

**Birth Cohort and the Black-White Achievement Gap:  
The Roles of Access and Health Soon After Birth**

Kenneth Y. Chay  
Jonathan Guryan  
Bhashkar Mazumder

April 2009

We thank ... for helpful comments.

Birth Cohort and the Black-White Achievement Gap: The Roles of Access and Health Soon After Birth  
Kenneth Y. Chay, Jonathan Guryan, and Bhashkar Mazumder  
April 2009

### **ABSTRACT**

One literature documents a significant, black-white gap in average test scores, while another finds a substantial narrowing of the gap during the 1980's, and stagnation in convergence after. 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: 1) the 1980's convergence is due to relative improvements across successive cohorts of blacks born between 1963 and the early 1970's and *not* a secular narrowing in the gap over time; and 2) the across-cohort gains were concentrated among blacks in the South. We then demonstrate that the timing and variation across states in the AFQT convergence closely tracks racial convergence in measures of health and hospital access in the years immediately following birth. We show that the AFQT convergence is highly correlated with post-neonatal mortality rates and not with neonatal mortality and low birth weight rates, and that this result cannot be explained by schooling desegregation and changes in family background. We conclude that investments in health through increased access at very early ages have large, long-term effects on achievement, and that the integration of hospitals during the 1960's affected the test performance of black teenagers in the 1980's.

Kenneth Y. Chay  
Department of Economics  
Brown University  
Box B  
Providence, RI 02912  
and NBER  
[kenneth\\_chay@brown.edu](mailto:kenneth_chay@brown.edu)

Jonathan Guryan  
Booth School of Business  
University of Chicago  
5807 S. Woodlawn Avenue  
Chicago, IL 60637  
and NBER  
[jonathan.guryan@chicagoGSB.edu](mailto:jonathan.guryan@chicagoGSB.edu)

Bhashkar Mazumder  
Federal Reserve Bank  
230 S. LaSalle Street  
Chicago, IL 60604  
[bmazumde@frbchi.org](mailto:bmazumde@frbchi.org)

## I. Introduction

By most measures, there is a significant gap in skills between blacks and whites in the United States. One such measure that has received much attention from social scientists and the public 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.<sup>1</sup>

In this paper we present evidence on the source of the convergence in the measured black-white test score gap during the one period in which the gap fell significantly – the 1980's.<sup>2</sup> Our analysis uses two datasets of test scores: National Assessment of Educational Progress-Long Term Trend (NAEP-LTT) scores from 1971 to 2004, and Armed Forces Qualifying Test (AFQT) results for the *universe* of applicants to the U.S. military between 1976 and 1991. While the former is a nationally representative random sample, it is relatively small and lacks the detail needed to make comparisons at narrowly-defined geographic levels while plausibly differentiating age, year and birth cohort effects. The latter is ideal for addressing these issues, but only includes those who applied for potential induction into the military.

We correct for the potential selection bias in the AFQT sample by: i) conditioning on a large and rich set of fixed effects, effectively differencing out several sources of selection across time, demographic groups and geographic areas; and ii) adjusting for any remaining selection by using Inverse Probability Weighting (IPW), in which each AFQT observation is weighted by an estimate of the probability of selection into the sample within unrestricted state-race-age-year cells. In our context, these probabilities are credible and easy to construct since we have the universe of those selected; and therefore we only need to estimate the population size for a particular cell, which we do by using Census and Natality data. Furthermore, we can examine patterns in the selection probabilities to assess the appropriateness of using fixed effects to control for selection in the regression models. We find similar results from the NAEP-LTT and (corrected) AFQT samples along several dimensions, suggesting that our models deal effectively with selection into AFQT test taking.

---

<sup>1</sup> Fryer and Levitt, Neal, Card and Rothstein

<sup>2</sup> Hanushek, Dickens and Flynn, Cook and Evans

Both datasets 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, we find that for the cohorts born in the 1950's and early 1960's, the racial gap in NAEP scores is large (1.3 to 1.4 standard deviations) and exhibits no convergence across cohorts. Beginning with those born in the mid-1960's, however, there are striking across-cohort improvements in black relative test scores that continue up to those born in the early 1970's, with the NAEP gap narrowing by 0.6 standard deviations. Also, the across-cohort reductions in the gap are much larger among students in the South than for their Northern counterparts.

The AFQT data – which allow for a more detailed differentiation between age, year and birth cohort effects – also show a large reduction in the racial test score gap that is concentrated in the South. Southern black and white AFQT scores show no convergence between cohorts born in the late 1950's and early 1960's; however, the AFQT gap is 40 percent smaller by the early 1970's cohorts. Further, this cohort-based convergence explains all of the narrowing of the AFQT gap in the South during the 1980's – that is, the racial convergence across the calendar years of the 1980's appears to have been the result of factors related to the year and place in which the test taker was born.

Having established the importance of cohort effects, we propose and test a specific hypothesis for the cohort-related convergence in the test score gap; that it was caused by relative improvements in black health in the years immediately after birth. As Almond, Chay and Greenstone (2008) demonstrate, black relative infant mortality – particularly in the post-neonatal period 28 days to one year after birth – fell dramatically in the United States between the mid-1960's and mid-1970's. Further, the improvements varied widely across states, with the greatest convergence occurring in the South. They argue that these patterns, as well as their concentration in causes of death sensitive to hospital admission (pneumonia and diarrhea), were largely the result of the forced integration of Southern hospitals in the 1960's. Consistent with this, they find strong evidence of increased access and admission of black infants to hospitals in the South following the integration.

In this paper, we hypothesize that this intervention(s) led to improved postnatal health among blacks born between the early 1960's and early 1970's, which in turn led to long-term improvements in

the academic and cognitive “skills” of these cohorts as teenagers (aged 17 and 18). The neuroscience literature has found that the most critical and rapid period of human brain development occurs within the first three years of life; this development is vulnerable to postnatal experience; and these effects are long lasting.<sup>3</sup> For example, recent medical research has found an association between diarrheal disease burden in the first two years of life (in Brazilian shantytowns) and impaired cognitive development and school performance later in childhood.<sup>4</sup>

In the absence of a perfect measure of latent health in infancy and early childhood, we use the post-neonatal mortality rate (PNMR) as a proxy. Previous work of ours has shown the strong association between PNMR and postnatal access to hospital care (Almond, Chay and Greenstone 2008); thus, we also view it as a proxy for hospital access. The caveats of using PNMR as a latent health measure are discussed as well.<sup>5</sup>

Graphical and regression analyses show a remarkable correspondence between the racial gaps in AFQT and PNMR by one’s place and year of birth. The timing and variation across states in the AFQT convergence closely tracks PNMR convergence in the years immediately following birth; with falling PNMR’s explaining 50 to 80 percent of the across-state variation in cohort-to-cohort reductions in the AFQT gap. On the other hand, the AFQT convergence has little to no correlation with low birth weight (LBW) and neonatal mortality (NMR) rates, family background measures, and migration rates.

The AFQT gap is most highly correlated with the PNMR gaps that prevailed one and two years after the cohort was born. This result suggests that an improvement in health in the first two to three years of life for black children may be the cause of the narrowing of the test score gap in the 1980’s. The weak correlations of the AFQT gap with LBW and NMR suggest that the *root causes* of the black test score gains were postnatal factors that affected health, rather than *in utero* conditions.

The post-birth factor that we focus attention on is the increased admission of black infants and children to hospitals following the desegregation efforts of the 1960’s. Using newly available data from

---

<sup>3</sup> Johnson 2001

<sup>4</sup> Oriá, et al. 2005 and 2007. Also Currie, et al.

<sup>5</sup> For example, infant health can improve with little effect on infant survival rates, and mortality rates are inherently linked with potential selection bias in who survives to the ages at which the tests are administered.

the *National Health Interview Survey* on hospital discharge rates, we show that hospital admissions of black children up through the age of four increased significantly more in the South than in the North after desegregation. Taken literally, the magnitudes imply that a black child who gained admission to a hospital early in life had, on average, a 0.7 to 0.9 standard deviation gain in their AFQT score relative to a counterpart who was denied admission. We use these numbers to estimate the costs of narrowing the black-white test score gap under the assumption that the narrowing resulted solely from the racial integration of Southern hospitals.

Finally, we investigate a number of competing hypotheses for the racial convergence in test scores. We note that while there are plausible alternatives to hospital integration as a root cause, several of these stories share the feature that black health improvements are the mechanism for the narrowing of the test score gap – for example, the expansions of AFDC, Medicaid, Food Stamps and Head Start. Further, we discuss how the roll-outs of many of these programs do not match the across-state patterns in AFQT convergence as well as PNMR. The stories that do not rely on early health as a mechanism – in particular, school desegregation – also fail to match the cohort-based convergence in test scores. We conclude that investments in health at very early ages have large, long-term effects on achievement, and that the integration of hospitals during the 1960's affected the test performance of black teenagers in the 1980's. Future research, however, should compile more evidence on the potential role of each alternative story, as well as examine additional human capital outcomes.

The next section presents background on infant mortality trends in the United States for the key cohorts. In Section III, we show results from the NAEP-LTT data, which match the PNMR trends. Section IV describes the military applicant dataset that contains AFQT scores, and Section V describes and presents evidence on the models used to correct for selection into the AFQT sample. In section VI, we present the AFQT results, which also match the regional trends in PNMR. Section VII states and tests the early health hypothesis, and shows the results comparing the roles of various markers of early life health. Section VIII presents evidence on hospital integration as a root cause and provides cost-benefit estimates, while Section IX discusses alternative root causes. We conclude in Section X.

## II. Aggregate trends in infant mortality, 1950 to 2000

Below, we find that improvements in black relative test scores accrued to cohorts born between the early 1960's and early 1970's, and that these gains are concentrated among blacks in Southern states. Here we briefly present background on trends in infant mortality rates in the United States after 1950 since we hypothesize that the test score gains are linked to cohort health soon after birth. Though we do not have data on an ideal measure of latent health in infancy and early childhood, we use mortality rates of infants in the first year of life as proxies for the early health of cohorts. Later in the paper, we discuss the caveats with using these proxies and formally lay out the conditions under which they are useful.

The second half of the 20<sup>th</sup> century saw a remarkable improvement in these indicators for blacks in the United States. Panel A of Figure 1 plots the black-white difference in the infant mortality rate (IMR) – defined as number of deaths within the first year of life per 1,000 births – from 1950 to 2000. During the 1950's and early 1960's, there was a fairly stable black-white gap in infant mortality of about 20 per 1,000 births. Following 1964, this gap began to narrow dramatically. Within ten years, the infant mortality gap had closed to about 12 in 1,000.

Panel B separately plots the racial gaps in neonatal (NMR, deaths within one month of birth per 1,000) and post-neonatal (PNMR, deaths between one month and one year following birth per 1,000) mortality rates for 1950 to 1990. It shows that nearly three-quarters of the decline in the IMR gap between 1964 and 1972 is attributable to PNMR convergence. This suggests that for these cohorts of blacks, post-neonatal health improved substantially more than neonatal health. After the mid-1970's, however, the relatively small declines in the IMR gap are driven mostly by NMR convergence.

For the South and Rustbelt, respectively, Panels C and D of Figure 1 show the trends in the racial gaps in PNMR and NMR, as well as the gap in the percent of infants born at low birth weight (LBW).<sup>6</sup> The patterns are substantially different across regions. The sharp, national decline in the PNMR gap between the mid-1960's and early 1970's is concentrated in the South, where the decline in the NMR gap

---

<sup>6</sup> The South consists of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia; the Rustbelt of Illinois, Indiana, Michigan, Missouri, New York, Ohio and Pennsylvania. These are the state groupings that we use when we examine AFQT scores below, where we also examine the Border states of Delaware, Kentucky, Maryland, Texas, and West Virginia. According to the U.S. Census, 46 (29) percent of all blacks lived in the South (Rustbelt) in 1960 by this definition of regions.

is comparably small. In the Rustbelt, the racial convergence in NMR is larger than that of PNMR. Overall, the decline in the IMR gap is much larger in the South, and about 80 percent of the Southern decline is due to PNMR convergence. The comparable figure for the Rustbelt is 30 percent. In both regions, there is little improvement in the LBW gap between 1955 and 1975; indeed the gap widens slightly for blacks born in the South.

### **III. NAEP test score results**

We begin the analysis of cognitive outcomes 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.<sup>7</sup> We analyze microdata of students' math and reading scores for all available years in which the tests were administered, for both boys and girls.<sup>8, 9</sup> We present the results from the scaled scores, which we further standardize by the standard deviation of scores by subject, student's age and year of the exam.<sup>10</sup>

Panel A of Figure 2 plots the black-white gap in the standardized reading and math scores by the calendar year in which the test was taken. These estimated gaps are derived from subject-specific

---

<sup>7</sup> 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. While the NAEP-LTT consists of random samples of enrolled students, some selection bias may be induced in the 17-year old sample by high-school dropouts. However, we have confirmed 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.

<sup>8</sup> 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 unable to obtain the 1973 math scores; but scores from all other tests listed above are included in the analysis.

<sup>9</sup> Below, we restrict our analysis of AFQT scores to men only. We make this restriction in the military sample since we are primarily concerned with non-random selection and 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 boys and girls are selected into the NAEP test-taking sample in the same way.

<sup>10</sup> 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)).

regressions that include race-specific age effects.<sup>11</sup> Consistent with the previous literature, the figure highlights a marked convergence in the black-white test score gap during the 1980's. From 1971 to 1980, the racial gap in NAEP-LTT scores remained fairly constant at slightly above 1.2 standard deviations. Between 1980 and 1988, however, the gap fell by about 0.4 standard deviations. This convergence halted abruptly in 1990, and for the next 15 years the gap shows no convergence.

In Panel B of Figure 2, we plot the standardized NAEP gaps separately for each age cohort. These are derived from regressions that pool math and reading scores and adjust for race-specific subject effects by age. When the 9-, 13- and 17-year-old series are plotted by year of the exam, an interesting pattern emerges. The black-white convergence seen in Panel A 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 racial convergence in NAEP-LTT scores begins at some point before 1974 for 9-year-olds; starts between 1978 and 1980 for 13-year-olds; and begins between 1980 and 1982 for 17-year-olds.

This pattern implies that the racial convergence in NAEP scores during the 1980's shown in Panel A was not a secular time phenomenon, but rather occurred at different points in time for different age groups. It further suggests that the 1980's convergence is better understood as having accrued to successive birth cohorts of blacks, beginning with those born in the early-to-mid 1960's. Consistent with this interpretation is the fact that while there are significant age effects in the racial NAEP gaps between 1971 and 1980 – with the gap increasing with age – these effects disappear by the early 1990's.

Panel C of Figure 2 directly examines the possibility that the test score convergence is linked to birth year instead of calendar year. It plots white NAEP scores and the black-white gap by the year of birth of the student, which are derived from regressions that adjust for race-specific age effects that vary by subject. White scores show a general trend of improvement across successive cohorts born between 1953 and 1989.

The most striking pattern, however, is in the racial gap in NAEP scores. For blacks born between 1953 and 1964, there is a 1.3 to 1.4 standard deviation gap that shows no improvement across successive

---

<sup>11</sup> Put specification here – also weighted by sampling weights.

cohorts. However, this gap narrows by 0.6 standard deviations between the 1964 and 1973 birth cohorts, with no racial convergence for the cohorts born between 1973 and 1989. These patterns mimic the patterns in the racial gap in PNMR shown in Figure 1B, with about a one-year lag. That is, the sharp convergence in NAEP scores between the 1964 and 1973 birth cohorts roughly match the PNMR convergence that occurs between 1965 and 1974.<sup>12</sup>

An ideal analysis would explicitly distinguish between the test score convergence that can be attributed to a student's year of birth and convergence that can be explained by secular improvements that affected black children of all ages. Unfortunately, the design of the NAEP-LTT leads to the well-known problem of perfect collinearity between age, birth year, and year the exam is taken. Further, the relatively small sample sizes and the fact that the tests are not administered annually (and not for more age groups) make it difficult to use flexible, parametric restrictions on the various effects and still recover reasonably precise estimates.

However, if the race-by-year effects do not vary by geographic region, then comparing cohort-to-cohort convergence across regions (while adjusting for race-specific age effects that vary by region) allows for identification of the relative cohort effects in the test score gap. Given the different historical experiences in the U.S. South and North (as shown in Figure 1), it is also natural to ask whether black-white test score convergence followed different patterns in these two regions.

Panel D of Figure 2 provides evidence on these questions by plotting the racial gap in NAEP scores by birth cohort, separately for the South and North.<sup>13</sup> These are derived from regressions that control for race-specific age effects that vary by subject and region. It is clear that the between-cohort racial convergence was substantially larger among students in the South than for their Northern counterparts. For the 1953 to 1964 birth cohorts, the test score gap is about 0.3 to 0.4 standard deviations

---

<sup>12</sup> 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.

<sup>13</sup> In the NAEP-LTT, the South consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Southern Virginia, and West Virginia. We define the North to be the combined regions of East (Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Northern Virginia) and North Central (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin). Additional details are provided in the Appendix.

greater in the South than in the North with no pattern toward convergence. However, by the early 1970's birth cohorts, this regional difference has been completely eliminated, and the racial gap is actually slightly smaller among students in the South. Further, the patterns in the figure have a strong (inverted) relation to the patterns in the PNMR gaps in Figures 1C and 1D, but little resemblance to those in the NMR and LBW gaps.

It appears that the geographic place and year of birth is highly predictive of the racial gap in test scores. These findings are unchanged after we further control for family background variables, which are available for a subset of the data. This suggests that changes over time in the characteristics of the parents of black and white students are not driving the cohort-based convergence.

Next, we use the data for 17-year-olds to estimate differences between the South and North in the across-cohort racial convergence in reading and math scores, separately. These results provide a point of comparison for the AFQT results below, which are based on a sample of 17- and 18-year-olds. Columns (1a) to (2b) of Table 1 are based on the 1971, 1980 and 1990 Reading scores data. They show that for the 1953 to 1954 birth cohorts [column (1a)], the reading score gap at age-17 was 1.3 standard deviations (s.d.'s) in the South – roughly 0.1 s.d.'s greater than the gap in the North. The gap rises by 0.22 s.d.'s for the 1962 to 1963 cohorts in the South [column (1b)]; a greater divergence than the one in the North. This results in a reading score gap that is 0.24 s.d.'s greater in the South than North as of the early 1960's cohorts [column (2a), bottom row]. Between the 1962-1963 and 1972-1973 birth cohorts, however, the Southern gap falls by 0.83 s.d.'s, which is 0.37 s.d.'s greater than the Northern convergence. Thus, by the early 1970's cohorts, the reading score gap is 0.13 s.d.'s *smaller* in the South than North.

Columns (3a) and (3b) of Table 1 show similar results using the Math score data in the 1978 and 1990 surveys. For the 1961 birth cohort, the math score gap in the South was 1.28 s.d.'s, which is 0.13 s.d.'s larger than the Northern gap. The Southern gap falls by 0.7 s.d.'s in the 1972 to 1973 cohorts; a convergence that is 0.41 s.d.'s greater than the one in the North.

These results are highly (statistically) significant and provocative. They show that while the racial gap in test scores grew between the early 1950's and early 1960's birth cohorts – even more so in the South – there was a striking reduction in the gap for the early 1970's cohorts that was much larger in

the South than North. This suggests that events linked to the year-of-birth improved the relative standing of 17-year-old blacks born between 1963 and 1973; these events did not affect children of all ages in a given year; they did not negatively affect whites; and they affected blacks in the South more than their counterparts in the North.

That said, the NAEP-LTT data are neither large nor detailed enough to: i) examine more refined comparisons that would allow us to distinguish between potential causes; and ii) attempt to differentiate (region-specific) time effects from the effects of birth year during the critical period. As a result, we turn to a much larger dataset that has test scores for both blacks and whites for the birth cohorts of interest.

#### **IV. AFQT data for the universe of applicants to the U.S. Military, 1976 to 1991**

For any study examining changes in outcomes over time or the life-cycle, or across cohorts, it is critical to plausibly distinguish between time, age and birth year effects. As is well-known, it is impossible to perform this decomposition without assumptions since these effects are perfectly collinear at an appropriate level of detail (e.g., detailed age, day at which the outcome is measured, day of birth). Indeed, in most survey designs – such as the one used in the NAEP-LTT – the *year* of birth is equal to the survey *year* minus an individual's age (in years) at the time of the survey. An additional limitation of the NAEP-LTT is that tests are not administered on an annual basis.

While not “solving” this identification problem, a major improvement over typical survey designs would be to measure outcomes for large, random samples *on a rolling basis*. For example, one would collect test scores for samples of blacks and whites of similar ages at multiple points within a calendar year, and for a long and continuous set of calendar years. In this case, survey year, age-in-years and birth year would not be perfectly collinear since, for example, there can be 17-year-olds born in the same year who happen to take the exam in different calendar years. Of course, completely unrestricted effects at a fine enough level of detail – exact birthday, exact age at and date of exam – would still be collinear. However, this information is almost never available, and a restriction of common effects across the days of a calendar year does not seem innately implausible.

