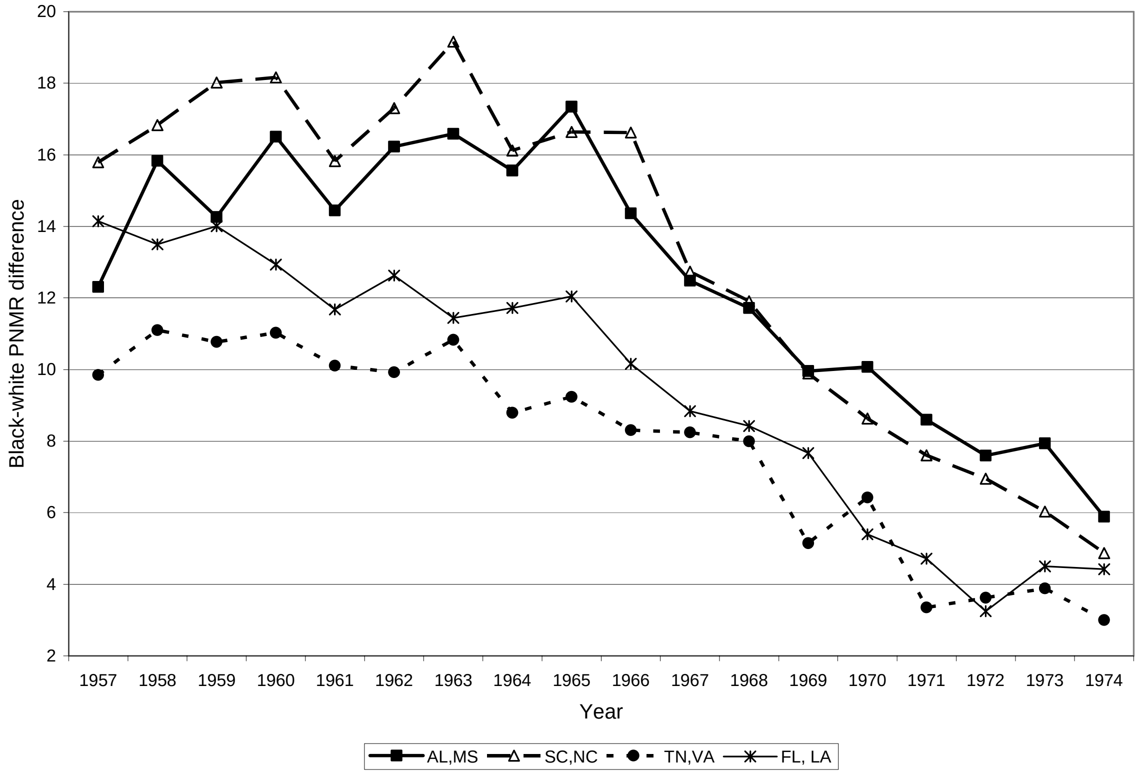
B. AFQT scores by year of birth



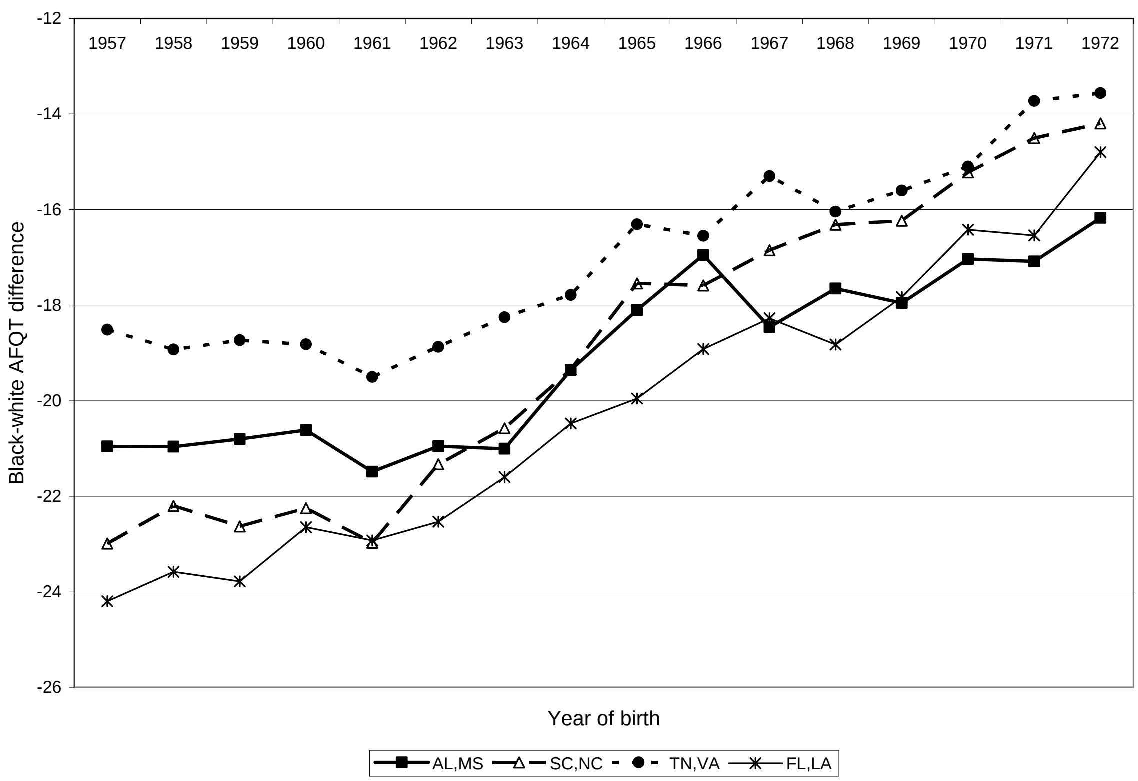
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age and year effects interacted with race, and unrestricted education effects interacted with race – except in the "educ fe's" model, which specifies common education effects by race.

