NBER WORKING PAPER SERIES

BIRTH COHORT AND THE BLACK-WHITE ACHIEVEMENT GAP: THE ROLES OF ACCESS AND HEALTH SOON AFTER BIRTH

Kenneth Y. Chay
Jonathan Guryan
Bhashkar Mazumder

Working Paper 15078
http://www.nber.org/papers/w15078

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
June 2009

We thank seminar participants at Baylor, Brown, BU/Harvard/MIT joint Health Economics seminar, Cornell, Federal Reserve Bank of Chicago, London School of Economics, NBER Summer Institute, New York Health Economics seminar, Princeton, Tinbergen Institute, University of Chicago, University College London, University of Michigan National Poverty Center, and especially Richard Blundell, David Card, Kerwin Charles, Angus Deaton, Michael Grossman, Ted Joyce, Jens Ludwig, Alan Manning and Steve Pischke for helpful comments. We are grateful to Daeho Kim and Kyung Park for outstanding research assistance and to Jens Ludwig, Douglas Miller and Sam Peltzman for providing their data. All errors are our own. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

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
NBER Working Paper No. 15078
June 2009
JEL No. I12,I18,J15,J24

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

Jonathan Guryan
University of Chicago
Booth School of Business
5807 S. Woodlawn Ave.
Chicago, IL 60637
and NBER
jguryan@chicagobooth.edu

Bhashkar Mazumder
Federal Reserve Bank of Chicago
230 S. LaSalle Street
Chicago, IL 60604
bmazumder@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.[1]

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.[2] 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.

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.

[1] Fryer and Levitt (2004, 2006), Neal (2006), Card and Rothstein (2007), Dobbie and Fryer (2009).

[2] Hanushek (2001), Dickens and Flynn (2006), Cook and Evans (2000), Jencks and Phillips (1998), Neal (2006).

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 these interventions 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.[3] 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.[4]

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 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 conditions under which PNMR will be a useful measure of latent health of a cohort are discussed below, as are the caveats.[5]

[3] See e.g. Johnson (2001).

[4] Neihaus, et al. (2002), Oriá, et al. (2005, 2007). See also Currie, et al. (2008), Currie (2009), and Mendez and Adair (1999). Malluccio et al. (2006) also find long-term effects of a nutritional intervention on cognitive skills

[5] 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.

Consistent with the idea that improved infant health played an important role in the narrowing of the measured racial skill gap, 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.[6] 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 cause of post-birth health improvements on which we focus attention is the increased admission of black infants and children to hospitals following the desegregation efforts of the 1960's. Using newly available data from 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. If hospital integration and increased access are the sole causes of the improved cognitive scores, the magnitudes imply that a black child who gained admission to a hospital early in life had, on average, a 0.7 to one standard deviation gain in their AFQT score as adults 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 at early ages 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 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

[6] As we explain more formally below, a shock to health that affects children ages 0 to A and has long-term effects on cognitive skill development should show up in test scores for birth cohorts born A−1 years before changes are seen in PNMR, which measures the health environment of 0 to 1 year olds. This is the case since the health and AFQT scores of cohorts between 1 and A years old at the time of the health shock are affected, but it is too late for the health shock to affect those cohorts' PNMR's.

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. Section III shows results from the NAEP-LTT data, which match the PNMR trends. Section IV describes the military applicant data that contain AFQT scores, and Section V presents the models used to identify age, cohort and time effects and correct for selection into the AFQT sample. Section VI shows AFQT scores by region, race and birth year, which also match the trends in PNMR. Section VII formally states and tests the early health hypothesis, and shows 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. The final section concludes.

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, as we hypothesize that the test score gains are linked to cohort health soon after birth. While 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 formally lay out the conditions under which these proxies are useful and discuss their caveats. The Appendix provides additional details on the *Vital Statistics of the United States* data used to construct these proxies.

The second half of the 20th 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 over 20 per 1,000 births. After 1964, however, this gap narrowed dramatically – falling over 40 percent to 12-in-1,000 by 1972.

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” regions, 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).[7] 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 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.[8]

Below, we adapt a model from Cunha and Heckman (2007) that formalizes the relationship between each marker of the early health of cohorts and eventual human capital formation. In the model, we argue that PNMR is an especially useful proxy for early health relative to NMR and LBW. Panel E of Figure 1 plots the percent of all infant death that occurs in the neonatal (versus post-neonatal) period, by race and region. As this percentage increases, a lower share of infant death occurs in the post-neonatal period, consistent with an improvement in postnatal health. The figure shows that the white percentage is the same in the South and Rustbelt and is stable over time at 75 to 77 percent – the percentages that prevail today in most of the industrialized world. The black percentage in the Rustbelt region is slightly lower, but also stable over time at roughly 72 percent.

By sharp contrast, the black percentage in the South is significantly lower between 1955 and 1965; hovering between 56 and 58 percent with no improvement over time. Between 1965 and 1972, however, the percentage increases by 12 points and roughly reaches the percentage for blacks in the Rustbelt by 1975. By this metric, there was a significant improvement in the postnatal health of black infants in the South after 1965 relative to their counterparts in the Rustbelt.

III. NAEP-LTT 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

[7] 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. We use these regional groupings when we examine AFQT scores below, where we also examine the “Border states” of Delaware, Kentucky, Maryland, Texas, and West Virginia. In the 1960 U.S. Census, 46 (29) percent of all blacks lived in the South (Rustbelt). Below we also present results separately for each state.

[8] *White* levels of PNMR, NMR and LBW, and their patterns over time, are very similar in the South and Rustbelt. Some have attributed the increase in the LBW gap between 1955 and 1964 to improved reporting as more black infants are born in hospitals. It is instructive, however, that the LBW gap grows in the Rustbelt region, where in 1955 over 95-percent of black infants were born in hospitals with a physician present. As discussed in more detail in the Appendix, the worsening over this period in the characteristics of black mothers relative to whites (in maternal age and marital status) – even more so in the South – is a more plausible explanation.

administered by the U.S. Department of Education aimed at documenting 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. We analyze microdata of students' math and reading scores for all available years in which the tests were administered, for both boys and girls.[9] 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.[10] The Appendix provides further detail on the NAEP-LTT data.

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 regressions that include race-specific age effects.[11] 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 showed no convergence.

In Panel B of Figure 2, we plot the standardized NAEP gaps separately for each age group. 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

[9] 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 (and to maximize sample size).

[10] We show results from scaled scores rather than some other transformation since these scores are reported in public releases of the data, and are used in much of the literature (e.g., Dickens and Flynn 2006, Cook and Evans 2000).

[11] Specifically, we estimate separate regressions for math and reading scores, each of which includes a full set of year effects interacted with race and a full set of age effects interacted with race. The regressions are weighted by sampling weights. Further details are provided in the Appendix.

birth year instead of calendar year. It plots white NAEP scores and the black-white gap by the year of the student's birth (from regressions that adjust for race-specific age effects that vary by subject). White scores show a trend of improvement across successive cohorts born between 1953 and 1989.

The most striking pattern 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 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.[12]

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 the year in which 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.

If the race-by-year effects do not vary by geographic region, however, 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 of blacks in the U.S. South and North 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.[13] 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 greater in the South than in the North with no pattern toward convergence. However, by the early 1970's

[12] 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.

[13] 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.

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.[14]

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 by the early 1970's cohorts that was much larger in the South than North. This suggests that events linked to 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

[14] Applying this specification to Figure 2D for the 1962-64 and 1972-73 birth cohorts results in an initial NAEP gap [standard error] in the South (relative to the North) of –0.353 [0.039] and a relative regional convergence of 0.448 [0.102]. In this comparison the 1962-64 cohort consists of reading and math scores, respectively, for 17- and 13-year olds; the 1972-73 cohort consists of math and reading scores for 9-, 13- and 17-year olds. Applying the model to 13-year olds only – using reading (math) and math (reading) scores, respectively for the 1961 (1972) and 1964 (1974) cohorts and controlling for race-specific subject effects that vary by region – results in an initial regional difference in gaps of –0.160 [0.041] and relative regional convergence of 0.302 [0.061].

counterparts in the North. Further, these findings are inconsistent with a school desegregation narrative as the slight relative loss in scores for 17-year-old Southern blacks occurred in the decade (1970's) in which desegregation should have had its largest impact, and the relative gains occurred for a black cohort whose exposure to white students was lower (relative to the earlier cohort) due to white flight.

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, which is made in most analyses, does not seem innately implausible.

For the purposes of this study, it 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*. The Appendix provides additional details on the military applicant data and the norming of AFQT scores.

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.[15] 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 one's state of birth, this paper presents results from the restricted sample of men aged 17 or 18 at the time of application (see Appendix for further discussion). The estimation results from these two samples are qualitatively the same (and available from the authors).

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 three regions. Due to their relative youth, most 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 data cover the key birth cohorts (and years) in which the NAEP-LTT 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 and across states within a region.

The primary weakness of the AFQT data is that it includes only those individuals who chose to apply to the military. This results in two main sources of selection in the sample. First, the group of

[15] The states included in each geographic region are the same as used in Figure 1 (see footnote 7).

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 sampling.

Second, we observe an applicant's residence at the time of application, but not his 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 effects, 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:

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

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

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 ε is an error term.

Our models allow each effect to vary by race, $r \in (B,W)$. So, $\delta_t^r = \left(\delta_{1976}^r,\ldots,\delta_{1991}^r\right)$ 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 $\gamma_c^r = \left(\gamma_{1957}^r,\ldots,\gamma_{1973}^r\right)$ are the race-specific birth-year dummies. The parameters of interest, $\left(\gamma_c^{(B)} - \gamma_c^{(W)}\right)$, 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 and 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, 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).[16] 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 men of all ages in particular calendar years.

Figure A1 provides evidence on this restriction. It plots the black-white, average AFQT gap by birth year in the South, separately for men aged 17 and 18. These are estimated from regressions similar to (1a) and (1b) that include additive year and education fixed effects interacted with race and are inverse probability weighted (see below). Also plotted are the 95-percent confidence intervals of the estimated race-age-cohort interactions. The figure shows across-cohort convergence in AFQT scores that is very similar by age (after adjusting for race-specific year effects), and implies that restricting the race-by-cohort effects to be the same by age is legitimate.

We also estimated the (race-specific) cohort effects under the assumptions typically used in the literature for contexts unlike ours, where additive age, birth year, and calendar year effects are collinear (see, for example, Deaton 1997). Panel A of Figure A2 shows the racial AFQT gaps by birth year in the South (from inverse probability weighted regressions) from three different models: 1) our preferred model with unrestricted race-time effects; 2) a model that constrains the race-time effects to be the same in 1985 and 1986; and 3) a model that restricts the race-time effects to follow a quartic polynomial. Panel B shows the difference in the racial AFQT gaps between the South and Rustbelt regions, in which each

[16] 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).

model is estimated separately for the two regions. As can be seen in the figures, our results are largely insensitive to the restrictions placed on the race-time effects.[17]

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 in 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. Furthermore, as we show in the next sub-section, the racial gap in the fraction of 17 and 18 year olds who apply to enter the military (i.e., who are in the AFQT sample) track each other closely year to year.

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 of the gap in NAEP-LTT scores illustrated in Figure 2D. Nevertheless, we now describe how we attempt to correct for any remaining sample 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) $$T_{icat} = I_{icat} \cdot T^*_{icat}$$

[17] We also estimated a model, suggested by Deaton (1997), in which the race-time effects vary but are constrained to have no trend. While this model led to similar results, it was statistically rejected in favor of the three models presented, as were models that constrained the race-time effects to be equal in pairs of years other than 1985 and 1986 (though they all led to similar findings).

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

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 (sample 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 $p(\cdot) = \Pr\left(I_{icat} = 1\right)$ 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 sample selection within cells (see, for example, Hirano, Imbens, and Ridder 2003 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 race, cohort and state of birth from the *Vital Statistics of the United States*; and two based on counts of residents by race, cohort and state of residence from the decennial Censuses of the United States – 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 Appendix. Since the results do not vary by the choice of any of these three population counts, we report primarily the results using the births data.

To construct the sample selection probabilities, we divide the number of military applicants in each state-race-cohort-age-year cell by the population size of the cell. This is equivalent to estimating a probability model (e.g., probit or logit) that includes unrestricted state-race-cohort-age-year indicators. We then estimate equations (1a) and (1b), for example, weighting the regressions by the inverse of these probabilities.[18] These probabilities vary along the full interactions of (state, race) cohort, age and time, and will sweep out sampling bias that varies along these dimensions. Table 2 presents the mean and quantiles of AFQT scores, weighted by the inverse sampling probabilities. Compared to the unadjusted means, the IPW-means are higher (more so in the South), which implies slight negative selection in the pool of military applicants. The two sets of means also suggest that this selectivity is partially driven by the overrepresentation of blacks in the applicant pool.

Panels A through D of Figure 3 show the black-white gaps in the probability of being in the sample over time for various comparison groups. In Panel A, which contains plots by age and region, blacks are more likely to apply to the military than whites – even more so in the South – and application rates are more countercyclical for blacks than for whites. Importantly, the racial gap in application rates has similar patterns over time for 17- and 18-year olds, in both regions. Thus, the probability of selection does not exhibit race-age-time interactions within a region; and regression equations (1a) and (1b), which sweep out selection that is additive in race-time and race-age, may be suitable.

Panel B shows a sharp drop in the cumulative application rates of 17- and 18-year old black men in 1982, which may be partially driven by the re-norming of the AFQT (see Angrist 1998 and the Appendix for a discussion). However, the relative drop was similar in the South and Rustbelt regions; black application rates rebounded during the mid-1980s; and black (relative) application rates increased slightly more in the South than in the Rustbelt between 1976 and 1991. This last point implies that the greater racial convergence in AFQT scores in the South found below is not due to a relative decline in the military application rates of black men in the South. Furthermore, the relative drop in applications in 1982 was similar for 17- and 18-year old blacks (Panel A); thus, this episode can be attributed to a secular year effect and not a birth year effect.

One concern for the validity of our estimates would be selection on unobservables. For example, if the application rates of high and low "ability" blacks changed in differential ways *within cells* – and conditional on the state-race-cohort-age-year selection probabilities – then the estimates of the cohort-specific convergence in AFQT scores could be biased. It is difficult to envision an alternative explanation that fits these narrow requirements. Nevertheless, although we cannot provide direct evidence on unobservable characteristics, we can examine application rates by educational attainment.

Panel C of Figure 3 shows the (cumulative) gap in the selection probabilities for 17- and 18-year olds with two years or less of completed high school education. While the application rates of less-educated blacks (relative to whites) fall slightly between 1979 and 1982, they rebound by the end of the

[18] Weighting by the inverse probabilities of taking the AFQT effectively leads to an evaluation of AFQT scores at the same probability of taking the exam (of one) across groups within each cell.

1980's.[19] More importantly, the application rates of less-educated blacks grew more in the South than in the Rustbelt over the period of interest; implying that the greater AFQT convergence in the South shown below is not due to a decline in the application rates of less-educated blacks in the South.

Finally, Panel D shows the differences in the (cumulative) selection gap between Alabama-Mississippi and two other pairs of states: Tennessee-Virginia and Illinois-New York. These comparisons are motivated by comparisons made below in the racial convergence in AFQT scores and PNMR across these pairs of states. There is little difference in the path of selection over time between Alabama-Mississippi and Tennessee-Virginia. Also, the application rates of less-educated blacks in Alabama-Mississippi increase more over time than those of their counterparts in either pairs of states.

VI. Cohort-based racial convergence in AFQT scores

We begin by showing the convergence in the black-white AFQT gaps by birth year and region. Panel B of Figure 4 presents estimates of $\left(\gamma_c^{(B)} - \gamma_c^{(W)}\right)$ for cohorts born between 1957 and 1973 in the South, "Border" and Rustbelt states. The results are from IPW-regressions using equations (1a) and (1b), estimated separately for each region. They are largely insensitive to the states included in each region (e.g., including Texas in the South and Missouri in the Border states).[20]

In the South, the AFQT gap increases slightly between the cohorts born in 1957 and 1963 – from 22 to nearly 24 percentile points (roughly one standard deviation). After the 1963 cohort the gap falls sharply, declining by 50 percent by the 1972 birth cohort. In the Rustbelt, the gap is 18-percent smaller for the late 1950's and early 1960's cohorts. Also, the racial convergence across later cohorts is much smaller in magnitude than in the South, and much more gradual, especially between the 1963 and 1968 birth cohorts. The AFQT convergence across the 1960's cohorts is greater in the Border states than in the Rustbelt, but significantly less than that for the South.

Panel D of Figure 4 plots the South-Rustbelt and Border-Rustbelt differences in the black-white AFQT gap, along with the South-Rustbelt difference in *white* AFQT scores, by cohort. For whites, the levels and across-cohort trends in AFQT scores are similar in the South and Rustbelt, with Southerners scoring roughly 0.5 to one percentile points less. By contrast, the South-Rustbelt difference in the racial AFQT gap follows the pattern described above. The gap is four percentile points larger in the South with no trend toward convergence among those born between 1957 and 1962. Between the 1963 and 1968 cohorts, however, the AFQT scores of Southern blacks increase sharply and are three percentile points

[19] Figure A3 shows the (cumulative) *white* selection probabilities by region and education level. The patterns over time are nearly identical in the South and Rustbelt. Consistent with the re-norming of the AFQT, application rates for the less-educated fall sharply in 1982, though overall military application remains relatively stable (see the Appendix for more discussion).

[20] The estimated black-white differences in the variable effects underlying the figure are provided in Table A2.

higher than their Rustbelt counterparts by the 1971 and 1972 birth cohorts. Blacks in the Border states experience a more general trend of relative improvement in their AFQT scores across birth cohorts, scoring two percentile points lower (higher) than Rustbelt blacks in the 1958 (1972) cohort.

Before proceeding, we discuss the importance of controlling for these year-of-birth effects when examining changes across calendar years in the test score gap between blacks and whites. We find that the overwhelming majority of the racial convergence exhibited during the 1980's (e.g., Figure 2A) is attributable to one's year-of-birth instead of events during the decade. In particular, we estimated the black-white differences in the calendar year effects in AFQT scores, both unadjusted and adjusted for race-specific cohort effects, using IPW-regressions of equations (1a) and (1b).

For the twenty-two states in all three regions, the AFQT gap narrowed by 7.51 percentile points between 1979 and 1989 when birth-cohort driven composition effects are not controlled for. Further, the plotted gaps look very similar to the NAEP-LTT gaps for 17-year-olds shown in Figure 2B. After adjusting for cohort effects, however, the 1980's convergence falls by 91 percent to 0.68 percentile points. Thus, the racial convergence in test scores during the 1980's is almost completely driven by composition effects.

To gauge the magnitudes and statistical significance of the across-cohort convergence in AFQT scores, we estimate the following IPW-regression models:

(3a)
$$\begin{aligned} T_{icat}^{(r),S} = {} & \theta_{pre}^{(r),S} \cdot 1\left(c = 1960 - 62\right) + \theta_{post}^{(r),S} \cdot 1\left(c = 1970 - 72\right) + \gamma_{c}^{(r),S} + \delta_{t}^{(r),S} + \alpha_{a}^{(r),S} + \\ & X_{icat}^{(r),S} \beta^{(r),S} + \varepsilon_{icat}^{(r),S} \end{aligned}$$

(3b)
$$\begin{aligned} T_{icat}^{(r),S} = {} & \theta_{pre}^{(r),S} \cdot 1\left(c = 1960 - 62\right) + \theta_{post}^{(r),S} \cdot 1\left(c = 1970 - 72\right) + \gamma_{c}^{(r),S} + \delta_{t}^{(r),S} + \alpha_{a}^{(r),S} + \\ & \lambda_{at}^{(\cdot),S} + \lambda_{at}^{(r),\cdot} + X_{icat}^{(r),S} \beta^{(r),S} + X_{icat}^{(r),S} \pi_{t}^{(r),S} + \varepsilon_{icat}^{(r),S} \end{aligned}$$

where (r) indexes race, S indexes region (South, Rustbelt), $1(\cdot)$ is an indicator function equal to one if the individual is born between 1960 and 1962 (or 1970 and 1972), and the error terms allow for heteroskedasticity and state-level clustering.

Equation (3a) fits early 1960's and early 1970's cohort averages to the regressions underlying Panel B of Figure 4. Relative to equation (3a), equation (3b) also includes region-specific and race-specific age-by-time effects and race-by-region-by-time effects in the education indicators. The parameters of interest are $\left[\left(\theta_{post}^{(B),2} - \theta_{post}^{(W),2}\right) - \left(\theta_{pre}^{(B),2} - \theta_{pre}^{(W),2}\right)\right] - \left[\left(\theta_{post}^{(B),1} - \theta_{post}^{(W),1}\right) - \left(\theta_{pre}^{(B),1} - \theta_{pre}^{(W),1}\right)\right]$, where $S = 2$ (1) for men in the South (Rustbelt) – that is, the difference-in-differences-in-differences (DDD) estimates of the between-cohort convergence in AFQT scores in the South relative to the Rustbelt.