For the purposes of this study, it is also critical to have a data source large and geographically detailed enough to allow for these narrow comparisons across regions and even states. This paper presents results from a unique dataset that satisfies many, though not all, of these criteria. In particular, we have obtained data on the test scores of the *universe* of applicants to the United States military between 1976 and 1991. The data include the birth year of the applicant and his age-in-years, as well as completed education, and zip-code of residence; all measured 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 to 23 year olds from the *Profile of American Youth 1980*.<sup>14</sup>

The AFQT data are summarized in Table 2. Columns (1a) to (1c) are based on the sample of men aged 17 to 20 at the time of application, who were born between 1957 and 1973 and took the exam in either the South, Rustbelt, or Border states – over four million observations.<sup>15</sup> They show: i) two-thirds of these men applied to the military at age 17 or 18; ii) nearly 90 percent had completed at least three years of high school at the time of application, with 46 percent having graduated high school, earned a GED, or started college; and iii) black applicants had slightly more completed education, on average, than white applicants, but their AFQT scores were over 20-percentile points lower at the mean and median.

We estimated all of our models using this sample. To minimize the effects of migration from state of birth, this paper presents results from the restricted sample of men aged 17 or 18 at the time of application.<sup>16</sup> The estimation results from these two samples are qualitatively the same (and available from the authors).

---

<sup>14</sup> The *Profile of American Youth* is a sample from the 1979 National Longitudinal Survey of Youth (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 1980 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 misnormed 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.

<sup>15</sup> Full sample contains male and female applicants between the ages of 16 and 28. The states included in each geographic region are the same as used in Figure 1.

<sup>16</sup> Discuss Census migration findings.

Columns (2a) to (2c) of Table 2 are based on the restricted sample. There are 1,977,118 white males and 725,480 black males, born between 1957 and 1973, who took the test at age 17 or 18 in the South, Rustbelt, or Border states. Due to their relative youth, the majority have not graduated high school at the time of application to the military; but the black applicants still have higher education levels than their white counterparts, on average. Even so, there is a racial gap in AFQT scores of 20-percentile points at the mean. The bottom rows show the percentages of the relevant populations who applied to the military and took the AFQT (calculations described below). Over 14 percent of all men in these birth cohorts took the AFQT at either age 17 or age 18; with black men applying to the military at a much higher rate than white men (21.8 percent v. 13.4 percent). Among 18 year-olds, military application rates for those with no more than two years of completed high school education are low for both races – that is, AFQT test taking rates are higher among men with more completed education.

The 1976 to 1991 AFQT testing data cover the key birth cohorts (and years) in which the NAEP test score gap narrowed; as well as the cohorts (and years) preceding the convergence. Unfortunately, military testing data are not available for the cohorts born after the convergence stopped. On the other hand, the large sample size allows us to compare the precise timing of test score convergence across regions (e.g., South v. Rustbelt), across parts of the South (e.g., Southern v. Border states), and across states within a region (e.g., Alabama and Mississippi v. Tennessee and Virginia).

A weakness of the AFQT data is that it only includes individuals who chose to apply to the military. This results in two main sources of selection in the sample. First, the group of applicants is not a representative sample of all U.S.-born 17 and 18 year olds. For example, military applications tend to be countercyclical, and blacks are more likely to apply than whites. To obtain unbiased estimates of black and white average test scores for a given cohort in a given year, we must therefore correct for this nonrandom selection.

Second, we observe applicants' residence at the time of application and not their place of birth. Since a goal of the analysis is to test for links between conditions in early infancy and outcomes in young adulthood, the ideal dataset would include the location of birth. As mentioned above, we restrict the sample to 17 and 18 year-olds, who are the most likely to still live in their state of birth. Table 2 shows

that the AFQT scores in this sample are similar to those in the larger sample of 17 to 20 year-olds.

Below, we find that the results are unaffected by direct controls for state migration rates.

## V. Distinguishing age, cohort and time, and correcting for selection in the AFQT sample

Here, we discuss our models for differentiating age, cohort and time effects (by race) in AFQT scores and for correcting for nonrandom selection in who applies to the U.S. military. Simply put, our approach is to control for as detailed a set of fixed effects as allowed by the data, and to correct for any remaining selection within cells by weighting the analyses by the inverse probability that an individual within a cell applies to the military.

### A. Regression models

The parameters of interest in this study are the racial differences in the birth cohort effects in average AFQT scores. We estimate models of the following form:

$$(1a) \quad T_{icat}^{(W)} = \gamma_c^{(W)} + \delta_t^{(W)} + \alpha_a^{(W)} + X_{icat}^{(W)}\beta^{(W)} + \varepsilon_{icat}^{(W)}$$

$$(1b) \quad T_{icat}^{(B)} = \gamma_c^{(B)} + \delta_t^{(B)} + \alpha_a^{(B)} + X_{icat}^{(B)}\beta^{(B)} + \varepsilon_{icat}^{(B)},$$

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. The outcome variable,  $T$ , is the test score,  $X$  is a vector of controls – which include unrestricted education indicators – and  $\varepsilon$  is an error term.

Our models allow each effect to vary by race,  $r \in (B, W)$ . So,  $\delta_t^r = (\delta_{1976}^r, \dots, \delta_{1991}^r)$  are the race-specific calendar year fixed effects,  $\alpha_a^r = \alpha_{18}^r$  is the race-specific age effect in the case where there are only 17 and 18 year olds in the sample, and  $\gamma_c^r = (\gamma_{1957}^r, \dots, \gamma_{1973}^r)$  are the race-specific birth-year dummies. Thus, the parameters of interest,  $(\gamma_c^{(B)} - \gamma_c^{(W)})$ , measure the black-white gap in average AFQT scores by birth year. Below, we also allow all of the race-specific effects to vary by region or state, since we are interested in geographic variation in the cohort-specific AFQT convergence.

As noted above, it is impossible to nonparametrically identify unrestricted age, cohort and time effects since they are perfectly collinear at a detailed enough level. For example, an individual's birthday and calendar year/day of exam fully characterize his exact age. However, the design of the military

testing data do allow us to identify *additive* age, cohort and time effects measured *in years* (and not days) since the test is administered on a rolling basis throughout a calendar year – something that is not possible in most survey data sets.

Table A1 in the Appendix, which presents age at exam by birth year and year the test is taken, illustrates this. For example, some 17 (18) year-olds in the 1960 birth cohort take the exam in 1977 (1978), while others take it in 1978 (1979).<sup>17</sup> Thus, birth year effects can still be estimated even after adjusting for additive fixed effects in age at and year of exam. They cannot be identified, however, if unrestricted (race-specific) age-year interactions are included in the regression model.

As a result, while our basic analysis controls for unrestricted race-age, race-year and race-education fixed effects, we restrict the race-age profile of test scores to be the same within a calendar year, but allow this profile to shift up and down year-to-year. This restriction allows us to separate the remaining variation in test score convergence into the amount that accrued to successive cohorts, and the amount that accrued to all individuals of all ages in particular calendar years.

Two points need to be made before proceeding. First, the detailed set of fixed effects in the regression models will absorb any nonrandom selection in individual military application that varies at the level of the (race-specific) fixed effects. The model controls for selection that varies by race in ways that are different for 17 and 18 year-olds and different in each calendar year; but the evolution over time in the selection of (black relative to white) 17 year-olds is not allowed to be different than that of 18 year-olds. If the selection of *black versus white* applicants changes over time in different ways for 17 and 18 year-olds, then the regressions will not remove all of the selection bias in test taking.

While we examine and correct for this possibility below, the patterns in Figure 2B suggest that this restriction may be plausible. Specifically, for these *random samples* of students, in which test taking is not selective, the NAEP test score gap does vary by age and time between 1971 and 1990; but in a way that is systematically related to birth year, as shown in Figure 2C. During the 1990's, however, the NAEP gap does not vary by age or time, and certainly not by age over time.

---

<sup>17</sup> In the AFQT data, roughly one-third of 17 (18) year-olds in the 1960 cohort take the exam in 1977 (1978), and two-thirds in 1978 (1979).

Second, in much of the analysis below, we compare geographic differences in the cohort-specific convergence in AFQT scores and link these to geographic variation in convergence in early health. In such comparisons, it is possible to include full race-age-year interactions as long as they are not allowed to vary by region or state. Here, the form of selection that would lead to biased results is much more complicated (and probably, much less plausible). For example, the regression models can control for more sources of selection than are implicitly adjusted for in the South-North comparisons in NAEP score gaps illustrated in Figure 2D. Nevertheless, we now describe how we attempt to correct for any remaining selection bias *within* these narrowly-defined cells.

### B. Inverse probability weighting

We observe AFQT scores for only those men who applied to the U.S. military. Allow  $T_{icat}^*$  to represent the AFQT score for a randomly selected 17 or 18 year-old man from the population. Then, the military sample contains information on:

$$(2a) \quad T_{icat} = I_{icat} \cdot T_{icat}^*$$

$$(2b) \quad I_{icat} = 1(I_{icat}^* > 0),$$

where  $I_{icat}$  is an indicator variable equal to one if the latent process governing the decision to apply,  $I_{icat}^*$ , is greater than zero (e.g., the benefits minus the costs of applying). In conventional terms, equation (2a) is the (selected) outcome equation, and equation (2b) is the selection equation.

Several sources of “sampling” bias are controlled for by the fixed effects included in regression models (1a) and (1b). To address potential selection across men within these narrow cells, we weight the regression models by the inverse of an estimate of the probability that different men within the cell took the test – also known as Inverse Probability Weighting (IPW). Define  $p(\cdot) = \Pr(I_{icat} = 1)$  to be the true likelihood that a given individual will take the AFQT, and  $\hat{p}(\cdot)$  to be an estimate of that likelihood. Then weighting the regression equations by the weight,  $w_i = 1/\hat{p}(\cdot)$ , will remove any remaining selection bias, as long as the observables used to estimate the probabilities account for all nonrandom selection within cells (see, e.g., Hirano, Imbens, and Ridder and Wooldridge 2002).

Thus, we estimate the probability that each observation, or group of observations, is selected into the AFQT sample, and then weight the analyses by the inverse of that probability. In most settings of this type, researchers must 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 our context, however, because we know the universe of military 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.

We estimate these denominators in three ways, one based on counts of births by cohort and state of birth from the natality files of the National Vital Statistics System, and two based on counts of residents by cohort and state of residence from the U.S. Census, one of which adjusts for variation in the distribution of completed education across states and over time. We describe each of these population estimates in detail in the Data Appendix at the end of the paper. Since the results do not vary by the choice of any of these three weights, we report primarily the results using the natality data (denoted *IPW\_natality* in the Data Appendix).

- Divide # AFQT takers in (state-race-birth cohort-age-year) cells by population size of cell.
- Denominators from: i) births (*Vital Statistics*); and ii) cell population around test year (*Censuses*). Same findings.
- Same as Probit/Logit with state-race-cohort-age-year dummies.
  
- Evaluate scores across cells at same probability (one) of taking exam (“identification at infinity”).
- “Remove” selection bias across cells.
- Varies along full interaction of cohort-age-time – sweeps out added bias over above equations.

In figures 3a-d, we show the patterns in the probability of being in the sample (i.e. the selection probability) over time, for various comparison groups. As we describe just below, our data include the universe of applicants to the military, so estimating the selection probability only requires estimating the population size. As can be seen in figure 3a, blacks are more likely to apply to the military and black application rates are typically more countercyclical than for whites. However, the black-white gap in application rates appears to have followed a very similar pattern in the South and Rustbelt.<sup>18</sup> The figure also shows a sharp drop in applications by blacks in 1982 that was driven by the renorming of the AFQT

---

<sup>18</sup> 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.

(see e.g. Angrist 1998 for a discussion). The drop in applications appeared to affect 17 and 18 year old blacks similarly, and as can be seen in figure 3b the drop appears very similar in the South and Rustbelt.

The fact that selection probabilities changed similarly for 17 and 18 year olds means that this episode was a year effect rather than a cohort effect, and our estimates therefore account for such changes by conditioning on year fixed effects. We also note that the racial gap in the probability of taking the test rebounded during the mid-1980s suggesting that the relative improvement in tests scores for blacks during the 1980s was not due to a sustained decline in the relative rate of test-taking by blacks.

Figure 3c shows the trend in the black-white selection probability gap for less-educated men. Though there was a drop in application rates between 1980 and 1982, this change appears to have been about the same in the South and Rustbelt. Since our estimates examine changes within region, the effect of this selection will effectively be differenced out.

Finally, figure 3d shows the relative changes in black-white selection probabilities between Alabama and Mississippi and two other pairs of states, Tennessee and Virginia and Illinois and New York. This comparison is motivated by a comparison we make below between cohort effects in AFQT scores and PNMR across these pairs of states. There does not appear to be a big difference in the path of selection between Alabama and Mississippi and Tennessee and Virginia, suggesting that comparisons across these pairs of states within the South are not driven by differential selection. The overall trend is also similar when compared with Illinois and New York, though there does appear to be a small dip and then increase in selection of higher educated blacks in Alabama and Mississippi compared with Illinois and New York.

The most serious threat to the validity of the estimates is selection on unobservables. To be a problem, differential patterns of high and low-ability blacks in selection into military application would have to occur even conditional on the large set of fixed effects we include. However, it is further reassuring that trends in application rates, shown in appendix figure A1, are very similar by completed education level in both the South and Rustbelt.

## VI. Cohort-based Patterns in AFQT Scores

- Estimate separately by region (South, Rustbelt, Border) or state group (ALMS, TNVA, ILNY).
- Plot  $(\gamma_c^{(B)} - \gamma_c^{(W)})^S$ ,  $S = \text{state/region}$
- Not sensitive to how we do the regional groupings – e.g., including Texas in South and Missouri in Border states.

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.

We show that the patterns in black-white cohort effects are stable after controlling for various interactions of year, race, region and education fixed effects. Thus, for selection to be driving the results we find, it must be that selection is different, for example, across region and race (i.e. for blacks in the South, but not for whites in the South or for blacks in the Rustbelt).

In figure 4b, we present estimates of  $\gamma_c^B - \gamma_c^W$ , the black-white difference in AFQT scores, for birth cohorts 1957 to 1973, separately for the South, “Border” states, and the Rustbelt. The cohort effects

are estimated conditional on year effects, age effects and completed education effects, all fully interacted with race.

As can be seen in panel b, the estimated black-white test score gap remained fairly constant in the South at about 21 to 23 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 1963 cohort, black scores began to rise sharply relative to white scores. Black AFQT scores converged towards white scores at a fairly constant rate for at least the next 10 cohorts. By the 1973 birth cohort, the 21-point gap had shrunk to about 10 percentile points, more than a 50 percent decline. A similar, though muted, pattern is seen in the Border states. The initial black-white AFQT gap was smaller in Border states, and the improvement was slightly smaller.

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 and much more gradual.

Figure 4d plots the South-Rustbelt and Border-Rustbelt differences in the black-white gap, along with the South-Rustbelt difference in white AFQT scores by cohort. Among whites, trends over cohorts in AFQT scores were quite similar in the South and Rustbelt. There was some variation year to year, but the difference remained between 0.5 and 2 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 began to improve sharply. This convergence continued for the next 10 cohorts until for those born in the late 1960's, the black-white test score gap was actually larger in the North than the South.

## **VII. An Explanation: Relative Improvements in Black Infant Health**

Put “Heckman” model here.

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 the beginning of 1966 that improves the health of blacks between the ages of zero and 2 years. If the improvement in early health has long-term effects on measured cognitive skills, we should expect to see improvements in AFQT scores for blacks born in 1964 and 1965 relative to those born in 1961 and 1962. If a similar intervention affected the health of children up to 3 years of age, we would then expect to see improvements in AFQT scores for those born after 1962.

#### *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 4a. This figure shows three series, the black-white gap in post neonatal mortality rates (PNMR) by year in the South, Border states, 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 because 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 we hypothesize that those factors do not appear to have caused the specific convergence in black-white skills that we documented in the previous section. Recall, that the vast majority of the improvement in the black-white infant mortality rate gap in the South was due to improvements in PNMR rather than NMR. We test later whether PNMR is more closely associated with AFQT scores than neonatal mortality (NMR) and low birthweight, each of which might be more directly related to conditions *in utero*.

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 decreases 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.

White PNMR was declining fairly gradually during this period in both the South and Rustbelt. As can be seen in figure 4a, 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 1958 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 1975. The late-50's racial PNMR gap in Border states was about two-thirds as large as in the South and there was a comparably smaller, though significant decline. In contrast with the South, the timing of the beginning of the decline is harder to date based on the figure.

The analogous series for AFQT scores are plotted in figure 4b and the regression coefficients underlying this table can be found in appendix table A1. 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 1963, Southern blacks began to improve relative to Southern whites. This relative improvement continued for the next ten birth cohorts.

These two mirroring patterns can be seen clearly by comparing figures 4a and 4b. The decline in black PNMR is matched by an increase in AFQT scores when those cohorts reach 17 and 18 years of age. The largest declines in PNMR and increases in AFQT are seen in the South, and the slight improvement in black PNMR in the Rustbelt is matched by an increase in black AFQT scores in the Rustbelt in the late 1960's and early 1970's cohorts.

This mirror relationship can be seen perhaps even more clearly in figures 4c and 4d, where another level of differencing is introduced. Here, the South-Rustbelt and Border-Rustbelt differences in black-white gaps are shown along with the South-Rustbelt gap for whites. Again, 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 1963. By 1972, the South-Rustbelt gap in PNMR was virtually erased. The more gradual and smaller decrease in the black-white PNMR gap in the Border states relative to the Rustbelt is matched by a more gradual increase in the relative black-white AFQT gap. And, as a comparison, there is no change in either PNMR or AFQT among whites in the South relative to the Rustbelt. The black South-Rustbelt difference (not shown) for both PNMR and AFQT are remarkably similar to the black-white South-Rustbelt difference-in-difference. This implies that the vast majority of the cross-region relative convergence in both PNMR and AFQT was driven by improvements among blacks. Taken as a whole, these patterns, estimated non-parametrically, are strongly suggestive that the mechanism described by the infant health hypothesis is at work in explaining the convergence during this period in the black-white AFQT gap.

Why however does the convergence in the black PNMR series begin at least a year before the convergence in black AFQT scores? Recall that we take PNMR to be a proxy for early 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.<sup>19</sup> We investigate one possible such intervention, the integration of Southern hospitals, in section VIII.

Figure xx shows a complementary interesting fact. Here we plot the year effects from the same regressions estimated to produce Figures 4a, along with the unconditional trend in black and white AFQT scores by year in the full sample. The open squares show the unconditional trend, an increase in black relative to white scores that was observed in the 1980's. The solid squares show that after conditioning on birth-year effects, this black-white test score gap narrowing during the 1980's is significantly less

---

<sup>19</sup> In addition, the fact that PNMR is measured by year of death rather than year of birth suggests that some of the deaths included in year  $t$  are of babies born in year  $t-1$ .

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 4. 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 17-18 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 specifically, 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.

Black-white time-effects in the South disappear after adjusting for cohort effects.

#### *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 the South across 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 three pairs that had similar patterns of black-white PNMR convergence: Alabama and Mississippi, South Carolina and North Carolina, and Tennessee and Virginia.<sup>20</sup> The patterns in black-white PNMR convergence can be seen for each of these three pairs of states in Figure 5a. The solid 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 began to decline sharply and by 1974 the gap narrowed to about 5. Compare this pattern with that of Mississippi and Alabama (solid squares). The racial PNMR gap tracked that of the Carolinas, though at a slightly lower rate until 1963. However, the gap did not converge in AL-MS in 1964. Instead it remained constant for another two years, until it began 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.

In figure 5b, we compare these cross-state patterns in racial PNMR gaps to those for AFQT scores. The correspondence is striking. 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 began to converge to whites while the convergence appears to have begun in 1963 in TN-VA. Remarkably, just as for PNMR, the racial convergence in AL-MS began two years after than in NC-SC. The AFQT gap appears to have narrowed about one year earlier in TN-VA.

Alabama and Mississippi experienced a quick and dramatic improvement in black PNMR beginning suddenly in 1966. The abruptness of this change makes these states an ideal pair to examine relative to other states both within and outside the South. With this motivation, figures 6c-d focus on the experience of AL-MS relative to TN-VA and a pair of northern states with large black populations and similar patterns in black-white PNMR trends, Illinois and New York. The figures show that black PNMR declined in AL-MS relative to TN-VA from 1966 to 1968. AL-MS also experienced an improvement in

---

<sup>20</sup> We leave out some states to keep the number of series plotted on the figure a readable amount.

black AFQT scores for the cohorts born two years earlier, from 1964 to 1966. This comparison within the South rules out any within-region trends in omitted variables.

The comparison of AL-MS to IL-NY shows a similar decline in the black-white PNMR gap beginning in 1966, but this improvement continues more consistently through the mid 1970's. Correspondingly, the improvement in black AFQT scores in AL-MS improved two years earlier with the 1964 cohort and this relative improvement continued on through the cohorts born in the early 1970's. Interestingly, even the initial level differences in black-white PNMR and AFQT gaps fit the hypothesis. In the very early 1960's the black-white PNMR and AFQT gaps were larger in absolute value in TN-VA than in IL-NY.