Figure 5: Black-white differences in state groups

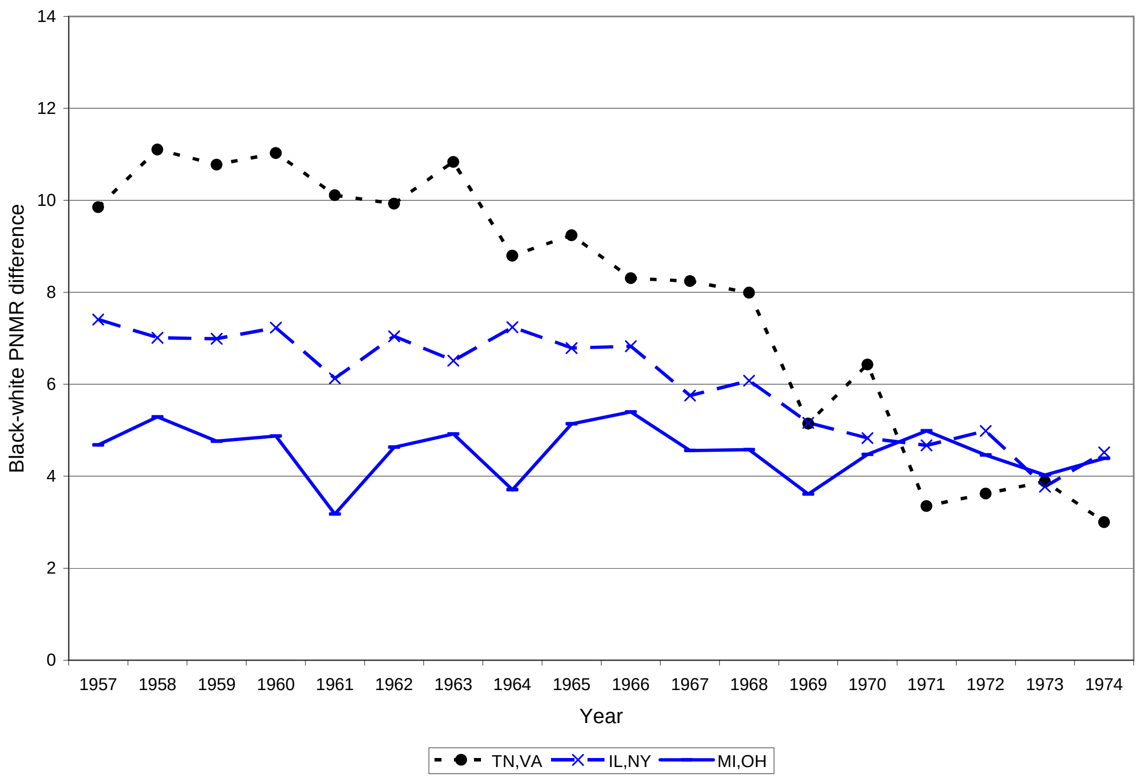
A. Post-neonatal mortality, South



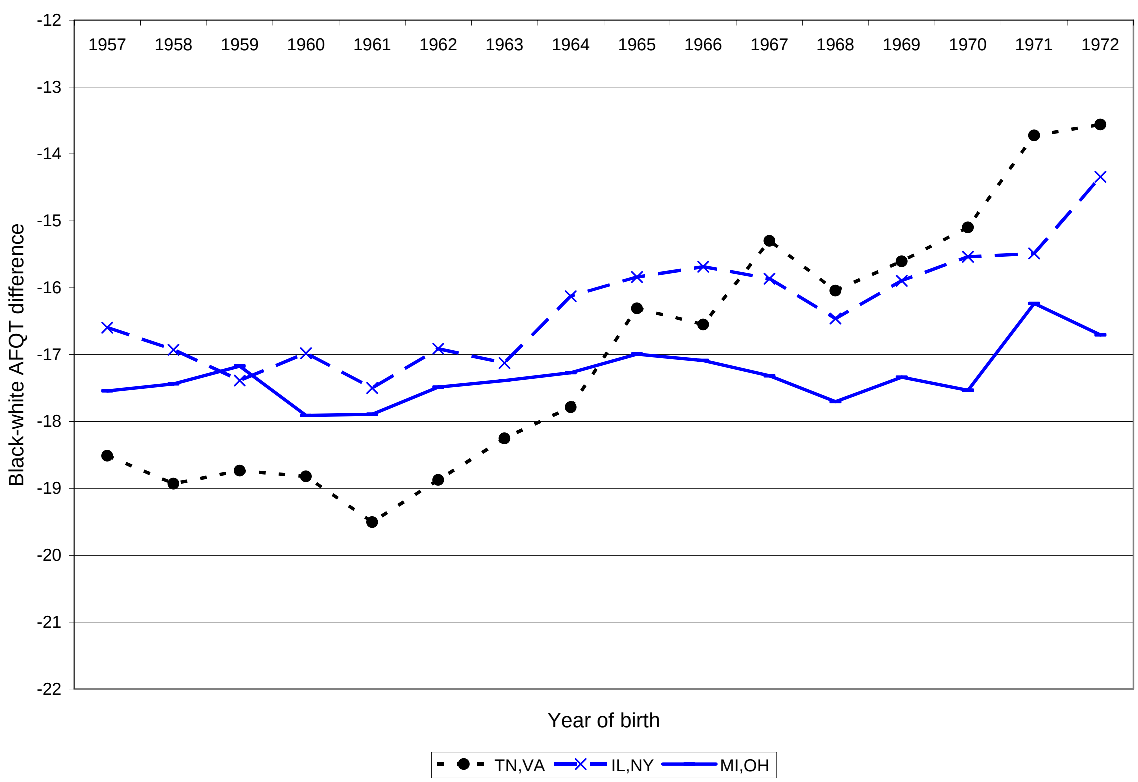
B. AFQT, South



C. Post-neonatal mortality, Rustbelt



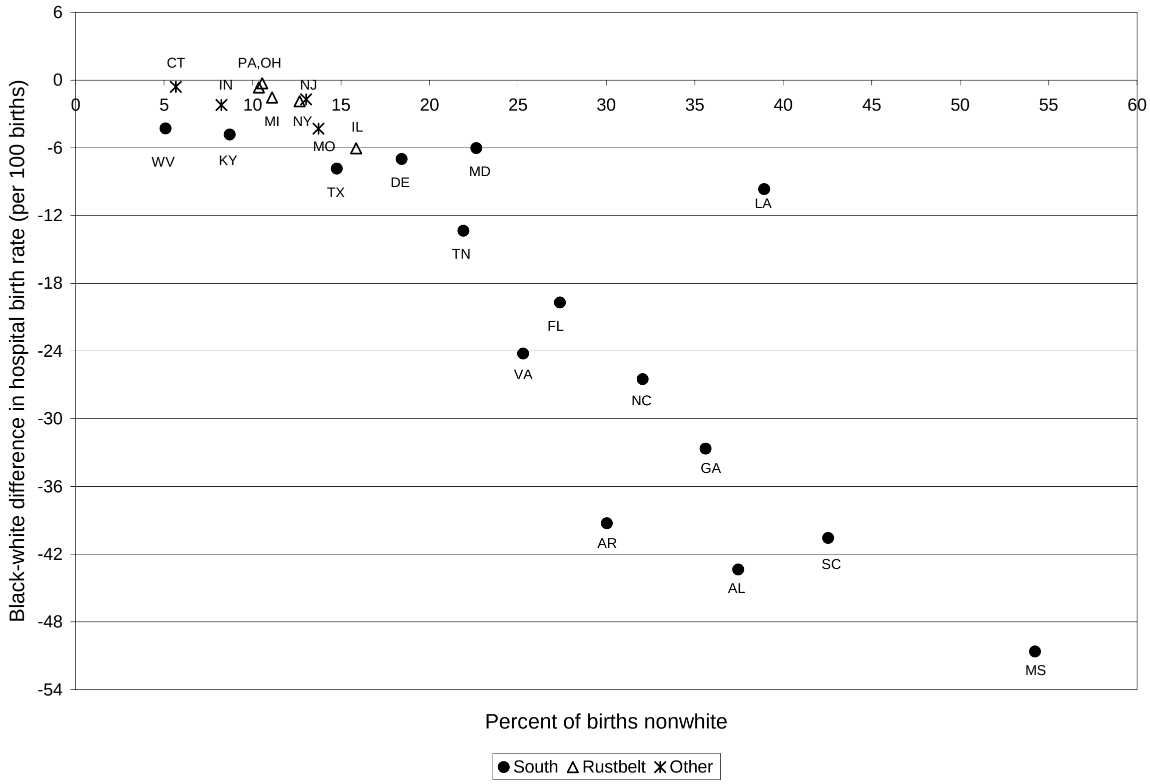
D. AFQT, Rustbelt



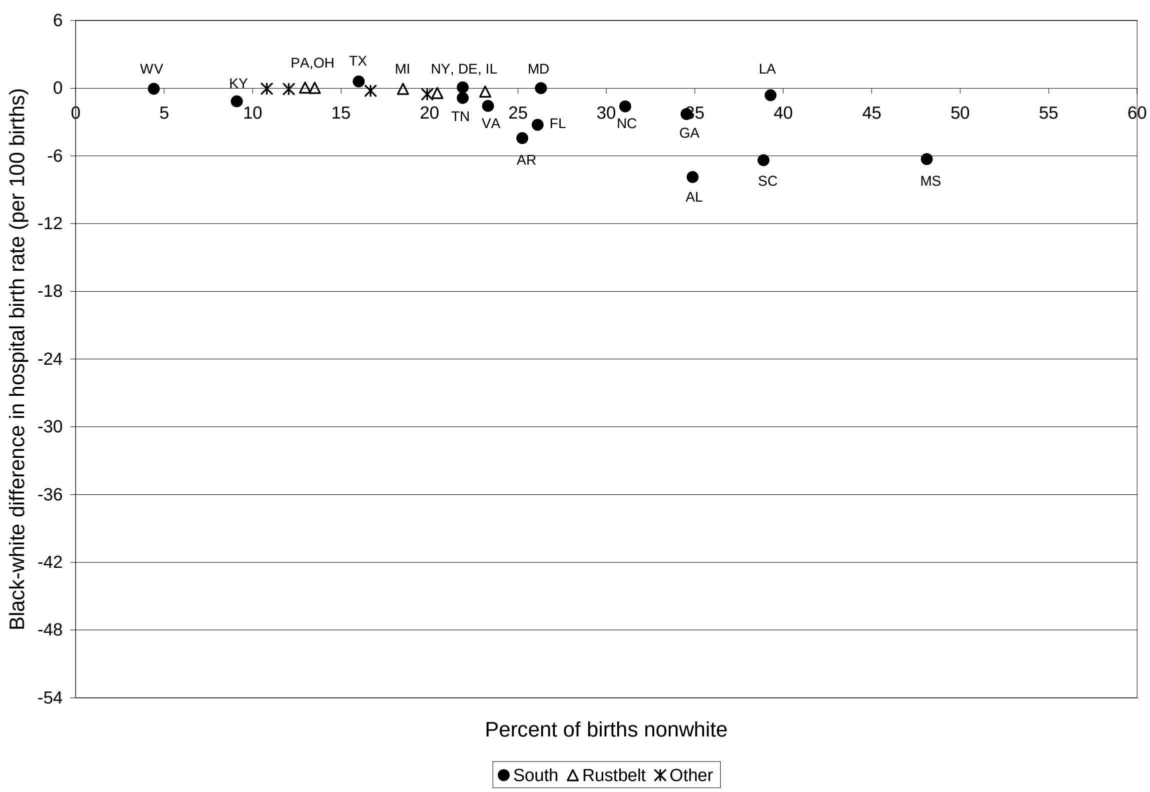
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age and year effects interacted with race, and unrestricted education effects interacted with race.

