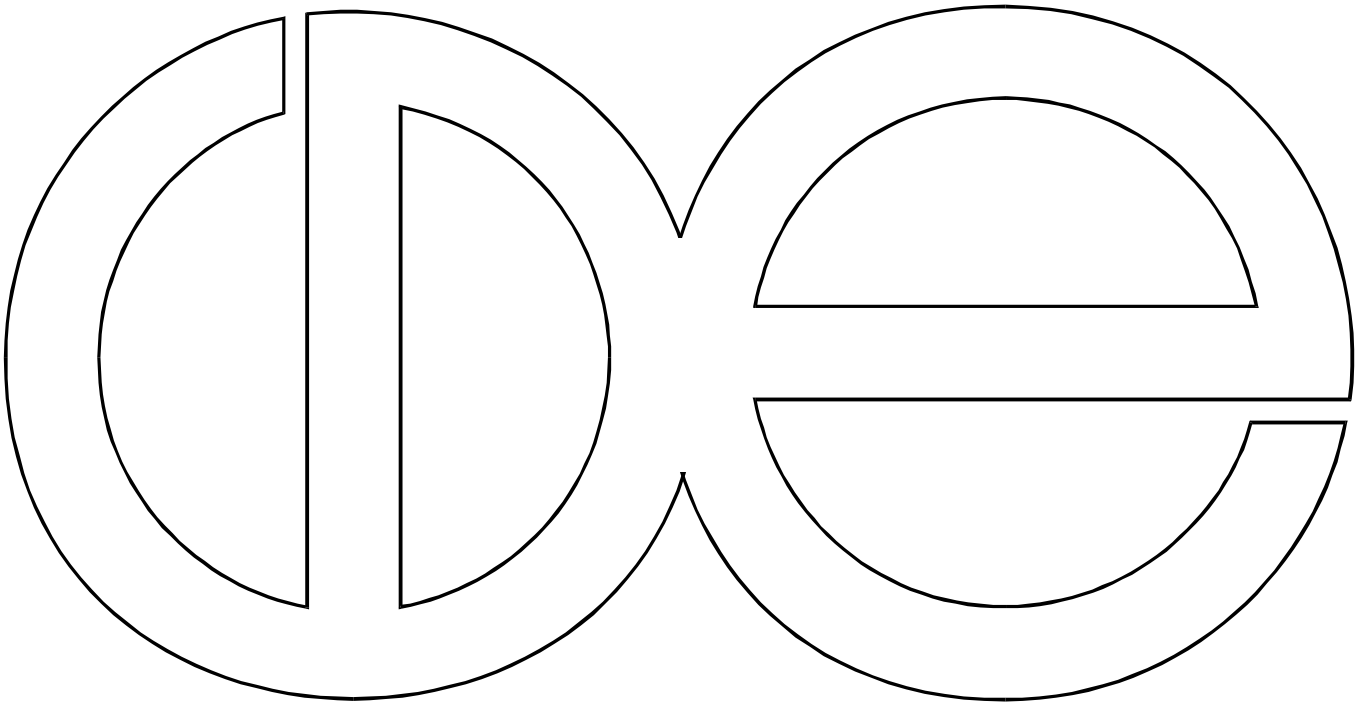
Center for Demography and Ecology

University of Wisconsin-Madison

Trends in Black-White Test Score Differentials: I. Uses and Misuses of NAEP/SAT Data

Robert M. Hauser

CDE Working Paper No. 96-29



Trends in Black-White Test Score Differentials:

I. Uses and Misuses of NAEP/SAT Data

Robert M. Hauser

Department of Sociology
Center for Demography and Ecology
The University of Wisconsin-Madison

Rev. November 21, 1996

Prepared for the conference, "Intelligence on the Rise: Secular Changes in IQ and Related Measures," Emory University, April 1996. Support for this research was provided by the National Science Foundation (SBR-9320660); the Office of the Assistant Secretary for Planning and Evaluation, Department of Health and Human Services; the Vilas Estate Trust; and the Center for Demography and Ecology at the University of Wisconsin-Madison. I thank Min-Hsiung Huang for helpful comments. The opinions expressed herein are those of the author.

Until the mid-1980s, there was not much good news for those who believed that ability and achievement test differences between Blacks and Whites in the United States were the malleable products of environment or culture. On IQ tests and other similar tests, there were typical and persistent mean differences between Blacks and Whites on the order of one standard deviation. If test scores were normally distributed, one would expect about 84 percent of Whites to exceed the mean score among Blacks and about 16 percent of Blacks to exceed the mean score among Whites. That is no longer the case on some cognitive tests, and there has been a substantial convergence in the performance of Blacks and Whites. This paper reviews evidence about the timing of the changes in test scores and the types of tests in which it has occurred. Herrnstein and Murray (1994) reviewed some of this evidence in their recent best-seller, *The Bell Curve*, but we shall see that their account was far from adequate. The data are incomplete but, by the standards of evidence routinely accepted in psychological research on test performance, I believe that the available evidence of change is highly significant and incontrovertible.

TRENDS BEFORE 1970

Loehlin, Lindzey, & Spuhler (1975, p. 140-41) offered two sources of trend data on Black-White differences in IQ. First, they assembled data from Shuey's (1966) extensive review. Even though they acknowledged problems with the quality of Shuey's data, they argued that "gross changes over time should be detectable, even in the absence of single studies making well controlled comparisons over substantial time spans." In a comparison of 259 studies of Black and White preschool children, elementary school children, and high school students, they found little evidence that test score differentials had declined between the pre-1945 period and that between 1945 and 1965. To be sure, the differential was unchanged (14 points) in individual tests administered to elementary school children, and it declined from 16 to 13 points in nonverbal

group tests of elementary school children. At the same time, the gap increased from 9 to 16 points among preschool children, from 13 to 16 points in verbal tests among elementary school children, and from 11 to 19 points among high school students. Loehlin, Lindzey, & Spuhler concluded that the data "fail to suggest much change over time in black-white differences in the groups for which there is the most data -- and probably the most representative data -- the elementary school children" (1975: p. 141).[1]

Second, Loehlin, Lindzey, & Spuhler (1975, p. 142-44) compared the test score distributions of Black and White military recruits in World War I, World War II, and the Vietnam War. These comparisons were hampered by changes in the tests (Army Alpha and Beta, Army General Classification Test, and Armed Forces Qualification Test), by changes in the intervals within which test scores were reported, and by changes in population coverage and definition. By assuming that the score distributions were normal, they estimated mean IQ score differences. These differences were 17 points in World War I, 23 points in World War II, and 23 points in the Vietnam War. They again concluded that the test score differences were large, but that no inference should be made with regard to trend: "Because there are a number of ways in which these samples are unrepresentative of the total U.S. male population, we are not willing to draw strongly the conclusion that the black-white gap in average measured ability has actually widened since the time of World War I" (1975: 144). And of course, these data provided no information about Black or White women.

Relative to typical standards in the population sciences, all of these findings are of questionable validity. The population coverage is haphazard; scientific sampling is the exception; scores on various tests are rendered nominally comparable, merely by assuming normality in the

trait distribution and expressing scores as deviations from the mean. In short, up to about 1980, data on trends in Black-White differences in IQ were probably of even worse quality than those available for global assessments of trends in test performance (Flynn, 1984; Flynn, 1987). All the same, the Coleman-Campbell report of 1966, *Equality of Educational Opportunity*, must have resolved any doubts about Black-White differentials in the early 1960s with its finding that test score differences among elementary and secondary students were roughly one standard deviation in reading and verbal tests within every region of the U.S. (Coleman, Campbell, Hobson, McPartland, Mood, et al., 1966).