### *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 3-5 we present regression estimates of these magnitudes. Table 3 begins with the South-Rustbelt comparison. Column 1a shows the black-white AFQT gap 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 black-white AFQT gap in each region, along with the double-difference, between the 1960-62 cohorts and the 1970-72 cohorts. The South-Rustbelt, black-white difference in difference is thus found in row C of column 1b. It shows that relative to the Rustbelt, the black AFQT scores in the South rose relative to whites by 7.60 percentile points, while the relative black-white PNMR gap fell by 6.78 births per 1,000. The right-hand panel (columns 2a-b) replicates this analysis while allowing education effects to vary by race. The relative AFQT improvement is slightly smaller when racial differences in test-score-returns-to education are assumed to be controls rather than outcomes suggesting that the improvement in infant health increased the test-score return to education.

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 6 includes a comprehensive set of interactions: region-race-cohort, region-race-age, region-race-year, region-education, race-education, region-race-education, age-time, region-age-time, race-age-time, education-time, region-education-time, race-education-time. This model yields an estimated improvement of 7.13 percentile points, which is significant at the 1-percent level and about the same as the uncontrolled estimate.

The set of interacted fixed effects is about as comprehensive as possible. It would not be possible to include region-race-age-time effects because this is the level of the variation of the cohort effects. The various included fixed effects restrict one or more of these dimensions. It is reassuring that the estimated change in AFQT across cohorts is not appreciably affected by conditioning on these various combinations of interacted fixed effects. Any alternative omitted variables explanation of these results, including a selection-based interpretation would have to operate within each of the cells enumerated by the various fixed effects.

In Table 5, we present regression estimates of comparing AL-MS to TN-VA and IL-NY. Because the drop in black PNMR was so sharp in Alabama and Mississippi, we make comparisons over the short window where this improvement occurred, between the 1961-63 cohorts and the 1969-71 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 3.13 percentile points to 3.54 percentile points (t-ratios are shown in square brackets). This change corresponded to a relative change of 2.02 per 1,000 in black-white PNMR. Comparisons of AL-MS to IL-NY yield estimated improvements of 5.59 to 6.85 percentile points. Notably, the size of the relative racial PNMR improvement (5.25 per 1,000) is correspondingly larger.

#### *D. PNMR versus earlier measures of infant health*

The infant health hypothesis rules out a number of alternative causes of the observed improvement in black AFQT scores. Many potential alternative causes can be ruled out by the way the AFQT convergence lines up by cohort rather than year effects; we discuss some of the prominent alternatives below in section IX. Neither the cohort-based pattern nor the correlation with PNMR, however, rule out the long-run effect of improvements at even earlier ages, such as improvements in *in utero* conditions. In addition to PNMR, table 5 shows the difference in difference in the black-white gap in two other such markers of early life health status, the neonatal mortality rate (NMR, the number of deaths in the first month after birth, per 1,000) and the low birth weight rate. Whereas the relative change in the PNMR is significant and in the predicted direction, the relative change across states in the racial gap in NMR and low birth weight go in the perverse direction. In table 7, we further explore the possibility that the cohort-based AFQT convergence was correlated with these other measures of health as well as PNMR.

The regressions reported in table 7 are the second stage of a two-step procedure. In the first step, we estimate the basic AFQT specification separately for each of the 22 states in the analysis. The race-by-cohort effects, which are the conditional black-white AFQT gap by birth year in the state, are then taken to be the dependent variable for the second step. The results reported in table 7 come from a regression with 308 observations (22 states, 14 years) that is weighted by the inverse of the variance of the estimated cohort effects from the first step.

For each of the three health measures, these cohort effects are regressed against the black-white gap in the contemporaneous value of the respective health measure and four leads of that variable. In the first column, state fixed effects are included. The second column adds cohort fixed effects, and the third instead adds state-specific cohort trends. Recall that the dependent variable is already estimated from a regression that includes race-state-year and race-state-education fixed effects. The controls in the second step are included to partial out variation in the right-hand-side health measures.

The individual coefficients should be interpreted with care because of serial correlation in the health measures, but a few patterns are worth noting. First, the one-to-two year lead of the PNMR series relative to the AFQT series that was apparent in the cross-state comparisons in figure 5 is apparent in

columns 1a-c. In each of the three specifications, the two-year lead of PNMR has the largest point estimate and *t*-statistic among the five PNMR measures. In each case, the one-year lead has the second strongest correlation. To help think about interpretation, a marginal effect of the 1-year lead of PNMR is consistent with a shock to health that affects 0-24 month-olds, and then has lasting effects on AFQT. Second, PNMR appears to explain a large portion of the variation across states and time in the black-white AFQT gap. For each specification, the partial *R*-squared of the health measured is reported. This is the fraction of the variation that is explained by the health measures after conditioning on the fixed effects included in the model. For example, racial PNMR gaps can account for 77.7 percent of the variation in the black-white AFQT gap, after controlling for state fixed effects. Some of this explanatory power can be attributed to cohort trends in PNMR that affected all states similarly. It is not clear that this is variation that should be controlled for, but by conservatively doing so we estimate that more than 40 percent of the black-white test score convergence can be accounted for by relative changes across states in the black-white PNMR gap.

Columns 2a-2c and 3a-3c report the same specifications for NMR and low birth weight, respectively. In contrast with the PNMR results, the estimates for these other health measures are sensitive to the specification. Without cohort effects, trends in the black-white NMR explain 22.5 percent of the variation in black-white AFQT convergence. In this specification (column 2a), the two year lead is the most highly correlated. However, the inclusion of cohort fixed effects or state-specific linear cohort trends reduces the partial *R*-squared to less than 9 percent. In the specification with cohort fixed effects, only the three- and four-year leads of NMR are statistically significant. The low birth weight regressions show a different pattern. When cohort effects are not included, the partial *R*-squared of low birth weight is less than five percent. Including cohort effects raises this to 26.8 percent, but the significant coefficients are of the perverse sign, implying that increases in low birth weight among blacks leads to an improvement in black AFQT scores. Including state-specific linear trends switches the sign of the birth weight effects and raises the partial *R*-squared slightly higher. In this specification, again it is the two-year lead that has the strongest correlation.

In table 8, we include all three early measures of health in a single regression to see which appears most strongly correlated with the AFQT cohort effects. Motivated by the results in table 5, we include only the 1- and 2-year leads of the health markers, but now we include all three in the same regression. The specification in column 1 has no controls, and is included so that we can report an estimate of the constant. This result implies that if the racial gap in each of the three health markers were set to zero, the black-white AFQT gap would fall to 7.19 percentile points. We treat this estimate with care, however, because it is not conditional on state or cohort fixed effects. The specifications in columns 1-3 correspond to those reported in table 5.

The results tell a clear story. In each specification, the 1- and 2-year leads of the PNMR gap are strongly significant and of the predicted sign, with  $t$ -ratios ranging from 3.39 to 9.34 in absolute value. In contrast, neither the 1- nor 2-year leads of the NMR gap are significant in any of the three specifications once we control for the PNMR gap. The 1- and 2-year leads of the low birth weight gap are insignificant in the specifications in columns 2 and 3, but become significant once we control for state-specific linear cohort trends rather than cohort fixed effects. Interestingly, the effects of the PNMR gaps are not affected by this change in controls.

The comparison of PNMR and NMR can also be seen clearly in figures 6a-c. For each state in the analysis, the  $y$ -axis of figure 6a measures the change in black-white AFQT cohort effects between the 1961-63 cohorts and the 1967-69 cohorts. The  $x$ -axis measures the change in the black-white PNMR gap between 1962-64 and 1968-70. Southern states are labeled with solid black diamonds, Border states with blue X's, and Rustbelt states with open circles. The figure shows the strong negative relationship that was seen in figure 4, for example, and in the regression results in table 2. With 22 observations, the regression yields a  $t$ -statistic of 3.91 and an  $R$ -squared of 0.520. Notice also that the regression line crosses the  $y$ -axis near the origin. If one were to interpret the regression causally, the constant estimate of 1.37 would imply that in the absence of any change in PNMR, the black AFQT scores would only have increased by 1.37 percentile points relative to whites between the early and late 1960's birth cohorts.

Figure 6b shows a similar figure for NMR. The relationship between changes in the racial NMR gap and changes in the black-white AFQT gap is far less clear. The OLS estimate is insignificant, the  $R$ -

squared is 0.085 and the constant is estimated to be 4.27 with a very large standard error. Figure 6c plots the residual changes in AFQT and PNMR gaps after partialling out the effects of NMR and low birth weight. As in the unconditional plot in figure 6a, the conditional relationship appears strong. The *t*-stat remains above 3 and the estimate of the constant is near zero. In future analyses, we will include additional controls for changes migration patterns and social program spending and participation by state. However, as we discuss below a cursory examination in the patterns in many of these variables suggests the patterns across states are not likely to explain the changes in AFQT we see.

Finally, in figure 6d we show a plot that is similar to figure 6a, but where we break out the change in AFQT scores across the distribution. In particular, we show the change in the 75<sup>th</sup> percentile, median and 25<sup>th</sup> percentile. An interesting pattern emerges. The strongest relationship is between black PNMR improvements and improvements in the 75<sup>th</sup> percentile of black scores. The racial gap in median AFQT scores shows the next strongest relationship, and the gap in 25<sup>th</sup> percentile scores shows the weakest correspondence.

### **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 17-18 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. A possible interpretation is that PNMR is as a marker for access to hospital care in the first years of life. 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. If PNMR is a marker for access to hospitals, we would conclude that access to treatments available in hospitals in the first years of life – and not necessarily treatments for diarrhea and pneumonia in the post-neonatal period, *per se* – are the source of the later AFQT gains.

Recall that the strongest relationship between PNMR and AFQT was at a 1- and 2-year difference. These results, which are also seen in the 1- and 2-year differences in the timing of the beginning of PNMR and AFQT convergence, imply that the convergence in black-white AFQT scores was caused by access to hospital care in the first few years of life. If access to hospital care at later ages had long-term effects in AFQT scores, longer leads of PNMR would correlate strongly with AFQT.

If we make this assumption – that the correspondence between PNMR and AFQT is driven by access to hospital care – an interesting set of questions are raised. Do investments in health early in life have larger long-run returns to cognitive skills than health investments later in childhood? Or, was the improvement in black AFQT scores only caused by improvements in access to early healthcare because there was no simultaneous improvement in healthcare at older ages? To address this question, we examine data from various waves of the *National Health Interview Survey*. This survey asks respondents

whether they were admitted to a hospital during the past year.<sup>21</sup> Figure 8 shows the results from this survey for Southern black and white children of different ages. Panel A shows the Southern black-white gap in hospital admissions by age for three different time periods: July 1962 to June 1964, July 1965 to June 1967, and January 1971 to December 1972. The data show that the black-white gap ranged from 2 to 6 percent in the 1962-64 period with the largest gaps for those 1-4 years old. By the mid-1960's, however, the gap among those 4-years-old and younger began to close. By the early 1970's the black-white gap was virtually erased in the South among 1 to 4 year olds, and blacks less than 1 year old were actually more likely to be admitted to the hospital. Panel B shows the growth in hospital admissions by age. There were large increases in black hospital admissions for children up to the age of 4.

South (DE, MD, DC, VA, WV, NC, SC, GA, FL, KY, TX, TN, AL, MS, AR, LA, OK), Northeast (ME, NH, VT, MA, RI, CT, NY, NJ, PA), North Central (MI, OH, IN, IL, WI, MN, IA, MO, ND, SD, KS, NE)

If we interpret these changes in admissions as changes in access, then the AFQT/PNMR results discussed above imply larger long-term cognitive returns to hospital access at ages 0 to 2 than at ages 3 and 4. Essentially, there is evidence of improved access to hospital care for black 3 and 4 year olds, but the patterns in PNMR and AFQT imply that access at these ages did not have large effects on AFQT scores at ages 17 and 18.

Our data do not allow us to pin down the mechanism behind these age-dependent returns to healthcare. For example, it could be that the differential returns stem from an increased risk at very young ages of catching diseases that have long-term cognitive effects. Alternatively, it could be that brain development is more sensitive to the treatment of health conditions at ages 1 and 2 than ages 3 and 4. Distinguishing these two, and possibly other, hypotheses is important and should be pursued, but we leave it for separate investigation.

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

---

<sup>21</sup> Until 1968, the past year refers to the past fiscal year (July to June). After 1968, it refers to the past calendar year.

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 18 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.<sup>22</sup> 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 not 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

---

<sup>22</sup> Medicaid adoption in Rustbelt: IL (Jan. 66), NY (Oct. 66), PA (Jan. 66), OH (July 66), MI (Oct. 66), MO (Oct. 67), IN (Jan. 70). In South: AL (Jan. 70), MS (Jan. 70), NC (Jan. 70), AR (Jan. 70), FL (Jan. 70), VA (July 69), TN (Jan. 69), SC (July 68), GA (Oct. 67), LA (July 66). In Border states: KY (July 66), WV (July 66), MD (July 66), DE (Oct. 66), TX (Sept. 67).

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 compiled 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).<sup>23</sup>

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.

---

<sup>23</sup> 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.

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 find that 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.

The data allow us to estimate exposure rates for cohorts born between 1950 and 1978. 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 exhibit patterns in integration very similar to the unconditional ones. 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.

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.<sup>24</sup>

---

<sup>24</sup> 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.

## **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 18 years later, remarkably well. Comparisons across region (South versus North), and within region (Southern v. Border States, AL-MS v. NC-SC, AL-MS v. TN-VA) 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 that access to hospitals increased for

blacks children ages 0-4 in the South in the mid-1960's and early 1970's, implying that access to hospital care has larger long-term cognitive benefits for 0-2 year olds than for 3-4 year olds.

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.

## Data Appendix

We use three different sets of IPW weights, each based on a different estimate of the population size for each cohort. Here we describe each of those three weights:

*IPW\_natality*: 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  $\times$  state of residence  $\times$  year  $\times$  birth year  $\times$  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  $\times$  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  $\times$  state of residence  $\times$  year  $\times$  birth year is then multiplied by this probability to obtain an estimate of cohort size that varies by race  $\times$  state of residence  $\times$  year  $\times$  birth year  $\times$  completed education at time of application.

Table 1: Change between birth cohorts in black-white NAEP score gap of 17-year olds, South and North

	Black-white difference in NAEP scores (in standard deviations)					
	Reading scores by birth cohort (1971, 1980, 1990 surveys)				Math scores by birth cohort (1978, 1990 surveys)	
	Early 50s and 60s cohorts		Early 60s and 70s cohorts		Early 60s and 70s cohorts	
	Average in 1953-1954	Change by 1962-1963	Average in 1962-1963	Change by 1972-1973	Average in 1961	Change by 1972-1973
(1a)	(1b)	(2a)	(2b)	(3a)	(3b)	
<i>A. South</i>						
Black-white NAEP gap	-1.300*** (0.031)	-0.222*** (0.052)	-1.522*** (0.042)	0.828*** (0.084)	-1.281*** (0.030)	0.698*** (0.076)
Sample Size	9,966		5,020		7,164	
<i>B. North</i>						
Black-white NAEP gap	-1.201*** (0.035)	-0.086* (0.048)	-1.287*** (0.033)	0.460*** (0.073)	-1.154*** (0.030)	0.293*** (0.072)
Sample Size	20,762		11,122		16,573	
<i>C. South – North</i>						
Black-white NAEP gap	-0.099** (0.047)	-0.136* (0.071)	-0.235*** (0.053)	0.368*** (0.111)	-0.127*** (0.042)	0.405*** (0.104)
Sample Size	30,728		16,142		23,737	

*Notes:* The South consists of Alabama, Arkansas, Florida, Georgia, Kentucky, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee, Virginia (outside of Northern Virginia), and West Virginia. The North consists of the Northeast (Connecticut, Delaware, District of Columbia, Maine, Maryland, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, Vermont, Northern Virginia) and North Central (Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, Wisconsin) regions. Test scores have been normalized by their standard deviations by survey year, subject and age. Regressions are weighted by the NAEP sampling weights. The estimated standard errors are in (parentheses) and are corrected for heteroskedasticity.

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

Table 2: Summary statistics for AFQT sample

	Men born between 1957 and 1973 who took AFQT between 1976 and 1991					
	Entry ages 17 to 20			Entry ages 17 and 18		
	Total	Black	White	Total	Black	White
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)
<i>Percentile score on AFQT</i>						
Mean, unweighted	45.7	30.6	51.7	46.0	31.4	51.4
[standard deviation]	[24.3]	[19.2]	[23.5]	[23.8]	[19.2]	[23.1]
Mean, IPW (weighted)	48.0	31.4	51.7	47.8	32.7	51.2
[standard deviation]	[24.5]	[19.5]	[24.0]	[23.7]	[19.4]	[23.2]
1 <sup>st</sup> percentile (IPW)	3	1	5	4	2	5
25 <sup>th</sup> percentile (IPW)	29	16	33	29	17	33
Median (IPW)	47	29	51	47	30	50
75 <sup>th</sup> percentile (IPW)	67	43	71	66	44	70
99 <sup>th</sup> percentile (IPW)	97	85	98	97	85	97
Mean AFQT score						
South (IPW)	46.3	30.0	52.0	45.9	31.5	51.0
{unweighted}	{42.8}	{29.4}	{52.0}	{43.0}	{30.5}	{51.4}
Border states (IPW)	47.0	31.4	49.8	46.8	32.6	49.4
{unweighted}	{45.6}	{30.8}	{50.0}	{45.6}	{31.4}	{49.6}
Rustbelt (IPW)	49.6	33.3	52.1	49.4	34.2	51.9
{unweighted}	{47.8}	{32.3}	{52.0}	{48.3}	{32.8}	{52.0}
<i>Age distribution (percent)</i>						
17 years old	32.4	27.6	34.4	48.9	44.0	50.7
18 years old	33.9	35.2	33.4	51.1	56.0	49.3
19 years old	21.2	23.5	20.3			
20 years old	12.4	13.6	11.9			
<i>Education distribution (percent)</i>						
1 year or less of HS	3.8	2.1	4.4	4.4	2.5	5.1
2 years of HS	8.5	7.6	8.9	10.0	8.8	10.4
3-4 years of HS	42.0	42.5	41.8	55.0	55.1	54.9
GED	3.4	2.5	3.7	2.5	1.8	2.8
High school graduate	40.5	43.6	39.2	27.8	31.5	26.4
1+ year college	1.9	1.7	2.0	0.3	0.3	0.3
<i>Percent of population who take AFQT</i>						
Age 17				7.10	9.63	6.94
Age 18				6.90	12.18	6.48
Age 18, $\leq 2$ yrs of HS				4.90	6.04	4.77
Number of observations	4,071,283	1,154,348	2,916,935	2,702,598	725,480	1,977,118

Notes: Data come from the universe of men who were born between 1957 and 1973 and took the AFQT between 1976 and 1991 in the South, Rustbelt and Border states. The South consists of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia; the Rustbelt of Illinois, Indiana, Michigan, Missouri, New York, Ohio and Pennsylvania; the Border states of Delaware, Kentucky, Maryland, Texas, and West Virginia. 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. The weighted summary statistics for AFQT scores are based on the inverse of these probabilities (inverse probability weighting – IPW).

Table 3: Change in black-white AFQT gap between 1960-1962 and 1970-1972 birth cohorts, South and Rustbelt

	Black-white difference in AFQT scores			
	Education fixed effects		Race-education fixed effects	
	Average in 1960-1962 (1a)	Change by 1970-1972 (1b)	Average in 1960-1962 (2a)	Change by 1970-1972 (2b)
<i>A. South</i>				
Black-white AFQT gap	-25.76 <sup>***</sup> (0.82)	12.69 <sup>***</sup> (0.79)	-23.46 <sup>***</sup> (0.73)	9.08 <sup>***</sup> (0.62)
{PNMR gap}			{14.05}	{-8.27}
<i>B. Rustbelt</i>				
Black-white AFQT gap	-21.01 <sup>***</sup> (0.88)	5.10 <sup>***</sup> (0.86)	-18.99 <sup>***</sup> (0.75)	2.01 <sup>**</sup> (0.66)
{PNMR gap}			{5.95}	{-1.49}
<i>C. South – Rustbelt</i>				
Black-white AFQT gap	-4.75 <sup>***</sup> (1.17)	7.60 <sup>***</sup> (1.13)	-4.47 <sup>***</sup> (1.01)	7.06 <sup>***</sup> (0.88)
{PNMR gap}			{8.10}	{-6.78}
Region-race-cohort	Y	Y	Y	Y
Region-race-time	Y	Y	Y	Y
Region-race-age	Y	Y	Y	Y
Region-education	Y	Y	Y	Y
Region-race-education			Y	Y

Notes: Sample contains all black and white men born between 1957 and 1973, who took the AFQT test between 1976 and 1991 in the South and Rustbelt, with entry ages of 17 or 18. The sample sizes are 934,296 in the South; 1,346,036 in the Rustbelt; and 2,280,332 in the pooled regression of South and Rustbelt states. All analyses include unrestricted race-by-birth cohort, race-by-time, and race-by-age fixed effects – interacted with region. Columns (1a) and (1b) include unrestricted education-by-region fixed effects; columns (2a) and (2b) include interactions of the education-by-region effects with race. 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. Black-white gaps in post-neonatal mortality rates (per 1,000 births) 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: South-Rustbelt difference in changes in black-white AFQT gap,  
1960-1962 and 1970-1972 birth cohorts

	South-Rustbelt difference in black-white AFQT gap					
	(1)	(2)	(3)	(4)	(5)	(6)
1960 to 1962 average	-4.75 <sup>***</sup> (1.17)	-4.59 <sup>***</sup> (1.04)	-4.47 <sup>***</sup> (1.01)	-3.91 <sup>***</sup> (1.18)	-3.46 <sup>***</sup> (1.05)	---
1960-1962 to 1970-1972 Change	7.60 <sup>***</sup> (1.13)	7.04 <sup>***</sup> (0.81)	7.06 <sup>***</sup> (0.88)	6.36 <sup>***</sup> (1.18)	5.62 <sup>***</sup> (0.92)	7.13 <sup>***</sup> (1.22)
Region-race-cohort	Y	Y	Y	Y	Y	Y
Region-race-time	Y	Y	Y	Y	Y	Y
Region-race-age	Y	Y	Y	Y	Y	Y
Region-education	Y	Y	Y	Y		Y
Race-education		Y	Y		Y	Y
Region-race-education			Y			Y
Age-time				Y	Y	Y
Region-age-time				Y		Y
Race-age-time					Y	Y
Education-time				Y	Y	Y
Region-education-time				Y		Y
Race-education-time					Y	Y
Region-race-educ-time						Y

Notes: See notes to Table 3.