Figure 6: Black-white differences in hospital birth rates and post-neonatal mortality rates by percent of births that are black by state, before and after 1964 Civil Rights Act

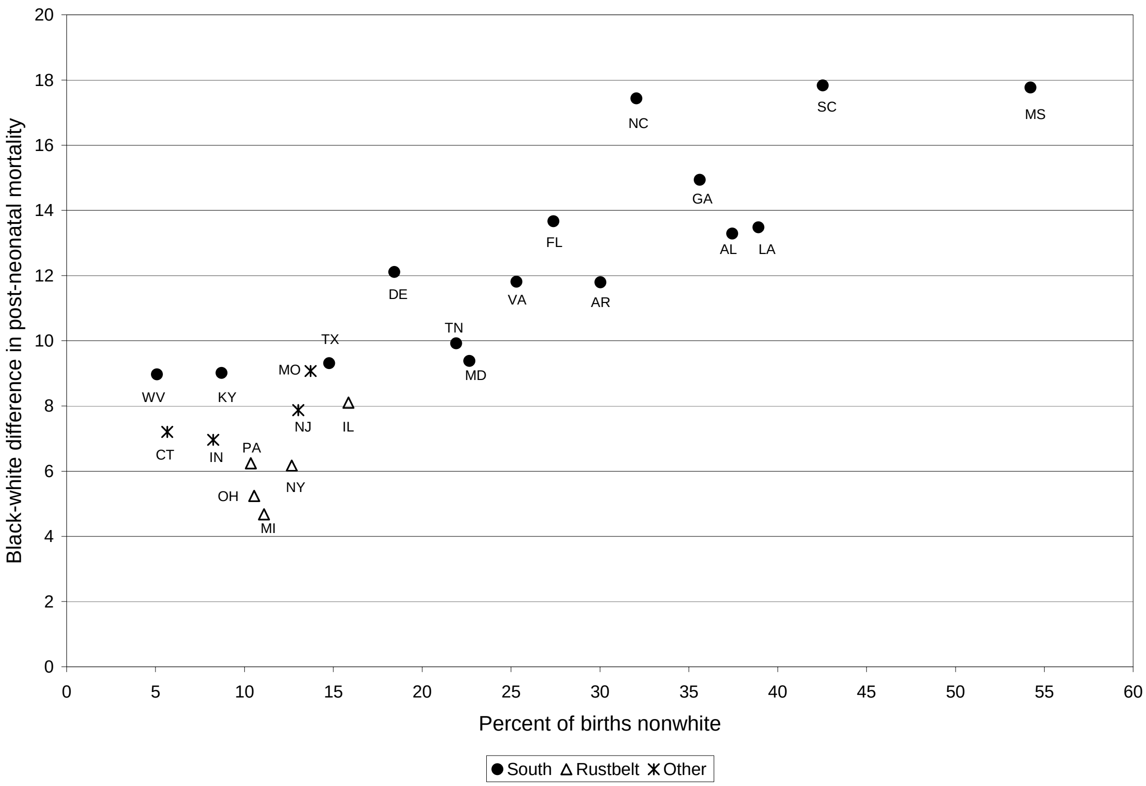
A. Black-white differences in hospital birth rates (per 100 births), 1958 to 1960



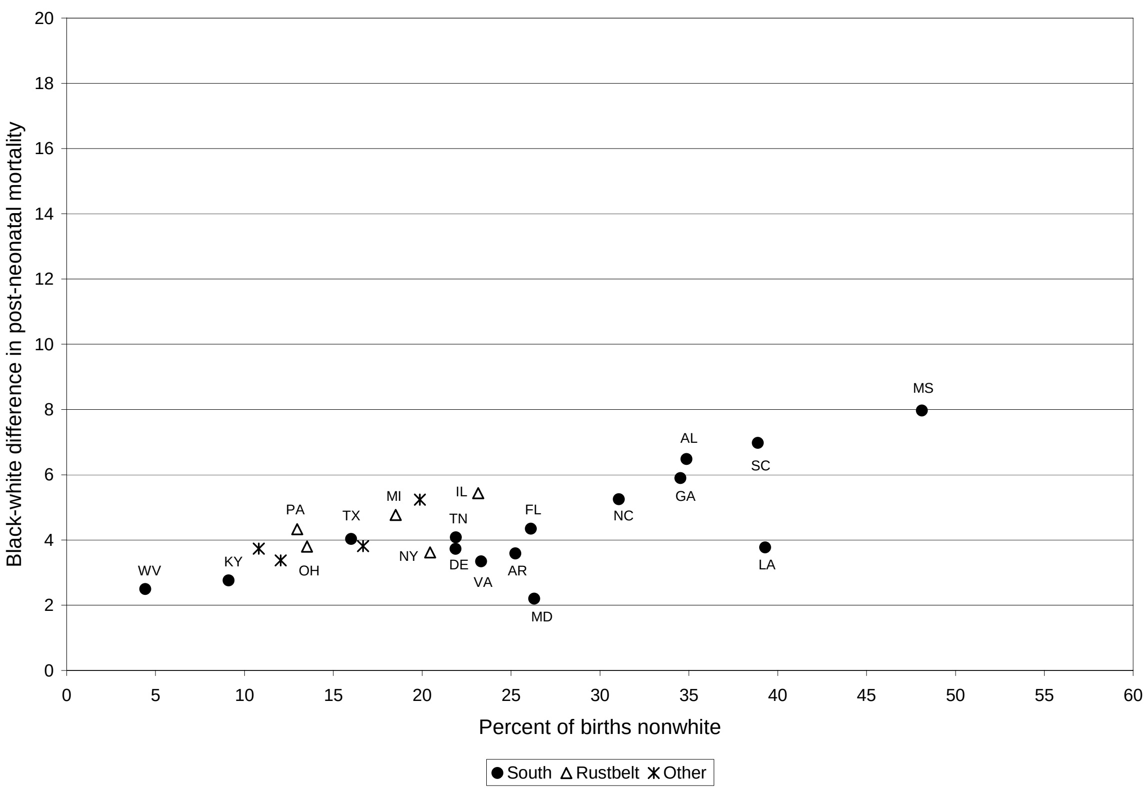
B. Black-white differences in hospital birth rates (per 100 births), 1972 to 1974