TRENDS SINCE 1970

The National Assessment of Educational Progress

In the 1980s, evidence of substantial aggregate change in Black-White test score differences began to accumulate. The primary source of these new data is the National Assessment of Educational Progress (NAEP), a large periodic national testing program with a complex sampling design (Zwick, 1992) that began in the early 1970s. Each participating student is asked to complete only part of each test, and scores for population groups are aggregated from the incomplete data. Until recently, only a few social background characteristics were associated with each student observation, but the 1988 redesign is much richer in variables than its predecessor. Originally, the NAEP samples were not representative at the state level, but this has begun to change as NAEP has become the vehicle for measuring progress toward national educational standards.

The NAEP testing program includes both grade-level tests in grades 4, 8, and 12, and age-specific tests at ages 9, 13, and 17. While the national NAEP samples are relatively large (though

decreasing in size) and well-designed, there are some problematic issues in population coverage. Some schools refuse to participate. Students in special programs are not covered, and student absence and drop-out create coverage problems by age 17. Moreover, there is some non-response among test-takers on racial or ethnic identification. At the same time, NAEP is plainly superior in design and coverage to previous mechanisms for monitoring children's academic performance at the national level and for specific age and population groups.

The age-specific NAEP tests, which cover youth in regular classrooms at every grade level, are designed to permit temporal comparisons of performance. NAEP tests are criterion-referenced, and they are administered on regular cycles of varying length, depending on the subject. Like most other investigators, I focus on the three tests that have been administered most frequently -- reading, science, and mathematics -- and which have gradually been shifted from four-year cycles to administration in every even-numbered year. NAEP uses a repeated cross-section design. It is not a longitudinal study of individuals, so one cannot follow the development of individual performance across time. However, one particular advantage of the NAEP design is that the two- or four-year testing intervals are commensurable with the four-year differences between age groups, so it is sometimes possible to follow the development of birth cohorts from ages 9 to 17 as well as to measure aggregate trends and differentials.

The Scholastic Aptitude Test

A secondary source of data on trends in Black and White test score performance is the Scholastic Assessment Test (SAT) of the College Entrance Examination Board (CEEB). Since the 1940s the SAT has been administered regularly to college-bound seniors (and some juniors) in U.S. high schools. The SAT has two components, verbal (SAT-V) and quantitative (SAT-M),

which are often used by elite colleges and universities in screening applicants for undergraduate admission. Perhaps because of the long decline in SAT-V scores that began in the early 1960s, there have often been well-publicized efforts to tie movements of the SAT to school or youth policies. Trends in SAT performance hit the front page of the *New York Times* each year, and they are often used as key indicators of trends in how schools are performing, as well as in comparisons among groups of students. However, the Preliminary Scholastic Aptitude Test (PSAT) has been administered since 1959 to a national sample of high school juniors (Solomon, 1983). The PSAT is just a shorter version of the SAT, and the problems of self-selection in these samples are limited to those implied in reaching the junior year of high school. In the aggregate, there has been no trend in PSAT performance in the past thirty-five years (Berliner & Biddle, 1995, p. 23-24).

The uses of SAT scores as social indicators are grossly disproportionate to their validity. Test-takers are self-selected from among high school students who plan to attend colleges that require SAT scores. Selection is known to vary across time with respect to academic performance (rank in class), sex, minority status, socioeconomic background, and geographic origin. Presumably, these variations are in part a consequence of variations in the entrance requirements of colleges and universities and of changes in the demand for college education among American youth. Typically, SAT coverage is lower in the central states than on either coast because of competition from less expensive tests of the American College Testing Program (ACT). As Wainer (1987, p. 2) states, "If we wish to draw inferences about all high school seniors, the possibly peculiar events that would impel someone to take the test or not makes these inferences difficult. These difficulties manifest themselves when we try to assess the significance

of changes observed over time. Is the change due to more poorly trained individuals, to a broader cross-section taking the test or merely to a different cross-section of individuals deciding to take the test?"

Problems in interpreting trends in the SAT have given rise to a minor industry of test-score adjustment and analysis. One major goal of the industry is to counter gross misinterpretations of trends and differentials in SAT scores, like meaningless state-to-state comparisons. For example, the highest-scoring states are typically those, like Wisconsin, in which most students take the ACT, which is required by the University of Wisconsin System, while a small minority of elite students take the SAT (Wainer, 1985). A second major goal is to find out what the SAT can actually tell us about trends and differentials in academic performance. This has yielded a lot of clever and careful statistical work, beginning with efforts to explain the long-term decline in SAT-V scores, but this work has provided few definitive answers about trends in academic performance (Wirtz, Howe, et al., 1977; Flynn, 1984; Zajonc, 1976; Zajonc, 1986; Menard, 1988; Alwin, 1991; Morgan, 1991; Murray & Herrnstein, 1992).

My favorite contribution to this literature is an elegant paper by Howard Wainer (1987). He shows that the uncertainty in SAT scores introduced by the average 12 to 14 percent nonresponse on the race-ethnicity question dwarfs the observed changes in minority SAT performance that occurred from 1980 to 1985. The average verbal and math scores of white, minority, and nonresponding test-takers are given. Wainer observes that, if the scores of nonresponding test-takers are the same as those of respondents of the same race-ethnicity, then it is possible to estimate the share of white and minority test-takers among nonrespondents. Depending on whether one uses the verbal or math scores to make the estimates, this estimation

procedure yields very different but rather high estimates of the share of minorities among non-respondents. From 1980 to 1985 the estimated share of minorities among nonrespondents is never less than half and ranges as high as 70 percent, while the share of minorities is always estimated to be higher when one uses mathematical than verbal scores. The discrepant estimates invalidate the assumption that respondents and nonrespondents of the same ethnicity perform equally well, and the resulting uncertainty in test scores is larger than the observed changes in average test performance among minority test-takers.

Black-White Test Score Differences in NAEP

I think the uncertainties of the SAT data are far greater than those of NAEP, and for that reason I focus mainly on trends and differentials in performance on NAEP. However, if one takes the scores at face value, there has also been a partial convergence in Black and White performance on the SAT. For the moment, I ignore the official reports of performance in NAEP and offer a brief review of their treatment in secondary sources. I also postpone discussion of Herrnstein & Murray's (1994) treatment of the NAEP data to a later section.

Jones (1984) was one of the first to examine the Black-White convergence in test scores. NAEP tests in 1971, 1975, 1980, and 1982 showed declining Black-White differences in the percentage of correct responses on the NAEP reading and mathematics tests for children who were born after 1965. He also analyzed differentials in mathematics scores, and, he suggested that the "difference between black and white students in algebra and geometry enrollment might be responsible for a large part of the white-black average difference in mathematics achievement scores" (1984, p. 1209-11).

As the evidence from NAEP accumulated, others noted the trends. A 1986 report of the Congressional Budget Office (CBO), *Trends in Educational Achievement*, reported -- with reference to the previous decline in academic achievement -- "the average scores of black students declined less than those of non-minority students during the later years of the general decline; stopped declining, or began increasing again, earlier; and rose at a faster rate after the general upturn in achievement began" (Koretz, 1986, p. 75-76). In reaching this conclusion, the CBO report relied mainly on trends in average proficiency scores during the first dozen years of NAEP, but it also found corroborating evidence in the SAT, in nationally representative samples of high school seniors of 1971 and 1979, and in several state or local studies. Similarly, Humphreys (1988, p. 240-41) reported substantial gains of Blacks relative to Whites at ages 9, 13, and 17 in the four NAEP reading assessments from 1971 to 1984.