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

Table 5: Comparison of between-cohort change in AFQT gap in Alabama and Mississippi with other states (1961-1963 and 1969-1971 birth cohorts)

	Comparison of black-white AFQT gaps in Alabama-Mississippi and					
	Illinois-New York			Tennessee-Virginia		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)
1961-1963 to 1969-1971 Change in AFQT gap	6.55 [9.94]	6.85 [5.80]	5.59 [5.13]	3.54 [10.33]	3.22 [2.85]	3.13 [3.16]
Change in black-white infant health gap						
PNMR (per 1,000)		-5.25			-2.02	
NMR (per 1,000)		1.80			-0.69	
LBW (per 100)		1.13			0.29	
State-race-cohort	Y	Y	Y	Y	Y	Y
State-race-time	Y	Y	Y	Y	Y	Y
State-race-age	Y	Y	Y	Y	Y	Y
Education fixed effects	Y	Y	Y	Y	Y	Y
State-education		Y	Y		Y	Y
Race-education		Y	Y		Y	Y
State-race-education		Y	Y		Y	Y
Age-time			Y			Y
Race-age-time			Y			Y
Education-time			Y			Y
Race-education-time			Y			Y
Sample size	591,646	591,646	591,646	304,469	304,469	304,469

Notes: Absolute values of t-ratios are in [square brackets] and are corrected for heteroskedasticity and unrestricted clustering at the state-level. The changes in the black-white infant health gaps are for the years 1962-1964 to 1970-1972.

Table 6: Across-state association of racial convergence from early to late 1960s birth cohorts in AFQT scores and infant health measures

	Racial convergence in AFQT score between 1961-1963 and 1967-1969 birth cohorts							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<u>Between cohort difference in racial gap in</u>								
PNMR (per 1,000)	-0.720 <sup>***</sup> [3.91]		-0.690 <sup>***</sup> [3.62]	-0.690 <sup>***</sup> [3.79]	-0.690 <sup>***</sup> [3.93]	-0.709 <sup>***</sup> [3.71]	-0.741 <sup>***</sup> [4.69]	
Birth in hospital with doctor present (per 100)								0.257 <sup>***</sup> [4.90]
NMR (per 1,000)		0.358 [1.37]	0.200 [0.95]			0.172 [0.68]	0.190 [0.64]	-0.048 [0.15]
LBW (per 100)							-0.139 [0.20]	-0.113 [0.16]
Migrate out of state (percent)				0.200 <sup>**</sup> [2.28]		-0.012 [0.06]	-0.030 [0.12]	-0.058 [0.26]
Mother HS dropout (percent)					-0.257 <sup>*</sup> [1.88]	-0.126 [0.53]	-0.132 [0.51]	-0.201 [0.92]
<u>Racial gap in Head Start spending per 4 year-old in</u>								
1968 (1/100)							0.051 [0.59]	0.010 [0.13]
1972 (1/100)						-0.235 [1.67]	-0.303 [1.26]	-0.363 <sup>*</sup> [2.05]
Constant	1.37 <sup>**</sup> [2.57]	4.27 <sup>***</sup> [7.32]					1.31 [0.80]	1.27 [1.02]
R-squared	0.520	0.085	0.546	0.567	0.571	0.664	0.669	0.689
Adj. R-squared	0.496	0.039	0.498	0.522	0.526	0.560	0.503	0.534
Number of states	22	22	22	22	22	22	22	22

Notes: Absolute values of t-ratios are in [square brackets] and are corrected for heteroskedasticity. See text for details on the construction of the variables and regressions.  
<sup>\*\*\*</sup> significant at 1-percent level, <sup>\*\*</sup> significant at 5-percent level, <sup>\*</sup> significant at 10-percent level

Table 7: State-level association between black-white AFQT and infant health gaps,  
1959 to 1972 birth cohorts

	Association between black-white differences in cohort AFQT scores and in infant health proxies								
	Post-neonatal mortality rate			Neonatal mortality rate			Low birth weight rate		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)
Lead, 4 years	-0.261 [3.09]	-0.154 [2.73]	-0.171 [2.35]	-0.077 [1.10]	-0.156 [1.74]	0.187 [3.10]	-0.205 [0.67]	0.136 [0.35]	0.740 [3.72]
Lead, 3 years	<b>-0.303</b> <b>[5.20]</b>	-0.158 [3.05]	-0.120 [2.19]	-0.113 [1.56]	-0.183 [2.77]	0.134 [2.22]	<b>-0.506</b> <b>[1.96]</b>	<b>-0.348</b> <b>[0.94]</b>	0.449 [2.72]
Lead, 2 years	<b>-0.299</b> <b>[4.44]</b>	<b>-0.332</b> <b>[6.02]</b>	<b>-0.232</b> <b>[3.78]</b>	<b>-0.169</b> <b>[4.06]</b>	<b>-0.279</b> <b>[5.02]</b>	0.050 [0.96]	<b>-0.362</b> <b>[2.07]</b>	<b>-0.343</b> <b>[1.44]</b>	0.280 [2.29]
Lead, 1 year	0.028 [0.50]	<b>-0.244</b> <b>[4.07]</b>	<b>-0.174</b> <b>[2.42]</b>	<b>-0.098</b> <b>[2.50]</b>	<b>-0.281</b> <b>[4.81]</b>	0.025 [0.52]	0.141 [0.71]	0.262 [0.95]	0.319 [2.19]
Contemporary	0.308 [4.03]	-0.189 [3.29]	-0.097 [1.34]	0.069 [0.99]	-0.216 [2.99]	0.046 [0.66]	0.821 [3.00]	1.080 [2.63]	0.322 [2.13]
R-squared	0.395	0.825	0.874	0.094	0.361	0.786	0.034	0.172	0.826
“Partial” R-squared		0.798	0.464		0.263	0.085		0.045	0.256
State fixed effects		Y	Y		Y	Y		Y	Y
Cohort fixed effects			Y			Y			Y

Notes: Stage-1 estimated cohort effects come from region-specific regressions that include race-by-education fixed effects and use IPW weights based on state births. Stage-2 regressions weighted by inverse of estimated variances of estimated cohort effects from Stage-1. Absolute value of t-ratios in [square brackets] and are corrected for heteroskedasticity and state-level clustering in the Stage-2 regression. The “Partial” R-squared is the fraction of the outcome variance, after adjusting for the respective fixed effects, that is explained by the infant health variables. There are 308 observations (22 states, 14 years) in the Stage-2 regression.

Table 8: State-level association between racial gaps in AFQT and in infant health,  
1959 to 1972 birth cohorts

	Association of racial gaps in cohort AFQT and in infant health								
	Pooled regression			State fixed effects			State and cohort fixed effects		
	(1a)	(1b)	(1c)	(2a)	(2b)	(2c)	(3a)	(3b)	(3c)
PNMR, 2 yr. lead	-0.486 <sup>***</sup> [6.85]	-0.456 <sup>***</sup> [5.31]	-0.434 <sup>***</sup> [5.89]	-0.553 <sup>***</sup> [9.49]	-0.537 <sup>***</sup> [8.14]	-0.466 <sup>***</sup> [5.30]	-0.356 <sup>***</sup> [5.79]	-0.382 <sup>***</sup> [5.28]	-0.370 <sup>***</sup> [4.54]
PNMR, 1 yr. lead	-0.066 [0.76]	-0.081 [0.85]	-0.119 [1.55]	-0.413 <sup>***</sup> [5.55]	-0.421 <sup>***</sup> [4.97]	-0.327 <sup>***</sup> [3.83]	-0.268 <sup>***</sup> [3.33]	-0.282 <sup>***</sup> [2.94]	-0.234 <sup>**</sup> [2.61]
NMR, 2 yr. lead	-0.197 <sup>**</sup> [2.64]	-0.215 <sup>***</sup> [2.87]	-0.055 [1.11]	-0.068 [1.50]	-0.065 [1.19]	-0.007 [0.09]	0.014 [0.21]	0.015 [0.20]	0.008 [0.12]
NMR, 1 yr. lead	-0.220 <sup>***</sup> [2.90]	-0.236 <sup>***</sup> [3.04]	-0.085 [1.73]	-0.110 <sup>*</sup> [2.02]	-0.099 [1.56]	-0.048 [0.75]	-0.012 [0.24]	-0.023 [0.38]	-0.043 [0.84]
LBW, 2 yr. lead	-0.350 [1.56]	-0.349 [1.43]	-0.135 [0.70]	-0.221 [1.15]	-0.217 [1.03]	-0.119 [0.60]	0.189 [1.22]	0.125 [0.76]	0.274 [1.49]
LBW, 1 yr. lead	-0.014 [0.07]	-0.073 [0.31]	0.075 [0.38]	0.127 [0.71]	0.156 [0.74]	0.159 [0.72]	0.240 [1.36]	0.173 [0.81]	0.201 [0.88]
Constant	-7.97 [4.36]	-7.38 [3.45]							
F(9, 18) for joint signif. illegit, age variables			14.62 <sup>***</sup> {0.000}			4.56 <sup>***</sup> {0.003}			2.35 <sup>*</sup> {0.059}
No. of observations	308	253	253	308	253	253	308	253	253
R-squared	0.489	0.492	0.710	0.810	0.810	0.829	0.866	0.860	0.870
Mother's marital status			Y			Y			Y
And age categories			Y			Y			Y
State fixed effects				Y	Y	Y	Y	Y	Y
Cohort fixed effects							Y	Y	Y

Notes: See above notes. Absolute value of t-ratios in [square brackets] and are corrected for heteroskedasticity and state-level clustering. There are 308 observations (22 states, 14 years). The F-test for the joint significance of the mothers' marital status and age variables has 9 (18) numerator (denominator) degrees of freedom. The p-values of the F-test are shown in {}

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

Table A1: Age of entry by year of birth and year of AFQT exam

<u>Year Of Birth</u>	<u>Year AFQT Test Taken</u>															
	1976	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990	1991
1957	18															
1958	17,18	18														
1959	17	17,18	18													
1960		17	17,18	18												
1961			17	17,18	18											
1962				17	17,18	18										
1963					17	17,18	18									
1964						17	17,18	18								
1965							17	17,18	18							
1966								17	17,18	18						
1967									17	17,18	18					
1968										17	17,18	18				
1969											17	17,18	18			
1970												17	17,18	18		
1971													17	17,18	18	
1972														17	17,18	18
1973															17	17,18

Notes: Approximately one-third of AFQT sample takes the test

Table A2: Estimates of black-white difference in AFQT effects,  
Inverse probability weighted regressions

	Black-white difference in coefficients	
	South	Rustbelt
	(1)	(2)
<u>Education effects</u>		
1 year of high school	9.20 <sup>***</sup> (0.23)	11.54 <sup>***</sup> (0.26)
2 years of high school	5.69 <sup>***</sup> (0.15)	7.94 <sup>***</sup> (0.14)
3-4 years of high school	---	---
GED	3.54 <sup>***</sup> (0.31)	5.42 <sup>***</sup> (0.31)
High school graduate	-4.24 <sup>***</sup> (0.10)	-3.71 <sup>***</sup> (0.11)
1 year of college	-9.84 <sup>***</sup> (0.89)	-7.10 <sup>***</sup> (1.04)
<u>Age effects</u>		
17 years old	---	---
18 years old	0.35 <sup>***</sup> (0.12)	1.62 <sup>***</sup> (0.13)
<u>Year effects</u>		
1976	-2.13 <sup>**</sup> (0.87)	-6.63 <sup>***</sup> (0.88)
1977	-1.51 <sup>*</sup> (0.78)	-4.93 <sup>***</sup> (0.79)
1978	-2.28 <sup>***</sup> (0.71)	-4.41 <sup>***</sup> (0.72)
1979	-2.40 <sup>***</sup> (0.63)	-4.16 <sup>***</sup> (0.64)
1980	-0.35 (0.54)	-1.28 <sup>**</sup> (0.56)
1981	-0.57 (0.47)	-2.75 <sup>***</sup> (0.49)
1982	0.45 (0.40)	-0.83 <sup>**</sup> (0.42)
1983	1.49 <sup>***</sup> (0.30)	0.15 (0.32)
1984	---	---
1985	-0.49 (0.30)	0.59 <sup>*</sup> (0.31)
1986	-0.68 <sup>*</sup> (0.41)	0.69 (0.43)
1987	-1.97 <sup>***</sup> (0.50)	-0.23 (0.53)
1988	-3.29 <sup>***</sup> (0.58)	-0.41 (0.63)
1989	-4.97 <sup>***</sup> (0.65)	-0.73 (0.71)
1990	-5.31 <sup>***</sup> (0.72)	-0.91 (0.79)
1991	-5.70 <sup>***</sup> (0.79)	-1.06 (0.89)

(Table A2 continued)

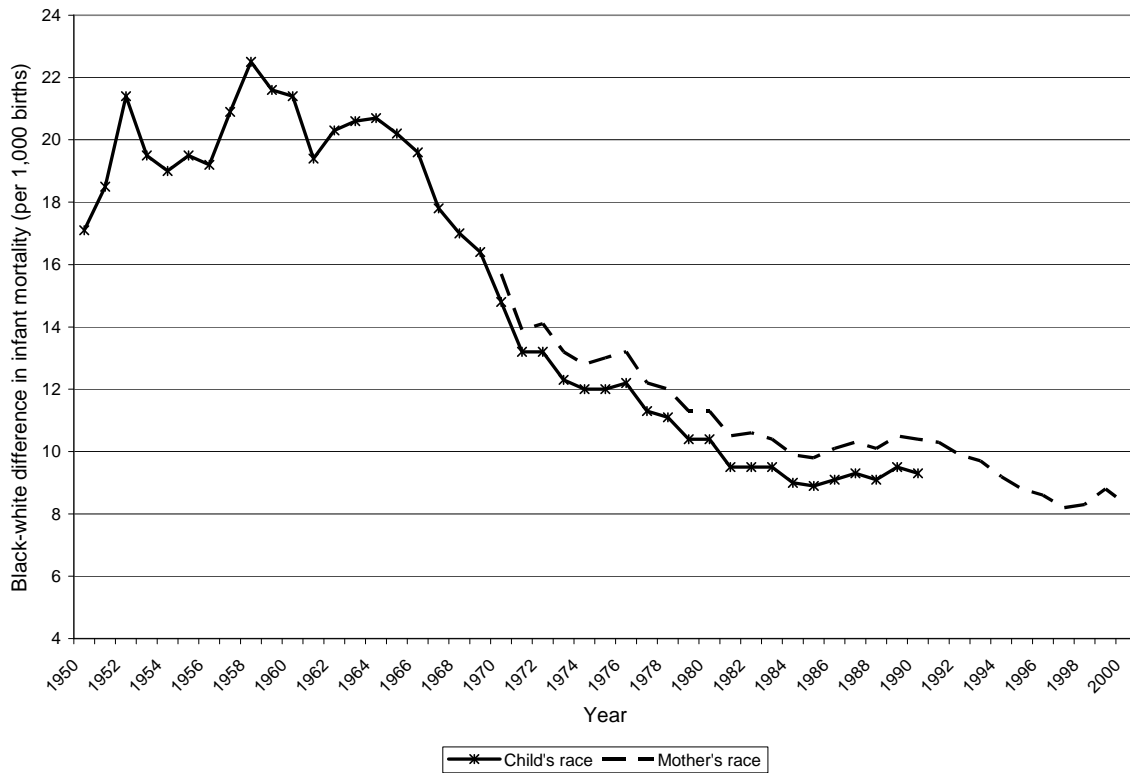
	Black-white difference in coefficients	
	South	Rustbelt
	(1)	(2)
<u>Birth cohort effects</u>		
1957	-21.23 <sup>***</sup> (0.95)	-17.07 <sup>***</sup> (0.97)
1958	-21.93 <sup>***</sup> (0.86)	-17.64 <sup>***</sup> (0.88)
1959	-22.17 <sup>***</sup> (0.79)	-17.66 <sup>***</sup> (0.81)
1960	-22.13 <sup>***</sup> (0.72)	-18.15 <sup>***</sup> (0.73)
1961	-23.40 <sup>***</sup> (0.63)	-19.12 <sup>***</sup> (0.65)
1962	-23.57 <sup>***</sup> (0.55)	-19.30 <sup>***</sup> (0.57)
1963	-23.16 <sup>***</sup> (0.48)	-19.37 <sup>***</sup> (0.50)
1964	-21.69 <sup>***</sup> (0.40)	-19.11 <sup>***</sup> (0.42)
1965	-19.49 <sup>***</sup> (0.27)	-18.70 <sup>***</sup> (0.29)
1966	-18.63 <sup>***</sup> (0.21)	-18.36 <sup>***</sup> (0.23)
1967	-17.25 <sup>***</sup> (0.32)	-18.30 <sup>***</sup> (0.34)
1968	-16.47 <sup>***</sup> (0.42)	-18.40 <sup>***</sup> (0.44)
1969	-15.40 <sup>***</sup> (0.50)	-17.64 <sup>***</sup> (0.54)
1970	-14.47 <sup>***</sup> (0.58)	-16.88 <sup>***</sup> (0.62)
1971	-13.00 <sup>***</sup> (0.65)	-15.85 <sup>***</sup> (0.71)
1972	-11.66 <sup>***</sup> (0.71)	-14.91 <sup>***</sup> (0.79)
1973	-9.87 <sup>***</sup> (0.78)	-14.24 <sup>***</sup> (0.88)
Sample size	934,296	1,346,036

Notes: Sample contains all black and white men born between 1957 and 1973, who took the AFQT test between 1976 and 1991 in the South and Rustbelt, with entry ages of 17 or 18. 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.