C. Black-white differences in PNMR (per 1,000 births), 1958 to 1960



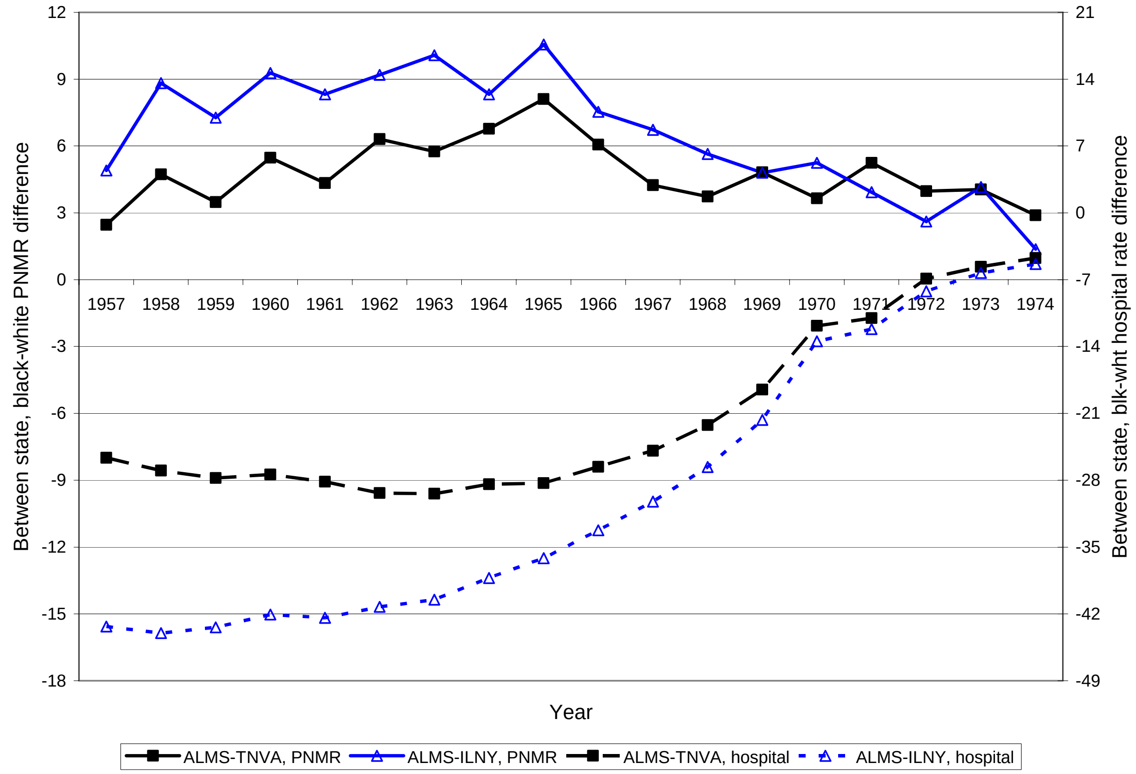
D. Black-white differences in PNMR (per 1,000 births), 1972 to 1974



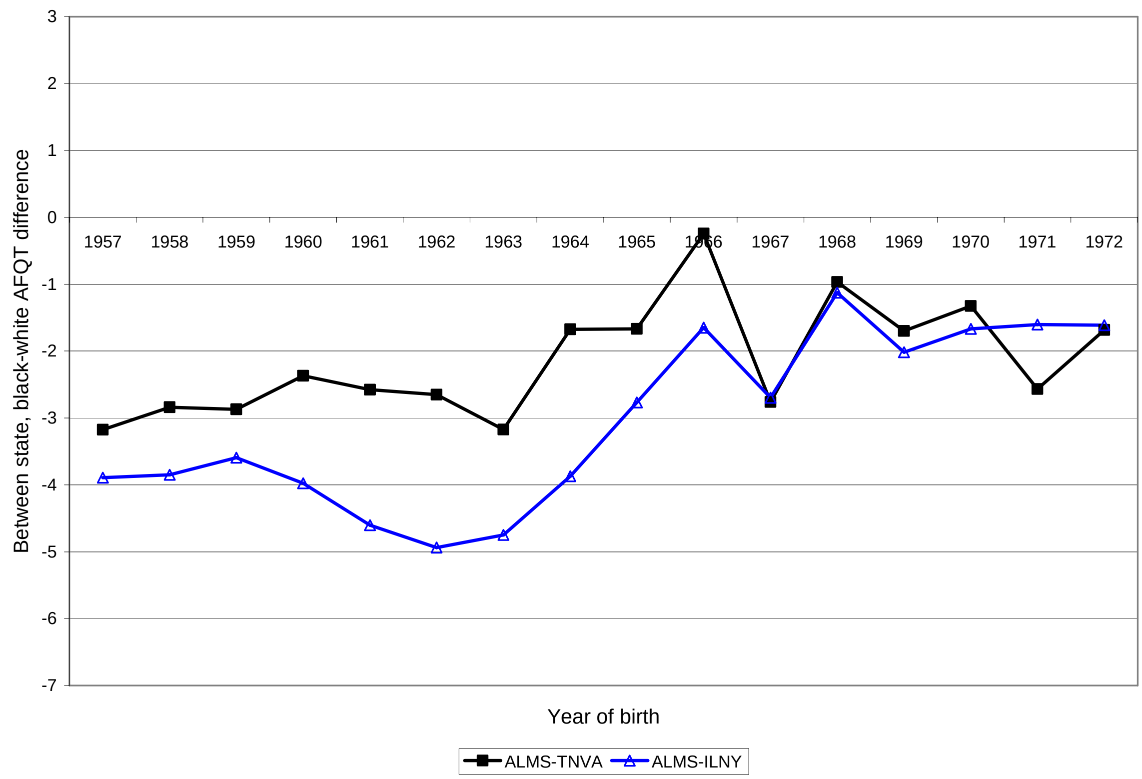
Notes: Data come from annual publications of the *Vital Statistics of the United States*. Only states in which over 95 percent of all nonwhite births are black (labeled "Negro" in the *Vital Statistics*) are included.

Figure 7: Comparisons between black-white differences in Alabama-Mississippi and in Tennessee-Virginia (Illinois-New York)