The National Research Council's (NRC) 1989 report, *A Common Destiny: Blacks and American Society*, also reported trends in Black-White gaps in the NAEP assessments of reading, mathematics, and science through 1986 at ages 9, 13, and 17 (Jaynes & Williams, 1989, p. 348-54).[2] Beyond finding signs of aggregate convergence, the NRC panel also disaggregated the Black-White differences by levels of proficiency and by region. For example, they found, "The broad pattern is one of improvement over time at each level of reading proficiency." Furthermore, they found that the same broad pattern of improvement occurred in each of four geographic regions. The NRC report also noted the large remaining cognitive gaps between Blacks and Whites, adding the evidence of a national literacy survey to that from NAEP and the SAT.

In a fascinating preview of *The Bell Curve*, the late Richard Herrnstein (1990a) wrote a review of *A Common Destiny*, which appeared in *The Public Interest*. His main theme was that "*A Common Destiny* suffers ... from one crucial failing: in assessing the gaps separating white and black Americans, it obstinately refuses to consider the evidence concerning racial differences at the individual level" (p. 4). Herrnstein claimed that the themes of the book were "rooted" in the discrimination model, that any Black-White differences were viewed as prima facie evidence of discrimination, while the book "ignores the alternative model, the 'distributional' model, which explains the overlapping of the populations and their differing averages by referring to characteristics of the populations themselves" (p. 6). Herrnstein mainly faulted *A Common Destiny* for failing to root its explanations in IQ differences between Blacks and Whites: "*A Common Destiny* says almost nothing about differences between blacks and whites on standardized tests of intelligence or cognitive aptitude; what little it says is mostly wrong. ... Notwithstanding some vague hints in the book, there is no clear evidence that the gap between the races has been closing recently or that it shrank when the economic gap between the races was shrinking" (p. 7). The remainder of Herrnstein's review argued the centrality of intelligence in accounting for racial differentials in economic status, crime, health, and housing. However, neither in his review nor in a subsequent exchange with the authors of the NRC volume (Hauser, Jaynes & Williams, 1990), did Herrnstein acknowledge the findings from NAEP of decreasing differences in achievement test performance between Blacks and Whites. Rather, he reiterated, "the differences *are* intractable, for we do not know how to eliminate them" (Herrnstein, 1990b, p. 125).

After the hubbub about *A Common Destiny* subsided, other scholars continued to draw on the NAEP test series. Smith & O'Day (1991, p. 72-77) reported declining test score differences at ages 9, 13, and 17 in cohorts tested in reading, mathematics, and science from 1971 to 1988. With respect to reading scores, they observed: "These are extraordinary data. By conservative estimate, they indicate a reduction in the gap between black and white students over the past twenty years of roughly 50 percent when the students are seventeen years old. Moreover, these reductions took place during the same time period as a striking decrease in dropout rates for black students" (p. 75). Smith & O'Day further estimated that the reduction in the Black-White gap in mathematics was on the order of 25 to 40 percent, while that in science was roughly 15 to 25 percent. Finally, Grissmer, Kirby, Berends, & Williamson (1994, p. 11-17) reported narrowing gaps between Blacks and Whites in reading and mathematics achievement using the NAEP data for the middle to late 1970s and for 1990 at ages 13 and 17, while Miller (1995, p. 45-59) offered a detailed review of the performance of Blacks and Whites at ages 9 and 17 for each administration of the reading and mathematics assessments since 1971.

Although these reviews covered various years of NAEP and differed, also in their coverage of specific ages, tests, and functions of test performance, the reviews were unanimous in reporting an overall trend toward reduced Black-White performance differentials. The works cited were also unanimous in drawing attention to remaining large gaps in performance. There was relatively little attention to the reasons for the gaps or their partial closure. As noted above, Jones (1984) had pointed to exposure to math courses as a remaining source of Black-White math score differences, and, following Jones, the NRC report also emphasized differential course taking as well as reduced segregation and compensatory education (Jaynes, et al., 1989, p. 350-52).

Smith & O'Day (1991, p. 79-84) offered no specific analyses of change in test score gaps, but suggested that they might be explained by improved social background and reduced poverty, increased access to preschool, reduced racial isolation -- especially in the South, and changes in instruction and curriculum -- especially increased emphasis on basic skills and minimum competencies. Grissmer, et al (1994, p. xxv-xxxi). carried out detailed analyses of the effects of changes in family background on test scores. They found that changes in family background composition, especially improved maternal schooling and fewer siblings, accounted for about one-third of the improvement in test scores among minority students from the 1970s to 1990.[3]

Achievement or Ability?

There is something schizophrenic in American opinion about cognitive ability and academic achievement. We think we value academic achievement and that it represents, to some degree, the kind of merit we want to see rewarded. We worry endlessly about trends and differentials in academic achievement. We spend a great deal of money to create and improve it in the public schools, and we blame the schools because we think that they have not produced enough of it. We think that if academic achievement were higher, we would do better economically and socially, as individuals and local communities and in the world economy. Yet we grow rigid with apprehension when someone applies terms like ability, intelligence, or -- worse yet -- IQ, rather than academic achievement, to what are usually rather similar and highly correlated measures. We fret about the fairness of standardized tests, though lack of statistical bias is long-established (Wigdor & Garner, 1982, p. 3), and we often disapprove -- both personally and legally -- of the mechanical use of achievement or ability test scores to make decisions about entry to jobs or to schools. Obversely, we have turned test preparation into a

minor industry. Among college admission tests, we prefer the ACT to the SAT because it focuses relatively more on achievement than aptitude, and we applaud the revision of the latter for shifting in the same direction, yet the ACT and old SAT were highly correlated, as are the new and old versions of the SAT.

It is a serious question whether NAEP assessments -- or the SAT for that matter -- are truly tests of achievement, scholastic aptitude, ability, intelligence, or IQ. As a non-member of the psychometric profession, I am inclined to join those who elide or ignore the distinction between achievement and ability (Jencks & Crouse, 1982). I do not believe that ability can be assessed without reference to past learning and opportunity to learn. Moreover, I think it is difficult to maintain sharp distinctions in test content between aptitude and achievement. Thus, while I will not ignore the specific content of tests, I also think that any test performance partly indicates overall levels of realized ability.

For example, I think there is wide agreement that scores on the Armed Forces Qualification Test (AFQT) can justifiably be interpreted much like performance on an IQ test, and there is ample precedent for this, both in the historical development of the test and its use by Loehlin, Lindzey, & Spuhler (1975) and others, including Herrnstein & Murray (1994). At the same time, there is a great deal of evidence that schooling raises scores on IQ tests (Ceci, 1991), and the best recent evidence is based on the AFQT, suggesting that each year of schooling raises IQ by about 2 to 3.5 points (Neal & Johnson, 1994; Korenman & Winship, 1995; Fischer, Hout, Sanchez-Jankowski, Lucas, Swidler, et al., 1996).

Before its recent renaming as the Scholastic Assessment Test, the SAT was called the Scholastic Aptitude Test, although it was based on the original Army Alpha Test of World War I

(Lemann, 1995). Because its purpose is to select among high school seniors, there are no age norms for the test. However, the eminent psychologist, Julian Stanley and his associates have for years applied a set of age norms to SAT scores to select gifted younger students for special summer enrichment programs. For example, the gifted 6th, 7th, and 8th graders who took the SAT in the Midwest Talent Search in 1987 had combined scores of 793 (male) and 656 (female), compared to the combined scores of 1986 college bound seniors of 938 (male) and 877 (female). Among the gifted younger students, average SAT scores increased regularly with age, from a combined score of 696 for those born in 1975 to 826 for those born in 1972 (Northwestern University, 1987).

If SAT scores rise regularly with age and exposure to schooling, do they not reflect achievement as well as aptitude or ability? Throughout *The Bell Curve*, Herrnstein & Murray (1994) play with the tensions and contradictions between our images of ability and achievement, and they repeatedly shift the line between the two to suit their rhetorical purposes. The SAT is at some times a measure of "achievement," whose downward trend shows our neglect of education among the cognitively gifted, while at other times it is a measure of "intelligence," whose use in college entry demonstrates both the establishment of a national cognitive elite and the defects of affirmative action.