Table 3 reports results from estimating equation (3a) with education fixed effects [columns (1a) and (1b)] and race-specific education effects [columns (2a) and (2b)] that vary by region. Column (1a) shows that the average racial gap in AFQT scores is 25.8 percentile points (p.p's) for cohorts born between 1960 and 1962 in the South, which is 4.8 p.p.'s greater than the gap for their Rustbelt counterparts. The gap is half as large – reduced by 12.7 p.p.'s, which is 0.54 standard deviations – among Southerners born between 1970 and 1972. This between-cohort convergence is 7.6 p.p.'s greater than that in the Rustbelt, which is a highly significant difference (t-ratio of 6.73).

When the education effects are allowed to vary by race, the results comparing the South and Rustbelt regions change very little. Columns (2a) and (2b) show that the black-white AFQT gap is 4.5 p.p.'s larger among Southern men born in the early 1960's and falls 7.1 p.p.'s more by the early 1970's cohorts. Interestingly, the between-cohort racial AFQT convergence is 3.6 p.p.'s lower in the South and 3.1 p.p.'s lower in the Rustbelt relative to column (1b). One interpretation of this difference is that there was a (regionally secular) convergence in black-white skills between the early 1960's and early 1970's cohorts. For example, a relative improvement in the quality of schools attended by blacks between the beginning and end of the school desegregation era could account for some portion of the AFQT gains in both regions, but for little of the difference in skill convergence between the two regions.

Table 4 presents estimates of the South-Rustbelt difference in the racial AFQT gap for the 1960-62 cohorts and the relative improvement in this double-difference by the 1970-72 cohorts for various specifications. The estimates in columns (1) and (3) are from the same specifications underlying Table 3. The estimates in column (2) are from a model that constrains the race-specific education effects to be the same in the South and Rustbelt; while column (6) presents the results from estimating equation (3b) – the most unconstrained model.

The estimated improvement in the AFQT gap across cohorts in the South (relative to the Rustbelt) is remarkably similar across specifications. Even after controlling for region-age-time and race-age-time effects and education effects that vary by region-race-time, the DDD estimate is 7.1 percentile points (0.30 s.d.'s) and highly significant (t-ratio of 5.85). The stability of the estimates suggests that the larger narrowing of the AFQT gap in the South was not driven by differential selection or compositional changes along observable dimensions. 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.

VII. An Explanation: Relative improvements in black infant health

A potential explanation for the sharp decrease in the black-white AFQT gap that accrued to Southern cohorts born during the 1960's is the large improvement in their early health. We hypothesize that the Southern convergence in the black-white infant health gap explains a significant portion of the cohort-based convergence in the achievement gap that we have shown in both NAEP and AFQT scores. If correct, this hypothesis implies that investments in early-life health have long-term effects on human capital accumulation.

A. Model of health and human capital formation

To organize ideas, consider a model of human capital formation based on Cunha and Heckman (2007). Assume successive cohorts of individuals are born each year. The stock of health and human capital of individuals born in year *c* measured at age *a* is determined according to $\theta_{ca} = f_a(\theta_{c0}, \theta_{ca-1}, I_{ca})$, where *I* is the investment in health and human capital at age *a*. Since investments in health and human capital affect both current and future stocks of θ, the stock of human capital is also a function of the full history of investments as well as the stock at birth – that is, $\theta_{ca} = m_a(\theta_{c0}, I_{c1}, \ldots, I_{ca})$.

Consider a shock to the health of children in their first year of life that begins in year c^* and continues in subsequent years. Due to this shock, an increase in θ_{c1} will occur for all cohorts born in years $c \geq c^* - 1$ (babies born in the year prior to the shock experience gains since some were less than one year old during year c^*). Cohorts born before $c^* - 1$ experience neither the early health improvement nor the human capital increases at older ages. Cohorts born in year $c^* - 1$ and later also experience increases in θ at older ages so long as human capital improvements are semi-permanent and/or they increase the productivity of future human capital investments.[21] This pattern – where one cohort experiences increases in human capital at all ages but a previous cohort experiences no change – is precisely what the estimated cohort effects, $(\gamma_{1957}, \ldots, \gamma_{1973})$, measure.

Now consider a shock to the health of 0 to *A* year-olds that begins in year c^* and continues in subsequent years. For human capital measured at any age $a^H \geq A$, increases in θ_{ca^H} will occur for cohorts born in years $c \geq c^* - A$. For human capital measured at any age $a_L < A$, increases in θ_{ca_L} will occur for cohorts $c \geq c^* - a_L$. These two observations imply that an estimate of *A* can be derived by comparing the earliest cohorts that experience effects of a common health shock on health and human capital measured at different ages. Denote the first cohort for which an increase in θ_a is seen resulting from a particular health shock $\kappa(a)$. We can infer *A*, for example, by comparing $\kappa(1)$ and $\kappa(a^H)$ since $\kappa(1) - \kappa(a^H) = (c^* - 1) - (c^* - A) = A - 1$; thus, $A = \kappa(1) - \kappa(a^H) + 1$. In our context, $\kappa(1)$ and

[21] Cunha and Heckman (2007) refer to these features of human capital formation, respectively, as self-productivity and dynamic complementarity.

$\kappa\left(a^{H}\right)$ are, respectively, the years in which PNMR and AFQT convergence begin. Such a comparison can allow identification of *A*, the oldest age affected by the health shock.

Comparing cohort patterns in AFQT to PNMR, neonatal mortality and low birth weight rates can also allow identification of the *youngest* age affected by the shock. Observing a strong association of $\theta_{ca^{H}}$ (AFQT) with θ_{c1} (PNMR) but not with θ_{c0} (e.g., NMR, LBW) implies that the AFQT gains were caused by a shock to health in the first year of life, but not, for example, *in utero*. Together, these analyses may pin down the ages affected by the early health improvement and the year in which this improvement began.

B. Post-neonatal mortality as an index of early health

In this paper, we use (black-white) post-neonatal mortality rates (PNMR) as a proxy for the underlying early health of a birth cohort (θ_{c1}). With respect to the above model, we use family background variables – such as maternal age, marital status and education – and low birth weight rates as measures of the initial health and human capital endowments (θ_{c0}) of each cohort. Below, we find that all of these measures are worsening more for blacks than for whites across the 1960's, even more so in the South.

Unfortunately, the entire sequence of *true* underlying health for each cohort of blacks and whites at each stage of early life (θ_{c1}, θ_{c2}, θ_{c3},…) cannot be observed. The average stock of health for each cohort at each stage is also not observed. Under certain conditions, however, PNMR can be a useful proxy for this sequence of average health stocks. Suppose a child dies if his latent health at a given age is below a survival threshold. Then the mortality rate of a cohort can decline for two reasons: i) the health distribution improves (shifts right); or ii) the survival threshold falls due to medical innovations or technology diffusion (e.g., neonatal intensive care units). The former dovetails with our model of human capital formation, while the latter will lead to a negative selection effect – i.e., if the additional black survivors are drawn from the lower tail of true health, then human capital measured later in life will decrease for these black cohorts.

In our context, reductions in PNMR may reasonably measure shifts in the health distribution, while declines in neonatal mortality rates (NMR) may reflect negative selection within a cohort. In the second-half of the twentieth century, NMR reductions were primarily driven by technological change and diffusion (Cutler and Meara, 2000). On the other hand, diarrheal and respiratory diseases – the primary causes of PNMR among Southern blacks in the early 1960's – are less likely to be selective on family background and initial health, especially in disadvantaged populations.[22] Also, the health shock that PNMR reductions seem to measure during the period of interest is not a decline in disease incidence, but rather improved health due to increased access to hospital care after disease has struck. For example, we document below significant growth in the hospital discharge rates of black children (and in their birth rates in a hospital with a physician present) during the 1960's.

Suppose the threshold for surviving the post-neonatal period did not change during the 1960's. Then racial convergence in the PNMR gap will reflect a relative shift in the black health distribution and an improvement in the average health of black cohorts.[23] If, for example, latent health is uniformly distributed, then a decrease in black PNMR maps directly into an increase in average cohort health. If health is normally distributed, then a PNMR decline still implies an increase in mean health. However, greater PNMR reductions from a higher initial level will reflect proportionally smaller gains in average health. Below, we find evidence consistent with this possibility.

There are two potential pitfalls with using PNMR as a proxy for average cohort health. First, it is possible that PNMR reductions can reflect negative selectivity within cohorts – for example, Bozzoli, Deaton and Quintana-Domeque (2008) argue that selective survival is more problematic at high PNMR. Second, PNMR could fall if, instead of a secular shift in the health distribution at every quantile, the lower tail (e.g., below the bottom quartile) shifted up. In this case, the average of the distribution would still increase, but by less than implied by a secular shift.

However, both pitfalls work against finding an association between black-white PNMR convergence and convergence in average AFQT scores. Indeed, it is fortunate that post-neonatal mortality rates are measured by race and state during the period of interest. For example, in their absence, analyses might assign test score convergence during the 1980's to health or human capital interventions that occurred in late childhood, without recognizing that PNMR improvements preceded these changes. In addition, black PNMR was relatively high in the early 1960's, suggesting that the subsequent declines pick up shifts in the location of the black health distribution.[24]

Below, we examine the association of PNMR convergence with convergence at different quantiles of the AFQT distribution across states. We find effects that are noticeably larger at the 75th than 25th percentiles, which is not consistent with a shift in the early health distribution at only lower quantiles.

[22] For example, pneumonia and diarrhea are by far the leading causes of child death in the developing world today; with pneumonia alone accounting for more deaths than malaria, AIDS, and measles combined.

[23] This is true even if there are heterogeneous survival thresholds in the population, as long as these "censoring" points are fixed over time and randomly distributed across black (and white) birth cohorts.

[24] We collected data on mortality by race for children between the ages of one and ten. They show that mortality rates for one-year-old black children are an order of magnitude lower than PNMR, and over three-times smaller for three-year-olds than for one-year-olds. Even so, the black-white mortality gap falls for children aged one to three during the 1960's; and in a way that is consistent with relative health improvements that begin after the 1963 birth cohort (results available from authors).

Figure A4 previews this finding by plotting the estimated black-white AFQT gaps by birth year in the South from various IPW quantile regressions. The racial convergence between the 1960-62 and 1970-72 cohorts is 16 and 13 percentile points, respectively, at the 75th percentile and median; but it is only 5 p.p.'s at the 25th percentile. With respect to the model above, this is consistent with a disease incidence that is (relatively) randomly distributed throughout the population of black children in the South and dynamic complementarity with later human capital investment or returns.

Below, we also show that hospital discharge rates grew more during the 1960's for black children *aged between zero and four* in the South than for their white counterparts and for Northern blacks. We argue that this finding is the result of improved access to hospital care for Southern blacks after desegregation. Suppose this greater access began in the second-half of 1966. In the model above, this would cause increases in (I_{c1}, I_{c2}, I_{c3}, I_{c4}) for blacks born in the South in 1966 and afterward; in (I_{c2}, I_{c3}, I_{c4}) for Southern blacks born in 1965; and in (I_{c3}, I_{c4}) for those born in 1964. Southern blacks born in 1962, on the other hand, would experience no increase in their utilization of hospital services.

Almond, Chay and Greenstone (2008) show that the increased hospital access for Southern blacks led to a striking decline in their relative PNMR after 1965, particularly in causes of death amenable to hospital care, such as diarrhea and pneumonia. In the above model, this will lead to improvement in θ_{c1} for blacks born in the South in 1965 and afterward. Given the relative growth in hospital discharge rates, θ_{c2} (θ_{c3}) would increase for Southern blacks born in 1964 (1963) and later. As a result, θ_{ca} at older ages would increase for these respective cohorts, due to either the permanent effects of these investments or dynamic complementarities of "future" input returns with these early investments.

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

We begin the investigation of our hypothesis with Panel A of Figure 4, which shows the black-white gap in post neonatal mortality rates (PNMR) by year in the South, Border, and Rustbelt regions. In the South, there is a sharp reduction in the racial PNMR gap after 1963 and particularly after 1965, with the gap falling from 14-per-1,000 births to roughly 4 by 1974. The PNMR gap in the Rustbelt is very stable between 1955 and 1966 and falls from 6-per-1,000 in 1966 to 4 by 1974. The Border states, on the other hand, show a general decline in the PNMR gap beginning in 1961 that is much smaller than the Southern convergence but larger than that in the Rustbelt.

The corresponding series for racial gaps in AFQT scores – shown in Panel B and described above – mirror the patterns in the PNMR gap. First, the sizes of the regional AFQT gaps in the late 1950's and early 1960's vary in direct proportion to those of the PNMR gaps, with Southern blacks having the largest gaps in both. Second, the patterns of declines in the PNMR gaps across regions are matched by increases in relative AFQT scores when those cohorts reach 17 and 18 years of age. The greatest reductions in the

PNMR and AFQT gaps are in the South; the next largest declines in both are in the Border states; and, in the Rustbelt, the small narrowing of the PNMR gap after 1966 matches the comparatively small increase in black relative AFQT scores for the late 1960's and early 1970's birth cohorts.

These "mirroring" relationships can be seen more clearly in Panels C and D of Figure 4, which plot between-region differences in the racial PNMR and AFQT gaps, respectively. The South-Rustbelt and Border-Rustbelt gap differences are shown along with the South-Rustbelt differences for *whites*. The South-Rustbelt differences in the PNMR gap hold steady between 1958 and 1963 at 8-per-thousand, and fall precipitously after 1963 to roughly no difference in 1974. The corresponding AFQT series remains steady at a four-percentile point disadvantage for Southern blacks born between 1957 and 1962, with a sharp decline after, particularly for those born between 1963 and 1968. Southern blacks born in the early 1970's have a 1-2 p.p. *advantage* relative to their counterparts in the Rustbelt.

The PNMR and AFQT patterns in the Border-Rustbelt differences are less sharp, but still exhibit a strong association. While the relative PNMR gap in the Border states falls systematically between 1961 and 1967, the relative AFQT scores of blacks in the Border states rise most between the 1959 and 1966 birth cohorts. The slower relative improvement in the PNMR gap between 1969 and 1975 matches the slower AFQT gains between the 1967 and 1973 cohorts. The PNMR and AFQT differences between whites in the South and Rustbelt show little change over the entire period; with Southern whites having barely higher PNMR and slightly lower AFQT scores across the birth years.

Taken together, these patterns imply that virtually all of the relative convergence in both the PNMR and AFQT gaps across regions was driven by improvements among blacks. Further, they strongly suggest the mechanism described by our hypothesis and model. Recall that we use a reduction in black PNMR as a proxy for an improvement in early cohort health. The fact that the racial convergence in AFQT scores begins for cohorts born approximately one to two years before convergence in PNMR starts implies that the intervention that led to the PNMR decline positively affected the health of infants between zero and 24 months in age.[25] Subsequent results will show roughly a two-year *lead* – i.e., AFQT gains start among those born two years before PNMR reductions begin – which implies that the driving intervention may have initially improved health for blacks aged 0 to 3 years-old. We investigate one candidate intervention – the integration of Southern hospitals – in detail in section VIII.

A potential concern is that the racial convergence in PNMR and AFQT 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 that in Figure 4 the plotted AFQT gaps have been adjusted for race-specific year and age effects and race-specific education effects that are all allowed to *vary by region*; and the

[25] 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 among babies born in year *t*–1.

AFQT scores are measured 17 to 18 years after the year of birth. Any omitted variable that explains the corresponding convergence in PNMR and AFQT between regions must therefore have three features: (1) it systematically affected the AFQT scores of 18 year-olds a year after it impacted the scores of 17 year-olds, and in a way that differed by race; (2) it caused a narrowing in the black-white AFQT gap 17 to 18 years after it caused a narrowing in the PNMR gap; and (3) it caused each of these features in the South but not in the Rustbelt. We believe that it is difficult to construct a story that satisfies each of these criteria without invoking an intervention that both affects early life health and has long-term consequences for human capital accumulation.

Before proceeding, we note that Table 3 provides a first look at the magnitude of the AFQT-PNMR association by also presenting the reduction in the PNMR gap from 1961-1963 to 1971-1973 in the South and Rustbelt. In column (2b) the ratio of the between-cohort racial AFQT convergence to the black-white PNMR change is –1.10 in the South (and –1.35 in the Rustbelt). The ratio of the AFQT to PNMR convergence in the South relative to the Rustbelt is –1.04. While we avoid giving these ratios a structural interpretation, we find similar magnitudes throughout the below analyses. In addition, we will calculate the ratio of black AFQT gains to the increased hospital admission rates of black children.

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

As demonstrated in Almond, Chay and Greenstone (2008), there was significant variation in the speed and timing of black PNMR reductions across states *within regions*, particularly in the South. The large size of the military applicant data allows for statistically meaningful comparisons of black-white AFQT changes across states. We now test whether the variation across states in the size and (cohort) timing of AFQT convergence matches the variation in PNMR convergence.

We first divided the South region into three pairs of states that had similar patterns of black-white PNMR convergence during the 1960's within each pair: Alabama and Mississippi (ALMS), South Carolina and North Carolina (SCNC), and Tennessee and Virginia (TNVA).[26] Panel A of Figure 5 presents the racial gap in PNMR between 1955 and 1975 for each pair. There are noticeable differences in the patterns across the state pairs. In the late 1950's, the PNMR gap was highest in SCNC, slightly lower in ALMS, and significantly smaller in TNVA. Between 1957 and 1963 the gap is relatively constant in TNVA, but increases in SCNC by roughly 3-per-thousand. While the gap falls substantially in TNVA after 1963, the reductions in SCNC are faster and larger, particularly after 1966. By contrast, the PNMR gap in ALMS rises steadily between 1957 and 1965 – reaching a level similar to that of SCNC – before narrowing precipitously after 1965.

[26] These pairs of states had the most visible differences in the size and timing of PNMR convergence. While this is somewhat ad-hoc, including other pairs of Southern states does not affect the visual impression left by the figures.

Panel B of Figure 5 presents the racial gap in AFQT scores for men born between 1957 and 1973 in each pair of states.[27] The correspondence of these patterns with those in Panel A is striking. For the late 1950's birth cohorts, the AFQT gap is highest in SCNC, followed by ALMS, and smallest in TNVA – patterns that mirror the PNMR gaps in Panel A. While the gap is relatively stable in TNVA for the 1957 to 1962 birth cohorts, it rises by 2.5 percentile points between the 1957 and 1961 cohorts in SCNC. In ALMS the AFQT gap increases by 3.5 p.p.'s through the 1963 cohort. The gap falls steadily in TNVA after the 1962 cohort, but decreases much more rapidly in SCNC after the 1961 cohort (particularly after the 1964 cohort). Just as with the PNMR gap, the sharp decrease in the AFQT gap in ALMS begins two cohorts after the decrease in SCNC.

The abruptness of the black PNMR improvement after 1965 in ALMS make these states an ideal pair to compare to other states both within and outside of the South. Panels C and D of Figure 5 present the differences between ALMS and TNVA and between ALMS and Illinois and New York (ILNY) in the racial gaps in PNMR and AFQT scores, respectively.[28] In Panel C ALMS has a slight widening in the PNMR gap relative to ILNY between 1958 (8-per-thosand) and 1965 (10), but then a sharp and continuous convergence after, reaching near parity by 1974 and 1975. The comparisons of ALMS to TNVA reveal very different patterns. In the late 1950's, the PNMR gap in ALMS relative to TNVA is half as large as when compared to ILNY. From 1958 to 1965, however, there is a greater divergence in the gap for ALMS relative to TNVA of 4-per-thousand. Between 1965 and 1968, there is a sharp relative convergence in the ALMS-TNVA gap of 4-per-thousand, but then none after. Relative to TNVA the PNMR gap for ALMS in the mid-1970's is only slightly below its level in the late 1950's

Panel D presents patterns in the relative AFQT gaps that are remarkably similar to the patterns in Panel C. For the late 1950's birth cohorts, the AFQT gap in ALMS is about half as large when compared to TNVA than to ILNY. Relative to ILNY, the ALMS gap grows slightly between the 1957 and 1963 birth cohorts, but then falls sharply and continuously between the 1963 and 1973 cohorts by 8 to 9 percentile points. Relative to TNVA, the ALMS gap diverges more from the 1957 to 1963 cohorts; falls by 3.5 p.p.'s between the 1963 and 1966 cohorts; and then shows no convergence for later cohorts – indeed, the relative ALMS-TNVA gap is only slightly lower for the early 70's cohorts than for the late 50's cohorts. As with the regional analysis in Figure 4, the black AFQT gains begin among cohorts born roughly two years before the PNMR convergence starts (i.e., a 2-year lead as defined above).