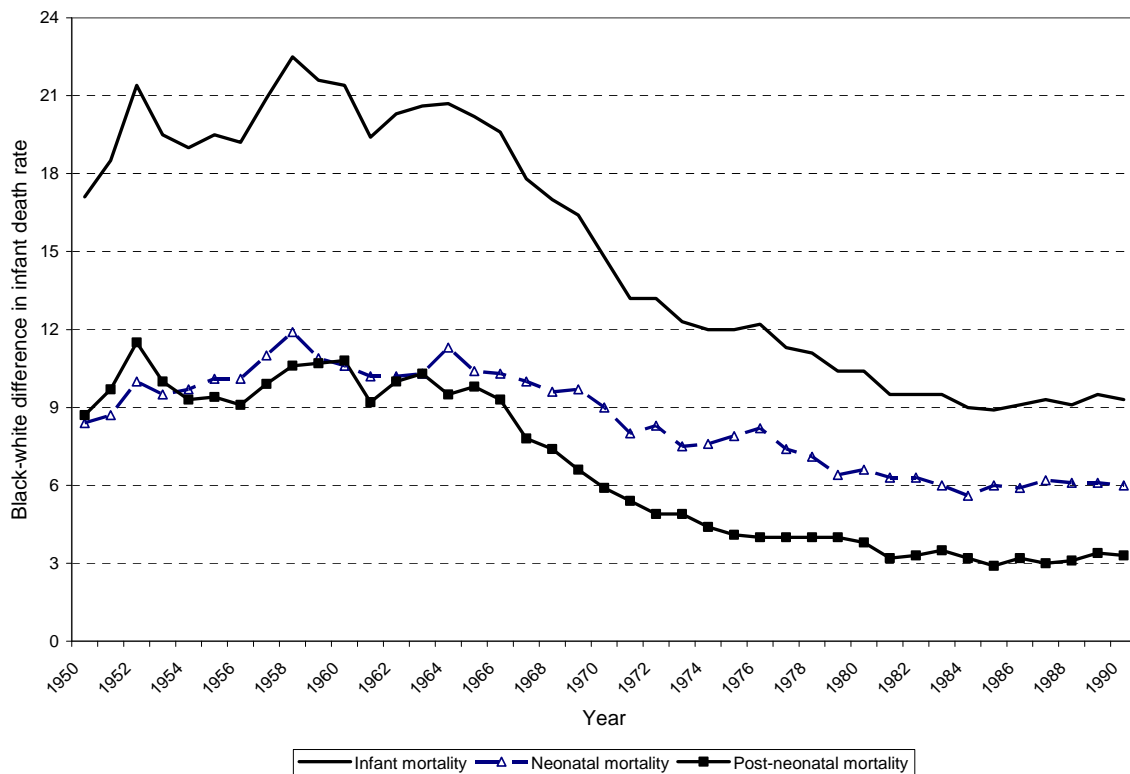
<sup>\*\*\*</sup> significant at 1-percent level, <sup>\*\*</sup> significant at 5-percent level, <sup>\*</sup> significant at 10-percent level

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

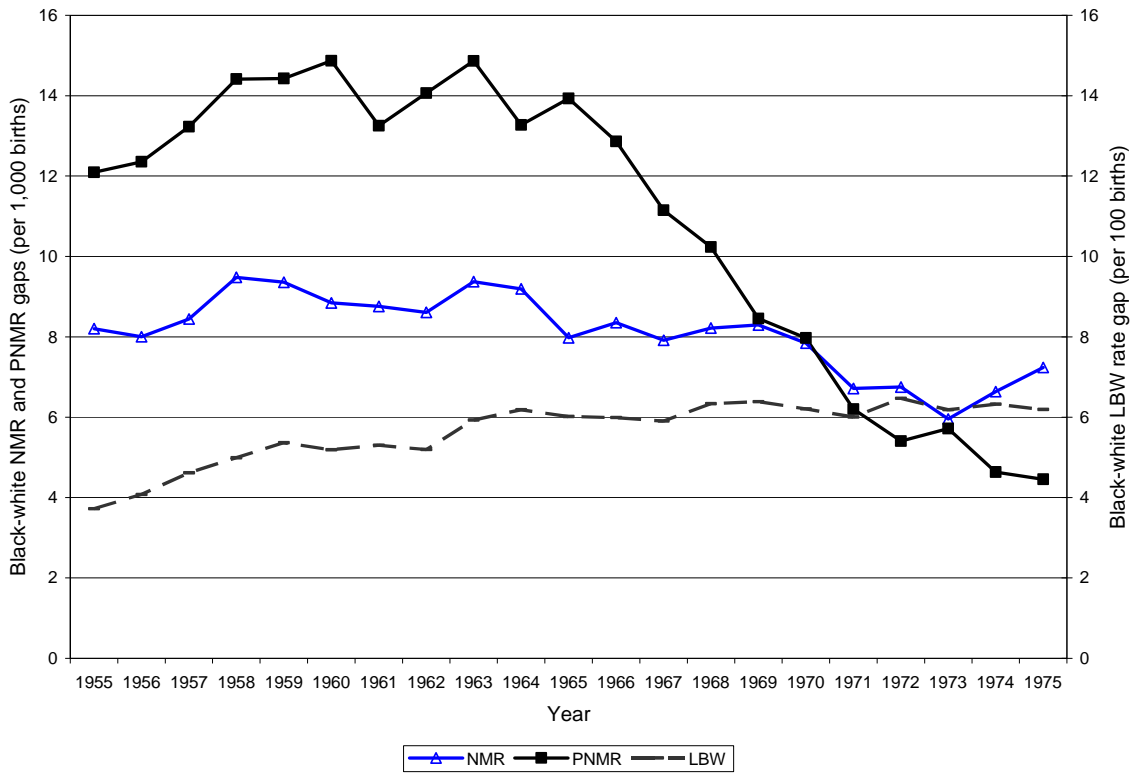
A. Infant mortality by child and mother's race in United States, 1950 to 2000



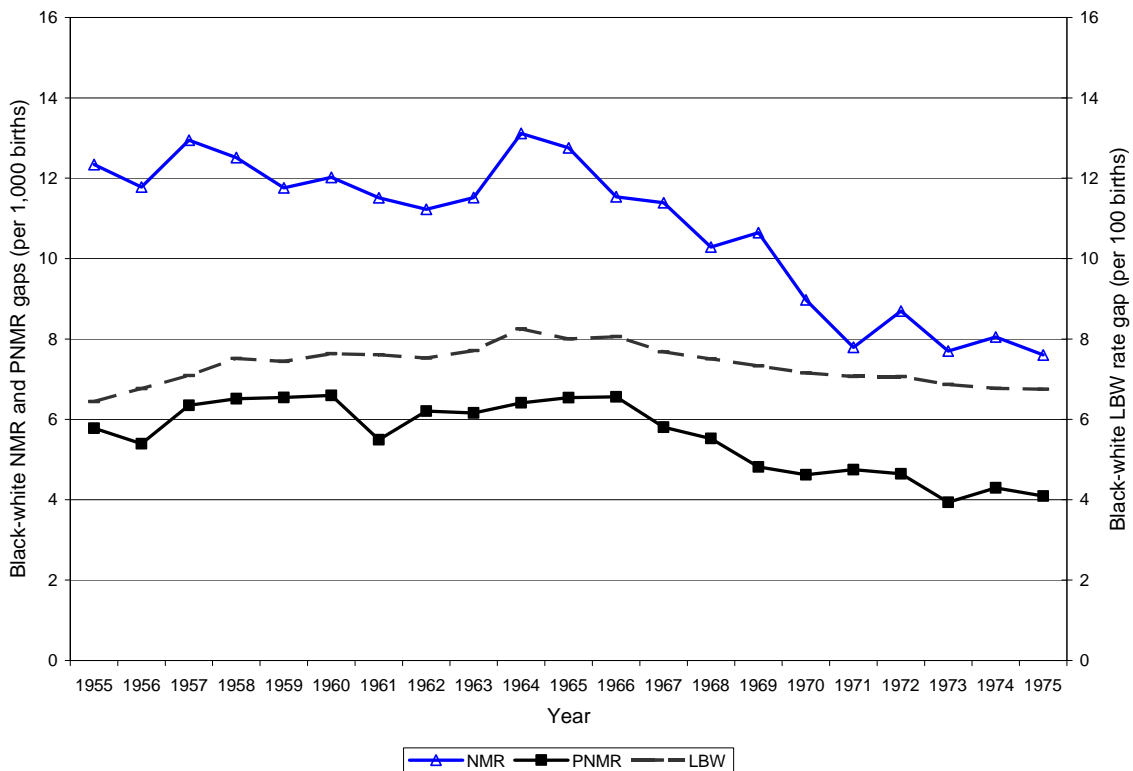
B. Black-white gaps in infant, post-neonatal and neonatal mortality rates (child's race), 1950 to 1990



C. Racial gaps in NMR, PNMR and low birth weight rates in South, 1955 to 1975



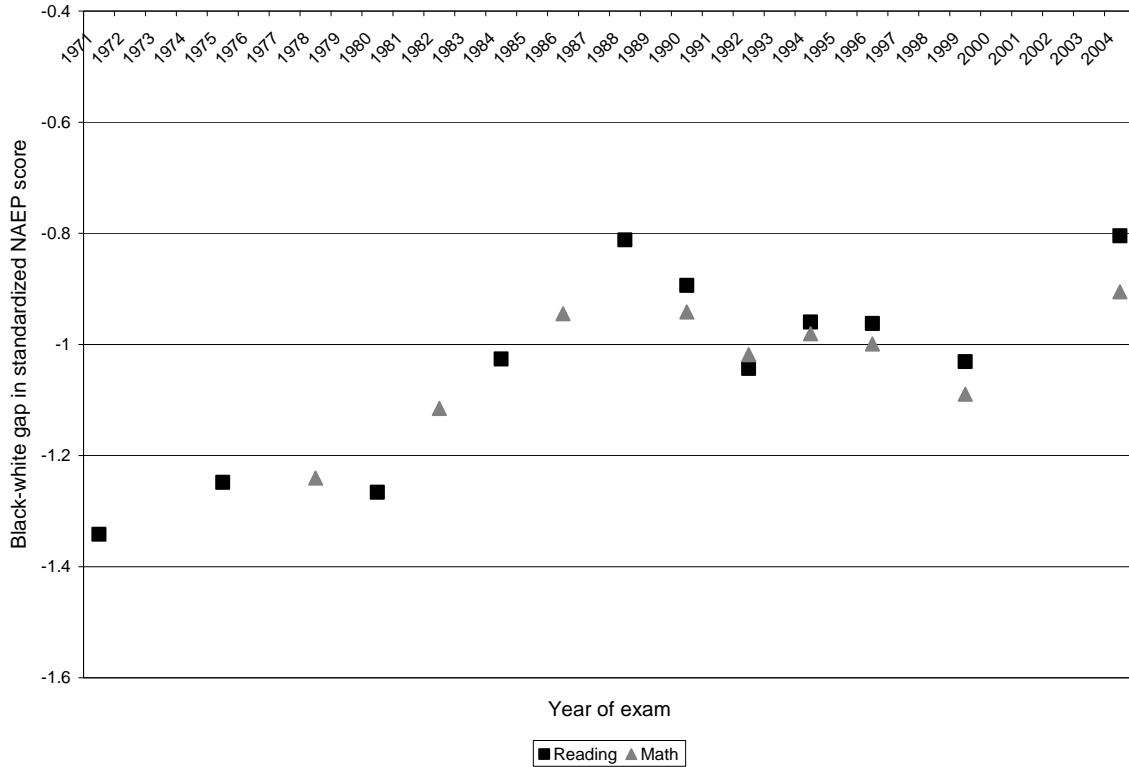
D. Racial gaps in NMR, PNMR and low birth weight rates in Rustbelt, 1955 to 1975



Notes: Data from the *Vital Statistics of the United States*. South consists of Alabama, Arkansas, Florida, Georgia, Louisiana, Mississippi, North Carolina, South Carolina, Tennessee and Virginia; Rustbelt of Illinois, Indiana, Michigan, Missouri, New York, Ohio and Pennsylvania

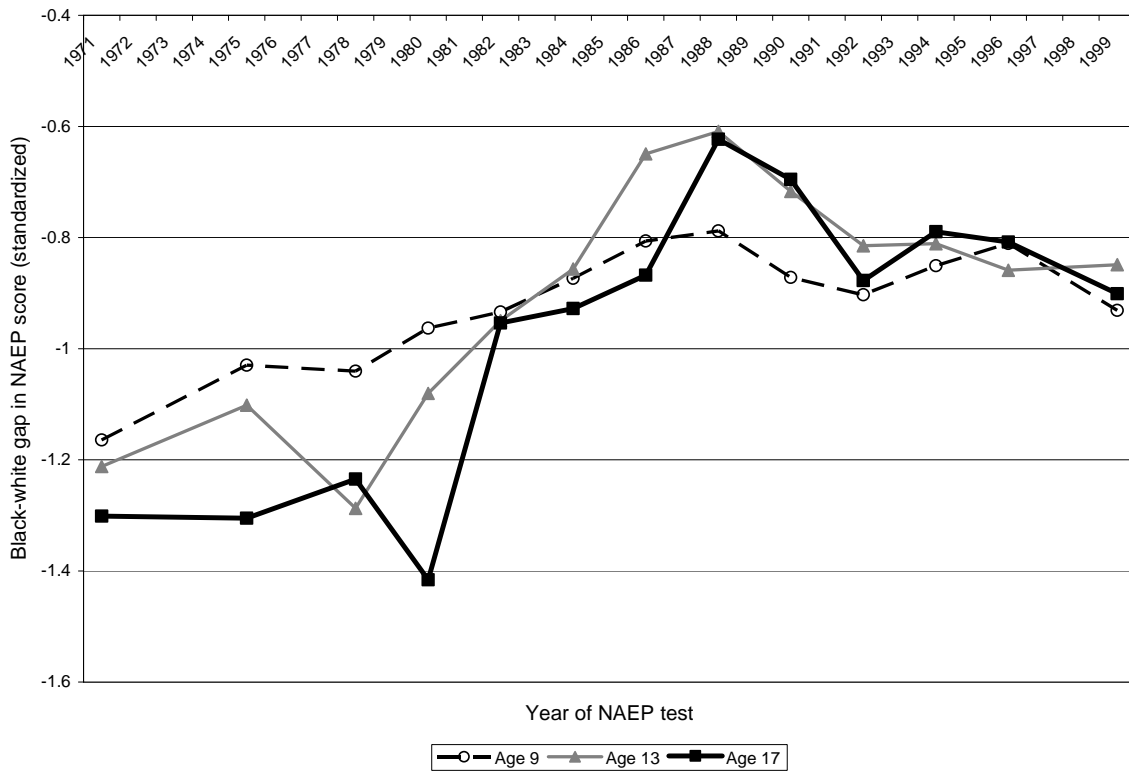
Figure 2: National Assessment of Educational Progress Scores

A. Black-white gap in standardized NAEP scores by calendar year of exam, United States



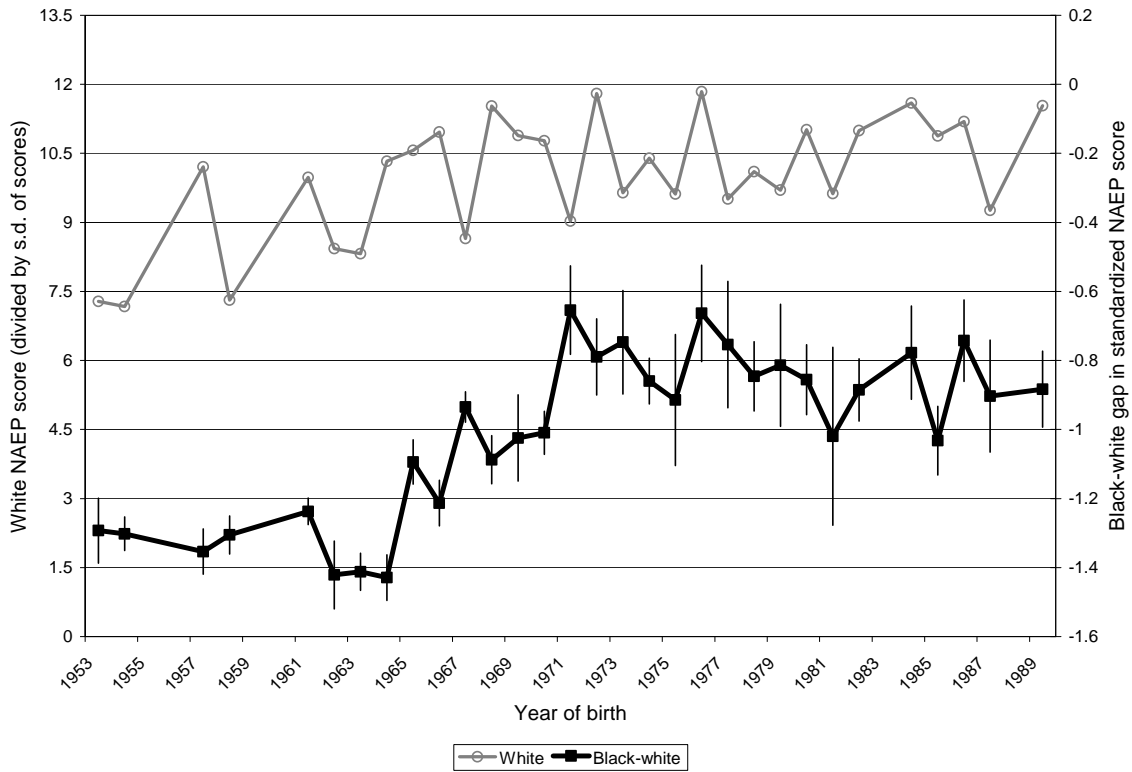
Notes: Figure plots racial differences in average scaled NAEP Math and Reading scores, normalized by the standard deviation of test scores by survey year, age, and subject. Subject-specific regressions adjust for race-specific age effects.

B. Black-white difference in standardized NAEP scores, by age cohort



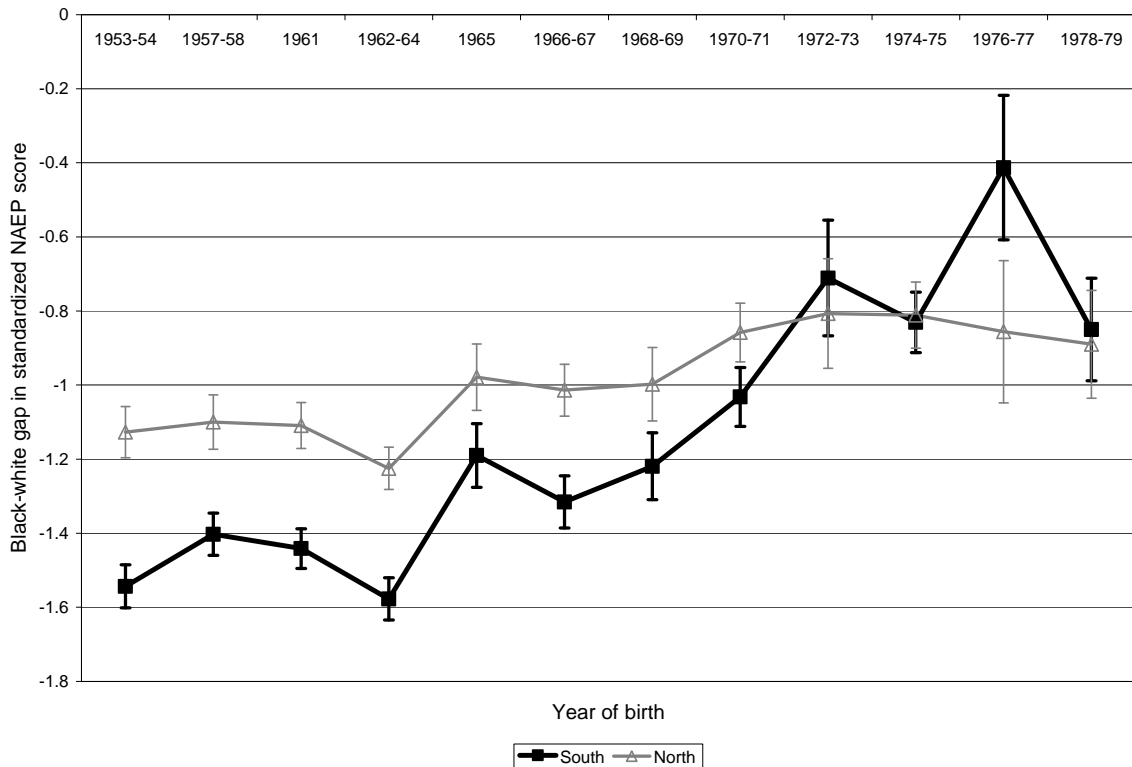
Notes: Figure plots racial differences in standardized NAEP score, separately for 9-, 13- and 17-year-olds. Regression adjusts for race-specific subject effects by age.

### C. Black-white differences in NAEP scores by year of birth, United States



Notes: Figure plots white levels of and racial differences in standardized NAEP scores by year of birth. Vertical lines represent ( $\pm$ ) twice the standard error of the estimate, corrected for heteroskedasticity. Regression adjusts for race-specific age effects that vary by subject.