Herrnstein & Murray on Black-White Test Score Trends

Although the NAEP tests would appear to be heavily loaded on the achievement end of the spectrum of test content, there is also some precedent for treating them as tests of ability. For this reason, I consider *The Bell Curve*'s treatment of the NAEP findings in detail. In my judgment, the changing test score differentials were given minimal attention, and much of what

was said about them was wrong. In all, about six pages of the main text of *The Bell Curve* (pp. 289-95) were devoted to the question, "Is the difference in Black and White test scores diminishing?" within the forty-six page chapter on "Ethnic Differences in Cognitive Ability." Most of the data on trends in test score differences were put into one of the book's many appendixes (pp. 637-42). One might compare this lack of emphasis on aggregate trend with the twenty-eight pages devoted to an essentially negative review of compensatory education programs.

On page 291 of the main text Herrnstein & Murray present the table, "Reductions in the Black-White Difference on the National Assessment of Educational Progress," which is based upon summary data from the early 1970s through 1990 (Mullis, Dossey, Foertsch, Jones & Gentile, 1991), and which I have reproduced here as Table 1. Across math, science, and reading examinations, and at ages 9, 13, and 17, Herrnstein & Murray report that the Black-White difference declined by an average 0.28 standard deviations between 1969 to 1973 and 1990. They describe these changes as presenting "an encouraging picture" (p. 291). After adding a summary of changes in the SAT, "from 1.16 to .88 standard deviation in the verbal portion of the test and from 1.27 to .92 standard deviation in the mathematics portion of the test," Herrnstein & Murray conclude that there has been a "narrowing of approximately .15 to .25 standard deviation units, or the equivalent of two to three IQ points overall" (p. 292). Apparently, Herrnstein & Murray temper their arithmetic with cautionary data from their fifth appendix when they decide that changes of 0.28, 0.28, and 0.35 standard deviations suggest a range of 0.15 to 0.25 standard deviation units. Then, in an endnote, they discount this range by a factor of 0.6 or 0.8 -- to account for the imperfect relationship between SAT or NAEP tests and IQ -- in order to come up

with the estimated change of two to three IQ points. They acknowledge that, if one relied on the SAT alone, the data would suggest a narrowing of 4 IQ points, "but only for the population that actually takes the test" (n. 57, p. 721). Even while acknowledging the trends toward convergence in test scores, Herrnstein & Murray were quick to point out that some of the trend was due to declining scores among Whites, rather than increasing scores among Blacks, and to add that it would be foolhardy to extrapolate the observed trends into the future. Whatever the specific estimate of test score convergence between Blacks and Whites, one would be hard-pressed to find any acknowledgment of it once Herrnstein & Murray started drawing conclusions and making recommendations.

The more I thought about their findings, the more curious seemed Herrnstein & Murray's treatment of the NAEP data, for it is the only set of test scores considered by them that consistently covers an unselected sample of the general population. If one applied their range of discount factors to their estimate of the test score convergence in the NAEP data alone, the estimated closure would lie between 2.5 and 3.4 points, which is not bad for aggregate change in an immutable quantity over a twenty year period. But there is more to the story than this, for the footnote in Herrnstein & Murray's table declares that they "assume a standard deviation of 50." I had recalled some variation in the standard deviations of the NAEP test scores across tests, from year to year, and between Blacks and Whites, so I went back to the source.

This proved a cautionary lesson in what the book jacket of *The Bell Curve* cited as the "relentless and unassailable thoroughness" of Herrnstein & Murray's analysis. To begin with, several of the numbers in the table are simply wrong. There are no fewer than five copying or multiplication errors in age- and test-specific entries in the body of the table, and these lead to

other errors in average differentials and in measures of change. In the end, the effect of these errors is small; the overall average change is 0.29, rather than 0.28.

But this is the least of their problems with the NAEP data. In their "relentless and unassailable thoroughness," Herrnstein & Murray evidently confined their reading to a one-page summary of change in the test score differences (p. 11), plus a footnote on page 1 of the source (Mullis, et al., 1991), stating that "each scale was set to span the range of student performance across all three ages in that subject-area assessment and to have a mean of 250.5 [sic] and a standard deviation of 50." However, a series of appendix tables provides details of the test score distributions for each population year by year, including their standard deviations, which are typically much less than the value of 50 adopted by Herrnstein & Murray. The difference is mainly due to the incorporation of variation by age in the larger overall value, whereas the Black-White comparisons should have been conditioned on age, just as Herrnstein & Murray attempted to condition on age in their regression analyses of the effects of the AFQT. The effect of choosing too large a standard deviation was to understate both the initial Black-White differences and the changes in test scores over time in standard deviation units.

Table 2 shows the change in test scores using the estimated standard deviations of the total population of each age in 1990 as the unit of measure. Mullis, et al (1991). did not provide standard deviations for test scores in science and mathematics in the period around 1970, and for this reason I based the comparisons on the population standard deviations in 1990. In science and mathematics, though not in reading, the variability of the tests declined across time. Both the initial differences between Blacks and Whites in science and mathematics and the changes in those

test score differences would be somewhat smaller if the changes had been normed on the standard deviations in 1970.

Using the revised standard deviations raises the overall average convergence from 0.29 to 0.39 standard deviations. Based on Herrnstein & Murray's assumptions, this raises the implied convergence in IQ between Blacks and Whites to a range between 3.5 and 4.7 points. I wonder whether Herrnstein & Murray would have waxed so eloquent about immutability and ineducability if they had acknowledged aggregate changes in test score differentials of this magnitude in the general population over the past two decades.

In one important respect, Herrnstein & Murray were surely right, for it is most dangerous to project trend lines unthinkingly. Yet another set of NAEP assessments -- for 1992 -- became available after *The Bell Curve* went to press, and they appear to confirm that the trend toward convergence in Black and White test scores was reversed after 1986 to 1988 (Mullis, Dossey, Campbell, Gentile, O'Sullivan, et al., 1994). For example, Figure 1 shows trends in the average (mean) NAEP scores of Blacks and Whites at age 13 in reading, science, and mathematics.[4] The years of greatest convergence are not entirely clear because there are no reading scores for 1986, nor science or math scores for 1988. It does appear that, sometime in the middle to late 1980s, the convergent trend ended, and Black-White gaps returned to levels of the early 1980s.

Was There No Change at All?

Immediately after the publication of *The Bell Curve* in October 1994, most commentary on Black-White test score differences focused either on the specious genetic arguments of the book or on its review of compensatory education programs. There was almost no reaction to the book's treatment of aggregate trend data. One significant exception was a letter from fifty-two

academics, which appeared on the editorial page of the *Wall Street Journal* as "Mainstream Science on Intelligence" (Arvey, 1994). The letter purported to outline conclusions "regarded as mainstream among researchers on intelligence, in particular on the nature, origins, and practical consequences of individual and group differences in intelligence." Among the letter's twenty-five conclusions, items 19 and 20 bear on change in Black and White test score differentials:

> 19. There is no persuasive evidence that the IQ bell curves for different racial-ethnic groups are converging. Surveys in some years show that gaps in academic achievement have narrowed a bit for some races, ages, school subjects and skill levels, but this picture seems too mixed to reflect a general shift in IQ levels themselves.
>
> 20. Racial-ethnic differences in IQ bell curves are essentially the same when youngsters leave high school as when they enter first grade. However, because bright youngsters learn faster than slow learners, these same IQ differences lead to growing disparities in amount learned as youngsters progress from grades 1 to 12. As large national surveys continue to show, black 17-year-olds perform, on the average, more like white 13-year-olds in reading, math and science, with Hispanics in between.