The comparisons in Panels C and D are particularly useful in starting to rule out competing hypotheses to the early health (and hospital desegregation) hypothesis. As we discuss below, the growth

[27] The estimates come from IPW-regressions that are run separately for each state group and include unrestricted age, year and education effects interacted with race.

[28] During the period of interest, Illinois and New York, respectively, contained black populations who disproportionately either migrated from Mississippi and Alabama or whose parents did.

of AFDC expenditures-per-capita and the roll-outs of Food Stamps, Medicaid and Head Start were faster and greater in Illinois and New York than in Alabama and Mississippi after 1964. Thus, there is little prima facie evidence that these programs can explain the sharp improvements in black PNMR after 1965 and in black AFQT scores after the 1963 birth cohort in ALMS relative to ILNY. Further, the between-state-pair comparison within the South rules out any within-region trends in omitted variables.[29] For example, the relative economic position of black adults in ALMS and TNVA improved similarly after 1964, which is inconsistent with the relative PNMR (AFQT) gains between 1965 and 1968 (between the 1963 and 1966 birth cohorts).

Finally, we find no evidence that schooling desegregation can explain either pattern, since: it also impacted cohorts born in the late 1950's and early 1960's; the changes in the "dissimilarity index" are no greater in ALMS than in the other state-pairs after 1964; and the relative black PNMR gains preceded the penetration of schooling integration. In the model presented above, it is possible that each program could magnify the between-cohort AFQT gains through a dynamic complementarity with the early health improvements. In section IX, we discuss each explanation in greater detail.

Table 5 presents difference-in-differences-in-differences (DDD) estimates from IPW-regressions of equations (3a) and (3b); where now the "pre-cohort" is born in 1961-1963 and the "post-cohort" in 1969-1971 – a smaller window than used for the regional DDD comparisons – and separate regressions are fit for each state pair. Columns (1a) to (1c) show that the black AFQT gap narrowed (from pre- to post-cohort) by 5.6 to 6.9 percentile points more in ALMS than in ILNY, which is highly significant (t-ratios between 5.1 and 9.9). The next rows show that PNMR is the only infant health measure that improved more for ALMS blacks relative to ILNY blacks, falling by 5.3-per-thousand more between 1962-64 and 1970-72. Black NMR and LBW increased slightly more in ALMS than in ILNY.

Columns (2a) to (2c) show that the AFQT gap narrowed by 3.1 to 3.5 p.p.'s more in ALMS than in TNVA (t-ratios between 2.9 and 10.3) across the cohorts. Again, PNMR is the only infant health measure that shows a noticeable relative improvement in ALMS. The ratios of the AFQT-to-PNMR convergence are 1.30 and 1.59 in columns (1b) and (2b).

E. PNMR versus earlier measures of infant health

The root cause of the black AFQT convergence affected men born in 1964 and later, particularly in the "Deep" South, which rules out explanations affecting black children of all ages in particular years. Further, the results in Table 5 suggest that the root cause affected black PNMR but not NMR or LBW.

[29] It should be noted that the AFQT gaps in Panel D have been adjusted for *state-pair-specific* time effects that vary by race, as well as race-age and race-education effects that also vary across state pairs. The between-state-pair differencing further removes any race-age-time and race-education-time interactions that vary commonly between the state pairs.

We now directly examine the strength of the association between PNMR and AFQT convergence relative to these two proxies of *in utero* health.

We start with a state-level analysis of the correlation of the relative gains in AFQT scores – from the 1961-63 to 1967-69 birth cohorts – with between-cohort changes in other variables (from 1962-64 to 1968-70). This analysis is based on a two-step procedure. First, we estimate equation (3a) separately for each of the three regions (containing a total of 22 states), and interact the between-cohort differences with state. Second, we regress the state-level DD estimates on between-cohort, black-white differences in the other variables, using the inverse of the estimated variances from the first step as weights.[30]

The results are presented in Table 6 and Figure 6. The second-stage specifications in Table 6 include various explanatory variables across the columns. Column (1) shows a strong association of between-cohort AFQT convergence and PNMR improvements (coefficient of –0.720) that is highly significant (t-ratio of 3.9). PNMR convergence alone explains 52 percent of the variation across states in AFQT convergence, and the estimated constant implies that the AFQT gap fell only 1.4 p.p.'s in states with no PNMR convergence between 1962-64 and 1968-70 (and 5 p.p.'s in states with a PNMR convergence of 5-per-thousand).

Panel A of Figure 6 presents the scatter plot underlying the specification in column (1). It shows a systematic increase in the between-cohort AFQT convergence as the PNMR gap decreases by more. This relation also holds across states within a region with one exception. The Southern states have both greater AFQT gains and larger PNMR reductions for blacks relative to states in the Border and Rustbelt regions; however, the relation across the Southern states is flat.

As we alluded to above, this could be the result of a potential nonlinear relationship between PNMR reductions and improvements in average cohort health. The three Southern states with the largest reductions in black PNMR had higher initial levels in the early 1960's. If the latent health distribution has declining probability densities in the tails, then greater declines in PNMR from a higher initial level will reflect proportionally smaller gains in average health. For example, suppose that post-neonatal health for blacks in 1962 is normally distributed with a mean of zero and a standard deviation of one. Then a PNMR of 2.1 percent implies a survival threshold of 2.31 s.d.'s below the mean, while one of 1.5 percent implies a threshold 2.43 s.d.'s below the mean. Now suppose that black PNMR falls by 9-per-thousand by 1970 in the former case, and by 6-per-thousand in the latter case. This would imply a secular location (and mean) shift in the health distribution of 0.20 and 0.18 s.d.,'s in the former and latter cases, respectively. Thus, while the decrease in black PNMR's is 50 percent greater in the former than latter case, the increase in mean cohort health is only 11 percent higher (and it is easy to construct examples that are less or more extreme).

[30] The estimated variances in both steps are corrected for heteroskedasticity.

Column (2) of Table 6 shows that the bivariate relation between AFQT and NMR convergence has the "wrong" sign as cross-cohort relative gains for blacks in AFQT are associated with *increases* in NMR. Further, NMR explains an order of magnitude less of the variation in AFQT gains than PNMR. The estimated constant implies that states with no reductions in the NMR gap still had AFQT convergence of 4.3 percentile points. Columns (3) through (5) show that the estimated PNMR coefficient is unaffected by controlling for changes in relative NMR, out-migration rates, and the relative high school dropout rates of the mothers of men born in the years of interest; though the latter two variables are marginally significantly related to AFQT convergence and divergence, respectively. Column (6) shows that the PNMR coefficient (and significance) is also unchanged when all of these variables, as well as the state-level racial gaps in Head Start spending per 4-year-old in 1972, are simultaneously included.[31] The control variables are insignificant both individually and jointly.

Panel B of Figure 6 plots the scatter of the "residual" changes in the AFQT and PNMR gaps, each adjusted for the same variables used in column (6) of Table 6. It shows that the residualized changes are even more systematically correlated than the raw relations (R-squared of 0.58 compared to 0.52). Also, the (negative) relations exist across states within each region, including in the South, which suggests that the flatter slope across Southern states in Panel A was partially driven by changes in other factors.

Panel C of Figure 6 plots residual changes in the AFQT gap against residual changes in the racial gap in the probability of being born in a hospital with a physician present – the one measure of hospital access that we can construct at the state-race-year level. Almond, Chay and Greenstone (2008) show that convergence in this variable is highly associated with PNMR convergence in the 1960's. Unsurprisingly, it is highly collinear with changes in the PNMR gap.[32] The (residualized) association between AFQT and hospital birth rate gains for blacks is even stronger (R-squared of 0.61), and appears more consistent with a linear relationship. While not an ideal proxy for *post-neonatal* access to hospitals, the coefficient estimate implies that a 30-percentage point increase in black hospital birth rates from 1962-64 to 1968-70 (the average increase in ALMS) increases cohort AFQT scores by 7.5 percentile points.

Columns (7) and (8) of Table 6, respectively, present the PNMR and hospital birth rate coefficient estimates from specifications that include racial gaps in low birth weight rates (LBW) and

[31] State-of-birth to state-of-current residence migration rates and mother's education – by state, race and birth cohort – were computed using the 1960 to 1990 decennial Censuses (see Appendix for additional details). In-migration rates are highly (negatively) collinear with out-migration rates, and controlling for them has little effect. We adjust for migration rates since the AFQT data provide state-of-residence but not state-of-birth. The Head Start data come from Ludwig and Miller (2007). The Appendix describes how these data were aggregated to the state level.

[32] The cross-state correlation between changes in the racial gaps in PNMR and hospital birth rates is –0.78. A bivariate regression of the change in the AFQT gap on the change in the hospital birth rate gap results in an estimated coefficient (t-ratio) of 0.214 (4.93) and an R-squared of 0.47.

1968 Head Start gaps, in addition to the previous control variables.[33] The PNMR coefficient is even larger than before in both magnitude and statistical significance, and the hospital birth rate coefficient is highly significant as well. Indeed, these are the only two variables that have a systematic relationship with black-white AFQT convergence across cohorts, with nearly all of the other t-ratios below one in absolute value. Further, the estimated constants imply that states with no change in any of the variables had little systematic convergence in AFQT scores, and the adjusted R-squared in column (7) is little higher than the one from the bivariate regression in column (1).

Finally, in Panel D of Figure 6, we show scatter plots of the distribution of AFQT convergence for a longer window – from the 1961-63 to 1969-71 cohorts – against changes in the PNMR gap between 1962-64 and 1970-72. Here, IPW quantile regressions were applied to equation (3a) to estimate the AFQT convergence at the 25th and 75th percentiles. While there is an association at the 25th percentile, the (negative) relations are steeper at the mean and 75th percentile. The plots imply that a state with an 8-per-thousand decline in the PNMR gap had a 6-p.p. greater AFQT convergence at the 25th percentile than a state with no gap reduction, and, respectively, an 8-p.p and 12-p.p. greater convergence at the mean and 75th percentile.

These results imply that both the mean and variance of the black AFQT distribution increased more in states with greater PNMR reductions. They also suggest that PNMR changes may be reasonable proxies for changes in the stock of early health across the entire black population (and not only for those with especially low levels of early health), and that early health may complement later human capital investments (possibly by magnifying their returns).

In order to more precisely examine the correlated timing of the gains in black infant health and test scores, we next estimate the association between the black-white AFQT gaps and the gaps in each infant health measure for every state-year (cohort) observation. The analysis is again performed in two steps. First, we estimate equations (1a) and (1b) separately for each of the regions, with the race-cohort effects interacted with state. Second, we regress the state-by-cohort estimates of the black-white AFQT gaps on the other variables, using the inverse of the (robust) variances from the first step as weights and correcting the standard errors for heteroskedasticity and state-level clustering over time. We use three primary specifications for the second-stage analysis: i) pooled regression; ii) including state fixed effects; and iii) adjusting for state and cohort fixed effects.

[33] A bivariate regression of the change in the AFQT gap on the change in the LBW gap results in an estimated coefficient (t-ratio) of 1.86 (3.20), which is *perversely* signed and highly significant. The cross-state correlation between changes in the racial gaps in NMR and LBW is 0.65.

Table 7 reports the results from the second-stage analysis for a sample of 308 state-cohort observations.[34] Columns (1a) to (1c) present the results from the three specifications in which the same-year PNMR gap and four "leads" of the PNMR gap are simultaneously included as regressors. Column (1b) shows that, conditional on state fixed effects in the gaps, the PNMR coefficients sum to –1.08, with the one- and two-year leads the most significant (t-ratios of 4.1 and 6.0). The PNMR variables explain 80 percent of the variation across states in between-cohort changes in the AFQT gap. Column (1c) shows that after further conditioning on cohort fixed effects, the PNMR coefficients sum to –0.79, with the one- and two-year leads again the most significant. Even after attributing all of the secular between-cohort changes to other factors, the PNMR variables account for 46 percent of the variation in the AFQT gap. Here, the PNMR coefficients are identified off of state-level deviations from the average AFQT convergence across cohorts, which may plausibly be due to early health convergence as well.

Columns (2a) to (2c) and (3a) to (3c) report the results from the same specifications applied to the NMR and LBW gaps, respectively. In sharp contrast to the PNMR results, the coefficient estimates are highly sensitive to the specification. When conditioning on state fixed effects, the NMR coefficients [column (2b)] sum to –1.11 and the one-and two-year leads are the most significant. After further adjusting for cohort fixed effects [column (2c)], all of the NMR coefficients have a "perverse" *positive* sign, and the three- and four-year leads are statistically significant. Also, in these two models the NMR variables explain 26 and 9 percent of the (conditional) variance in AFQT gap changes – 3 and 5.5 times less than the analogous PNMR models. With some differences in the details, the patterns in the LBW estimates are largely similar; and the model that conditions on state and cohort effects leads to significant, *positive* coefficient estimates for each of the LBW variables.

Table 8 presents the results when the one- and two-year leads of each of the infant health measures are simultaneously included as regressors in the second-stage analysis. To examine the sensitivity of the estimates to family background, we also control for racial gaps in maternal marital status and distribution across nine age categories – also constructed from the *U.S. Vital Statistics* – for every state, race and year in which the data are available (see the Appendix for details). We show the estimates from three different regressions for the pooled [columns (1a) to (1c)], state effects [(2a) to (2c)], and state and cohort effects [(3a) to (3c)] models: i) the full sample of 308 state-year observations; ii) the reduced sample for which the maternal background variables are non-missing; and iii) the reduced sample with the maternal variables included as controls.

[34] We limit the analysis to the 1959 through 1972 birth cohorts in each of the 22 states. The results are unchanged if we use all cohorts between 1957 and 1973; but Table A1 shows that the 1959 to 1972 cohorts have the most complete coverage of 17- and 18-year-old military applicants in our sample.

Focusing first on the pooled analysis, the most significant coefficient in both samples is for the 2-year PNMR lead (t-ratios of 6.9 and 5.3) followed by the 1- and 2-year NMR leads (t-ratios of 2.6 to 3.0). The estimated constants in these regressions imply that a state with no racial gaps in the 1- and 2-year leads of PNMR, NMR and LBW would have a racial gap in AFQT scores of 7.4 to 8.0 percentile points, which is 2.5 times smaller than the average (IPW) gap in the sample (18.5 p.p.'s). Column (1c) shows that the maternal background variables are highly jointly significant, increasing the R-squared from 0.49 to 0.71. Further, while the PNMR coefficient is virtually unchanged in magnitude or significance, the sizes of the NMR coefficients fall by a factor of three to four and are no longer significant (though still precisely estimated). This finding arises from the fact that maternal background is highly predictive of the racial gap in NMR but not of the gap in PNMR, and is consistent with our use of PNMR as a proxy for early health rather than selective survival, whereas the opposite may hold for NMR.

The results in the remaining columns tell a clear and cohesive story. When conditioning on state or state and cohort fixed effects, the 1- and 2-year PNMR leads are highly significant – summing to –0.96 and –0.65 in the former and latter cases – and insensitive to adjusting for mother's characteristics. The NMR and LBW coefficients, by contrast, are virtually never significant and often have the perverse sign. Further, not only does adjusting for state and cohort fixed effects greatly reduce the magnitudes of the NMR coefficients, it also reduces the joint significance of the maternal variables – that is, state and cohort are highly correlated with racial gaps in family background but not in a way that is collinear with changes in the PNMR gap across states. Taken literally, the results imply that *post-neonatal* health, and not health *in utero*, drove the rapid racial convergence in AFQT scores across cohorts born during the 1960's.

It is noteworthy that adjusting for both state fixed effects and state-specific cohort trends (instead of secular cohort effects) leads to similar results, with the first and second leads of PNMR being the most significant (negative) predictors of cross-cohort changes in the AFQT gap. Adding cohort fixed effects to this model asks a great deal from the data – e.g., state and cohort FE's and state-specific cohort trends explain 94.2 percent of the variation in the AFQT gap – but the PNMR variables are the only infant health measures with negative coefficients that are jointly significant. There is almost no variation across states in AFQT changes for *whites* that can be separated from secular cohort effects – e.g., state and cohort FE's alone explain 98.8 percent of the variation in white scores. Thus, in models that adjust for both effects, none of the white infant health coefficients are both negative and significant. Further, in models that control for state effects, adjusting for the maternal background of white cohorts substantially reduces the estimated effects of the white infant health measures (by a factor of four).

VIII. A candidate cause of black gains in the South: Hospital Integration

We have established the strong relationship between declines in the black post-neonatal mortality rate in the South and gains in AFQT scores measured 17 to 18 years later for Southern blacks born in the impacted cohorts. A possible cause of both improvements is increased access to hospital care in the first years of life. Almond, Chay and Greenstone (ACG 2008) argue that the integration of segregated hospitals in the South – through a combination of Title VI of the 1964 Civil Rights Act and the financial incentives engendered by the 1965 Medicare Act – caused the decrease in black PNMR.

In segregated hospitals, there were separate waiting rooms and patient wings for blacks and whites; 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. ACG show that access to hospital care – measured for example by the fraction of births that took place in hospitals – increased significantly for Southern blacks after integration, and that this increase improved post-neonatal health. ACG find that reduced death due to diarrheal dehydration and pneumonia – both of which were easily treated at the time by hospital admission but extremely threatening conditions if left untreated – accounted for the vast majority of the black PNMR improvement in the South.

We have used PNMR as a proxy for early life health more generally. If access to the medical care available in hospitals is the root cause of the black infant health improvement proxied by a PNMR decline, it seems likely that the range of underlying mechanisms is larger than receiving hospital treatments for just diarrhea and pneumonia. In Panel C of Figure 6 and column (8) of Table 6, we demonstrated the strong association between cross-cohort reductions in the AFQT gap and convergence in one measure of the racial gap in hospital access – the black-white difference in the fraction of babies born in a hospital with a doctor present. The estimates implied that the average increase for Southern states in black hospital birth rates between 1960 and 1970 (25 percentage points) resulted in AFQT gains of 6.4 percentile points across entire cohorts of black men. Further, the hospital birth rate was the only variable other than PNMR that was highly associated with the AFQT convergence.

If one assumes that the correspondence between PNMR and AFQT is driven entirely 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 hospital care because there was no increase in hospital utilization at older ages? To address these questions, we examine data from newly released waves of the *National Health Interview Survey* (*NHIS*). As a consequence, we also show that the growth in hospital birth rates is related to greater admission to hospitals after birth.

The *NHIS* asks respondents whether they were admitted to a hospital during the past year (see the

Appendix for details). Figure 7 shows admission rates for black and white boys in the South and North (Northeast and North Central regions). Panel A shows the Southern black-white gap in hospital admissions by age for the periods just before and after the Civil Rights Act (*CRA*) – July 1962 to June 1964 and July 1965 to June 1967 – as well as the racial gap in the North averaged over both periods.[35] It shows that before the *CRA*, the Southern gaps in admissions rates were substantial. While they are similar to the Northern gaps for boys between the ages of five and 18, they are 3-, 4-, 7-, 5- and 4-per-100 lower for those aged between zero and four, respectively. In the first years after the *CRA*, the regional differences in the admissions gaps for these children have been nearly eliminated.

Panel B shows the age-specific convergence after July 1962 to June 1964 in the hospital admissions gap by July 1965 to June 1967 in the South and by January 1971 to December 1972 in both regions. There was significant growth in black admissions rates in the South just after the *CRA* for those aged four and under, and the Southern racial convergence for these ages continued through the early 1970's. By contrast, the racial gap in Northern admission rates changed little between the early 1960's and early 1970's, particularly for boys under the age of five.

These findings, when related to the AFQT-PNMR results above, imply larger, long-term cognitive returns to hospital access for children up to age three. While there is improved access to hospital care for blacks under the age of five, the PNMR and AFQT patterns imply that improved access at age four had little effect on black AFQT scores among 17- and 18-year olds. We found that gains in AFQT scores for Southern blacks (relative to Northern blacks) were particularly large between the 1963 and 1965 birth cohorts (Figure 4D). Based on Figure 7, however, Southern blacks born in 1962 utilized hospital care at a significantly greater rate as four-year olds than their counterparts born before 1962. It appears this greater hospital utilization did not translate into higher AFQT scores later in life.

Unfortunately, the available data cannot pin down the mechanism behind these age-dependent returns to hospital care. On one hand, the differential returns could stem from an increased risk at very young ages of catching diseases that have long-term cognitive effects. On the other, brain development may be more sensitive to the treatment of health conditions at these ages. Distinguishing between these two (and other) hypotheses is important and should be pursued in future research.