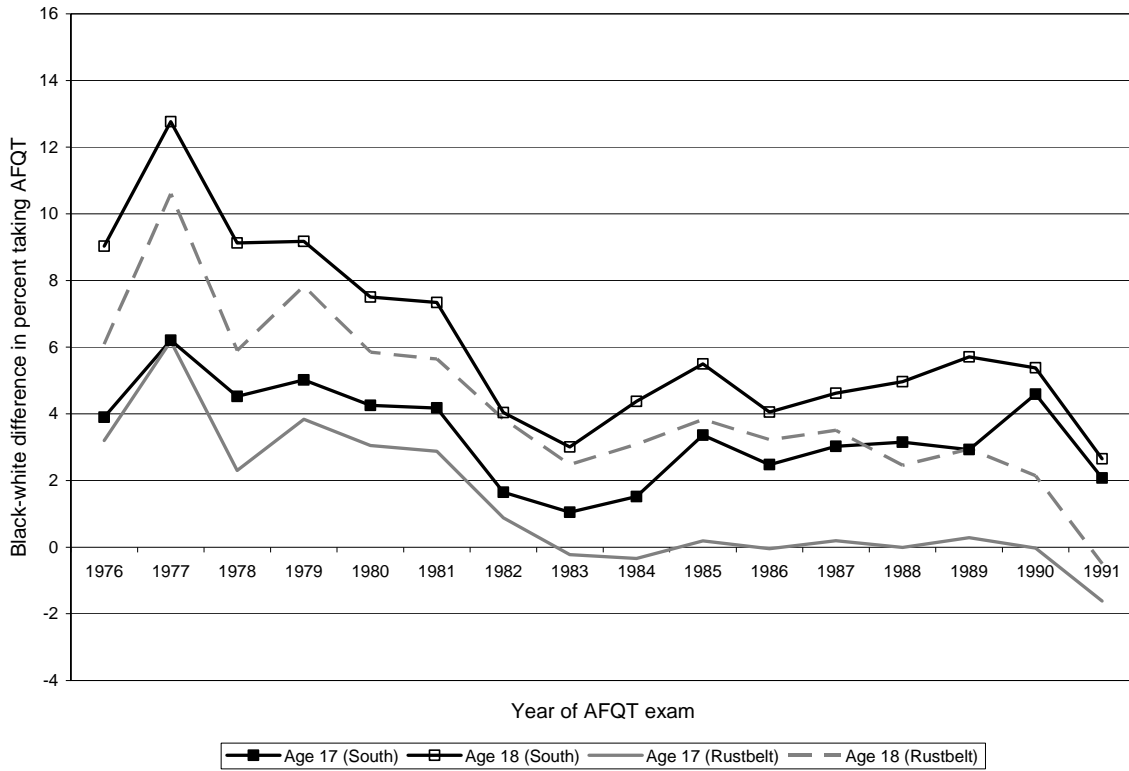
### D. Black-white differences in NAEP scores by year of birth, South and North



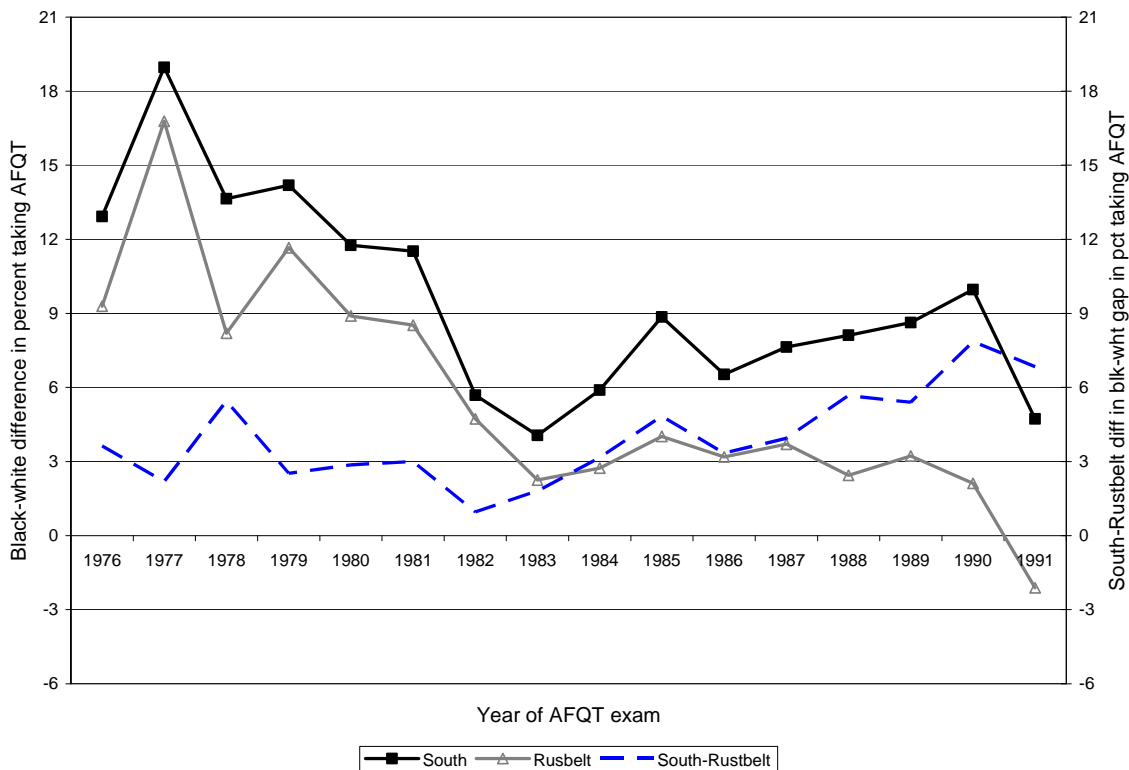
Notes: Figure plots racial difference in standardized NAEP scores, by year of birth, for the South and North using the 1971 to 1996 NAEP surveys. Vertical lines represent ( $\pm$ ) twice the standard error of the estimate, corrected for heteroskedasticity. Regression adjusts for race-specific age effects that vary by subject and region. See text for more details.

Figure 3: Probability in population of taking the AFQT, by year exam taken

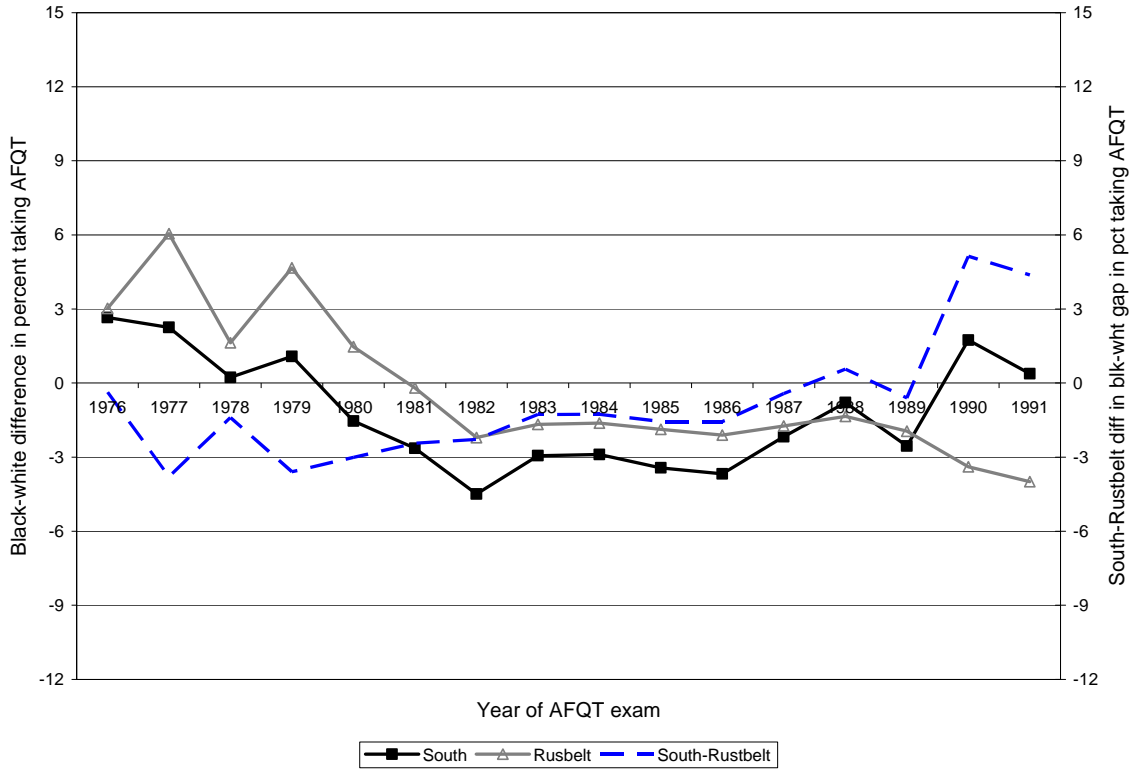
A. Racial gap in selection probabilities separately for 17 and 18 year olds, South and Rustbelt



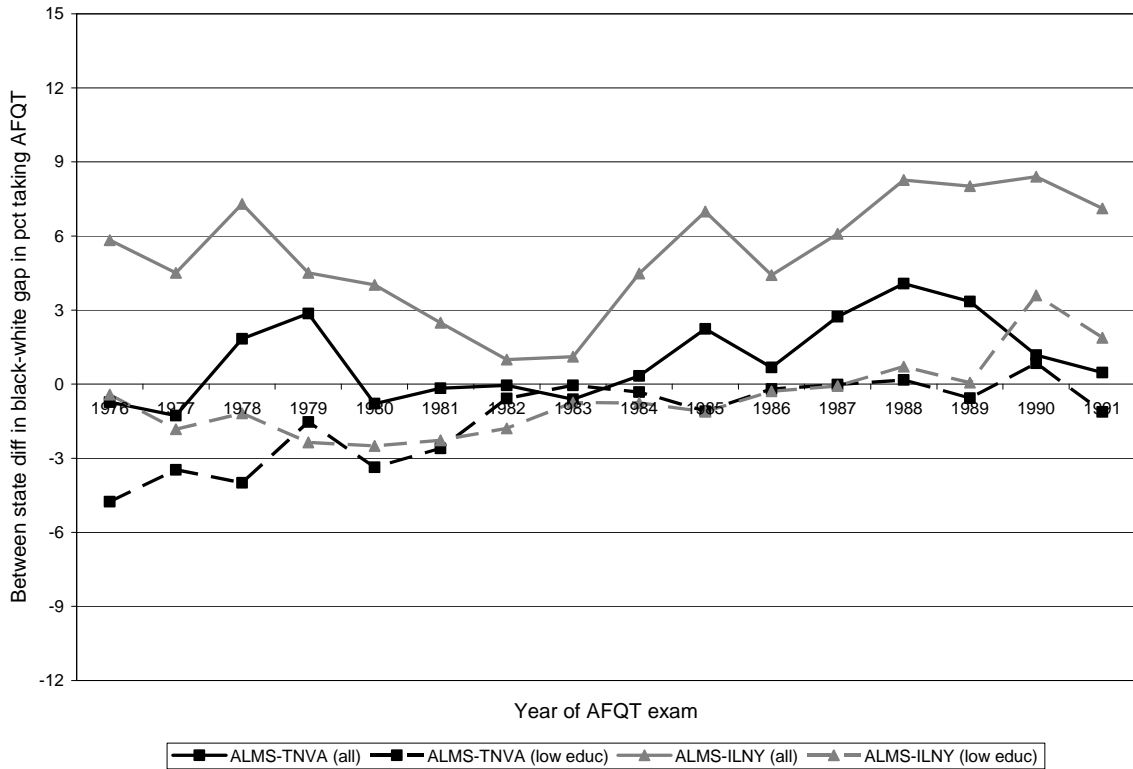
B. Racial gap in selection probabilities for 17 and 18 year olds combined



C. Racial gap in selection probabilities for men with two years or less of high school



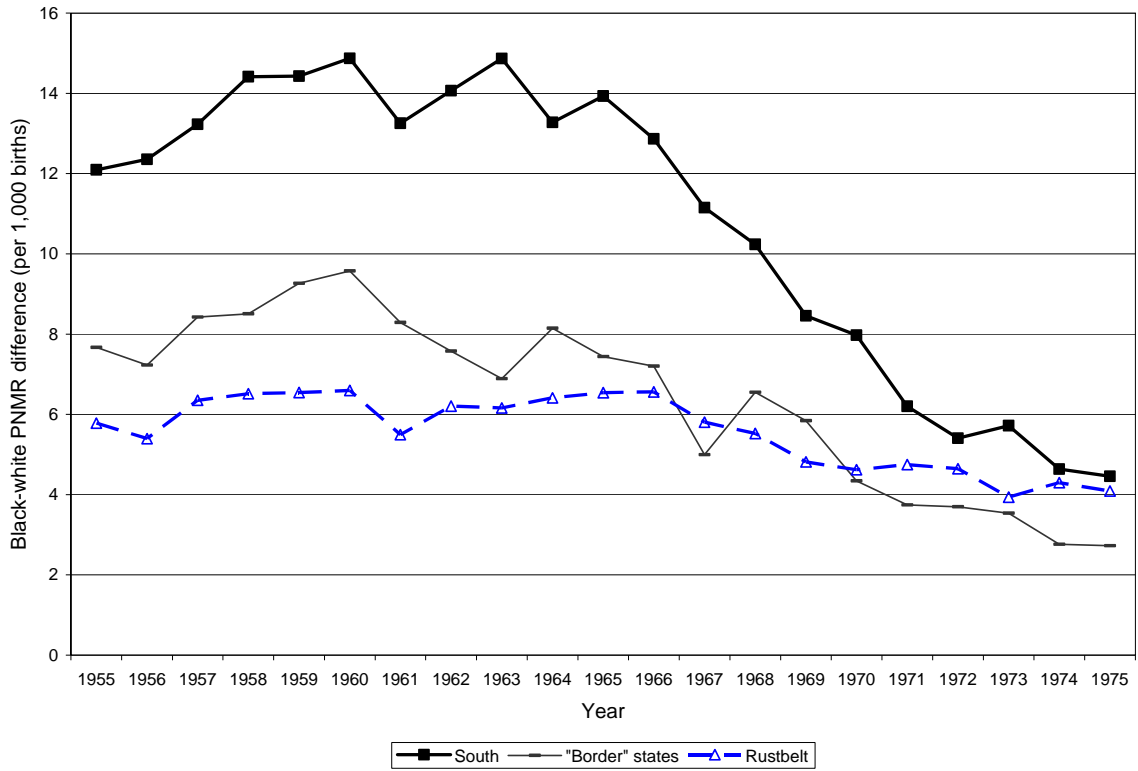
D. Difference in selection probability gap between Alabama-Mississippi and other state groups



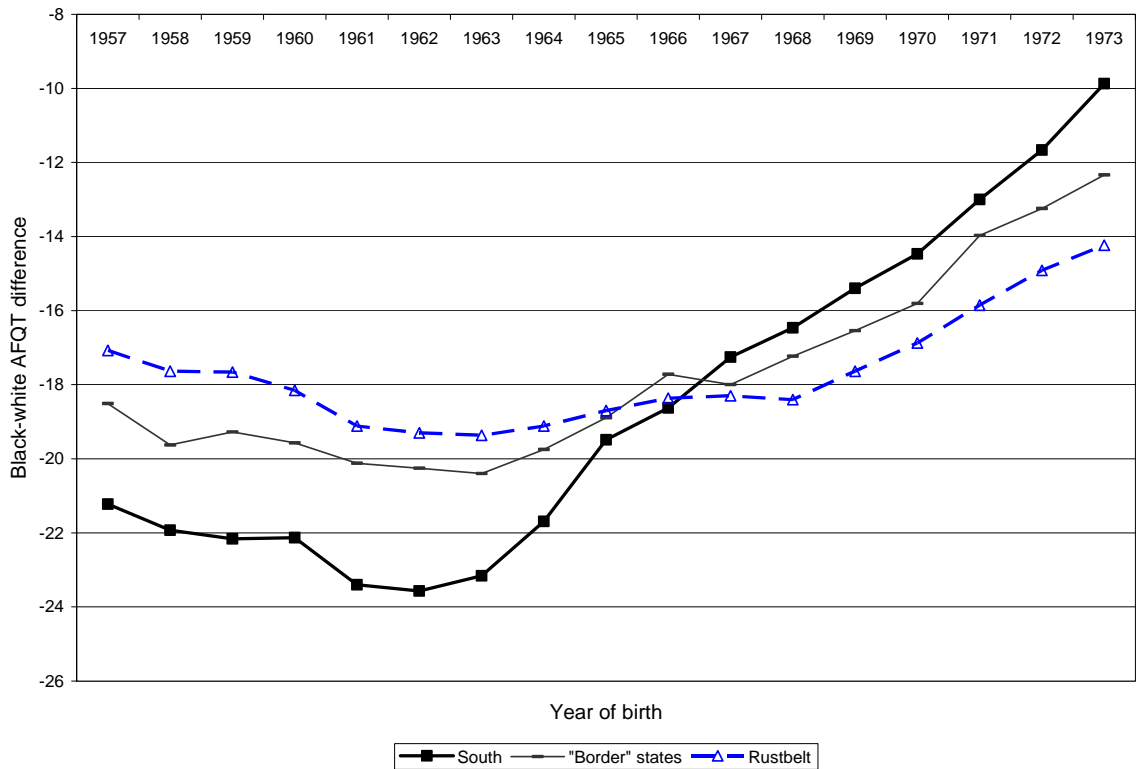
Notes: Population counts for each state-race-age-year (and education) cell come from the Decennial Censuses. In Panel D, the state groups are Alabama and Mississippi (ALMS), Tennessee and Virginia (TNVA), and Illinois and New York (ILNY); and "low educ" refers to men with two years or less of high school education.

Figure 4: Black-white differences in post-neonatal mortality rates and AFQT scores across regions

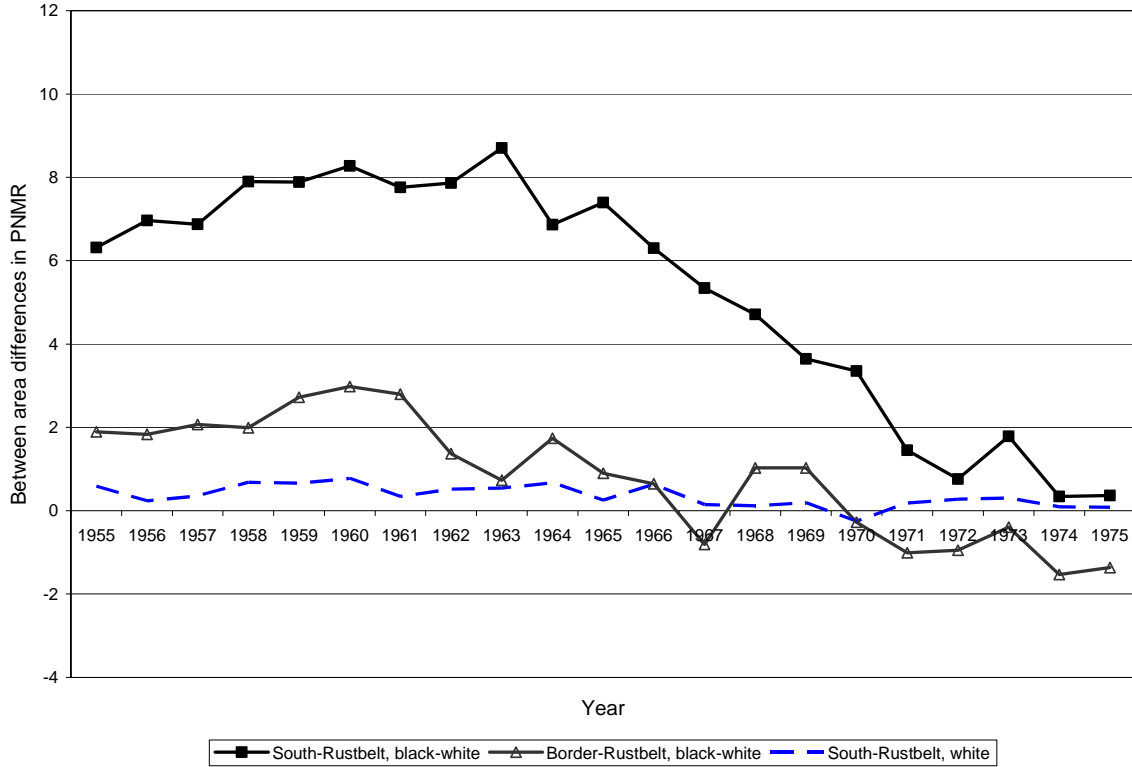
A. Black-white gaps in post-neonatal mortality rates by year



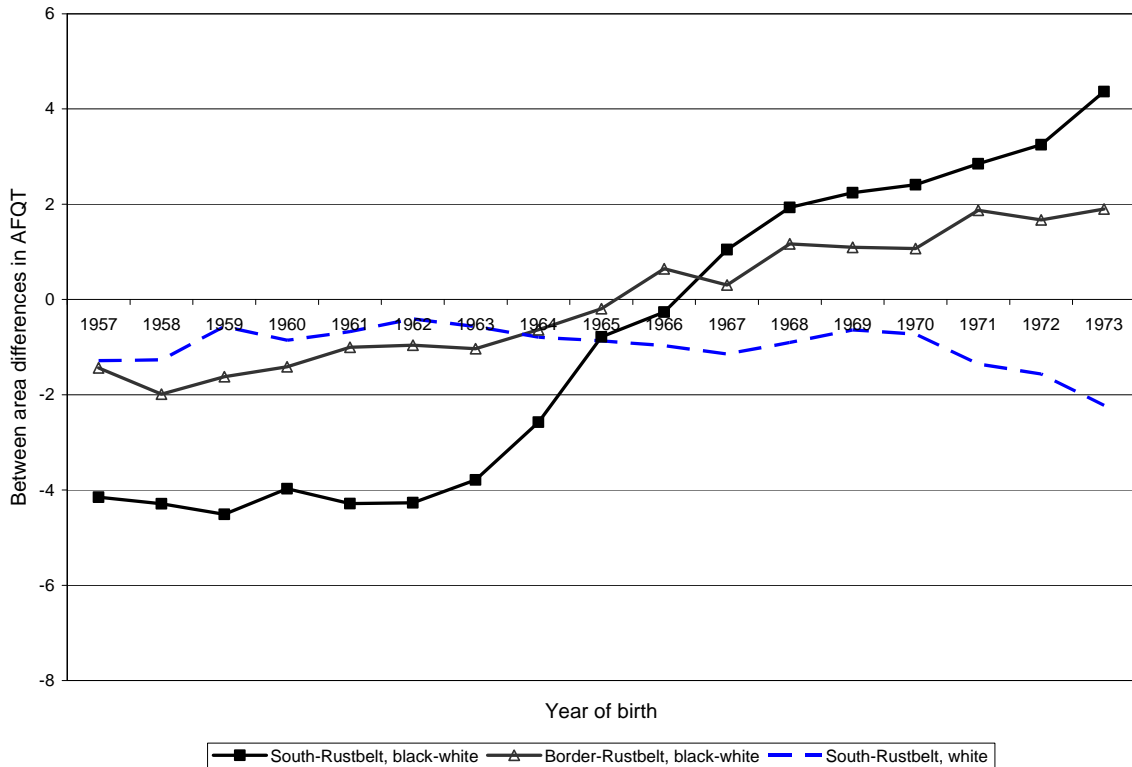
B. Black-white gaps in AFQT scores by year of birth