I thus looked further at the NAEP series to learn whether my reading of the trends -- and that appearing in published reviews -- might have been mistaken.

Figure 2 summarizes trends in White-Black differences in NAEP proficiency scores in the major subject matter series: reading, science, and mathematics. For each subject, I have used the same scale to show trend lines in mean test score differences by age, but I have arrayed the data

by birth year, rather than by year of assessment. For example, in the upper panel, the reading assessment covers the cohorts of 1954 to 1975 at age 17, the cohorts of 1958 to 1979 at age 13, and the cohorts of 1962 to 1983 at age 9. With this arrangement of the data, it is possible to compare White-Black differences in performance levels across ages by reading the graph vertically at a given birth year. For example, reading performance was assessed for the cohort of 1962 both at ages 9 and 13, and those performance levels might also be compared with that of the adjacent cohort of 1963, measured at age 17.

A first observation about the data in Figure 2 is that there is no distinct pattern to the within-cohort comparisons in any subject. That is, for members of the same (or adjacent) cohorts, the Black-White differences are only occasionally larger at age 17 than age 13 and at age 13 than at age 9. In other cases, the White-Black differences are largest at age 9 or at age 13. Thus, I find no substantial or consistent support in the NAEP data for the claim of Arvey, et al. that there are "growing disparities in amount learned as youngsters progress from grades 1 to 12." To be sure, one might observe smaller differences in the NAEP series at age 17 because low-scorers tend to drop out, but available data provide no evidence about the effects of attrition.

More important, Figure 2 clearly shows the major declines in White-Black differences in reading, science, and mathematics achievement. This arrangement of the data suggest that the declines may have occurred in particular cohorts, beginning before age 9. In reading, for example, the major gains for Blacks occurred in the cohorts of 1962 to 1971. In science, the gains appear, although not entirely consistently, for cohorts born after 1960. In mathematics, the Black gains appear for cohorts born from 1956 to 1973. In science, but not in reading or mathematics, Black gains were in part a consequence of declining test scores among Whites. I

would hesitate to connect these test score changes too closely to changes in IQ, although Herrnstein & Murray showed no hesitation in doing so. However, in my opinion, the trends have not been "too mixed" to reflect a general shift in test score differentials among cohorts born from the late 1950s to the early 1970s.

Mean achievement scores provide important -- but limited -- information about levels of achievement in the population. Such a measure of central tendency may be insensitive to changes in the shape or dispersion of the distribution. For this reason I have examined changes over time across the entire achievement distributions among Blacks. For example, Figure 3 shows selected percentile points of the NAEP mathematics distributions for Black children at ages 9, 13, and 17 from 1971 through 1992. The test scores are reported in the metric of "proficiency levels" for the 5th, 10th, 25th, 50th (median), 75th, 90th, and 95th percentiles at each age and year. I have used line patterns and symbols for the data points that emphasize the rough symmetry in the distributions; the same lines and marker shapes are used for the corresponding pairs of percentile points below and above the median: 5 and 95, 10 and 90, and 25 and 75. Thus, it is possible to follow changes both in the level and shape of the distributions across time. One should bear in mind that the writers of the *Wall Street Journal* letter could not have observed the final (1992) data points in each series.

Figure 3 shows that, at age 9, mathematics performance improved steadily throughout the distribution from 1978 to 1990, but especially between 1982 and 1990. At age 13, mathematics performance grew between 1978 and 1986, but leveled off thereafter.[5] Growth appears to have been more rapid in the lower half of the distribution than in the upper half. At age 17, there was steady growth from 1978 to 1990 throughout the distribution, but performance fell between 1990

and 1992 at the top and bottom of the distribution. During the 1978 to 1986 period, growth appears to have been faster at the bottom than at the top of the distribution.

I have also examined detailed displays of change in the distributions of reading and science proficiency, and the pattern of findings is similar to that in mathematics. There is much more to the NAEP trends than the statement by Arvey, et al. (1994) that "gaps in academic achievement have narrowed a bit for some races, ages, school subjects and skill levels." Almost all of the growth in Black scores led to convergence in the performance of Blacks and Whites. Although growth in Black performance did not occur between every assessment from the 1970s to 1990, growth did occur throughout most of this period and throughout the entire distributions of performance levels. Thus, I think that the statement in "Mainstream Science" substantially understates the extent and pervasiveness of change in Black achievement test scores. Indeed, I doubt that many readers would independently describe the NAEP series in the same terms as those of "Mainstream Science."

There is obviously a great deal of opportunity for research on the sources of Black-White test score convergence during the period from 1970 to the middle 1980s and on the sources of the subsequent slowdown or reversal. One might think, for example, of the reduced enthusiasm for compensatory education after the first Reagan administration took office in 1980 and of the length of time required for its effects to take hold. There also remains the possibility that some part of the convergence or of its reversal may be explained as methodological artifacts of the NAEP design. On the other hand, relative to the larger body of evidence on change in test scores, it seems hard to believe that the NAEP assessments are especially vulnerable to methodological error. I am more inclined to think that both the convergence and its subsequent reversal are real

and that both suggest the mutability of Black-White test score differences, even if the mechanisms of change are now poorly understood.

Summary and Discussion

An increasing array of evidence suggests that Black-White differences in cognitive tests have been reduced for cohorts born after the middle 1960s. While several test score series show some signs of convergence, the NAEP series in reading, science, and mathematics, which cover ages 9, 13, and 17, are more nearly representative of the general population than other testing programs. As Smith & O'Day summarized the findings, between 1970 and the middle to late 1980s, initial test-score differences in reading were reduced by 50 percent, those in mathematics were reduced by 25 to 40 percent, and those in science were reduced by 15 to 25 percent. However, there is cause for concern in the last two rounds of NAEP data, for the gains of the 1970s and early 1980s may have begun to erode. A preliminary report of NAEP findings from 1994 suggests that the recent divergent trend may have ended, but it has not been reversed. From 1992 to 1994 there were no significant changes in achievement differences between Blacks and Whites at ages 9, 13, or 17 (Campbell, Reese, O'Sullivan & Dossey 1996).

Of what importance is convergence in achievement test scores of Blacks and Whites? Herrnstein & Murray (1994) argue that IQ or *g* is the key source of variability in adult social and economic success. In so arguing, they follow a strong tradition in psychology. For example, referring to occupational standing, Jensen (1986, p. 318) writes

> Although *g* cannot account for all the variance in occupational level, it accounts for more than any other measurable sources of variance, independent of *g* that we have been able to discover.

If that were the case, I should be most concerned about the strength of the link of IQ with test series like the NAEP assessments, and I should also look for more direct evidence about trends in IQ differentials between Blacks and Whites.

On the other hand, there is increasing evidence that IQ or *g* is neither the sole nor necessarily the most important cognitive factor in adult success. Much of this evidence comes from new analyses of the National Longitudinal Study of Youth (NLSY), the same data analyzed by Herrnstein & Murray (1994) in *The Bell Curve*. For example, the Numerical Operations (NO) and Computational Speed (CS) components of the Armed Forces Vocational Aptitude Battery (ASVAB) are not closely related to the IQ factor measured by the four components of the ASVAB that make up the AFQT (Herrnstein & Murray, 1994, p. 580-83). Yet Goldberger (1995) and Heckman (1995) have found that NO and CS are at least as important as the AFQT in the determination of earnings. Also, Corcoran (1996) has found great variation in the importance of the several components of the ASVAB in determining educational, economic, and social success. That is, the several outcomes analyzed in *The Bell Curve* appear to respond differentially to the several components of the ASVAB, and the differential responses are not explained by the closeness of the components to a general ability factor. These findings, I believe, suggest the importance of the array of cognitive tests across which Black performance has begun to converge toward that of Whites, whether or not those tests may be said to reflect IQ or *g*. It is unfortunate that there are not more longitudinal data in which the effects of a full range of test performances can be assessed across a broad array of life outcomes.