We can calculate the cost of the AFQT gains that accrued to Southern blacks born in the 1960's, under the assumption that they were entirely the result of increased admission to hospitals. For example, when we fit equation (3a) using the 1961-63 and 1965-67 as the "pre" and "post" cohorts, we estimate that the Southern racial gap in AFQT scores fell by 3.6 percentile points between these cohorts. Figure 7B implies that the cumulative racial gap in (postnatal) hospital admissions for Southern blacks aged three and under fell by 16-per-100 soon after the *CRA* and by 20-per-100 by the early 1970's. These

[35] The racial gap in admission rates in the North exhibits little difference (by age) in the two periods.

numbers imply that a black child who gained admission to a hospital early in life had, on average, a 0.75 to 0.95 standard deviation gain in their AFQT score relative to a counterpart who was denied admission. These gains fall to 0.5 to 0.6 s.d.'s if the 10-percentage point convergence between these cohorts in the hospital births rates for Southern blacks are included in the denominator.

The *NHIS* data also contain information on the length of the hospital stay and the costs incurred during the stay. For the period of interest, the average length of stay for children under age-four was six days, and the average cost varied between $1,000 and $2,000 (in 1982-84 dollars) depending on the year. These numbers imply that children of these ages who were admitted to a hospital suffered from serious conditions that required treatment for one week, on average. They also imply that it cost roughly $1,500 to provide treatment to a black child that might directly or indirectly (through dynamic complementarity) raise his AFQT score by about 20 percentile points as a 17 or 18 year old.

IX. Alternative explanations for black gains in the South

We have presented evidence that improvements in early health caused test score gains later in life for Southern blacks, and that both were the result of the integration of Southern hospitals during the 1960's. However, there are several alternative hypotheses. The cohorts that experienced the racial convergence in test scores had childhoods in which a number of public policies may have benefited blacks relative to whites. Any explanation must reconcile: 1) the greater cross-cohort convergence in AFQT scores in the South than the North that was concentrated between those born between 1963 and 1968; and 2) the different timing across states within the South in the cohorts most affected.

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

As part of President Lyndon Johnson's War on Poverty, several social programs were initiated in the mid-1960's. Since the programs were aimed at helping the poor, many of them benefited blacks more than whites. If Food Stamps, AFDC, or Medicaid caused the narrowing of the black-white test score gap, an improvement in early health seems likely to be part of the mechanism. Each was aimed at either helping poor families buy sufficient supplies of food or subsidizing their medical care. Each also had an income effect, so it would be difficult to rule out direct effects of other expenditures.

The prima facie case for these programs, however, is weak. 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.[36] For Medicaid to be a valid hypothesis, only its health impact on 5- and 6-year olds

[36] Medicaid adoption in the Rustbelt: IL (Jan. 66), NY (Oct. 66), PA (Jan. 66), OH (July 66), MI (Oct. 66), MO (Oct. 67), IN (Jan. 70). In the 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 the Border states: KY (July 66), WV (July 66), MD (July 66), DE (Oct. 66), TX (Sept. 67).

could have long-term consequences, but not its health benefits for younger (or older) children.

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 counties 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).

Future research testing the long-term effects of these programs is warranted. Though the early evidence suggests that they played little role in the racial convergence in infant health and test scores, it is important to more thoroughly examine their potential contributions. For example, it is possible that the improved early health of Southern blacks born after 1963 interacted with Food Stamps, AFDC and Head Start in ways that increased their marginal effects on later test scores.

Changes in black relative earnings and parental investments in children

Another hypothesis is that the parents of black children born in the South during the 1960's earned more than their predecessors; leading to better nutrition, healthcare, and more human capital accumulation for their children. Some of the causes of black economic progress were too gradual to account for the sharp test score convergence among the mid-1960's cohorts.[37] There is evidence that the Civil Rights Act of 1964 had a more immediate effect (Chay 1995, Donohue and Heckman 1991, Heckman and Payner 1989). However, our own analysis based on the merged data of Social Security earnings records to the Current Population Survey used in Chay (1995) indicates no notable differences *across states within the South* in the size or timing of gains in the relative log-earnings of black men after 1964. Further, even if black earnings gains caused some of the AFQT gains, parental earnings could only have had a large impact on very young children. If earnings affected the human capital accumulation of older children, then AFQT gains should have accrued to blacks born before 1960 as well.

A relative improvement in black wages could have signaled to black parents that the return to investing in their children's human capital had increased (Neal 2006). While direct evidence on changes in investments by black parents during the 1960's is unavailable, Neal and Johnson (1996) find that racial differences in pre-labor-market conditions have important effects on earnings gaps. Such a story is hard to reconcile with the facts, though. The sharp reduction in the black-white test score gap across cohorts implies that parents changed their expectations of the investment returns and acted on it with implausible

[37] Black education levels had increased for decades prior to the 1960's, and the observable quality of black schools had been gradually improving since the early part of the century (Margo 1990, Card and Krueger 1992).

speed (and that investment only matters in the first two years of life). Further, it is not clear this could explain the difference across Southern states in the timing of the AFQT convergence.

School desegregation

School desegregation has been posited as an explanation for the racial convergence in test scores during the 1980's (e.g., Grissmer 1998). A cursory examination of the timing of school desegregation in large urban school districts, particularly in the South, seems to make a plausible case. 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, schools in the Rustbelt integrated both before and after 1968. If desegregation were the root cause of the black test score improvements we document in the South, we should have seen much larger increases in black test scores in the Rustbelt. Second, based on an analysis of U.S. Department of Education Office of Civil Rights data we find that districts either integrated all grades at once or started with the older grades and later moved to younger grades. Desegregation in 1968 therefore largely affected children in grades K-12, who were born between 1950 and 1963. Though it is reasonable that school integration's effects are cumulative, it is hard to believe that attending integrated schools starting in Kindergarten (the 1963 birth cohort) versus 1st grade (the 1962 birth cohort) have drastically different effects on cognitive skill accumulation.

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 it. This paper has documented an important set of facts that can guide future research in the search for root causes. We have shown that the narrowing of the black-white test score gap that occurred in the 1980's is better understood as an improvement by successive birth cohorts of blacks, rather than one that affected blacks of all ages during that decade. The cohort-based convergence began fairly suddenly with those born in 1963, and is most apparent in the South. This cross-cohort convergence opens the set of potential explanations to those that occurred well before the convergence in test scores was observed.

We test one such explanation – that an improvement in the early health of Southern blacks had long-term effects on the human capital accumulation for the cohorts that experienced this improvement. We use declines in post-neonatal mortality as a proxy for an improvement in the average (latent) health of a cohort. We show that the timing of black PNMR reductions matches the timing of black AFQT gains, measured 17 to 18 years later, remarkably well. The results imply that improved health in the first 3 years of life has long-term effects on human capital accumulation and explains a significant portion of the narrowing of the black-white test score gap during the 1980's.

We then turn to a possible explanation for the gains among Southern blacks: the racial integration of Southern hospitals during the 1960's. We demonstrate a strong relationship between black access to hospitals and the black-white test score gap. One set of calculations suggests that a black child who gained admission to a hospital early in life had, on average, a 0.75 to 0.95 standard deviation gain in their AFQT score relative to a counterpart who was denied admission.

Hospital integration was not the only change during the 1960's that may have disproportionately benefited Southern blacks. While additional research is warranted, we find little evidence in support of the most likely competing hypotheses. Further, to the extent that any of these explanations caused convergence in the test score gap, they likely worked through their effects on health and human capital accumulation at early ages. 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 that early investments 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). Moreover, Almond and Chay (2006) find that the racial convergence in post-neonatal health during the 1960's is associated with convergence in later health as adults and in the health of the infants in the next generation. Future research should examine the sequence of investments made in response to the early health gains of these cohorts and attempt to distinguish the effects of more investment, greater returns on investment and permanent health improvements on the improved human capital and health of black adults.

References

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

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.

Angrist, Joshua D. (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants," *Econometrica* 66(2): 249-288.

Bozzoli, Carlos, Angus Deaton and Climent Quintana-Domeque (2008). "Adult Height and Childhood Disease," *Demography* forthcoming.

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.

Card, David and Jesse Rothstein (2007). "Racial Segregation and the Black-White Test Score Gap," *Journal of Public Economics*, 91: 2158-2184.

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.

Cunha, Flavio and James J. Heckman (2007). "The Technology of Skill Formation," *IZA Discussion Paper No. 2550*. January.

Currie, Janet (2009). "Healthy, Wealthy and Wise: Socioeconomic Status, Poor Health in Childhood, and Human Development," *Journal of Economic Literature* 47(1): 87-122.

Currie, Janet, Mark Stabile, Phongsack Manivong and Leslie L. Roos (2008). "Understanding the Relationship Between Child Health and Long-Term Socioeconomic Status," mimeo. Columbia University.

Cutler, David M. and Ellen Meara (2000). "The Technology of Birth: Is It Worth It?" *Forum for Health Economics and Policy: Frontiers in Health Policy Research* 3(3): 33-67.

Deaton, Angus (1997). *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Baltimore: Johns Hopkins University Press.

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

Dobbie, Will and Roland G. Fryer, Jr. (2009). "Are High-Quality Schools Enough to Close the Achievement Gap? Evidence from a Bold Experiment in Harlem," Harvard University, http://www.economics.harvard.edu/faculty/fryer/files/hcz%204.15.2009.pdf.

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.

Fryer, Roland G. Jr. and Steven D. Levitt (2004). "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, 86(2): 447-464.

Fryer, Roland G. Jr. and Steven D. Levitt (2006). "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review*, 8(2):249-281.

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., Washington, DC: Brookings Institution Press.

Hanushek, Eric A. (2001). "Black-White Achievement Differences and Governmental Interventions," *American Economic Review Papers and Proceedings of the American Economic Association* 91(2): 24-28.

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.

Hirano, Keisuke, Guido Imbens and Geert Ridder (2003). "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score," *Econometrica*, 71(4): 1161-1189.

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

Jencks, Christopher and Meredith Phillips (1998). *The Black-White Test Score Gap*, Washington, DC: Brookings Institution Press.

Johnson, Mark H. (2001). "Functional Brain Development in Humans, *Nature Reviews Neuroscience* 2: 475-483.

Ludwig, Jens and Douglas Miller (2007). "Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design," *Quarterly Journal of Economics* 122(1): 159-208.

Maluccio, John A., John Hoddinott, Jere R. Behrman, Reynaldo Martorell, Agnes R. Quisumbing and Aryeh D. Stein (2006). "The Impact of Nutrition during Early Childhood on Education among Guatemalan Adults," Penn Institute for Economic Research Working Paper 06-026.

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

Mendez, Michelle A. and Linda S. Adair (1999). "Severity and Timing of Stunting in the First Two Years of Life Affect Performance on Cognitive Tests in Late Childhood," *Journal of Nutrition* 129(8): 1555-1562.

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.

Neal, Derek A. and William R. Johnson (1996). "The Role of Premarket Factors in Black-White Wage Differences," *Journal of Political Economy* 104(5): 869-895.

Neihaus, Mark D., Sean R. Moore, Peter D. Patrick, Lori L. Derr, Breyette Lorntz, Aldo A. Lima and Richard L. Guerrant (2002). "Early Childhood Diarrhea is Associated with Diminished Cognitive Function 4 to 7 Years Later in Children in a Northeast Brazillian Shantytown," *American Journal of Tropical Medicine and Hygiene* 66(5): 590-593.

Oria, Reinaldo B.,Peter D. Patric and H. Zhang (2005). "APOE4 Protects the Cognitive Development in Children with Heavy Diarrhea Burdens in Northeast Brazil," *Pediatric Research* 57: 310-316.

Oria, Reinaldo B., Peter D. Patrick, James A. Blackman, Aldo A.M. Lima and Richard L. Guerrant (2007). “Role of Apolipoprotein E4 in Protecting Children Against Early Childhood Diarrhea Outcomes and Implications for Later Development,” *Medical Hypotheses* 68: 1099-1107.

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

Appendix

A. United States Vital Statistics Data

The state-level infant health data come from the *Vital Statistics of the United States* (*VSUS*) publications from 1955 to 1975 (National Center for Health Statistics). Drawn from standard certificates of live birth, death, and fetal death, these data cover the universe of births and deaths in the United States.

From the *VSUS*, the primary data used are at the state level. For live births, we have white and nonwhite counts of total births; births by attendant (physician in hospital, physician not in hospital, midwife); and births of 2,500 grams or less. We also have white and nonwhite counts of deaths under one year of age (infant death); under 28 days (neonatal death); and fetal deaths. The infant and neonatal mortality rates are based on the ratio of the number of deaths to the number of live births; and the post-neonatal mortality rate is the ratio of the difference between infant and neonatal deaths to the number of births. The hospital birth rate is the ratio of births attended by a physician in a hospital to total births; and the low birth weight (LBW) rate is the proportion of births that are 2,500 grams or less.

In the *VSUS*, nonwhite births are available at the state level for the separate categories of "Negro," Native American, Chinese, Japanese, and "other" races; and by infant gender within racial category. Infant deaths (by age at death) are available at the state level for the categories white, "Negro" and "other" (also by infant gender). We use these data to calculate the fraction of nonwhite births in a state that are black. These data are also used to verify our findings – the results are unchanged if we use "Negro" births and post-neonatal deaths; the patterns for boy and girl "Negro" infants are similar to those for nonwhite infants. We focus on the 22 states in which 95 percent or more of nonwhite births are black over the period studied.

In order to control for mother's age and marital status, we construct measures of the black-white difference in these characteristics for each state for each birth cohort based on data derived from historical volumes from Vital Statistics. For mother's age we calculate the fraction of live births in each state to women in the following age categories: less than 15, 15 to 19, 20 to 24, 25 to 29, 30 to 34, 35 to 39, 40 to 44, 45 to 49 and 50 and over. For 1960 to 1963 the data are for "non-whites". For marital status we use the rate of illegitimate births (per 1000). This is reported directly for whites and non-whites from 1969 to 1973. For 1957 to 1968 we calculate these rates based on counts of the number of births and the number of illegitimate births for whites and non-whites. The racial gaps in LBW and in rates of birth for teenage and unmarried mothers grew during the 1960's, even more so in the South.

B. National Assessment of Educational Progress Data

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.

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. We analyze the scaled scores didvided by their standard deviation across the entire United States by age, subject, and year. The results in Table 1 and Figure 2 are insensitive to including state fixed effects and examining scaled scores (or the natural logarithm of scaled scores) instead of standardized scaled scores.

C. Armed Forces Qualifying Test Data

Our AFQT sample contains the percentile test scores of the universe of applicants to the United States military between 1976 and 1991. This data was previously used by Peltzman (1993) and Murphy and Peltzman (2004) and is described in detail by Peltzman (1993). The full sample contains male and female applicants between the ages of 16 and 28. In the analysis, results are reported for samples of applicants ages 17-20 and 17-18. The AFQT percentile scores are normed relative to the nationally representative sample called the *Profile of American Youth* 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. Consistent with the re-norming of the AFQT, application rates for the less-educated fall sharply in 1982, though overall military application remains relatively stable.

Peltzman, Sam (1993). "The Political Economy of the Decline of American Public Education." *Journal of Law and Economics*, 36 (April): 331-70.

Murphy, Kevin and Sam Peltzman (2004). "School Performance and the Youth Labor Market." *Journal of Labor Economics*, 22 (2): 299-327.

Estimates from the 5-percent samples of the 1980 and 1990 Censuses show that the likelihood that a person lives in a different state than he was born in rises sharply at the age of 19, and this increase varies by race. As a result, for most of the analysis we restrict our sample to those who were 17 or 18 years old when they took the AFQT.

<u>*D. Population counts by cell and constructing Inverse Proability Weight*</u>

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 × state of residence × year × birth year × completed education at time of application).

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

E. Construction of cohort-level variables from Decennial Censuses

Because the AFQT data indicate state of residence and not state of birth, a natural concern is that the results are driven by inter-state migration patterns. Motivated by increasing skills among blacks in the South, a second concern is that black children born in the late 1960's in the Deep South are born to mothers with more human capital.

In order to control for migration out of one's state of birth and whether the mother is a high school dropout, we created a measure of the cross cohort change in the black-white difference for each variable for each state. We start by pooling samples from the 1960, 1970, 1980 and 1990 IPUMS.[38] We restrict the sample to black and white children between the ages of 0 and 18 who were born between 1957 and 1973 in the 22 states of interest. For each child, we merge information on whether the mother is a high school dropout. We then run regressions analogous to what we do in our AFQT sample for each of our three broad regions (South, Rustbelt and Border). That is for each of the three regressions we control for race by Census year fixed effects, race by age fixed effects, and state by race by cohort fixed effects. The regressions are constructed with the appropriate interactions to yield the difference between the 1961-1963 cohorts and the 1967-69 cohorts in the black-white difference in migration and mother's high school dropout status for each state. These are used as controls in the results shown in Table 6.

F. Hospital Discharge rates, by race and region, from the National Health Interview Surveys

We use the 1963, 1964, 1966, 1967, 1971 and 1972 *NHIS* surveys to construct hospital admissions rates in the past year. Until 1968, the past year refers to the past fiscal year (July to June). After 1968, it refers to the past calendar year. The NHIS categorizes regions as follows: 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).

We checked the self-reported data in the 1971 and 1972 *NHIS* against the corresponding *National Hospital Discharge Surveys* (*NHDS*), which start in 1970. The hospital admission rates and length-of-stay in the *NHIS* were very similar to the hospital discharge rates and length-of-stay in the *NHDS* (by race, age and region).

[38] We use the one-percent sample for 1960. We combine the 1970 Form 1 and Form 2 one-percent samples to create a two percent sample. We use the 5 percent state samples for 1980 and 1990.