A. Post-neonatal mortality and hospital birth rates



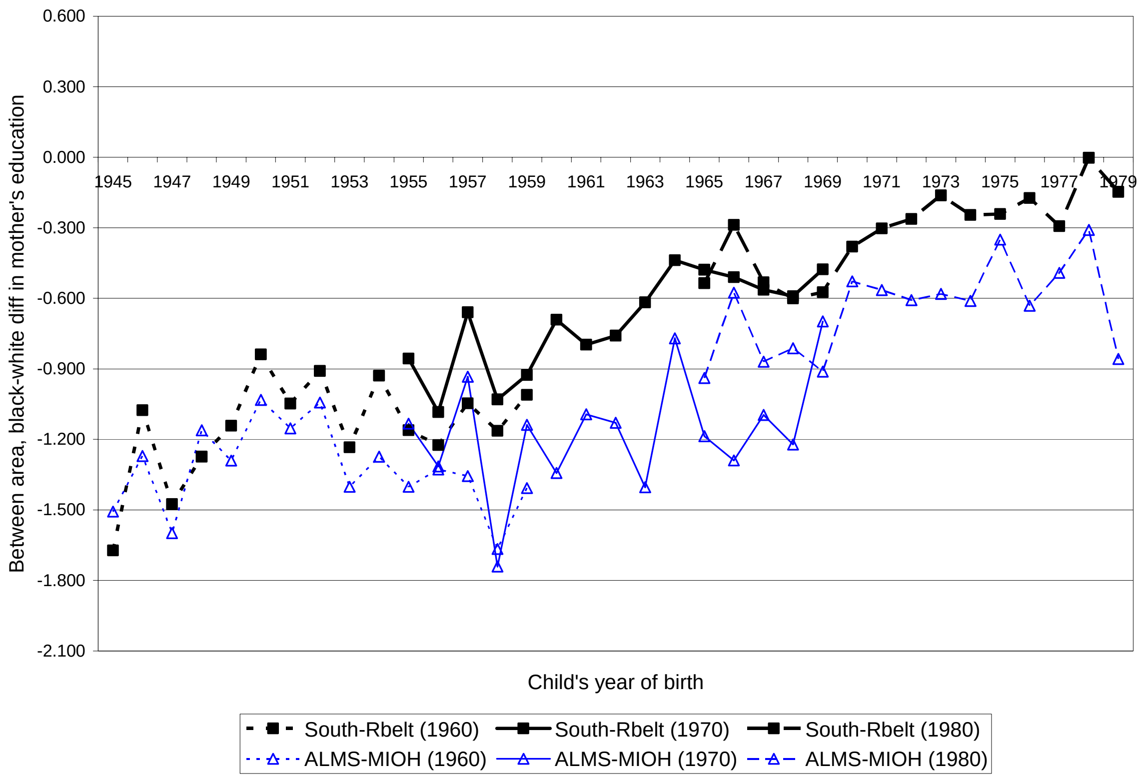
B. AFQT scores



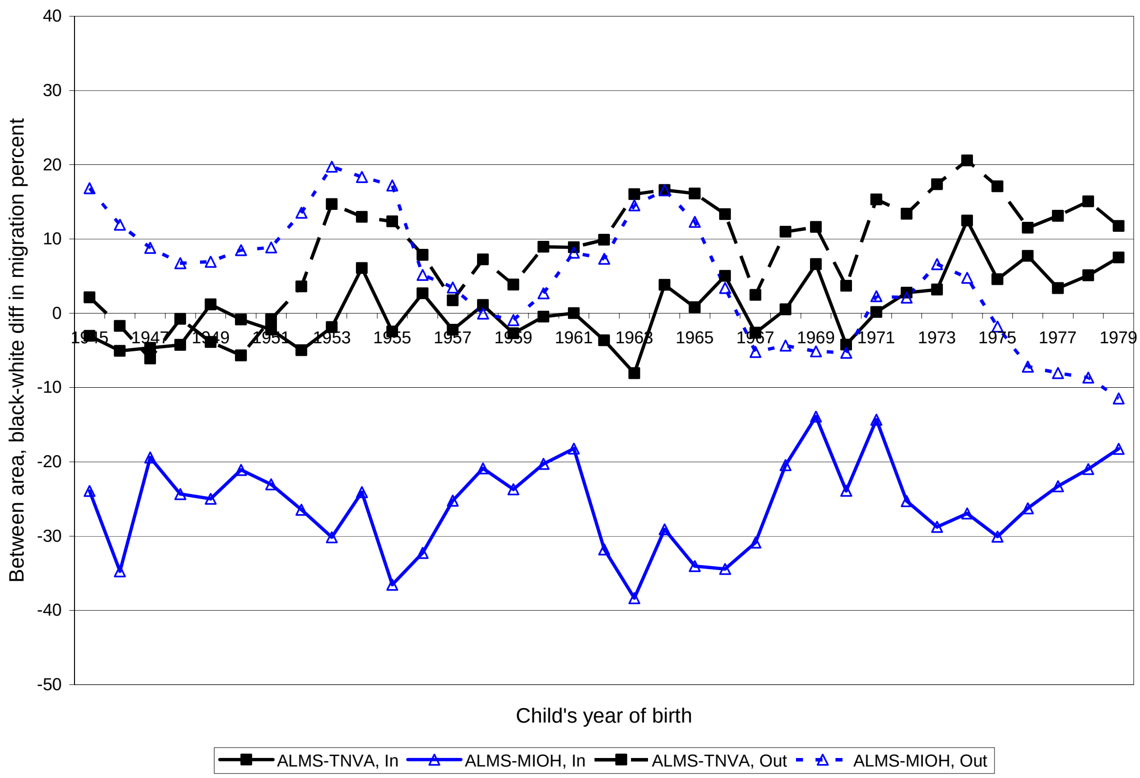
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age and year effects interacted with race and state group, and unrestricted education effects interacted with state group.

Figure 8: Competing hypotheses

A. Between area differences in black-white differences in mother's education

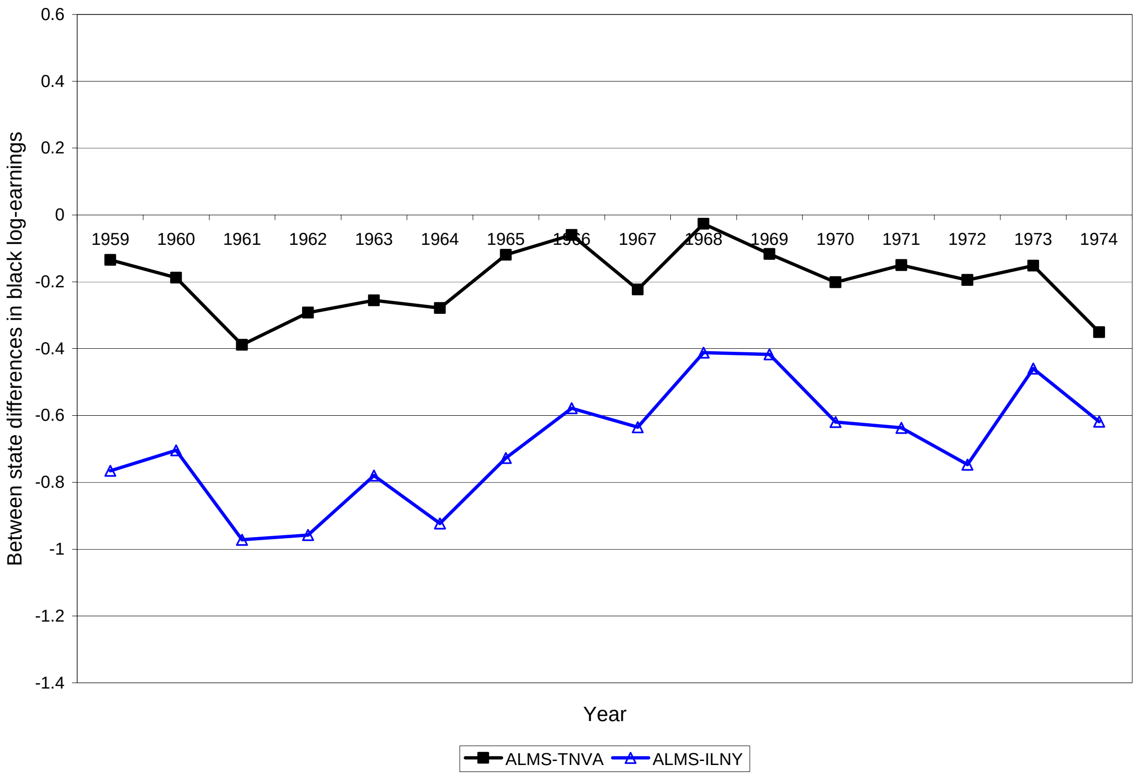


B. Between area differences in racial gap in probability of migration in and out



Notes: Panels A and B calculated from the 1960, 1970, 1980 and 1990 Decennial Censuses.

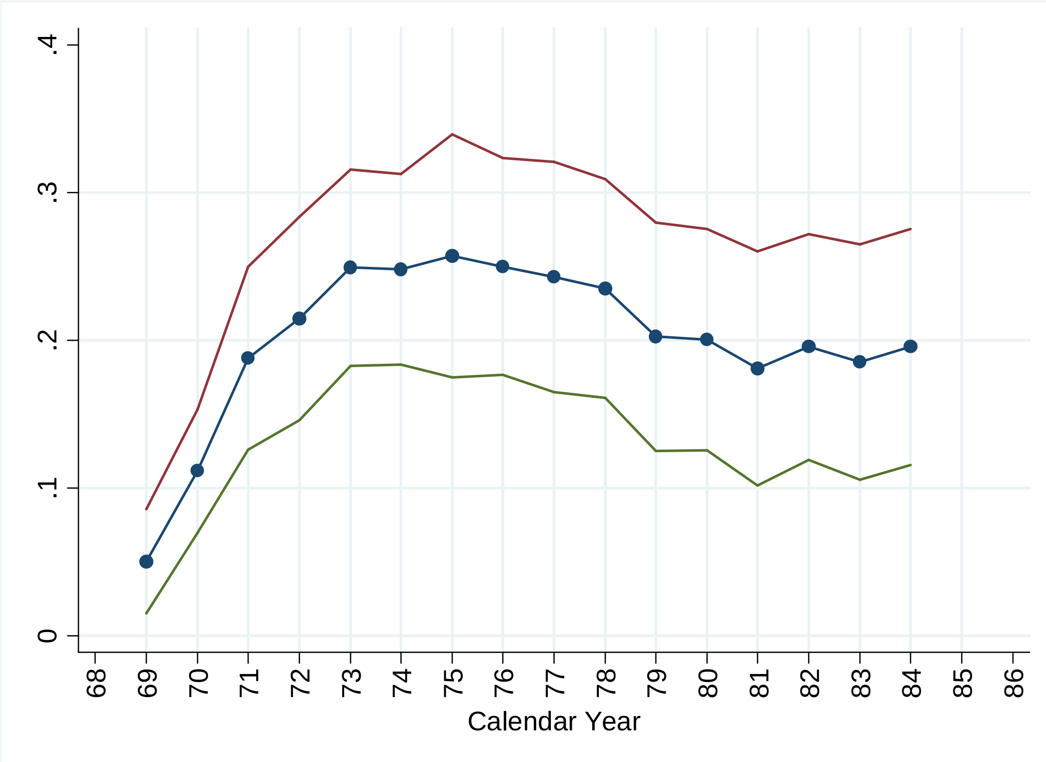
C. Between area differences in log-income of black men