### C. Between-region differences in post-neonatal mortality rate gaps



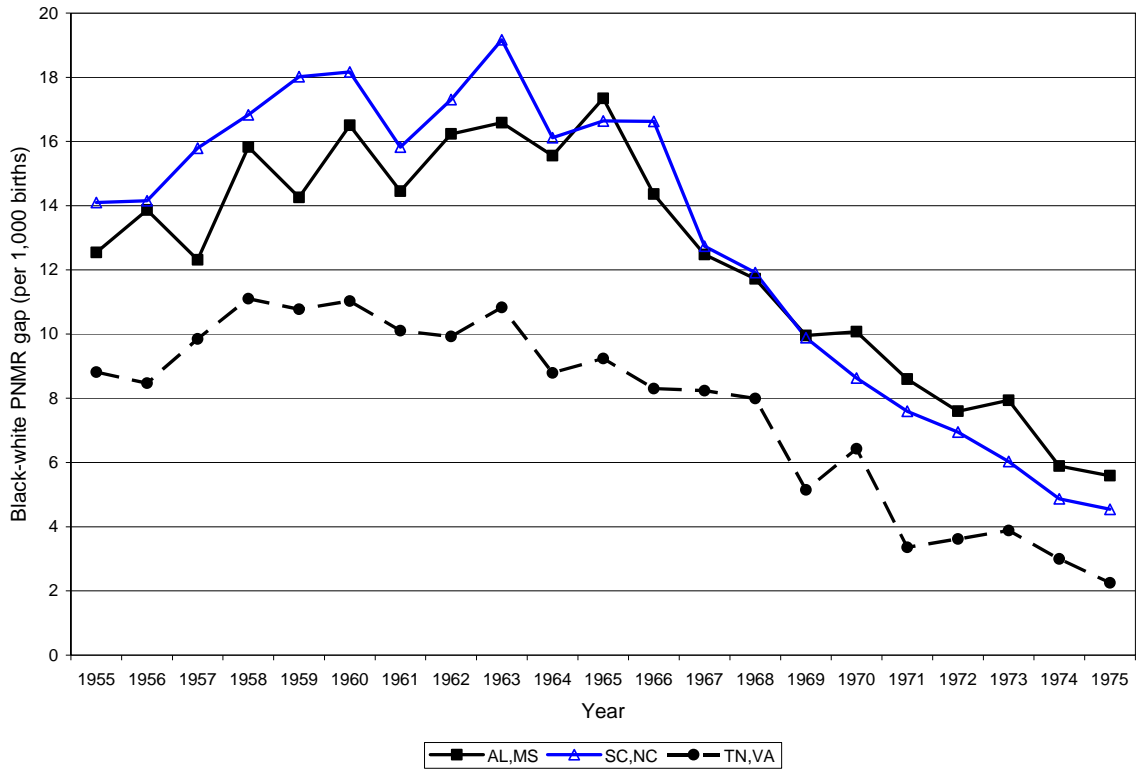
### D. Between-region differences in AFQT score gaps



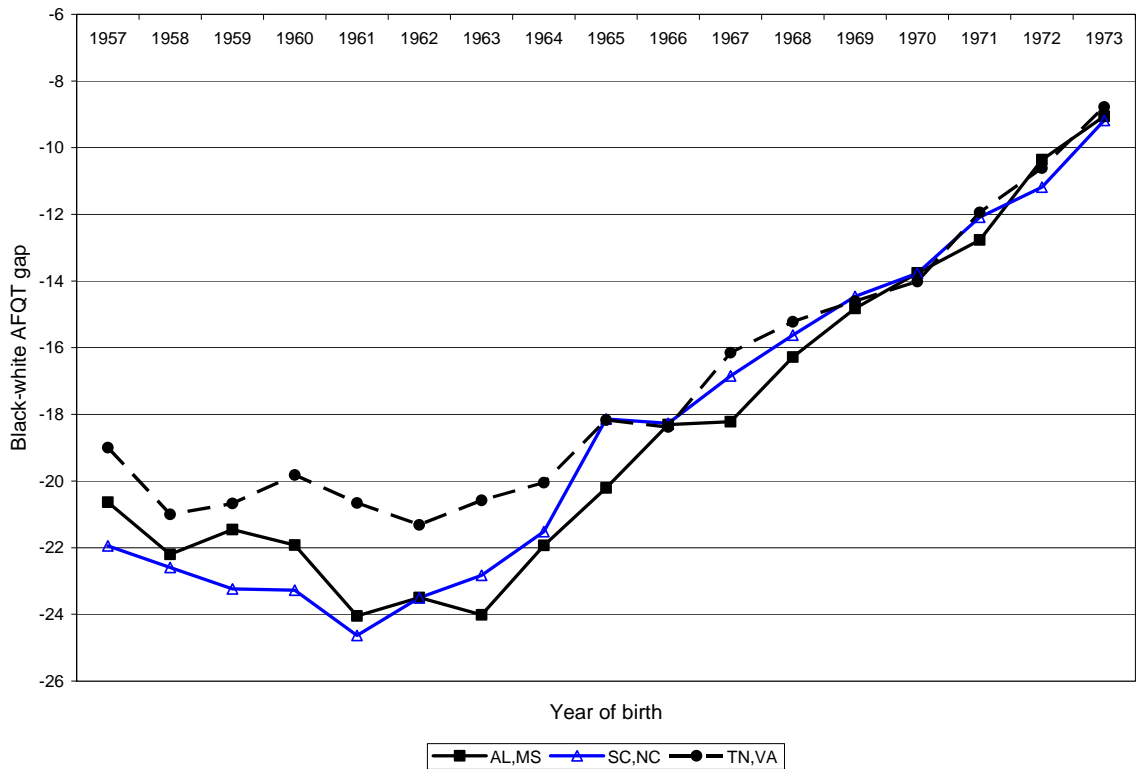
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age, year and education effects interacted with race; run separately by region. The baseline group is men with an entry age of 17 and 3 to 4 years of high school education (but not high school graduates) when they took the exam.

Figure 5: Across state-group differences in black-white gaps in PNMR and AFQT scores

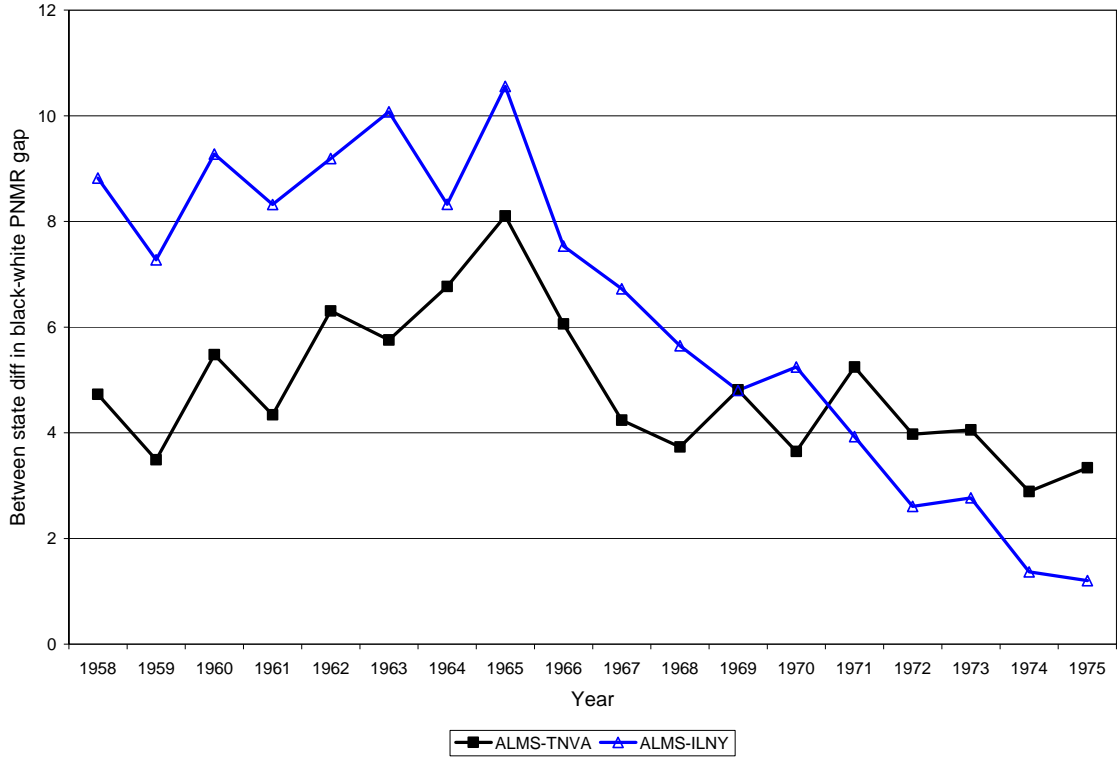
A. Post-neonatal mortality gaps



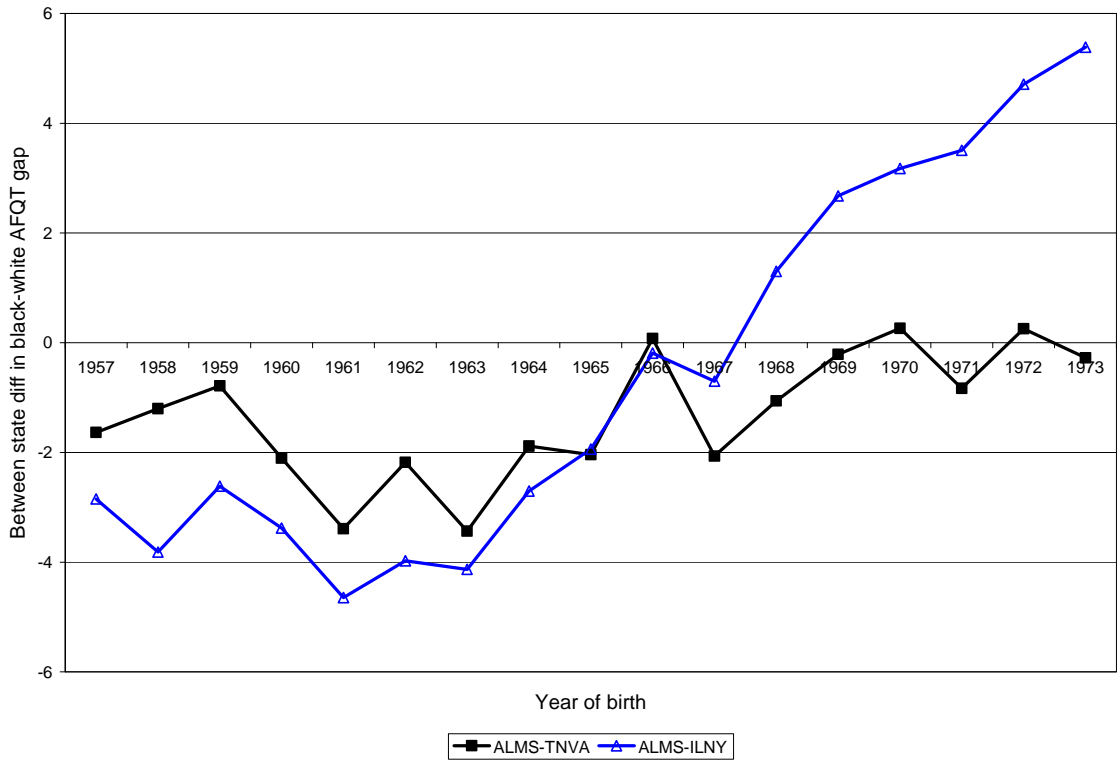
B. AFQT score gaps



C. Difference in PNMR gap between Alabama-Mississippi and other state groups



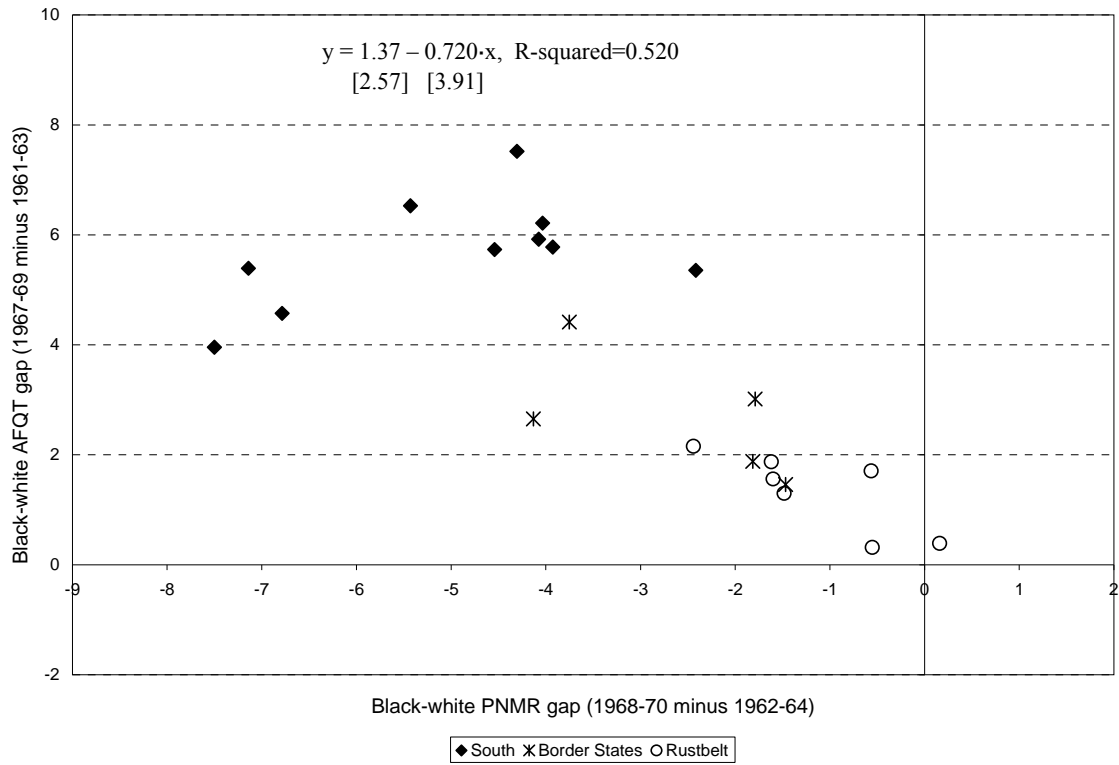
D. Difference in AFQT gap between Alabama-Mississippi and other state groups



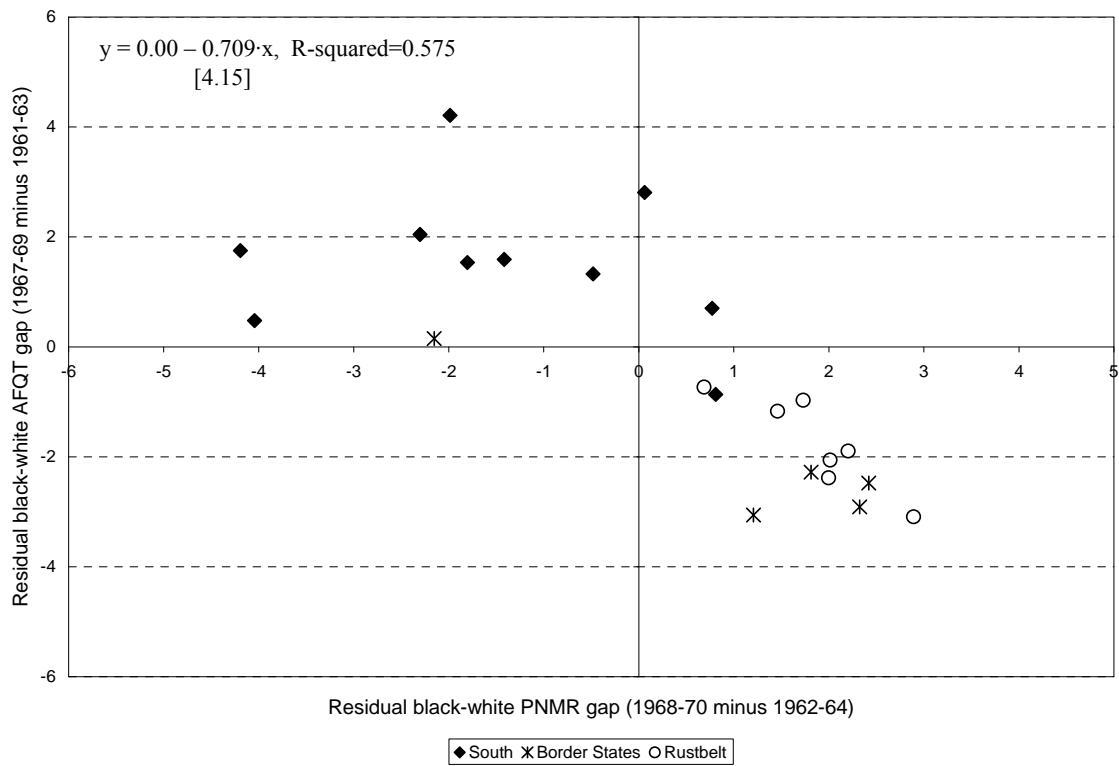
Notes: AFQT plots come from inverse probability weighted (by state births) regressions that allow for unrestricted age, year and education effects interacted with race; separately run for each state group – Alabama and Mississippi (ALMS); Tennessee and Virginia (TNVA); and Illinois and New York (ILNY).

Figure 6: Scatter plots of between-cohort changes in racial gaps in AFQT and infant health (22 states)

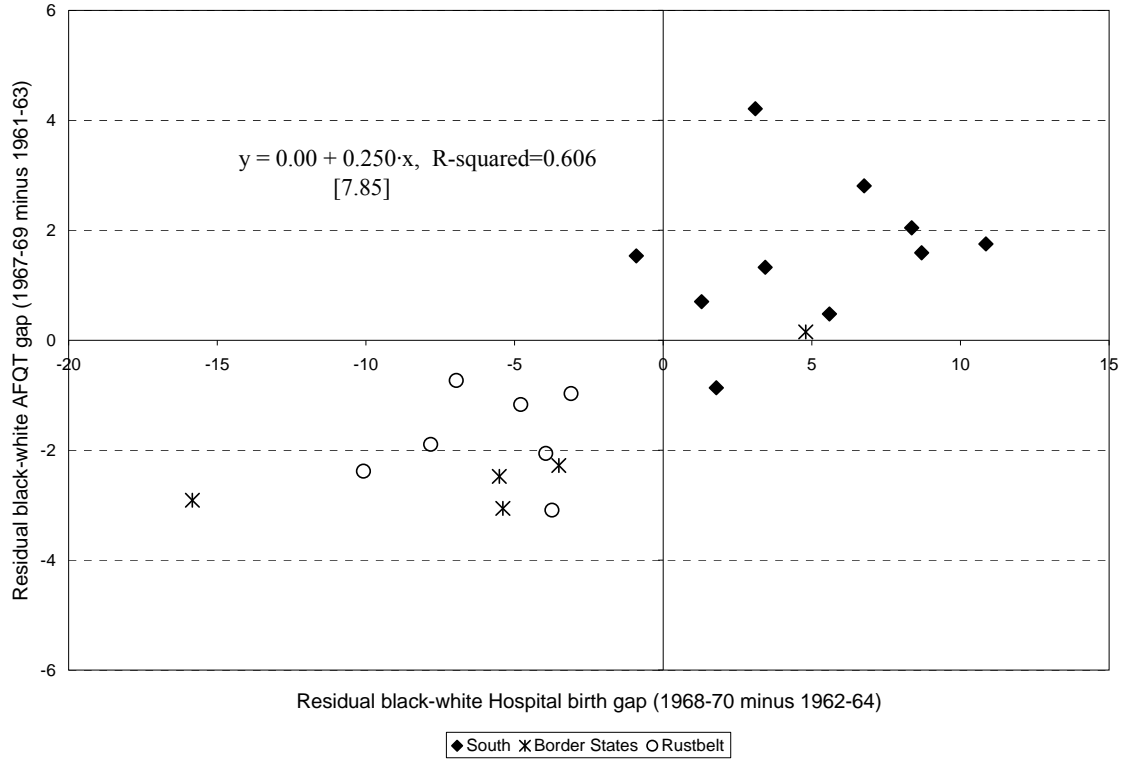
A. Changes in gaps in AFQT (1961-63 to 1967-69) and PNMR (1962-64 to 1968-70)