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)
A. South						
Black-white NAEP gap	-1.300***	-0.222***	-1.522***	0.828***	-1.281***	0.698***
	(0.031)	(0.052)	(0.042)	(0.084)	(0.030)	(0.076)
Sample Size	9,966		5,020		7,164	
B. North						
Black-white NAEP gap	-1.201***	-0.086*	-1.287***	0.460***	-1.154***	0.293***
	(0.035)	(0.048)	(0.033)	(0.073)	(0.030)	(0.072)
Sample Size	20,762		11,122		16,573	
C. South – North						
Black-white NAEP gap	-0.099**	-0.136*	-0.235***	0.368***	-0.127***	0.405***
	(0.047)	(0.071)	(0.053)	(0.111)	(0.042)	(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)
Percentile score on AFQT						
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]
1st percentile (IPW)	3	1	5	4	2	5
25th percentile (IPW)	29	16	33	29	17	33
Median (IPW)	47	29	51	47	30	50
75th percentile (IPW)	67	43	71	66	44	70
99th 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}
Age distribution (percent)						
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			
Education distribution (percent)						
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
Percent of population who take AFQT						
Age 17				7.10	9.63	6.94
Age 18				6.90	12.18	6.48
Age 18, ≤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	Change by 1970-1972	Average in 1960-1962	Change by 1970-1972
	(1a)	(1b)	(2a)	(2b)
A. South				
Black-white AFQT gap	-25.76***	12.69***	-23.46***	9.08***
	(0.82)	(0.79)	(0.73)	(0.62)
{PNMR gap}			{14.05}	{-8.27}
B. Rustbelt				
Black-white AFQT gap	-21.01***	5.10***	-18.99***	2.01**
	(0.88)	(0.86)	(0.75)	(0.66)
{PNMR gap}			{5.95}	{-1.49}
C. South – Rustbelt				
Black-white AFQT gap	-4.75***	7.60***	-4.47***	7.06***
	(1.17)	(1.13)	(1.01)	(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*** (1.17)	-4.59*** (1.04)	-4.47*** (1.01)	-3.91*** (1.18)	-3.46*** (1.05)	---
1960-1962 to 1970-1972 Change	7.60*** (1.13)	7.04*** (0.81)	7.06*** (0.88)	6.36*** (1.18)	5.62*** (0.92)	7.13*** (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	6.55	6.85	5.59	3.54	3.22	3.13
Change in AFQT gap	[9.94]	[5.80]	[5.13]	[10.33]	[2.85]	[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)
Between cohort difference in racial gap in								
PNMR (per 1,000)	-0.720***		-0.690***	-0.690***	-0.690***	-0.709***	-0.741***	
	[3.91]		[3.62]	[3.79]	[3.93]	[3.71]	[4.69]	
Birth in hospital with								0.257***
doctor present (per 100)								[4.90]
NMR (per 1,000)		0.358	0.200			0.172	0.190	-0.048
		[1.37]	[0.95]			[0.68]	[0.64]	[0.15]
LBW (per 100)							-0.139	-0.113
							[0.20]	[0.16]
Migrate out of state				0.200**		-0.012	-0.030	-0.058
(percent)				[2.28]		[0.06]	[0.12]	[0.26]
Mother HS dropout					-0.257*	-0.126	-0.132	-0.201
(percent)					[1.88]	[0.53]	[0.51]	[0.92]
Racial gap in Head Start spending per 4 year-old in								
1968 (1/100)							0.051	0.010
							[0.59]	[0.13]
1972 (1/100)						-0.235	-0.303	-0.363*
						[1.67]	[1.26]	[2.05]
Constant	1.37**	4.27***					1.31	1.27
	[2.57]	[7.32]					[0.80]	[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.
*** significant at 1-percent level, ** significant at 5-percent level, * 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	-0.154	-0.171	-0.077	-0.156	0.187	-0.205	0.136	0.740
	[3.09]	[2.73]	[2.35]	[1.10]	[1.74]	[3.10]	[0.67]	[0.35]	[3.72]
Lead, 3 years	***-0.303***	-0.158	-0.120	-0.113	-0.183	0.134	***-0.506***	***-0.348***	0.449
	[5.20]	[3.05]	[2.19]	[1.56]	[2.77]	[2.22]	***[1.96]***	***[0.94]***	[2.72]
Lead, 2 years	***-0.299***	***-0.332***	***-0.232***	***-0.169***	***-0.279***	0.050	***-0.362***	***-0.343***	0.280
	[4.44]	***[6.02]***	***[3.78]***	***[4.06]***	***[5.02]***	[0.96]	***[2.07]***	***[1.44]***	[2.29]
Lead, 1 year	0.028	***-0.244***	***-0.174***	***-0.098***	***-0.281***	0.025	0.141	0.262	0.319
	[0.50]	***[4.07]***	***[2.42]***	***[2.50]***	***[4.81]***	[0.52]	[0.71]	[0.95]	[2.19]
Contemporary	0.308	-0.189	-0.097	0.069	-0.216	0.046	0.821	1.080	0.322
	[4.03]	[3.29]	[1.34]	[0.99]	[2.99]	[0.66]	[3.00]	[2.63]	[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***	-0.456***	-0.434***	-0.553***	-0.537***	-0.466***	-0.356***	-0.382***	-0.370***
	[6.85]	[5.31]	[5.89]	[9.49]	[8.14]	[5.30]	[5.79]	[5.28]	[4.54]
PNMR, 1 yr. lead	-0.066	-0.081	-0.119	-0.413***	-0.421***	-0.327***	-0.268***	-0.282***	-0.234**
	[0.76]	[0.85]	[1.55]	[5.55]	[4.97]	[3.83]	[3.33]	[2.94]	[2.61]
NMR, 2 yr. lead	-0.197**	-0.215***	-0.055	-0.068	-0.065	-0.007	0.014	0.015	0.008
	[2.64]	[2.87]	[1.11]	[1.50]	[1.19]	[0.09]	[0.21]	[0.20]	[0.12]
NMR, 1 yr. lead	-0.220***	-0.236***	-0.085	-0.110*	-0.099	-0.048	-0.012	-0.023	-0.043
	[2.90]	[3.04]	[1.73]	[2.02]	[1.56]	[0.75]	[0.24]	[0.38]	[0.84]
LBW, 2 yr. lead	-0.350	-0.349	-0.135	-0.221	-0.217	-0.119	0.189	0.125	0.274
	[1.56]	[1.43]	[0.70]	[1.15]	[1.03]	[0.60]	[1.22]	[0.76]	[1.49]
LBW, 1 yr. lead	-0.014	-0.073	0.075	0.127	0.156	0.159	0.240	0.173	0.201
	[0.07]	[0.31]	[0.38]	[0.71]	[0.74]	[0.72]	[1.36]	[0.81]	[0.88]
Constant	-7.97	-7.38							
	[4.36]	[3.45]							
F(9, 18) for joint signif.			14.62***			4.56***			2.35*
illegit, age variables			{0.000}			{0.003}			{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

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

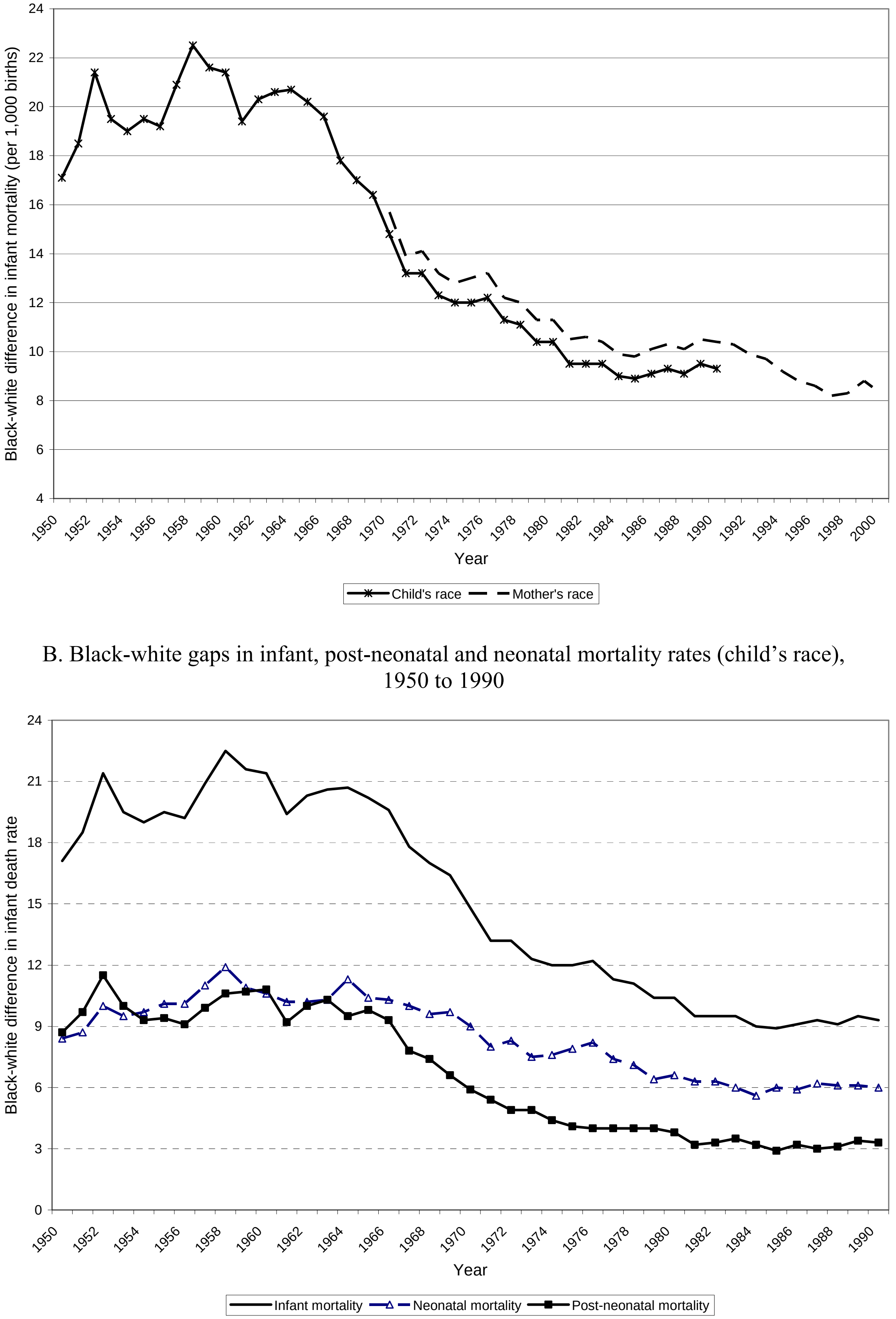
(Table A2 continued)

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

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

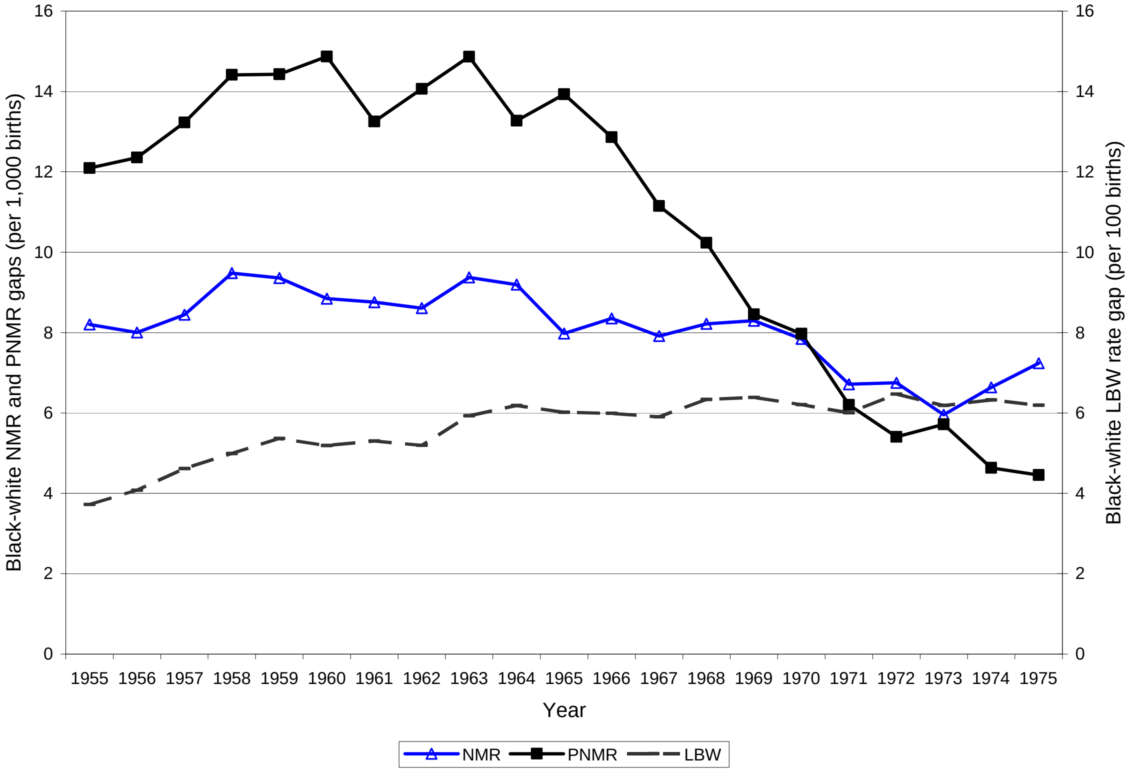
Figure 1: Black-white difference in infant mortality rates in the United States, 1950 to 2000



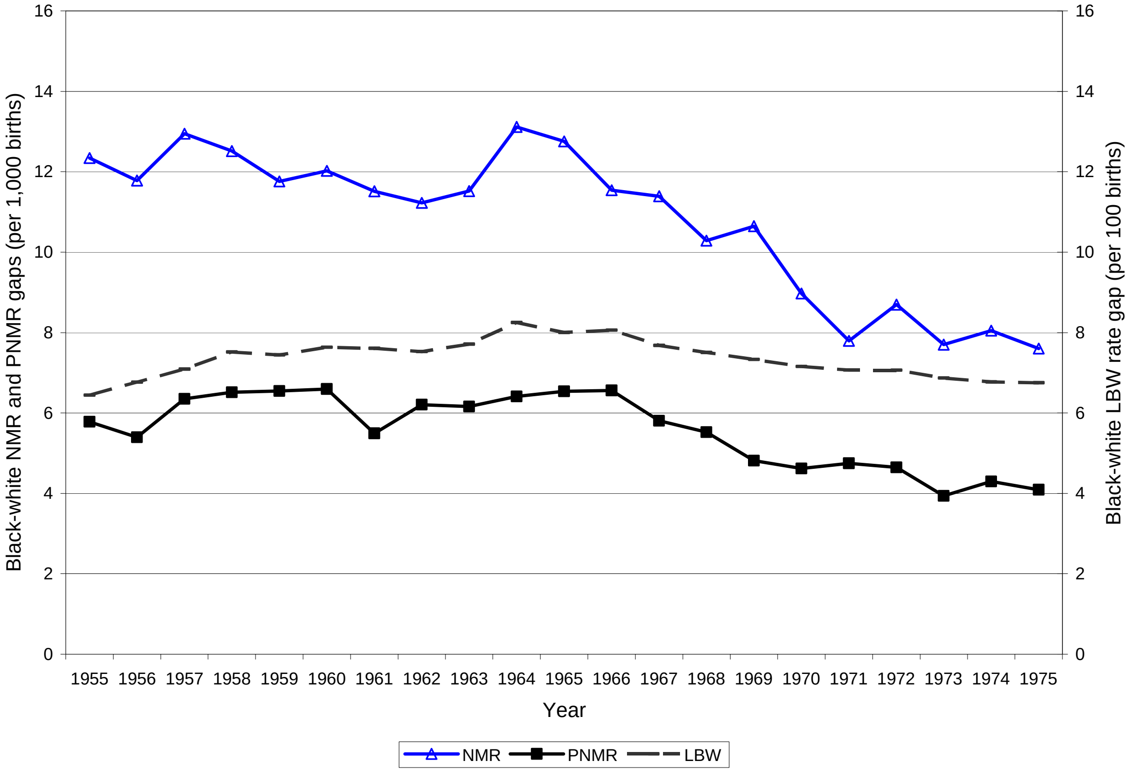
A. Infant mortality by child and mother's race in United States, 1950 to 2000

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

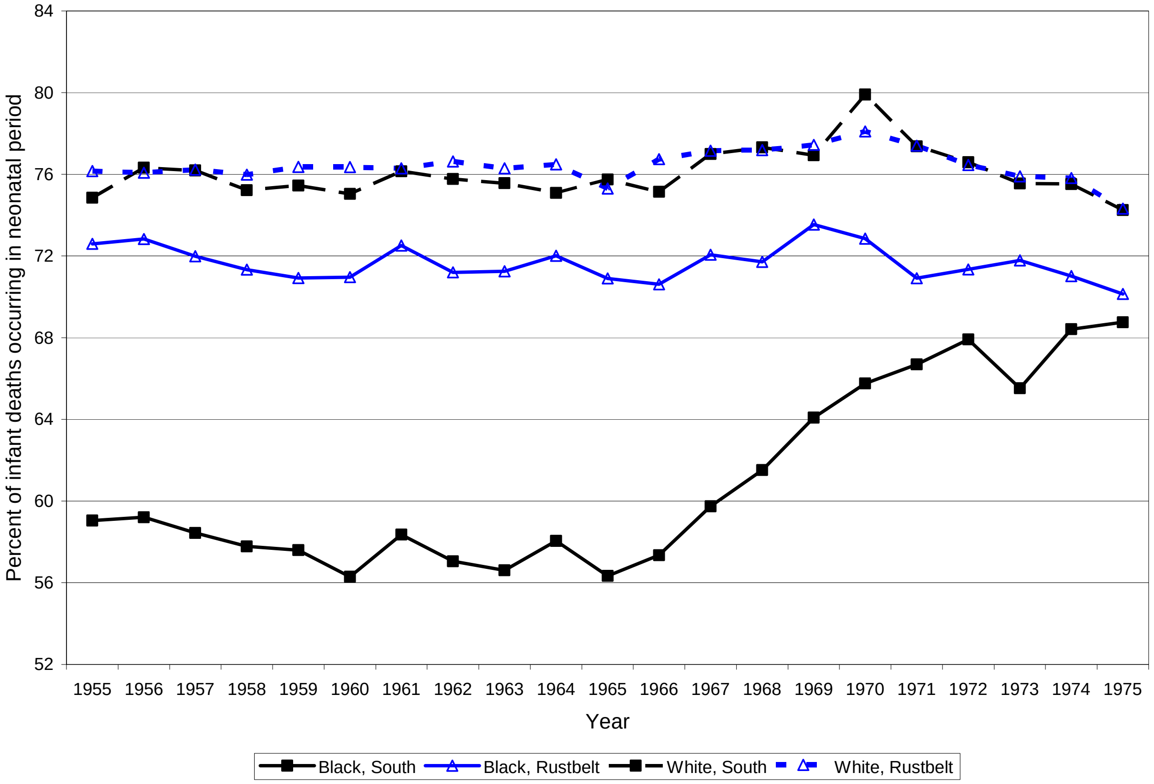
C. Racial gaps in NMR, PNMR and low birth weight rates in South, 1955 to 1975



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



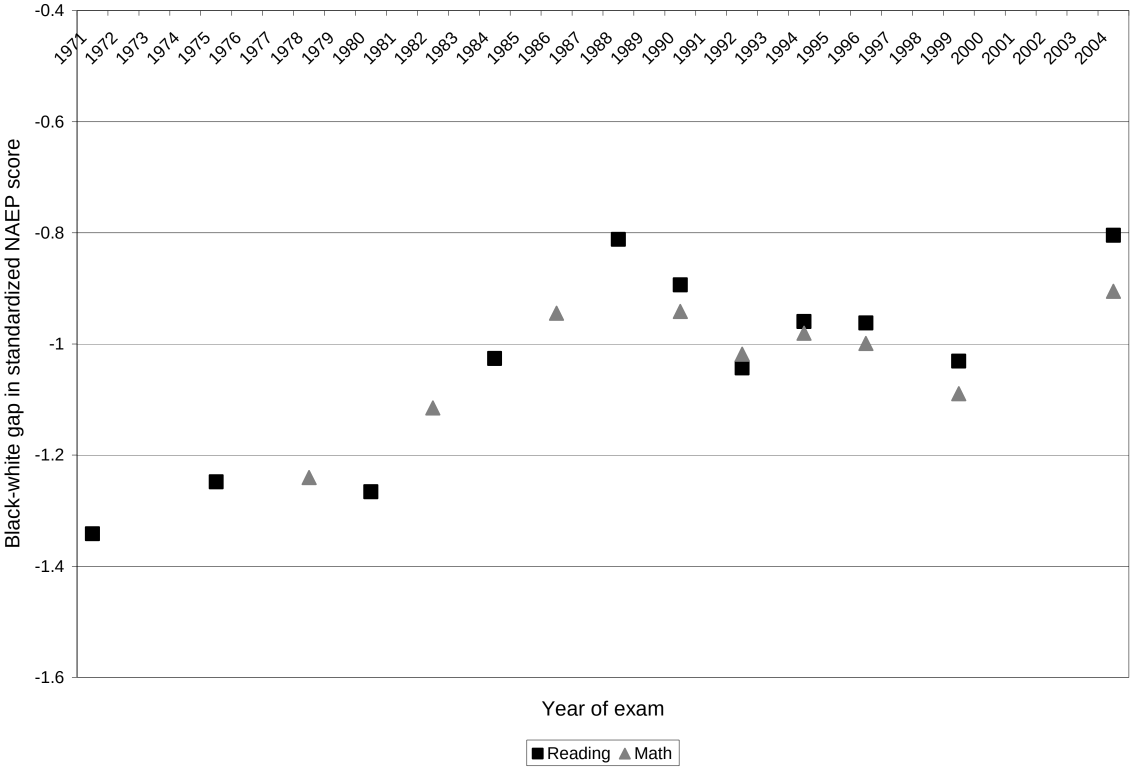
E. Percent of all infant death occurring in neonatal period, by race and region



Notes: Data from the *Vital Statistics of the United States*. After 1990 births and deaths are only recorded by the mother's race. 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

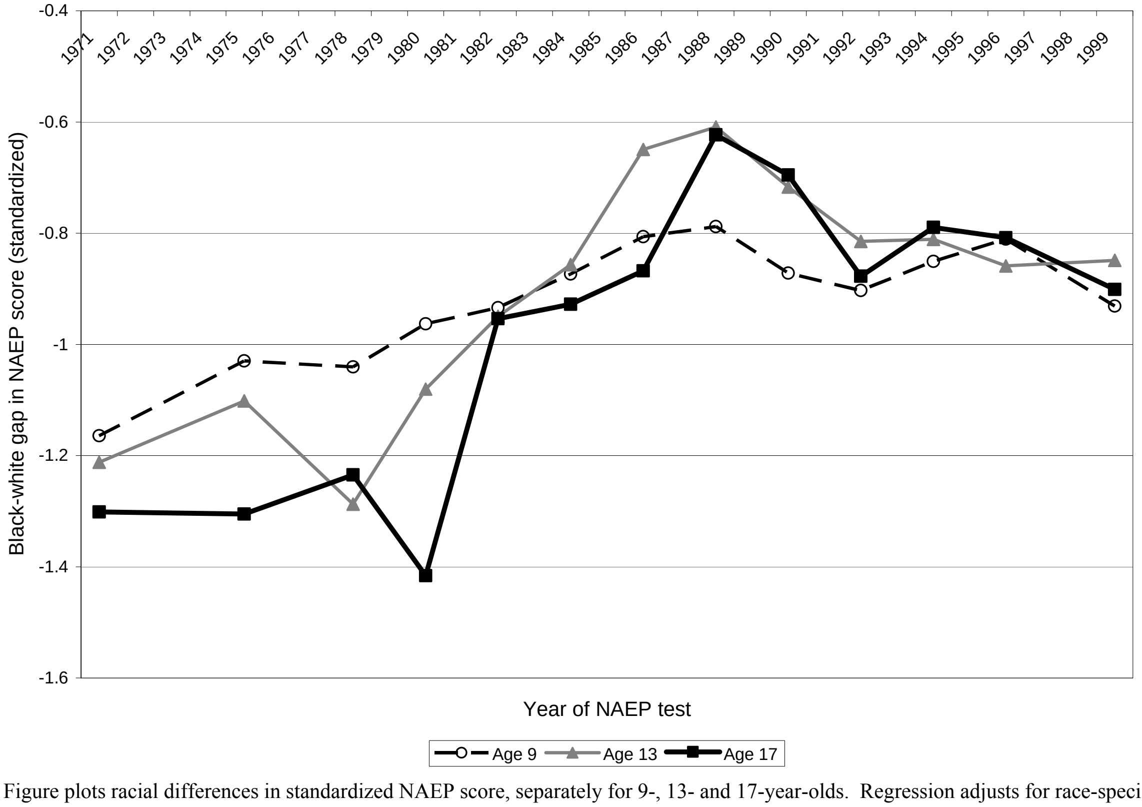
Figure 2: National Assessment of Educational Progress Scores