REFERENCES

Alwin, D. F. (1991, October). Family of Origin and Cohort Differences in Verbal Ability. *American Sociological Review, 56*(5), 625-38.

Armor, D. (1992, Summer). Why is Black Educational Achievement Rising? *The Public Interest, 108,* 65-80.

Arvey, R. D., et al. (1994, 13/December). Mainstream Science on Intelligence. *Wall Street Journal.*

Berliner, D. C., & Biddle, B. J. (1995). *The Manufactured Crisis: Myths, Fraud, and the Attack on America's Public Schools.* Reading, Massachusetts: Addison-Wesley.

Campbell, J.R., Rees, C.M., O'Sullivan, C., & Dossey, J.A. (1996). *NAEP 1994 Trends in Academic Progress* (National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education). Washington, D.C.: Government Printing Office.

Ceci, S. J. (1991). How Much Does Schooling Influence General Intelligence and Its Cognitive Components? A Reassessment of the Evidence. *Developmental Psychology, 27*(5), 703-722.

Coleman, J. S., Campbell, E. Q., Hobson, C. J., McPartland, J., Mood, A. M., Weinfeld, F. D., & York, R. L. (1966). *Equality of Educational Opportunity* (Office of Education, U.S. Department of Health, Education, and Welfare). Washington, D.C.: Government Printing Office.

Corcoran, J. (1996). Beyond The Bell Curve and g: Rethinking Ability and Its Correlates [Senior honors thesis]. Cambridge, Massachusetts: Department of Sociology, Harvard University.

Fischer, C. S., Hout, M., Sanchez Jankowski, M., Lucas, S. R., Swidler, A., & and Voss, K. (1996). *Understanding Inequality in America: Beyond the Bell Curve.* Princeton: Princeton University Press.

Flynn, J. R. (1984). The Mean IQ of Americans: Massive Gains 1932 to 1978. *Psychological Bulletin, 95*(1), 29-51.

Flynn, J. R. (1987). Massive IQ Gains in 14 Nations: What IQ Tests Really Measure. *Psychological Bulletin, 101*(2), 171-191.

Goldberger, A. S. (1995, December). Abilities, Tests, and Earnings. MacArthur Foundation Conference on Meritocracy and Inequality. Madison, Wisconsin.

Grissmer, D. W., Kirby, S. N., Berends, M., & Williamson, S. (1994). *Student Achievement and the Changing American Family.* Washington, D.C.: RAND Institute on Education and Training.

Hauser, R. M., Jaynes, G. D., & Williams, R. M., Jr. (1990, Spring). Explaining Black-White Differences. *The Public Interest, 99,* 110-119.

Heckman, J. J. (1995, October). Lessons from The Bell Curve. *Journal of Political Economy, 103*(5), 1091-1120.

Herrnstein, R. (1990a, Winter). Still an American Dilemma. *The Public Interest, 98,* 3-17.

Herrnstein, R. (1990b, Spring). On Responsible Scholarship: A Rejoinder. *The Public Interest, 99,* 120-27.

Herrnstein, R. J., & Murray, C. (1994). *The Bell Curve: Intelligence and Class Structure in American Life.* New York: The Free Press.

Humphreys, L. G. (1988). Trends in Levels of Academic Achievement of Blacks and Other Minorities. *Intelligence, 12,* 231-60.

Jaynes, G. D., & Williams, R. M., Jr. (Eds.). (1989). *A Common Destiny: Blacks and American Society* (Committee on the Status of Black Americans, Commission on Behavioral and Social Sciences, National Research Council). Washington, D.C.: National Academy Press.

Jencks, C. S., & Crouse, J. (1982). Aptitude vs. Achievement: Should We Replace the SAT? In W. Schrader (Ed.), *New Directions for Testing and Measurement, Guidance, and Program Improvement*. San Francisco: Jossey-Bass.

Jensen, A. R. (1986, December). G: Artifact or Reality? *Journal of Vocational Behavior, 29*(3), 301-331.

Jones, L. V. (1984, November). White-Black Achievement Differences: The Narrowing Gap. *American Psychologist, 39*(11), 1207-1213.

Korenman, S., & Winship, C. (1995, October). *A Reanalysis of The Bell Curve*. Unpublished paper.

Koretz, D. (1986). *Trends in Educational Achievement* [Congress of the United States, Congressional Budget Office]. Washington, D.C.: Government Printing Office.

Lemann, N. (1995, August). The Structure of Success in America: The Untold Story of How Educational Testing Became Ambition's Gateway -- and a National Obsession. *The Atlantic Monthly,* pp. 41-60.

Loehlin, J. C., Lindzey, G., & Spuhler, J. (1975). *Race Differences in Intelligence.* San Francisco: W.H. Freeman and Company.

Menard, S. (1988, September). Going Down, Going Up: Explaining the Turnaround in SAT Scores. *Youth & Society, 20*(1), 3-28.

Miller, L. S. (1995). *An American Imperative: Accelerating Minority Eduational Advancement.* New Haven: Yale University Press.

Morgan, R. (1991). Cohort Differences Associated with Trends in SAT Score Averages. (College Board Report, no. 91-1). Princeton, NJ: Educational Testing Service.

Mullis, I. V., Dossey, J. A., Campbell, J. R., Gentile, C. A., O'Sullivan, C., & Latham, A. (1994). *NAEP 1992 Trends in Academic Progress: Achievement of U.S. Students in Science, 1969-70 to 1992; Mathematics, 1973 to 1992; Reading, 1971 to 1992; and Writing, 1984 to 1992* (National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education). Washington, D.C.: Government Printing Office.

Mullis, I. V., Dossey, J. A., Foertsch, M. A., Jones, L. R., & Gentile, C. A. (1991). *Trends in Academic Progress: Achievement of U.S. Students in Science, 1969-70 to 1990; Mathematics, 1973 to 1990; Reading, 1971 to 1990; and Writing, 1984 to 1990* (National Center for Education Statistics, Office of Educational Research and Improvement, U.S. Department of Education). Washington, D.C.: Government Printing Office.

Murray, C., & Herrnstein, R. (1992, Winter). What's Really Behind the SAT-Score Decline? *The Public Interest, 106,* 32-56.

Neal, D. A., & Johnson, W. R. (1994, November 21, 1994). The Role of Pre-Market Factors in Black-White Wage Differences. University of Chicago.

Northwestern University. (1987). *Statistical Summary and Interpretation.* 1987 Midwest Talent Search. Evanston, Illinois: Center for Talent Development.

Osborne, T. R., & McGurk, F. C. (1982). *The Testing of Negro Intelligence, Volume 2.* Athens, Georgia: The Foundation for Human Understanding.

Shuey, A. M. (1966). *The Testing of Negro Intelligence.* New York: Social Science Press.

Smith, M. S., & O'Day, J. (1991). Educational Equality: 1966 and Now. In D. Verstegen & J. Ward (Eds.), *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook* (pp. 53-100). New York: Harper Collins.

Solomon, R. J. (1983). Information Concerning Mean Test Scores for the Graduate Management Admission Test (GMAT); Graduate Record Examination (GRE); Law School Admission Test (LSAT); Preliminary Scholastic Aptitude Test (PSAT); and Scholastic Aptitude Test (SAT) for the National Commission on Excellence in Education. Princeton, New Jersey: Educational Testing Service.

Wainer, H. (1985). Some Pitfalls Encountered While Trying to Compare States on Their SAT Scores: Page and Fiefs as an Example (85-62). Princeton, NJ: Educational Testing Service.