Notes: Based on Social Security tax records merged to the March 1978 Current Population Survey. Results come from series of annual cross-sections that use the Tobit model to correct for censoring due to top-coding at the taxable maximum. Sample is restricted to for 19-51 year-old black men.

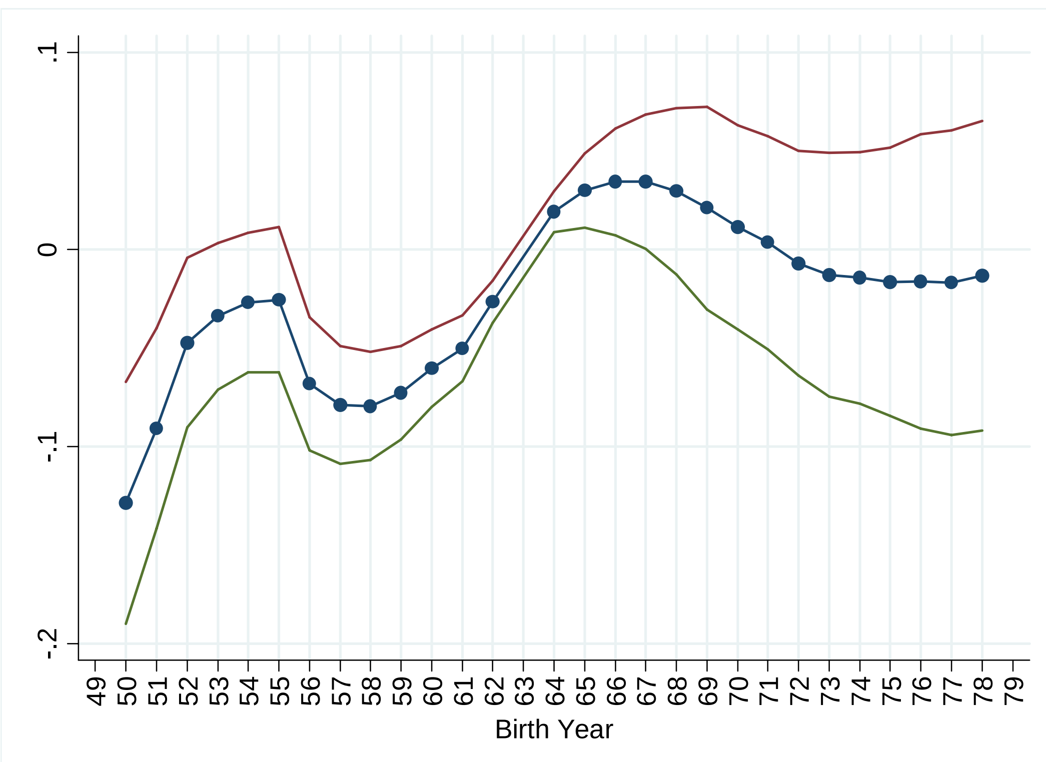
Figure 9: Desegregation Exposure

A: Black exposure to whites in school in the South by calendar year



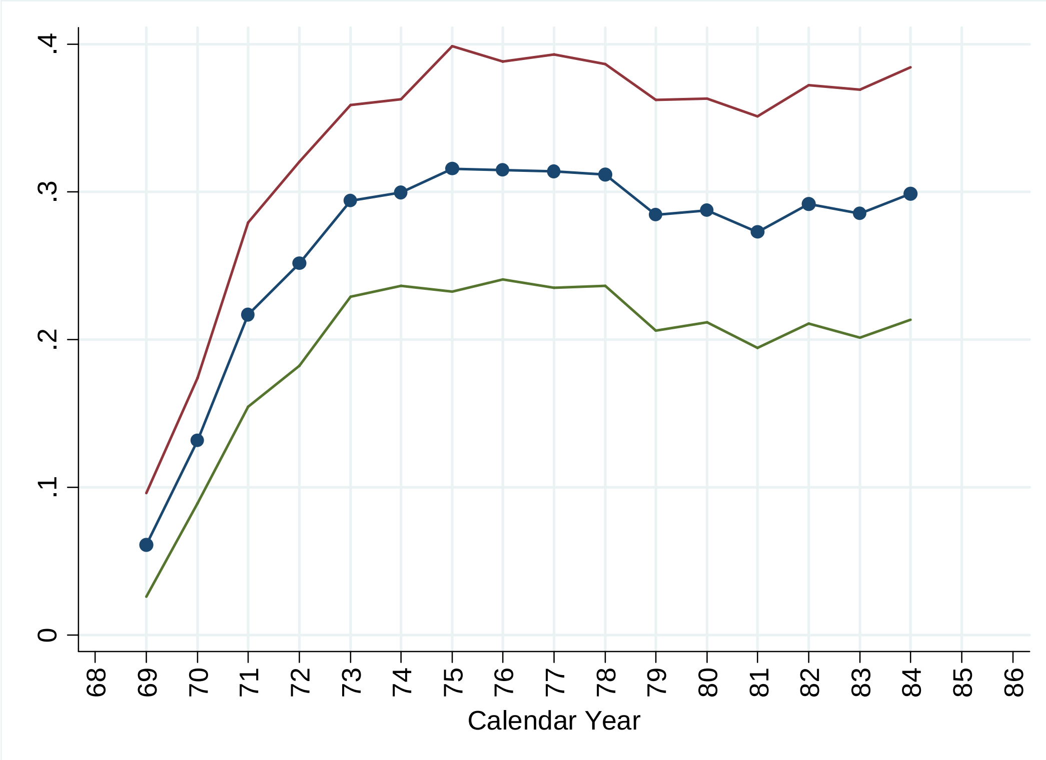
Notes: The figure plots the unconditional average black exposure to whites index by calendar year, for the sample of school districts examined in Welch and Light (1987).