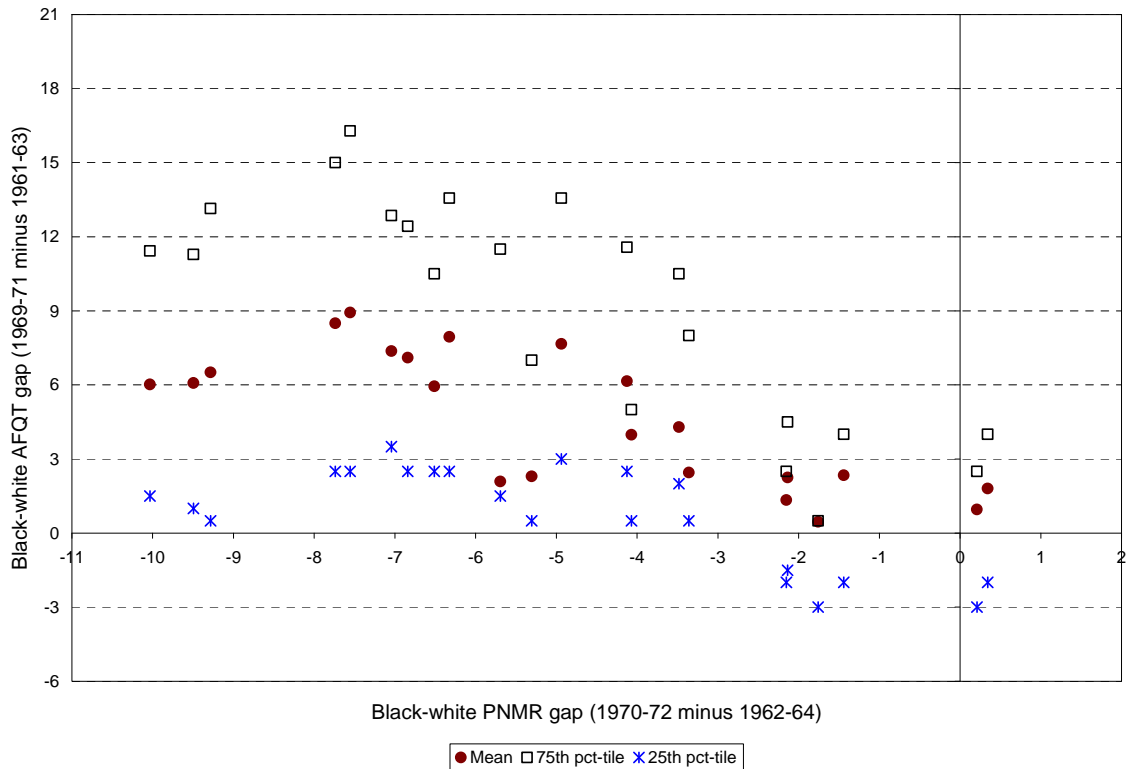
B. Residual changes in AFQT and PNMR gaps



C. Residual changes in AFQT and hospital birth rate gaps



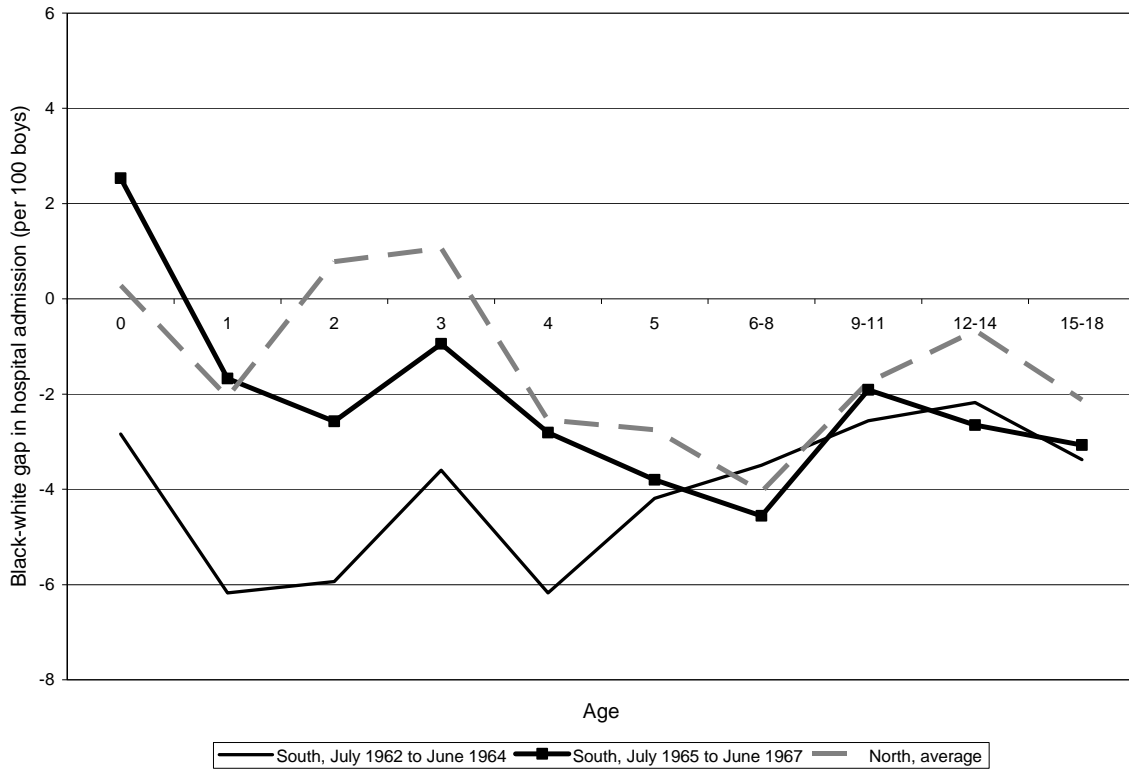
D. Changes in AFQT (1961-63 to 1969-71) and PNMR (1962-64 to 1970-72) gaps, Mean, 75<sup>th</sup> and 25<sup>th</sup> percentiles of AFQT scores



Notes: AFQT scores come from region-specific regressions that include race-by-education fixed effects and use inverse probability weights. Panels B and C plot the residualized between-cohort changes adjusted for the variables in column (6) in Table 6. Panel D plots between-cohort changes in racial gaps in AFQT scores estimated from OLS and quantile regressions.

Figure 7: Black-white hospital admission rate differences by age (boys)

A. Hospital admission gap (per 100 boys)

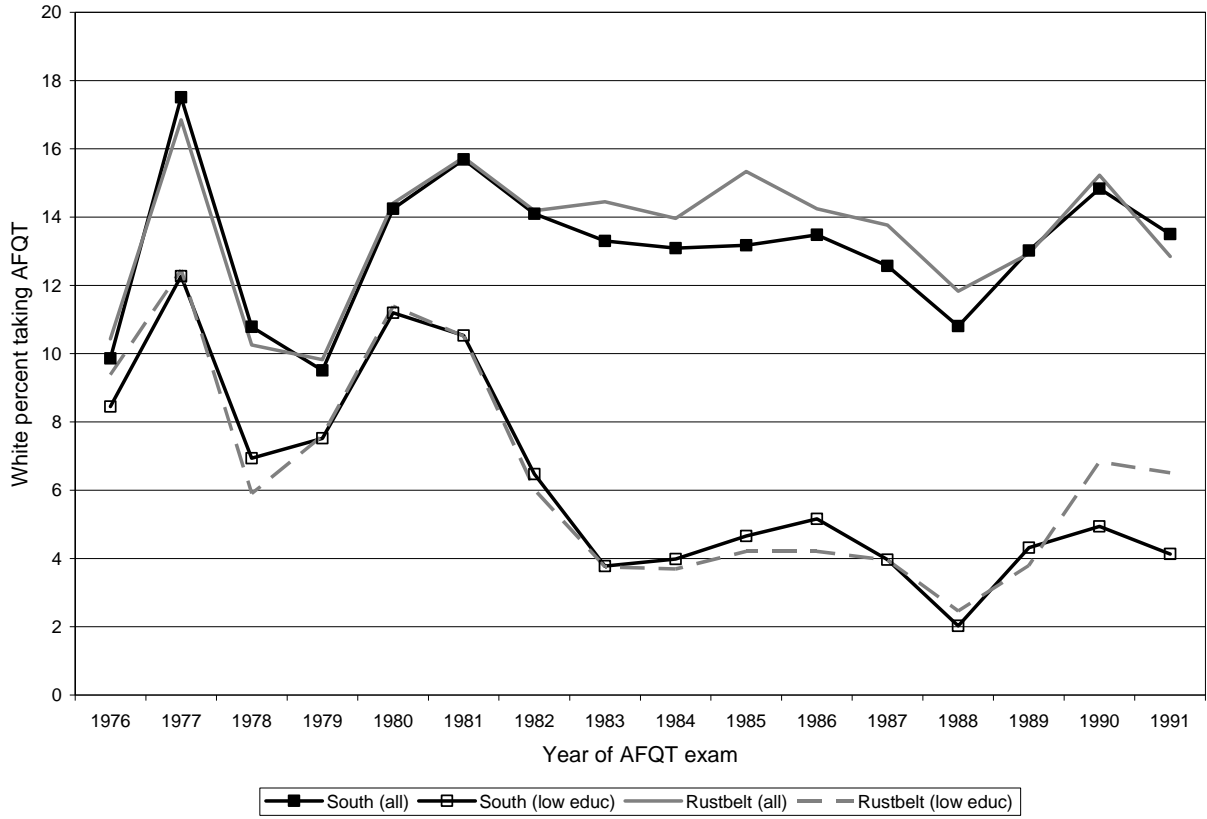


B. Convergence in hospital admission gap after July 1962 to June 1964



Notes: Data come from the 1963, 1964, 1966, 1967, 1971 and 1972 *National Health Interview Surveys*. South consists of Alabama, Arkansas, Delaware, D.C., Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Oklahoma, Texas, Virginia, and West Virginia. North consists of the Northeast ( ) and North Central ( ) regions.

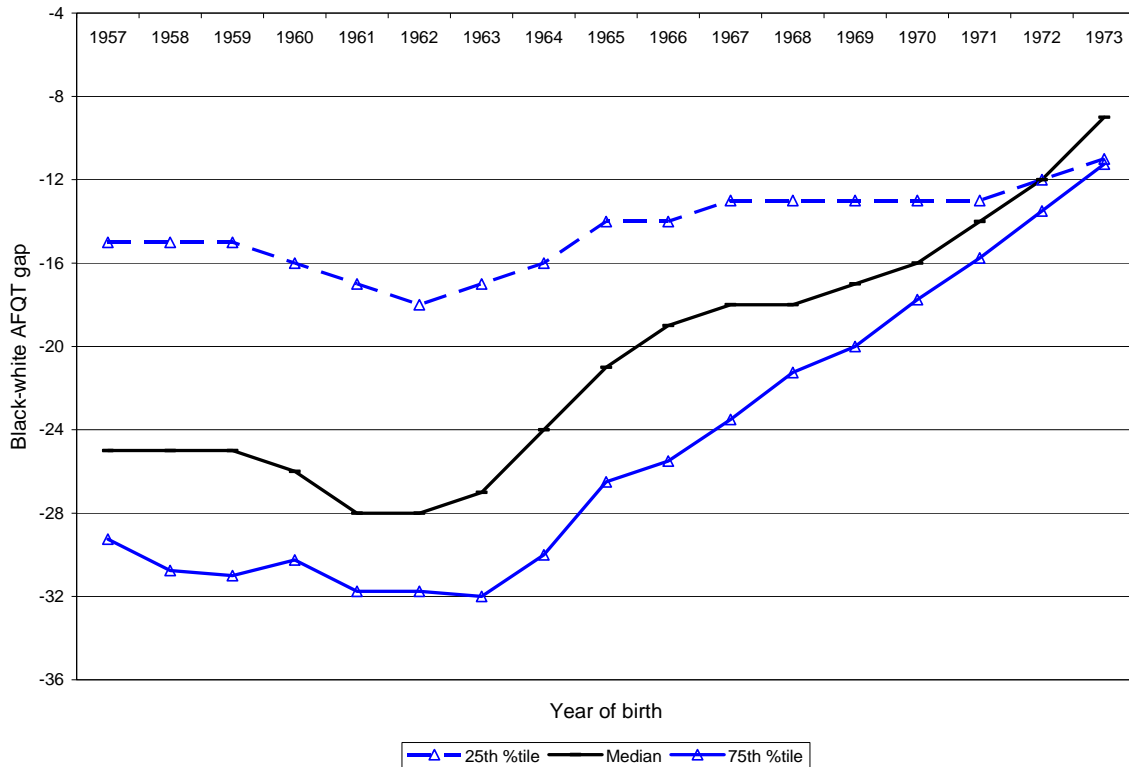
Figure A1: Selection probabilities for white men aged 17 and 18 combined



Notes: Population counts for each state-race-age-year (and education) cell come from the Decennial Censuses. "Low educ" refers to men with two years or less of high school education.

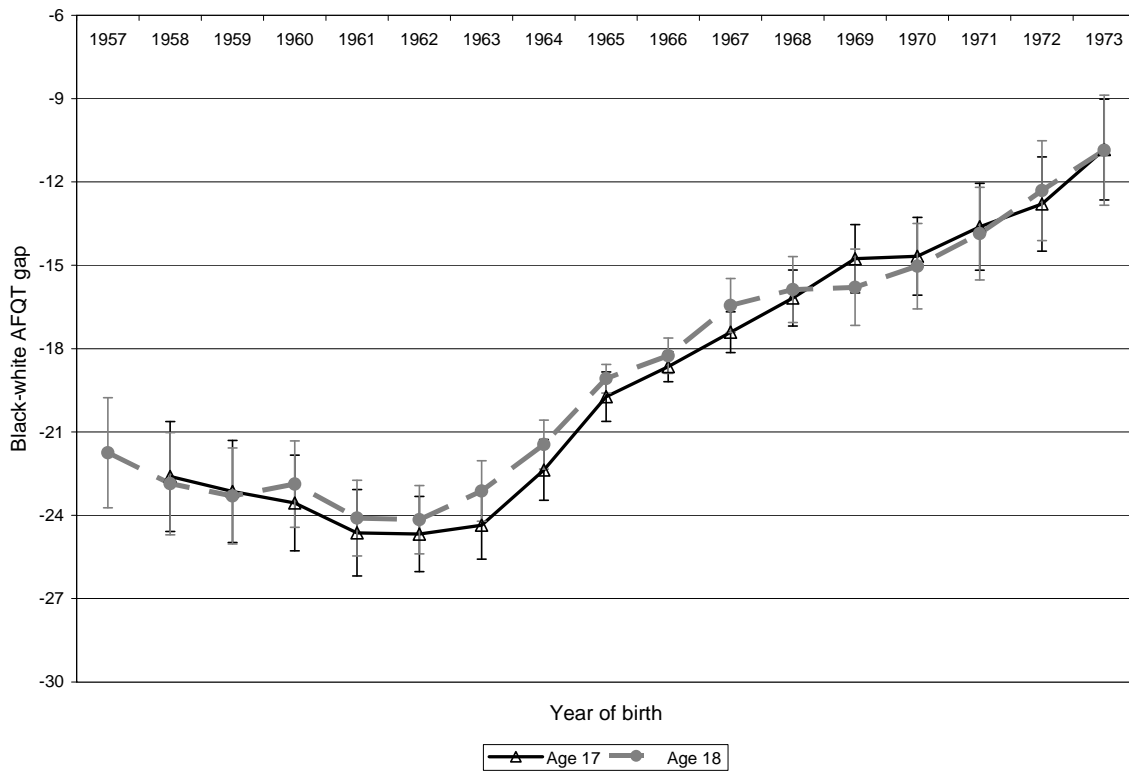
Figure A2: Black-white conditional quantile gap in AFQT scores in South, by birth year

A. Black-white conditional quantile gap in AFQT scores in South



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

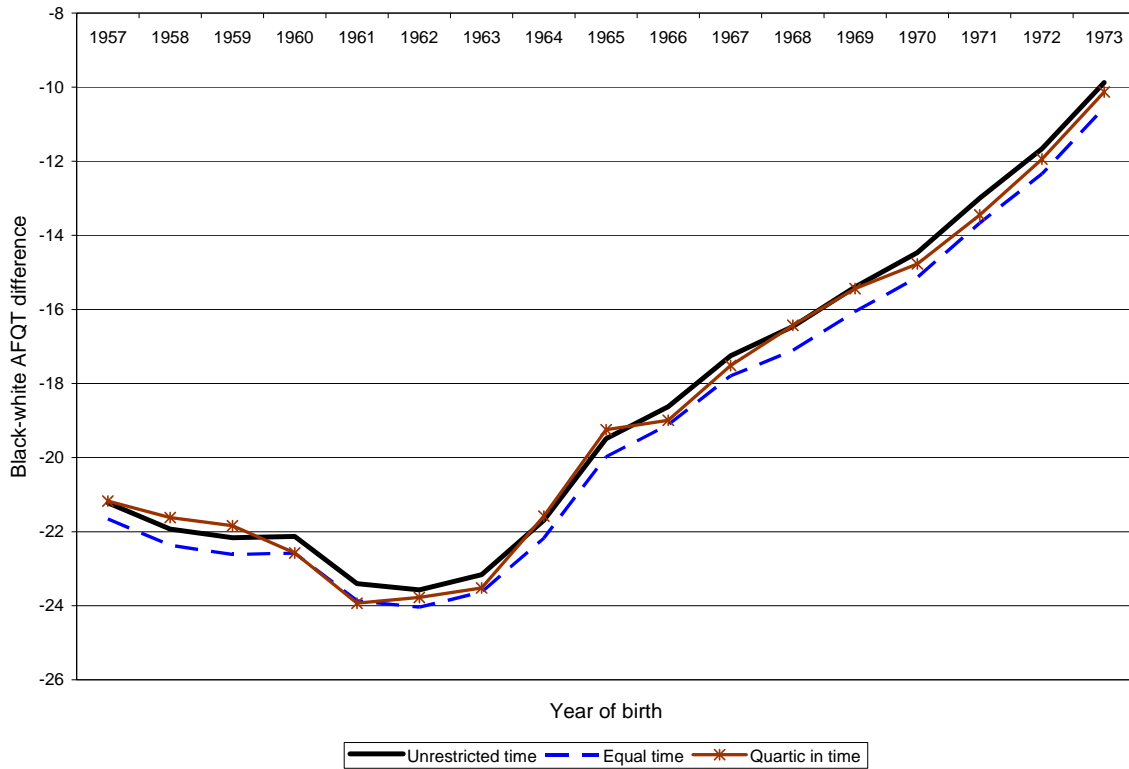
B. Black-white AFQT gap in South, separately for 17 and 18 year-olds



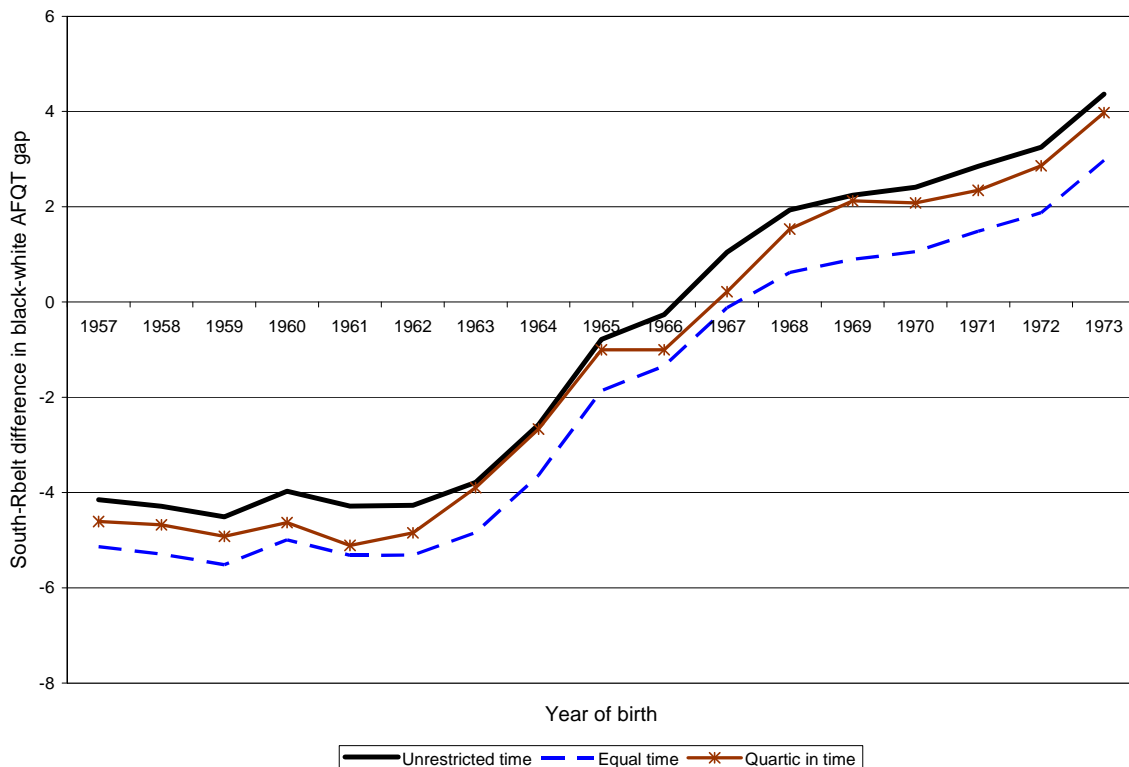
Notes: Plots come from inverse probability weighted (by state births) regressions that allow for unrestricted year and education effects interacted with race, and race-age-cohort interactions. Vertical lines represent  $(\pm)$  twice the standard error of the estimate, corrected for heteroskedasticity.

Figure A3: Estimated cohort-specific AFQT gaps under different time restrictions

A. Black-white gap in South



B. South-Rustbelt, black-white gap



Notes: Plots come from inverse probability weighted (by state births) regressions. “Unrestricted time” model includes unrestricted year effects interacted with race; “Equal time” model restricts the black-white time effect to be the same in 1985 and 1986; “Quartic in time” model imposes a quartic polynomial on the black-white year effects.