Wainer, H. (1987, January). Can We Accurately Assess Changes in Minority Performance on the SAT? Educational Testing Service.

Wigdor, A. K., & Garner, W. R. (Eds.). (1982). *Ability Testing: Uses, Consequences, and Controversies, Part I: Report of the Committee* (Committee on Ability Testing, Assembly of Behavioral and Social Sciences, National Research Council). Washington, D.C.: National Academy Press.

Wirtz, W., Howe, H., II, & and others. (1977). *On Further Examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline.* New York: College Entrance Examination Board.

Zajonc, R. B. (1976). Family Configuration and Intelligence. *Science, 192,* 227-36.

Zajonc, R. B. (1986). The Decline and Rise of Scholastic Aptitude Scores: A Prediction Derived from the Confluence Model. *American Psychologist, 41,* 862-867.

Zwick, R. (1992, Summer). Special Issue on the National Assessment of Educational Progress. *Journal of Educational Statistics, 17*(2), 93-232.

FOOTNOTES

1. Those conclusions would not change had I also considered Osborne & McGurk's (1982) continuation of Shuey's review, which covered the period from 1966 to 1979.

2. I was a member of the NRC Committee on the Status of Black Americans and contributed to Chapter 7, "The Schooling of Black Americans."

3. See Armor (1992) for similar findings.

4. Similar trends appear at age 17 and, to a lesser degree, at age 9.

5. Unfortunately, the published series do not include percentile points from the initial mathematics assessment of 1973. The convergence between the average (mean) performance levels of Black and White students in mathematics first appears between 1973 and 1978 at ages 9, 13, and 17 (Mullis, et al., 1994, p. A63-A71).

Table 1. Reductions in the Black-White Difference on the National Assessment of Educational Progress

	White-Black Difference, in Standard Deviations		
	1969-73	1990	Change
9 year olds			
Science	1.14	0.84	-0.30
Math	0.70	0.54	-0.16
Reading	0.88	0.70	-0.18
Average	0.91	0.69	-0.21
13 year olds			
Science	0.96	0.76	-0.20
Math	0.92	0.54	-0.38
Reading	0.78	0.40	-0.38
Average	0.89	0.57	-0.32
17 year olds			
Science	1.08	0.96	-0.12
Math	0.80	0.42	-0.38
Reading	1.04	0.60	-0.44
Average	0.97	0.66	-0.31
Overall average	0.92	0.64	-0.28

Source: Mullis, et al. (1991), as reported by Herrnstein & Murray (1994: 291), who stated, "The computations assume a standard deviation of 50."

Table 2. Revised Estimates of Reductions in the Black-White Difference on the National Assessment of Educational Progress

	White-Black Difference, in Standard Deviations		
	1969-73	1990	Change
9 year olds			
Science	1.42	1.04	-0.38
Math	1.06	0.82	-0.24
Reading	0.98	0.79	-0.19
Average	1.15	0.88	-0.27
13 year olds			
Science	1.28	1.01	-0.27
Math	1.48	0.87	-0.61
Reading	1.07	0.58	-0.49
Average	1.28	0.82	-0.46
17 year olds			
Science	1.17	1.04	-0.13
Math	1.29	0.68	-0.61
Reading	1.28	0.71	-0.57
Average	1.25	0.81	-0.44
Overall average	1.23	0.84	-0.39

Source: Author's computations from Mullis, et al. (1991), using standard deviations for each age group in 1990.

Figure 1. Trends in NAEP Scores at Age 13 among Blacks and Whites

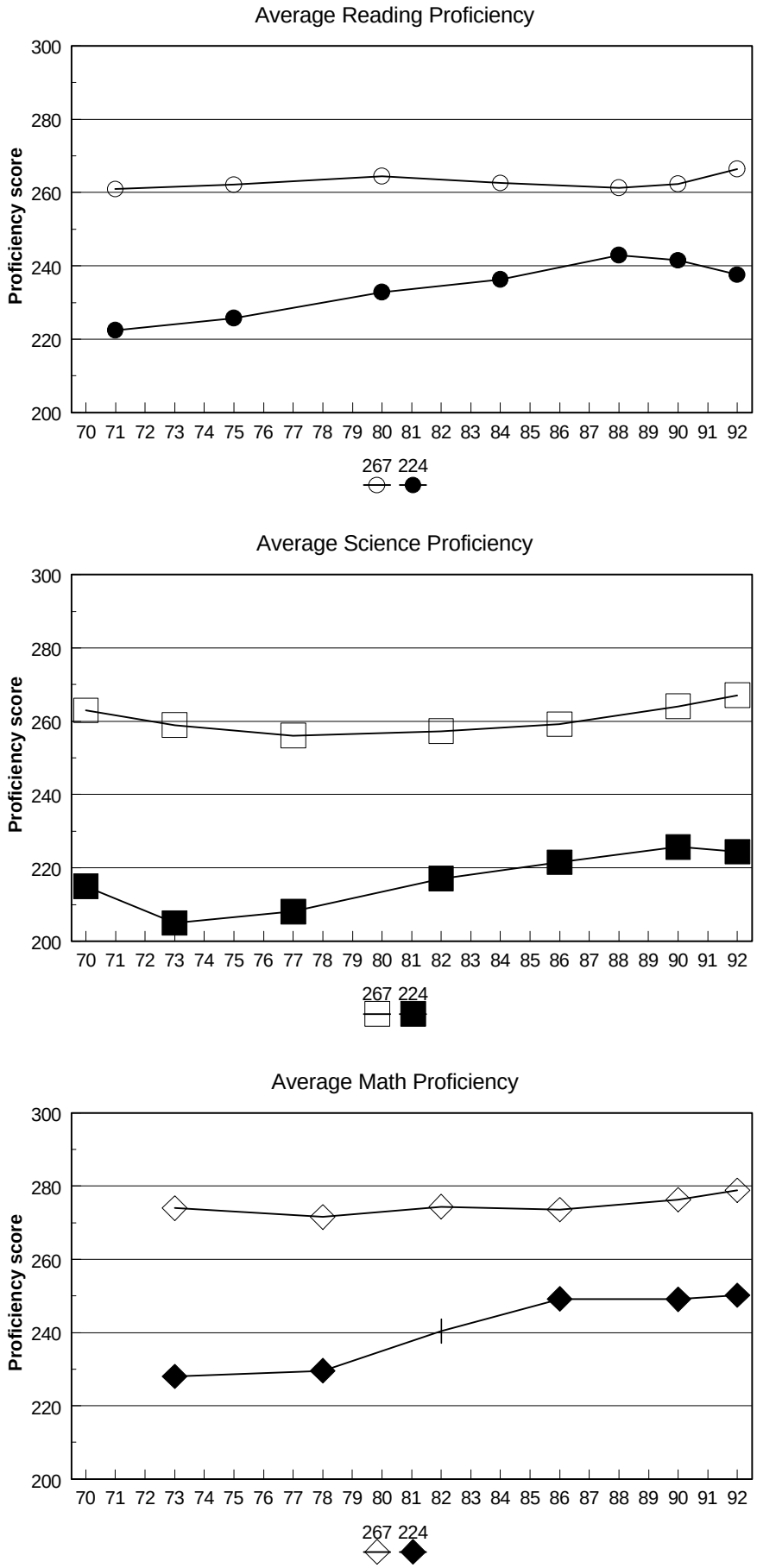


Figure 2. White-Black Differences in NAEP Proficiency Scores by Subject, Age, and Year of Birth