B: Black exposure to whites in school in the South by year of birth



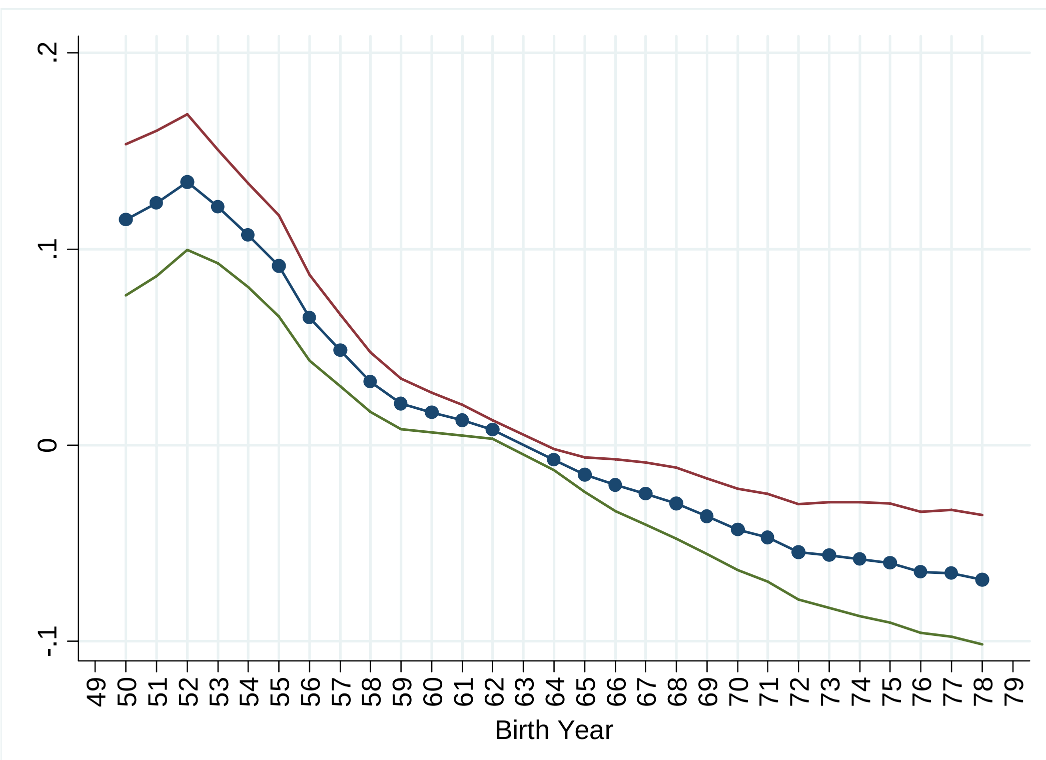
Notes: The figure plots the unconditional average black exposure to whites index by year of birth, for the sample of school districts examined in Welch and Light (1987). Exposure indices are computed by district×grade using information on which grades each school serves.

C: Black exposure to whites in school in the South by calendar year, conditional on birth year effects



Notes: The figure plots the average black exposure to whites index by calendar year, conditional on year of birth fixed effects, for the sample of school districts examined in Welch and Light (1987). Exposure indices are computed by district×grade using information on which grades each school serves.

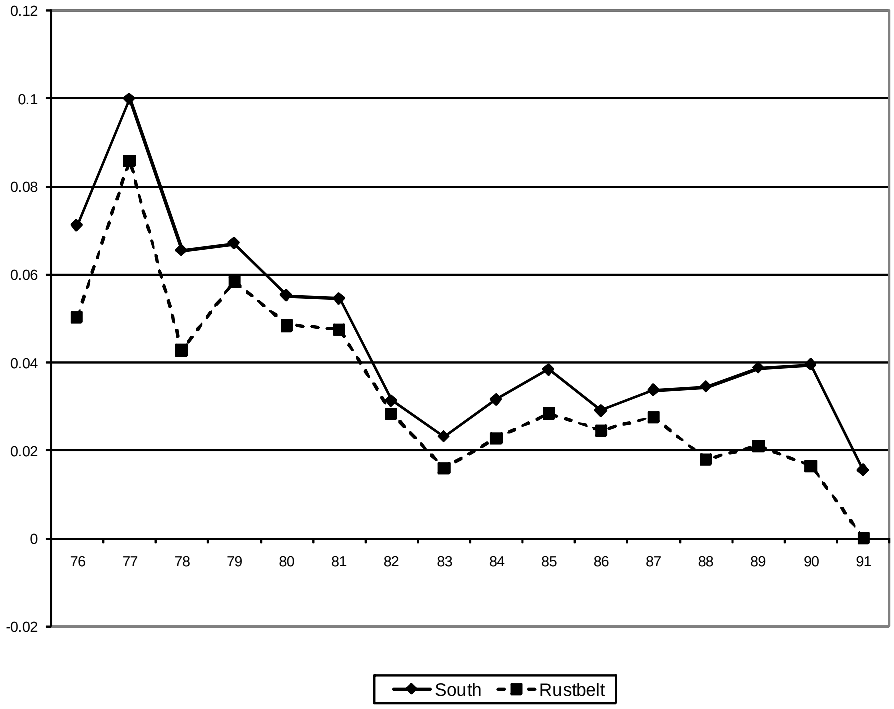
D: Black exposure to whites in school in the South by year of birth, conditional on calendar year effects



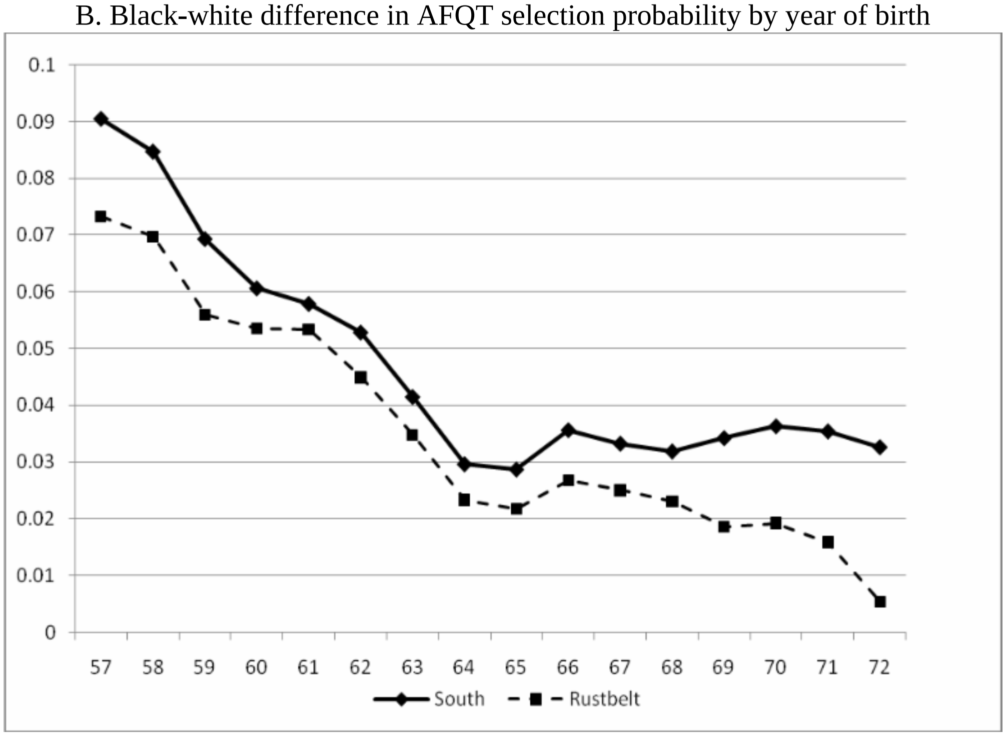
Notes: The figure plots the average black exposure to whites index by year of birth, conditional on calendar year fixed effects, for the sample of school districts examined in Welch and Light (1987). Exposure indices are computed by district×grade using information on which grades each school serves.

Appendix Figure 1: Racial differences in probability of taking the AFQT, South and Rustbelt