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



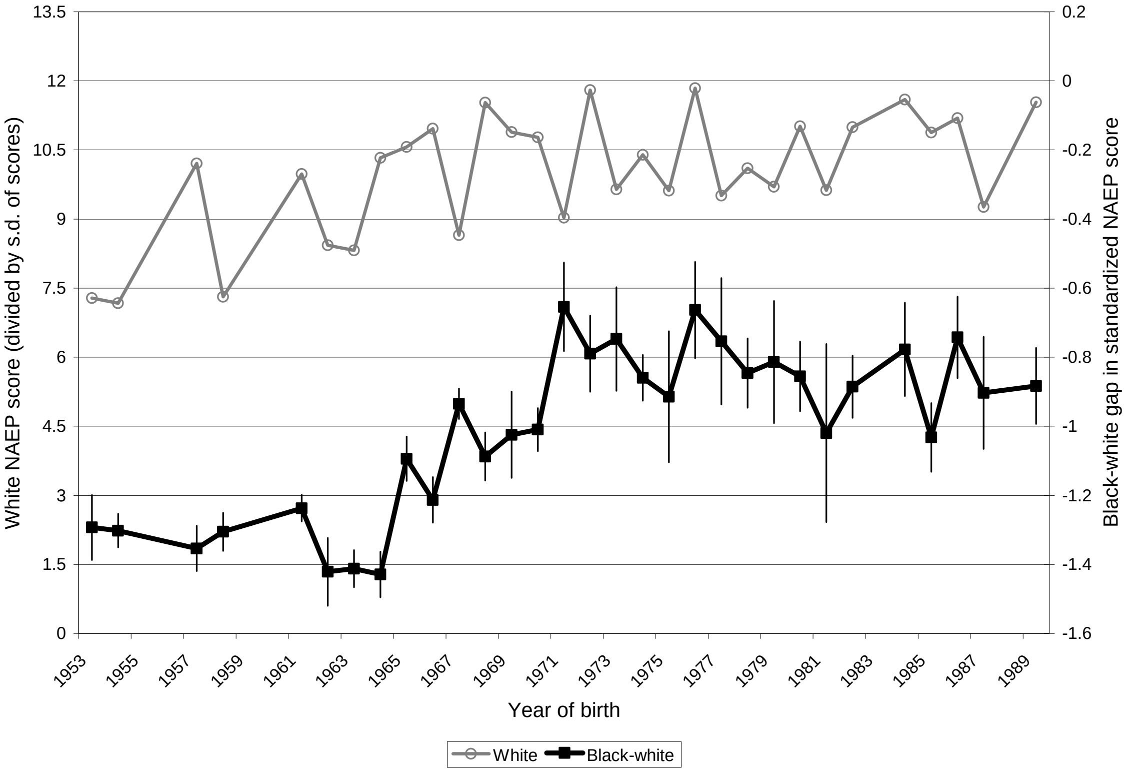
Notes: Figure plots racial differences in average scaled NAEP Math and Reading scores, normalized by the standard deviation of test scores by survey year, age, and subject. Subject-specific regressions adjust for race-specific age effects.

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



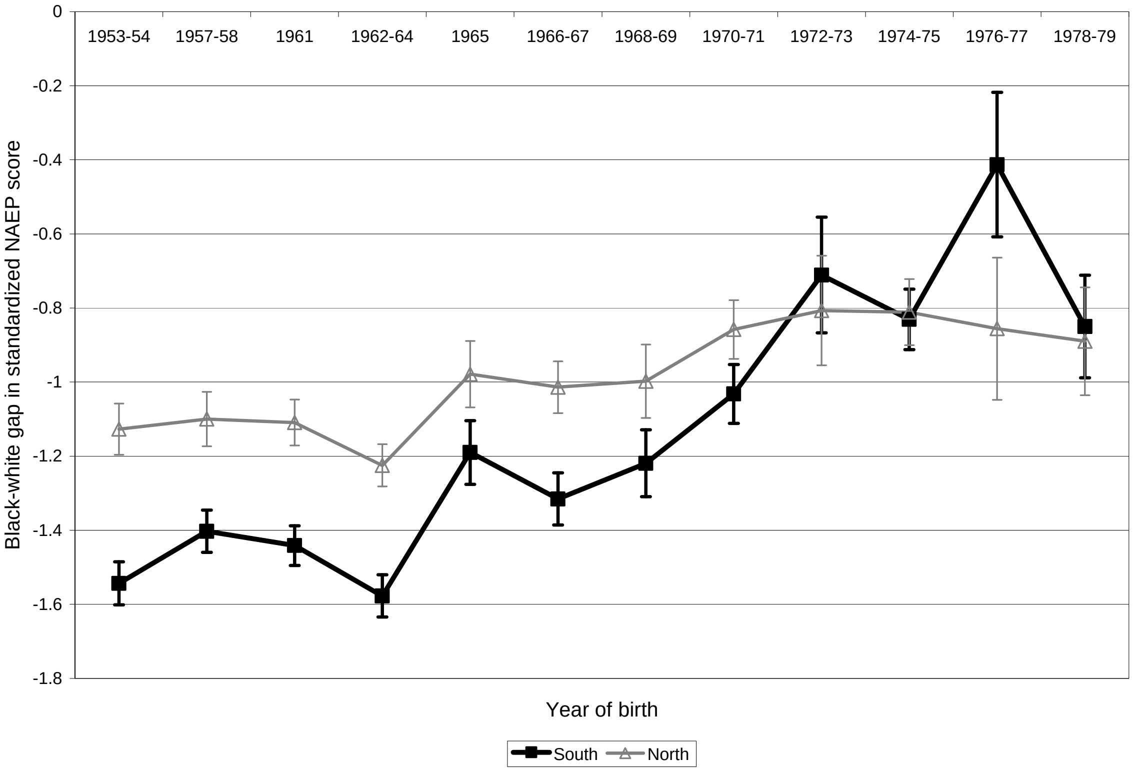
Notes: Figure plots racial differences in standardized NAEP score, separately for 9-, 13- and 17-year-olds. Regression adjusts for race-specific subject effects by age.

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



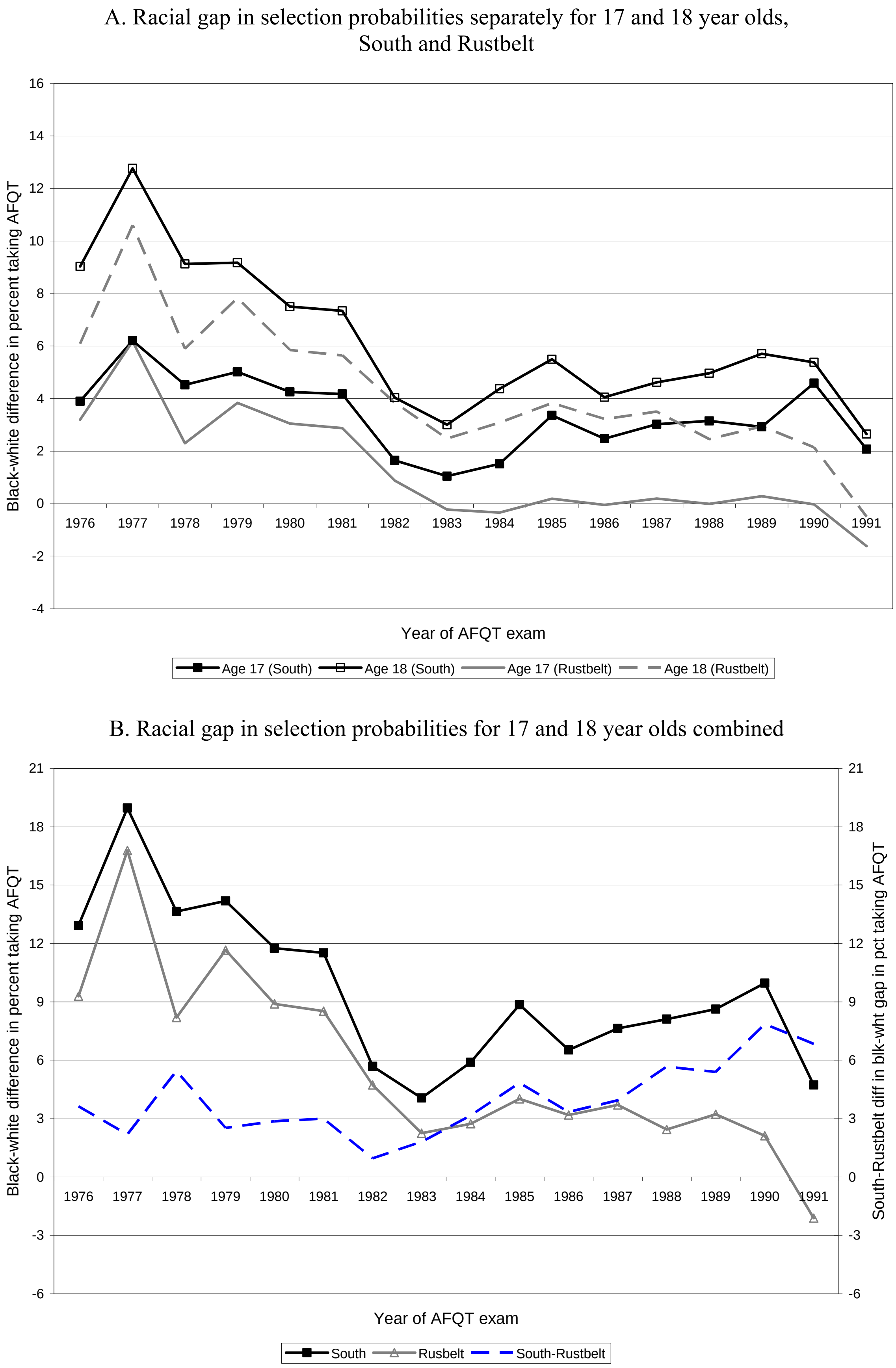
Notes: Figure plots white levels of and racial differences in standardized NAEP scores by year of birth. Vertical lines represent (±) twice the standard error of the estimate, corrected for heteroskedasticity. Regression adjusts for race-specific age effects that vary by subject.

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 (±) 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.

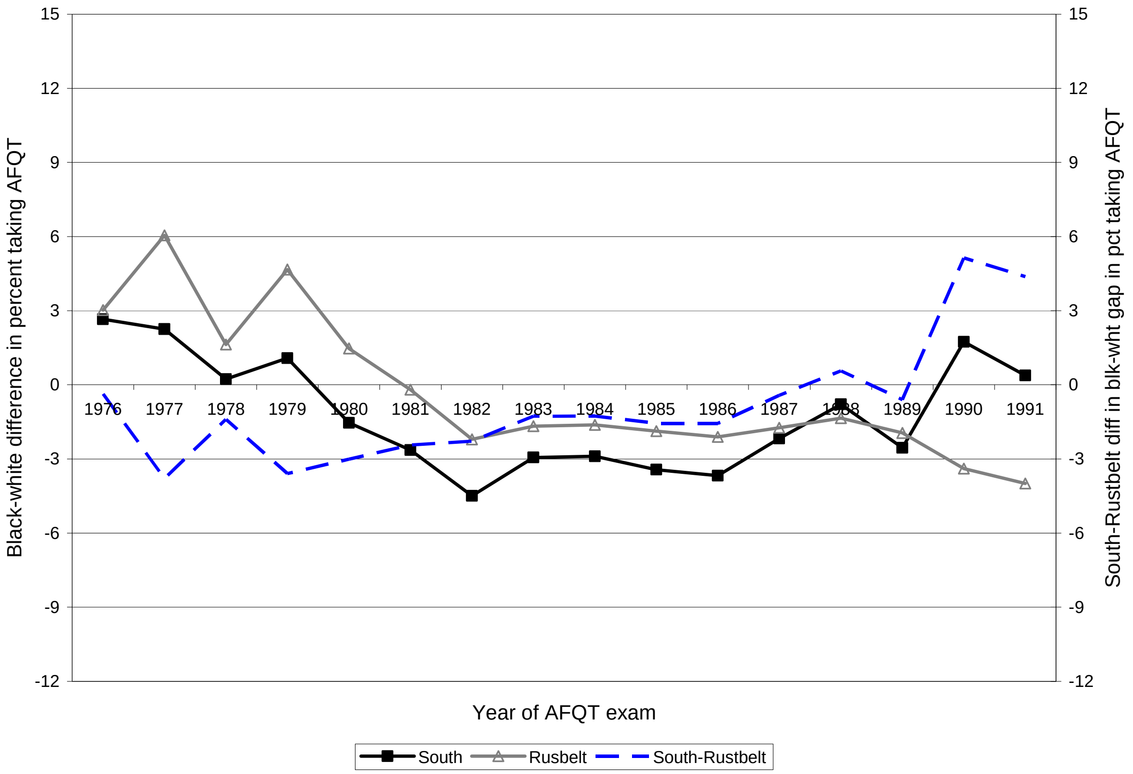
Figure 3: Probability in population of taking the AFQT, by year exam taken



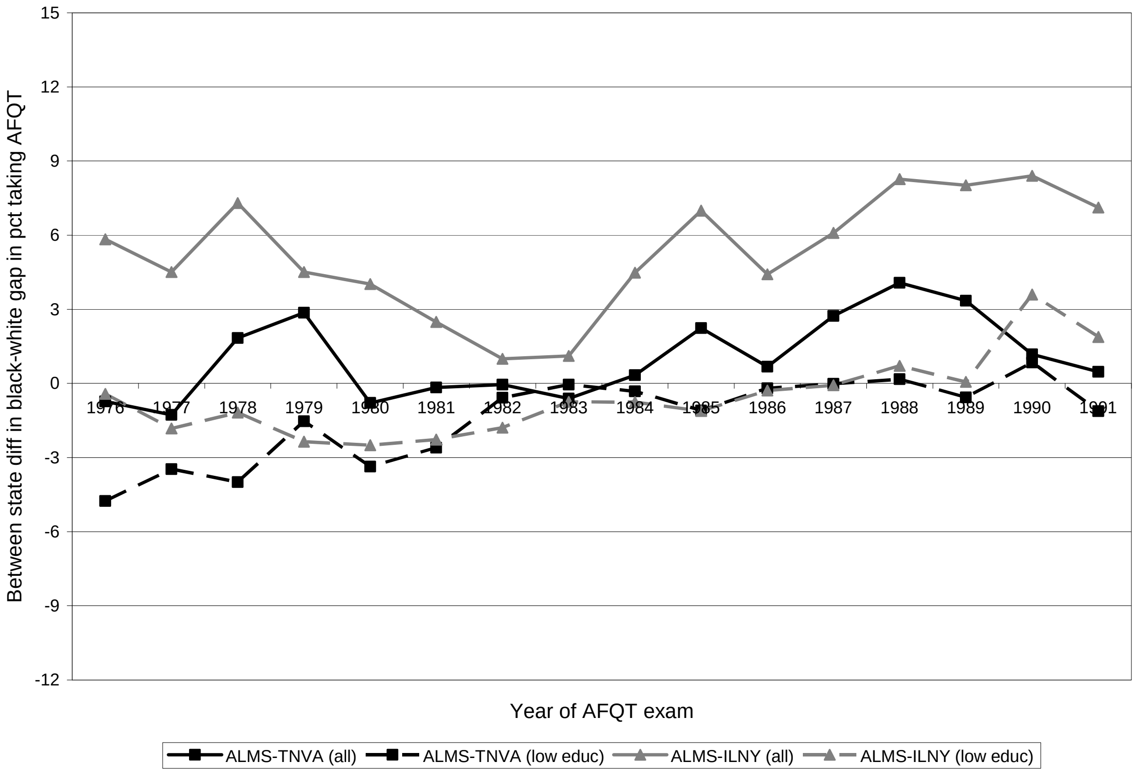
A. Racial gap in selection probabilities separately for 17 and 18 year olds, South and Rustbelt

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

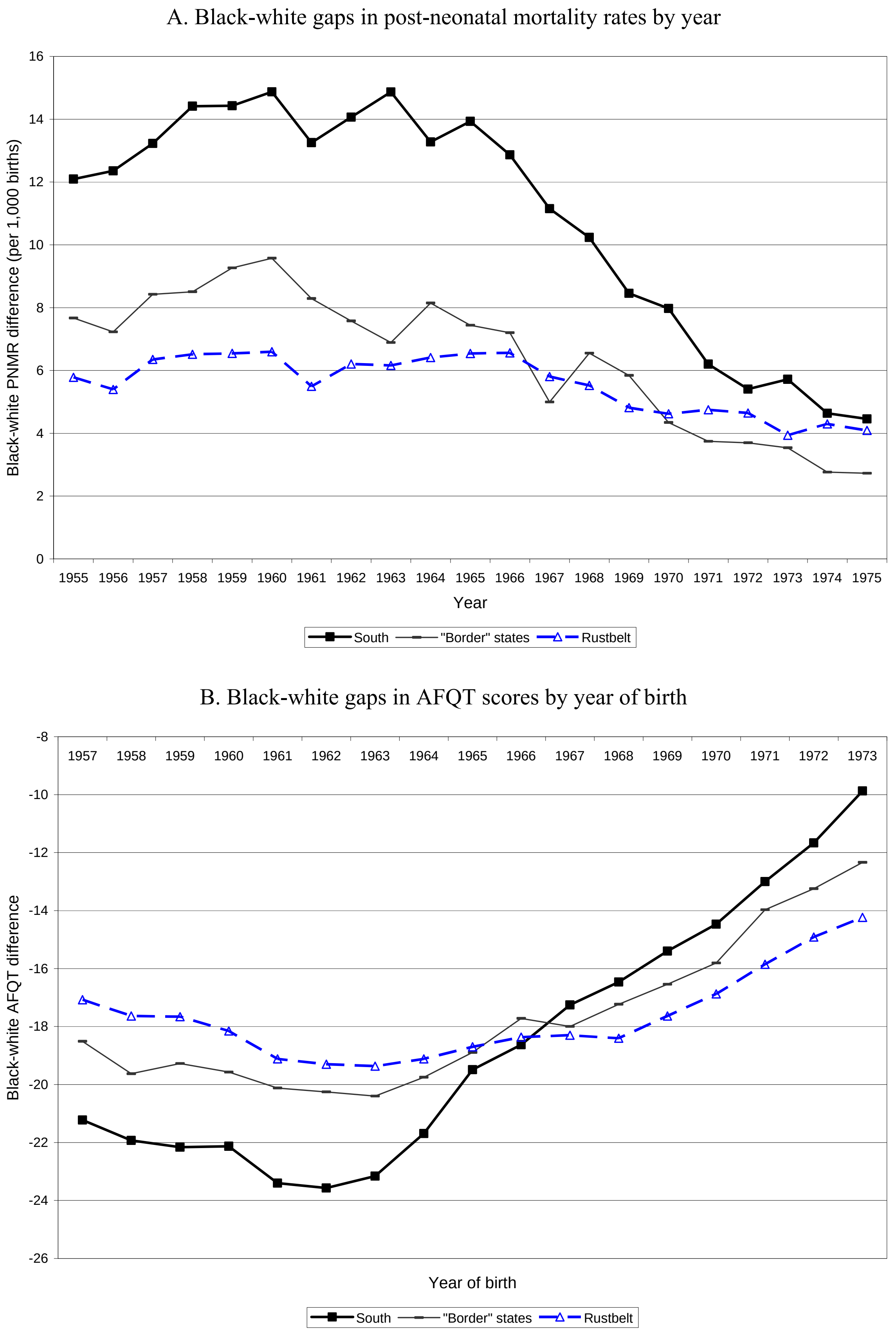


D. Difference in selection probability gap between Alabama-Mississippi and other state groups



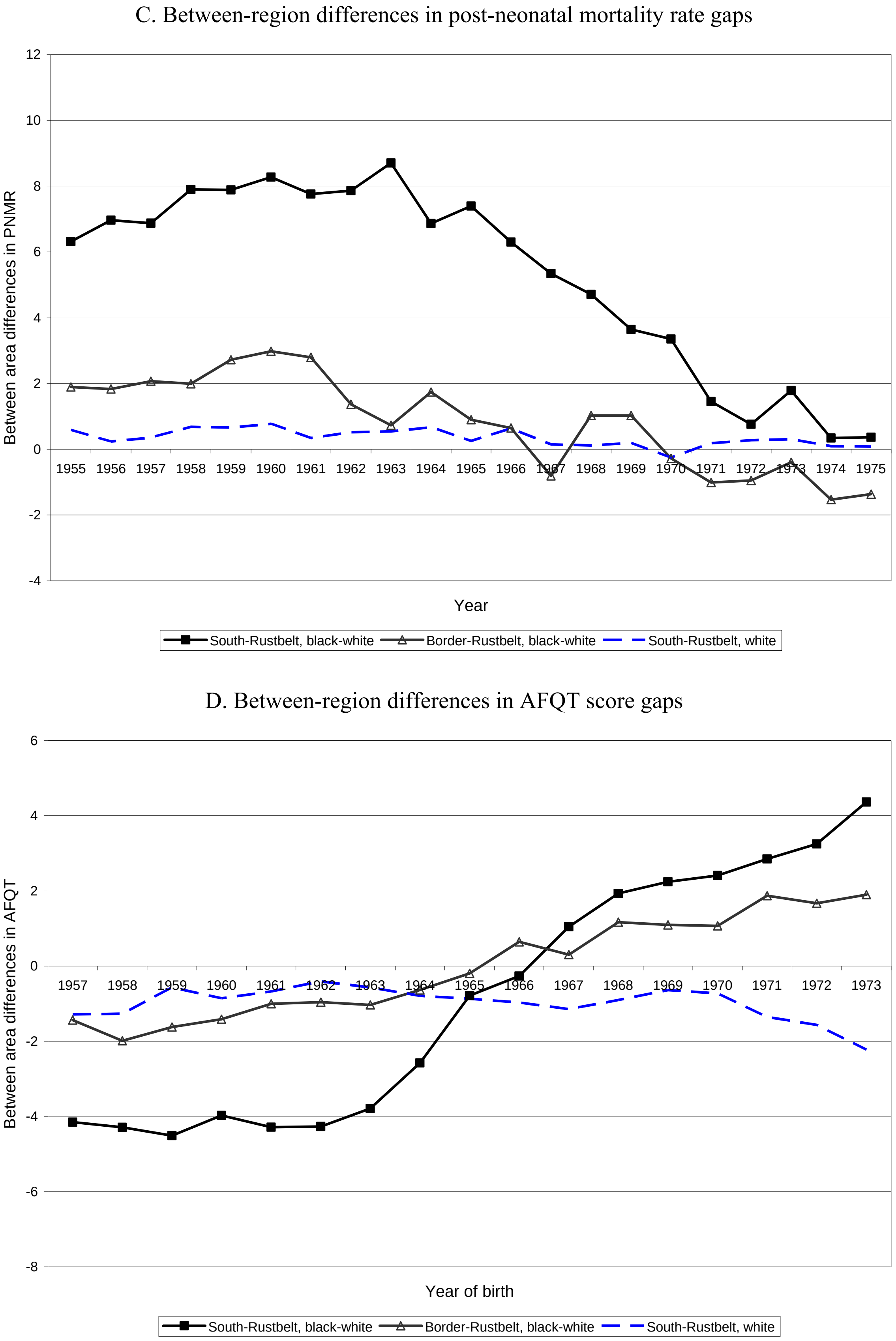
Notes: Population counts for each state-race-age-year (and education) cell come from the Decennial Censuses. In Panel D, the state groups are Alabama and Mississippi (ALMS), Tennessee and Virginia (TNVA), and Illinois and New York (ILNY); and "low educ" refers to men with two years or less of high school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

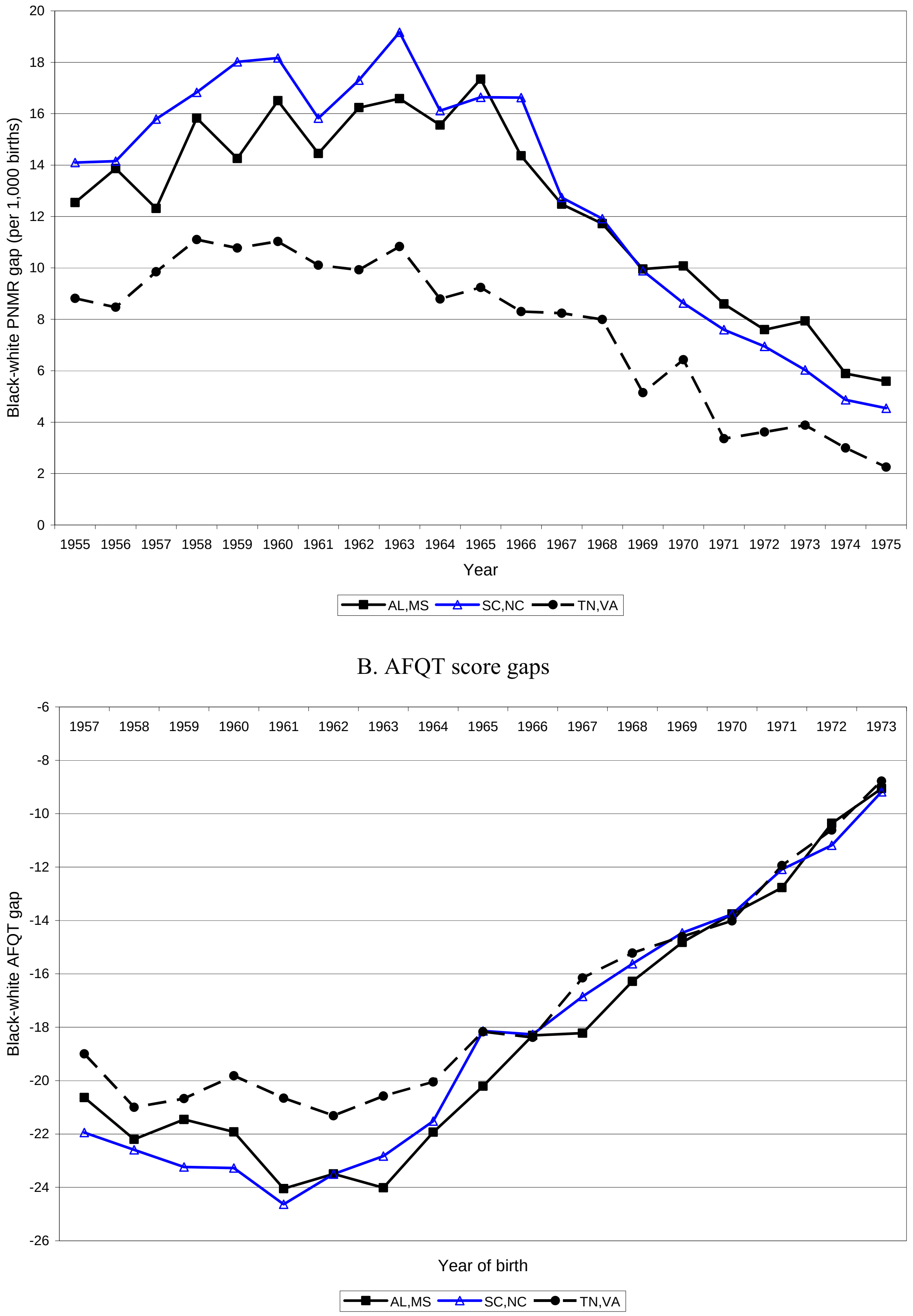


C. Between-region differences in post-neonatal mortality rate gaps

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.

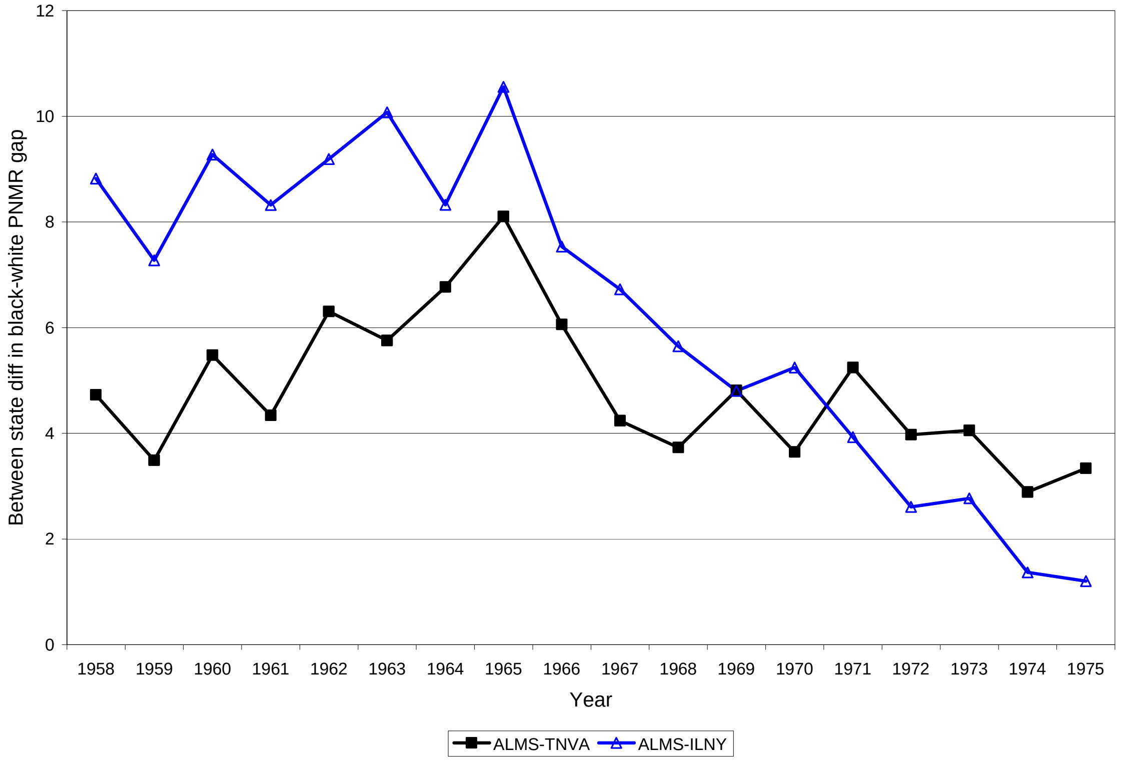
Figure 5: Across state-group differences in black-white gaps in PNMR and AFQT scores



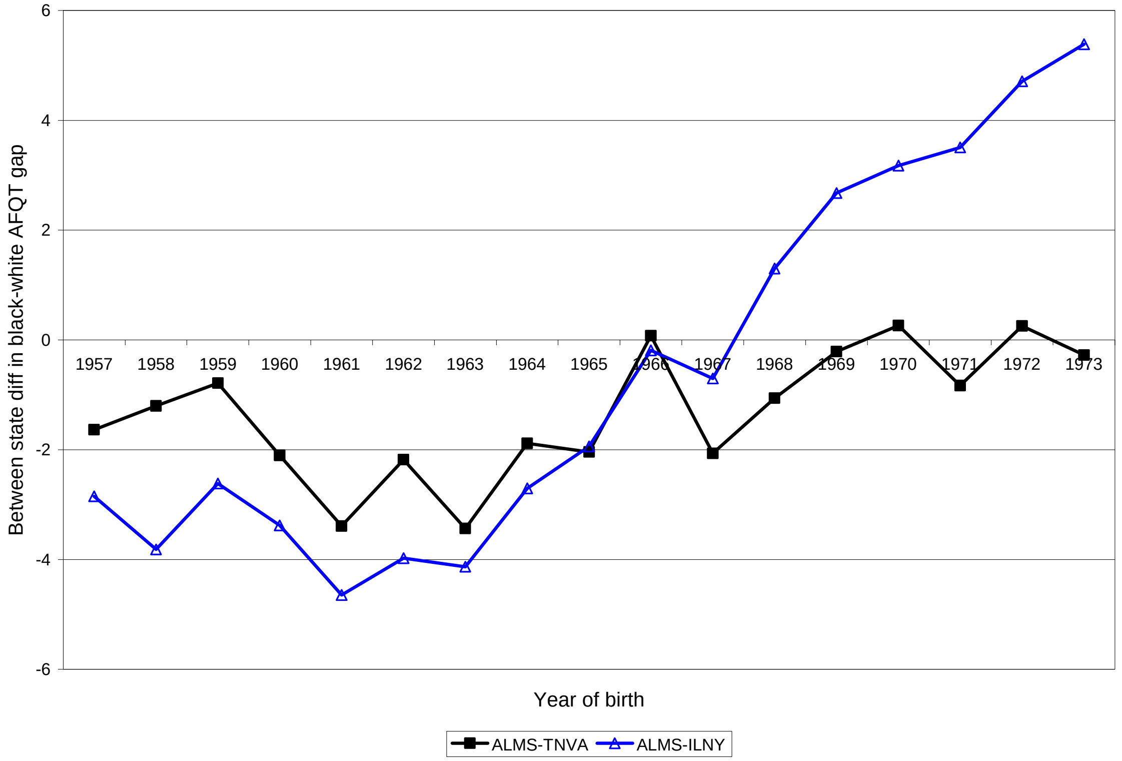
A. Post-neonatal mortality gaps

B. AFQT score gaps

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

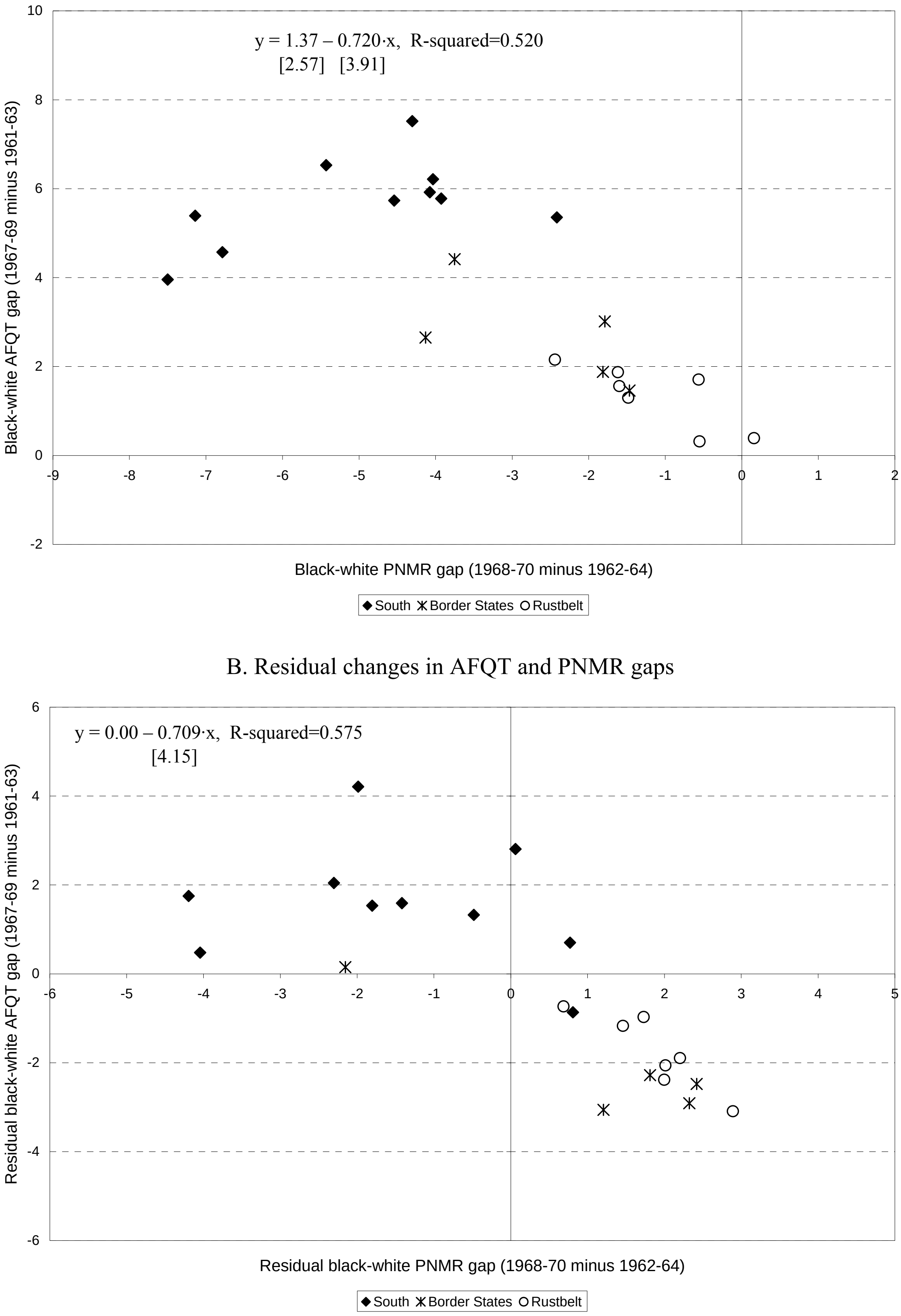


D. Difference in AFQT gap between Alabama-Mississippi and other state groups



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).

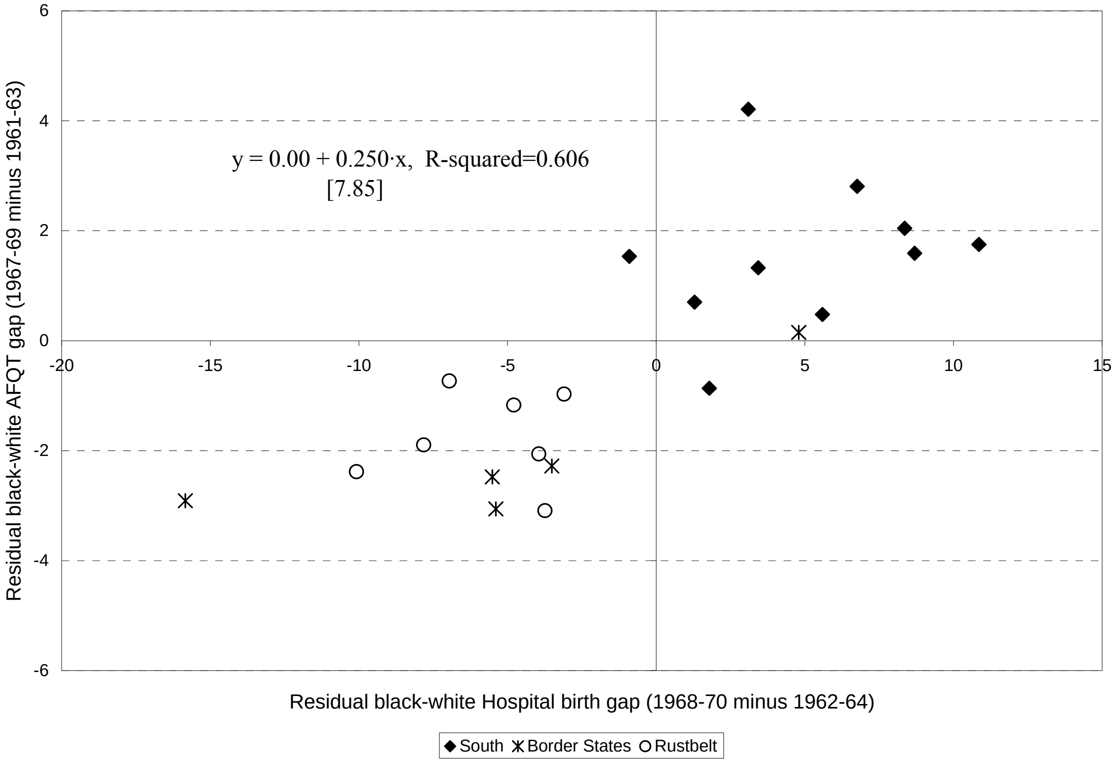
Figure 6: Scatter plots of between-cohort changes in racial gaps in AFQT and infant health (22 states)