A. Black-white difference in AFQT selection probability by test year



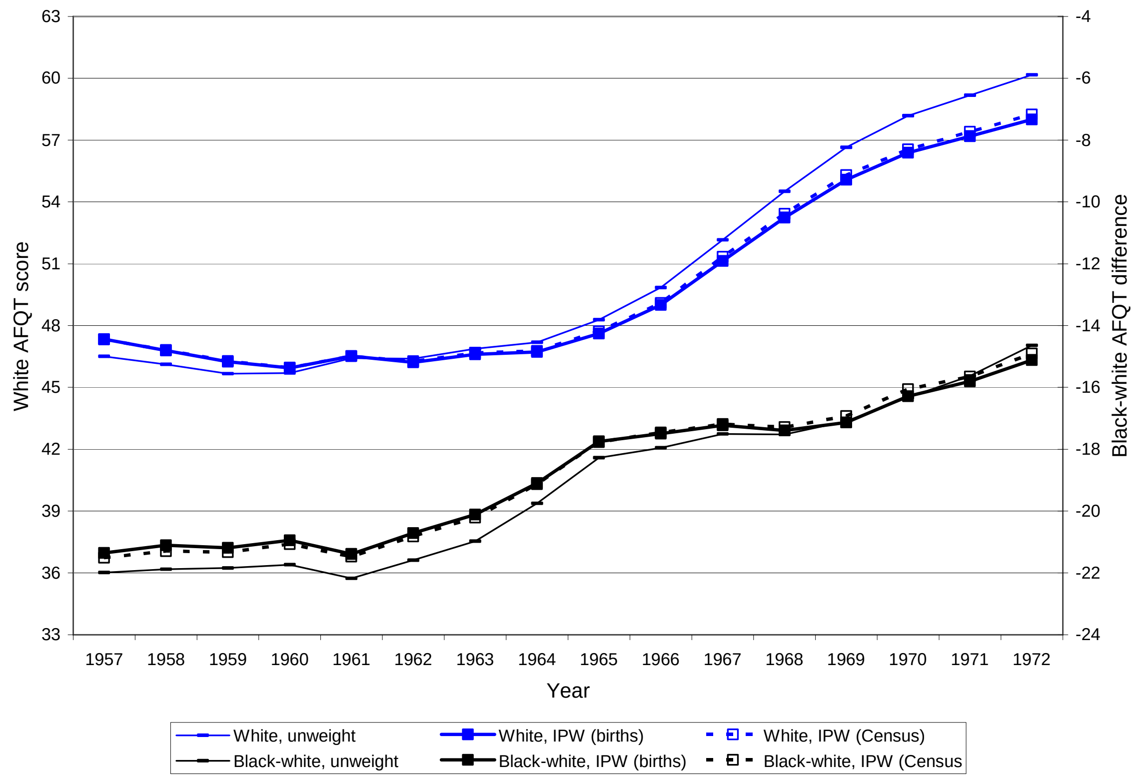
B. Black-white difference in AFQT selection probability by year of birth



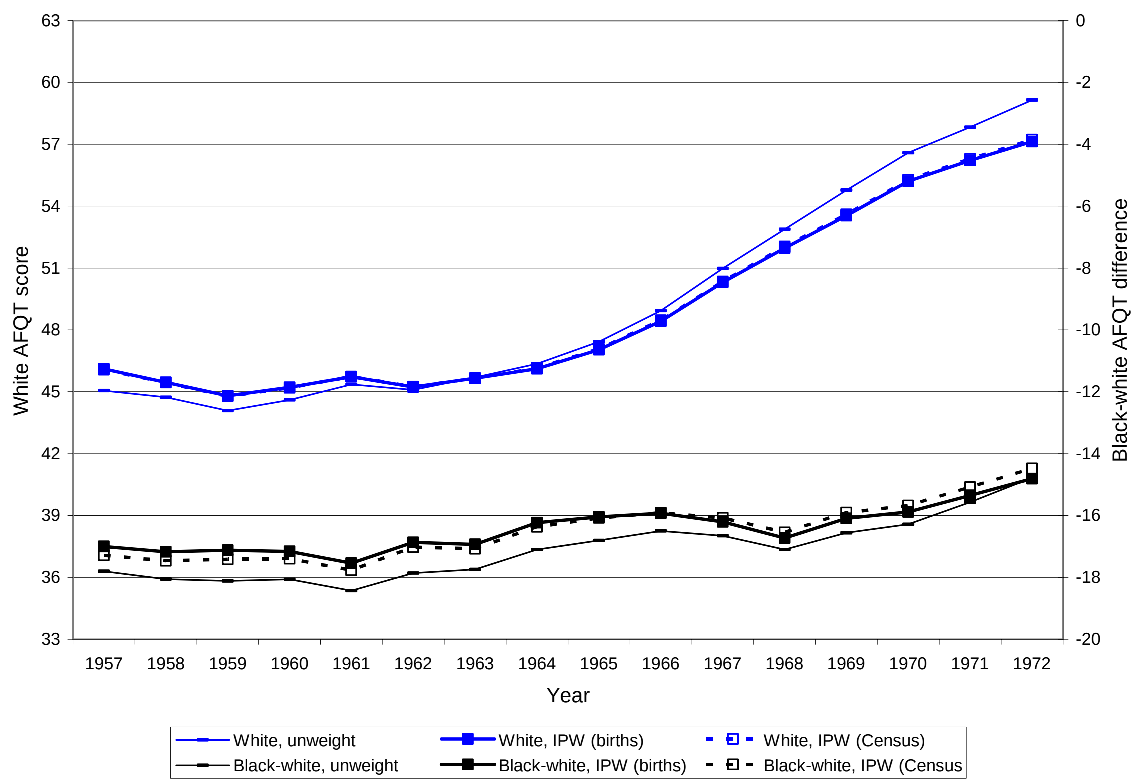
Notes: Figure plots the black-white difference in fraction of each cohort that took the AFQT test. This is also the estimated fraction of each cohort that is in the military applicant sample. In panel A this is plotted by year of the test; in panel B this is plotted by year of birth.

Appendix Figure 2: White and black-white differences in AFQT scores, uncorrected and corrected for sample selection

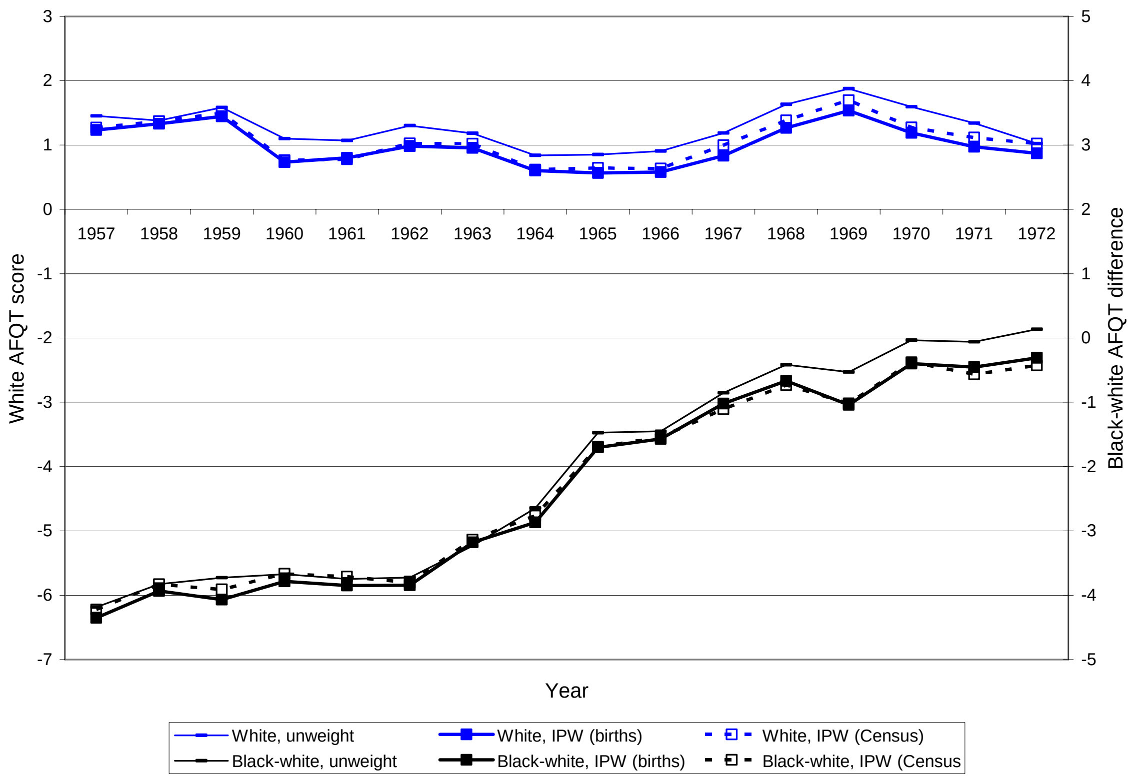
A. South



B. Rustbelt



C. South-Rustbelt difference



Notes: Regressions are run separately by region and allow for unrestricted age and year effects interacted with race, and unrestricted education effects interacted with race. "Unweight" results are from regressions that do not correct for sample selection in AFQT taking rates; "IPW (births)" are from regressions that use the inverse of the probability of taking the AFQT relative to state births (by race and year of birth) as weights; "IPW (Census)" use the inverse probability relative to the state Census (by race and age in the testing year) as weights.