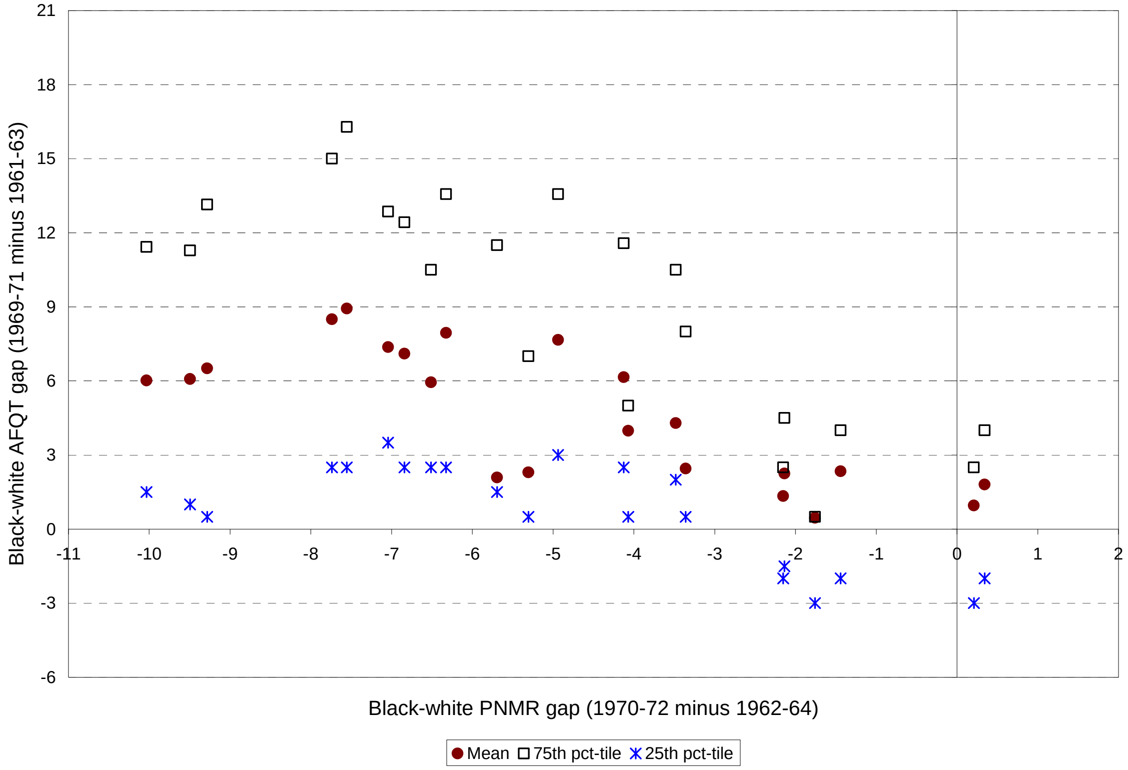
A. Changes in gaps in AFQT (1961-63 to 1967-69) and PNMR (1962-64 to 1968-70)

B. Residual changes in AFQT and PNMR gaps

C. Residual changes in AFQT and hospital birth rate gaps

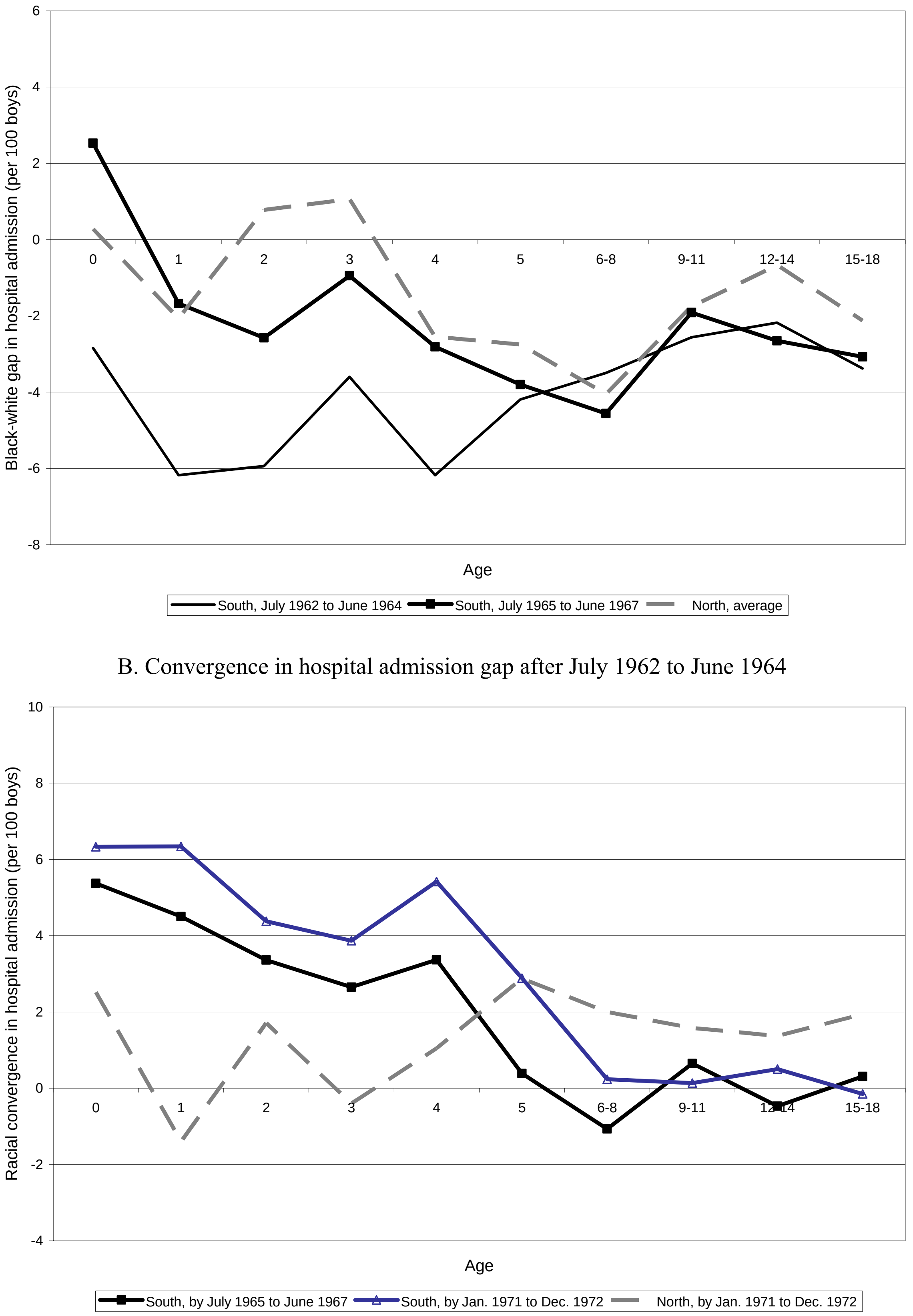


D. Changes in AFQT (1961-63 to 1969-71) and PNMR (1962-64 to 1970-72) gaps, Mean, 75th and 25th percentiles of AFQT scores



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)

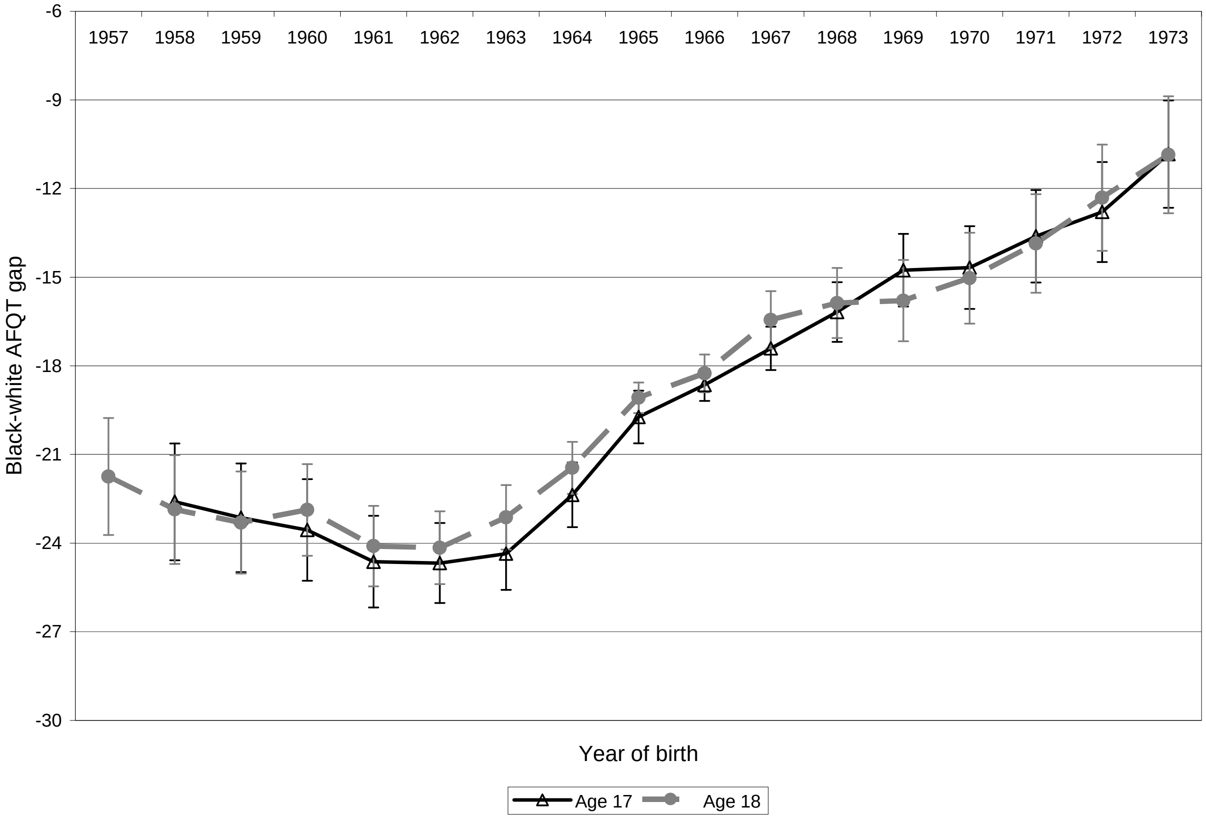


A. Hospital admission gap (per 100 boys)

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

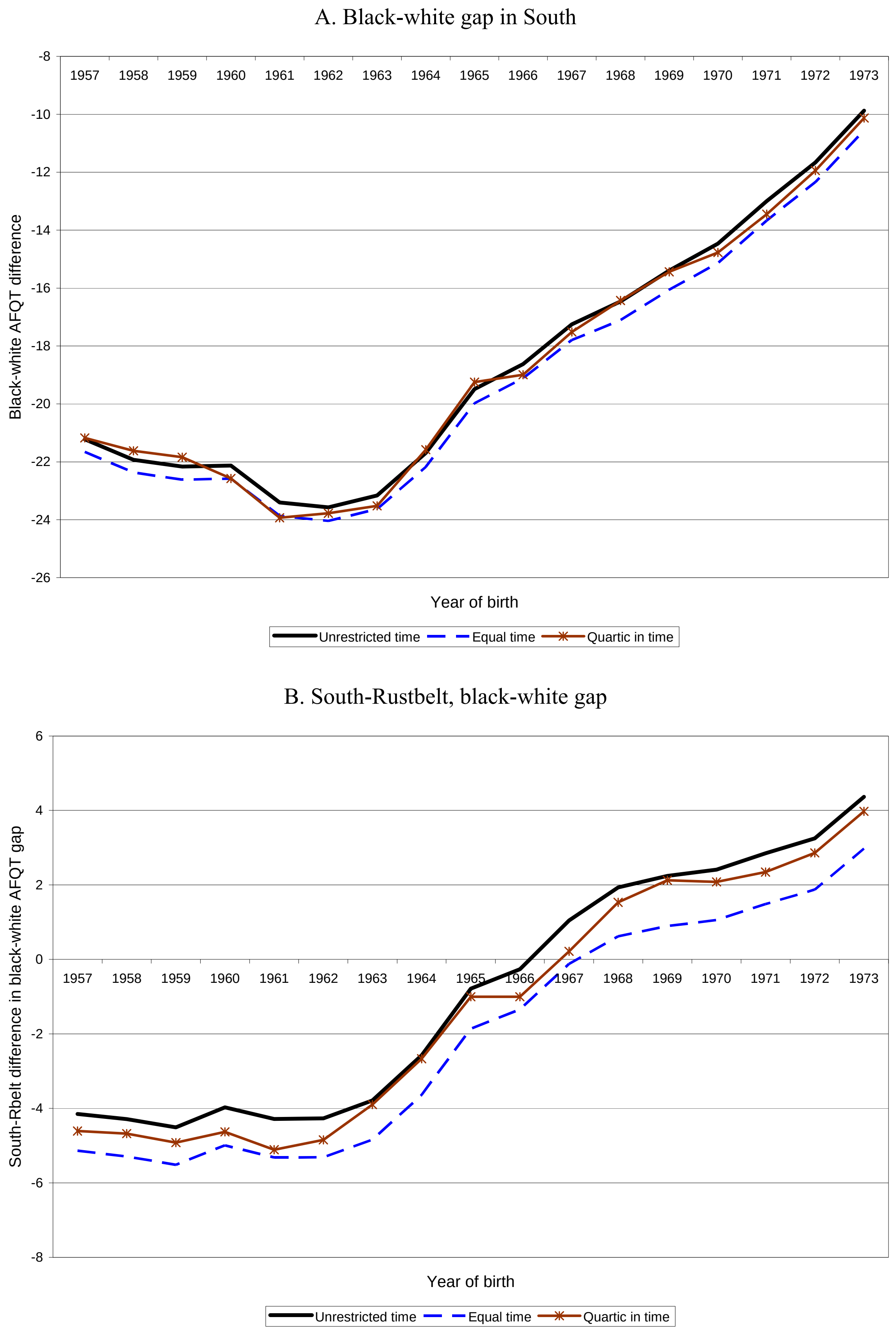
Notes: Data come from the 1963, 1964, 1966, 1967, 1971 and 1972 *National Health Interview Surveys*. South consists of Alabama, Arkansas, Delaware, D.C., Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, South Carolina, Tennessee, Oklahoma, Texas, Virginia, and West Virginia.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

Figure A1: 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 (±) twice the standard error of the estimate, corrected for heteroskedasticity.

Figure A2: Estimated cohort-specific AFQT gaps under different time restrictions

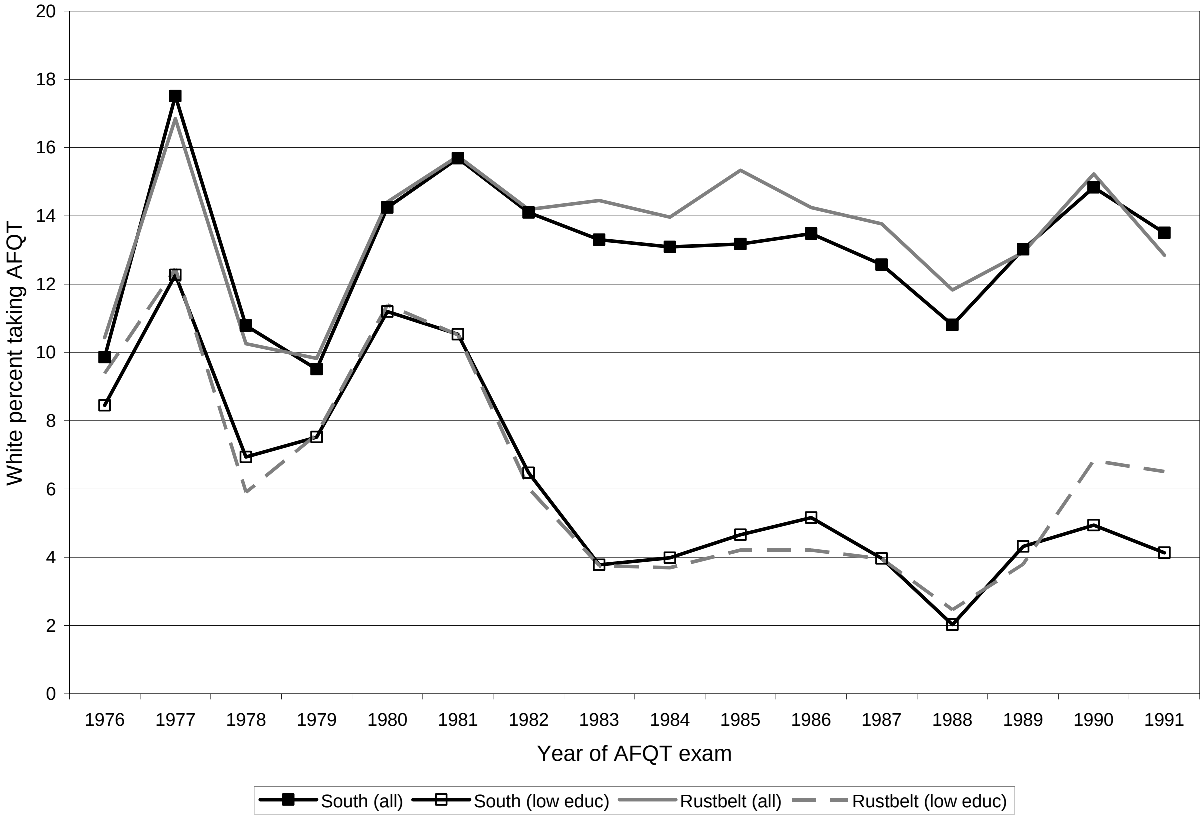


A. Black-white gap in South

B. South-Rustbelt, black-white gap

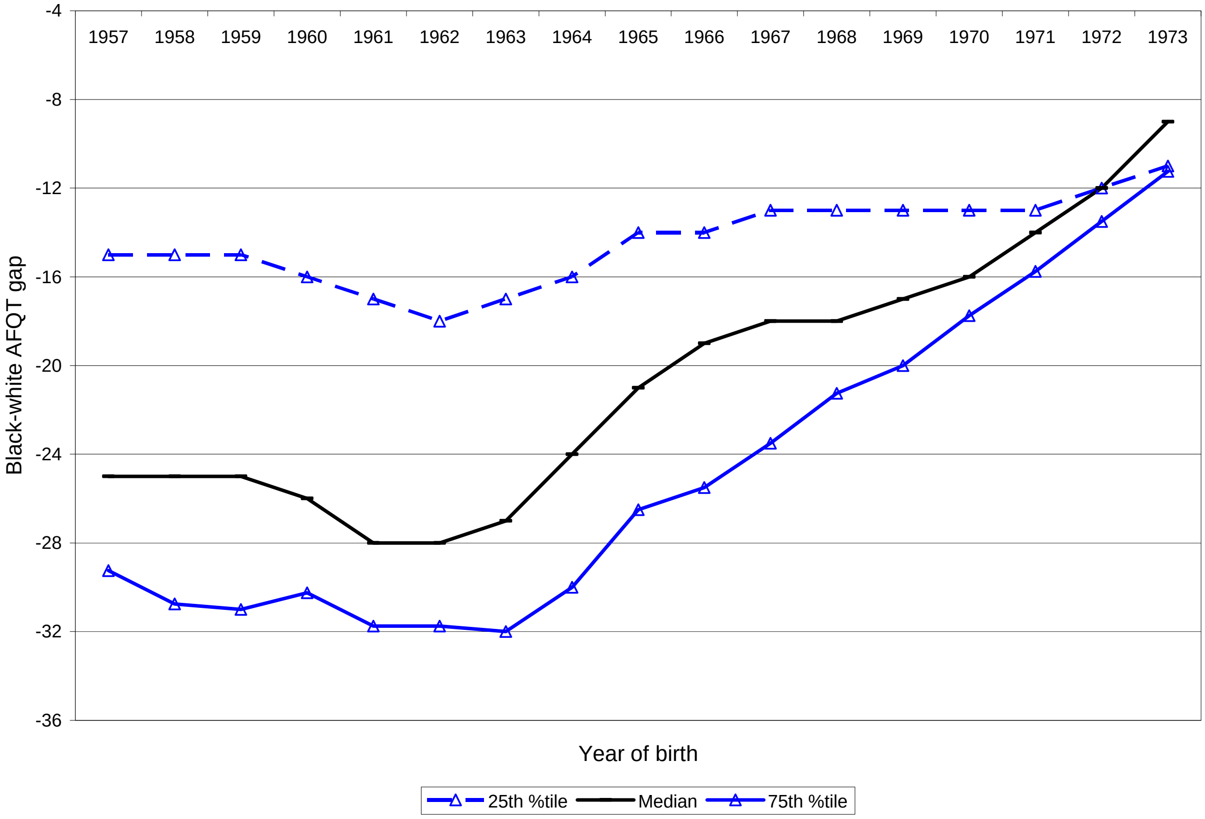
Notes: Plots come from inverse probability weighted (by state births) regressions. “Unrestricted time” model includes unrestricted year effects interacted with race; “Equal time” model restricts the black-white time effect to be the same in 1985 and 1986; “Quartic in time” model imposes a quartic polynomial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

Figure A3: 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 A4: Black-white conditional quantile gap in AFQT scores in South, by birth year



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