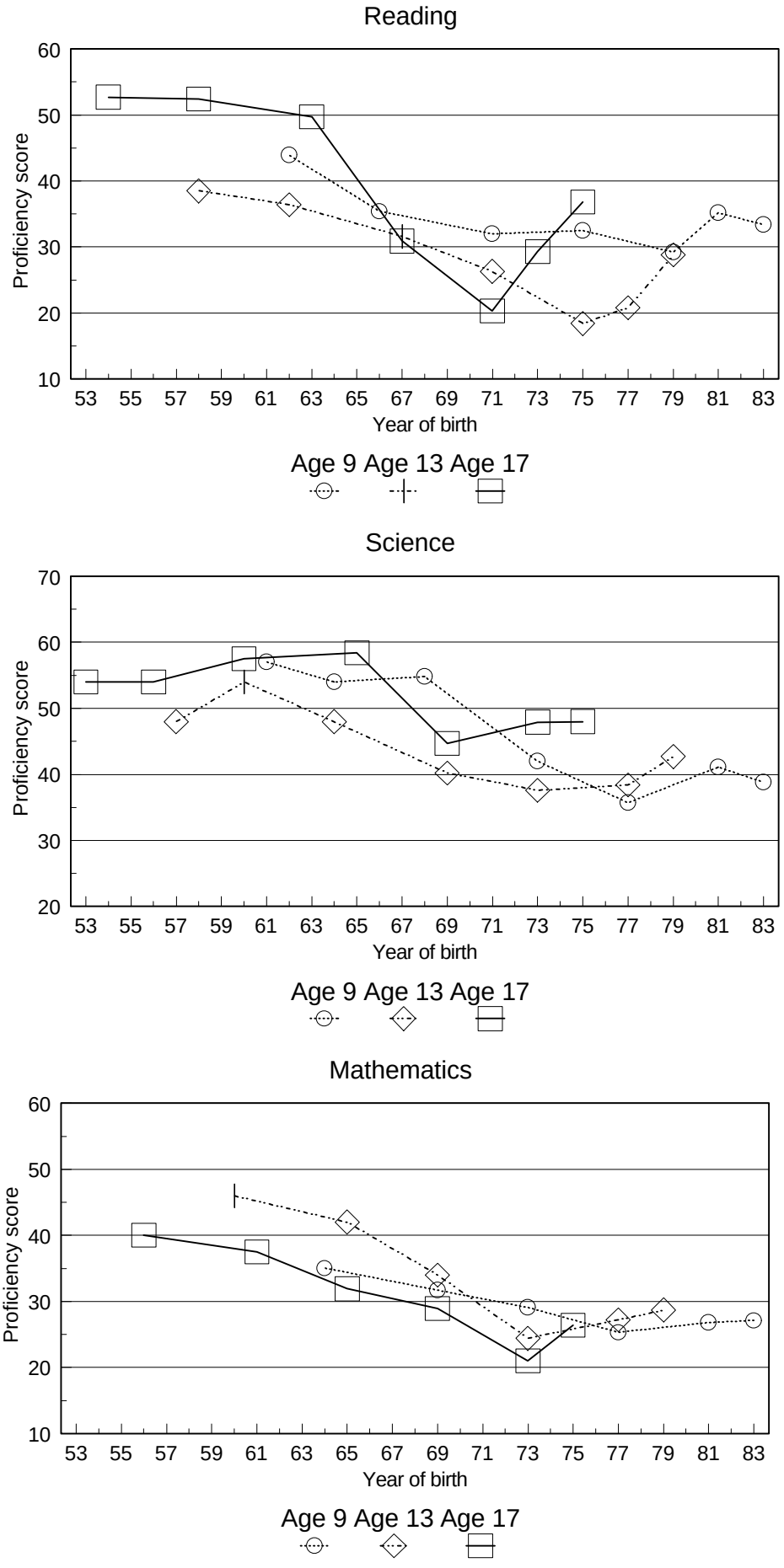


Figure 3. NAEP Mathematics Trend Assessment: Percentiles of the Mathematics Distribution among Blacks by Age, 1978 to 1992

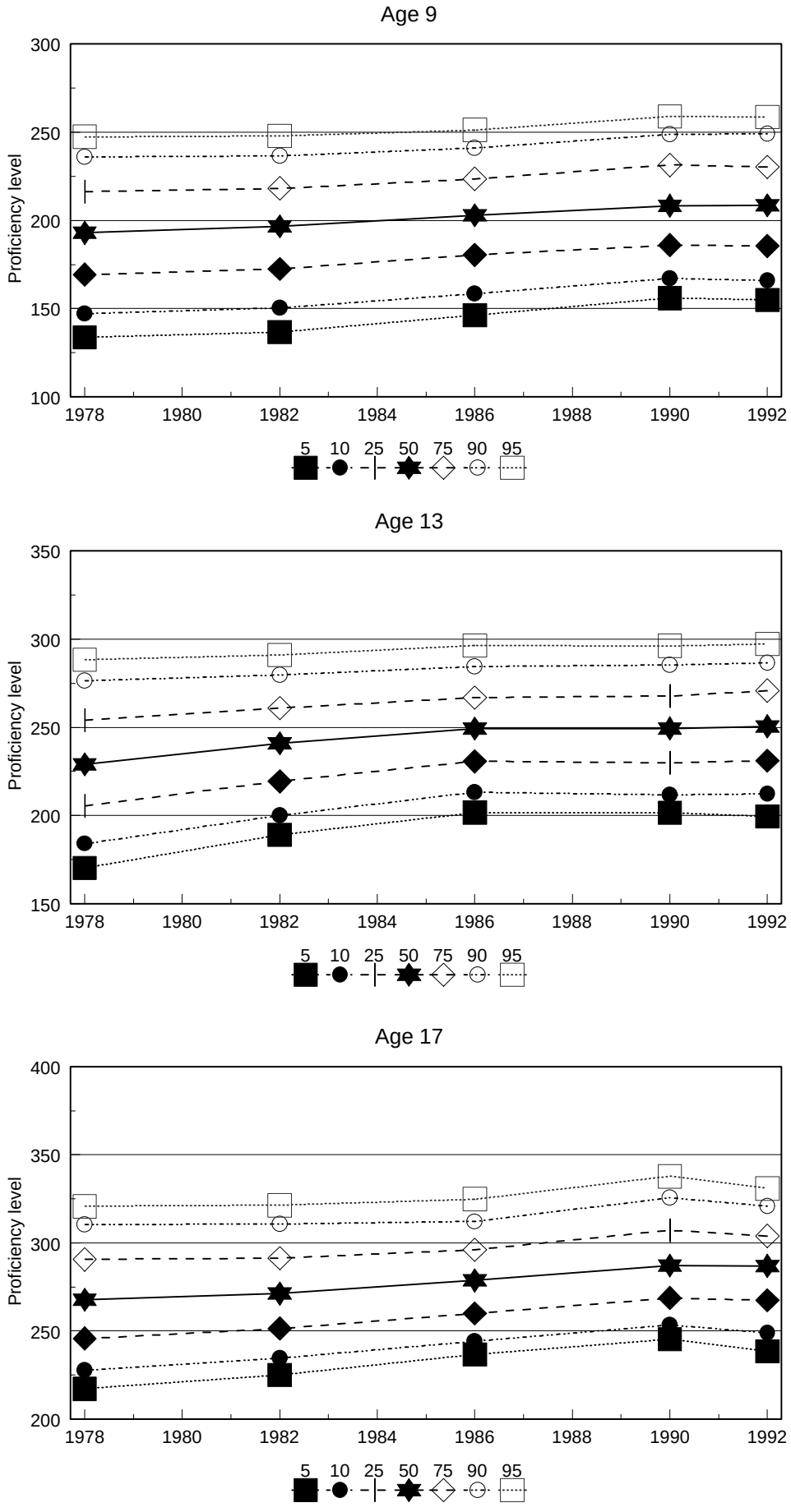
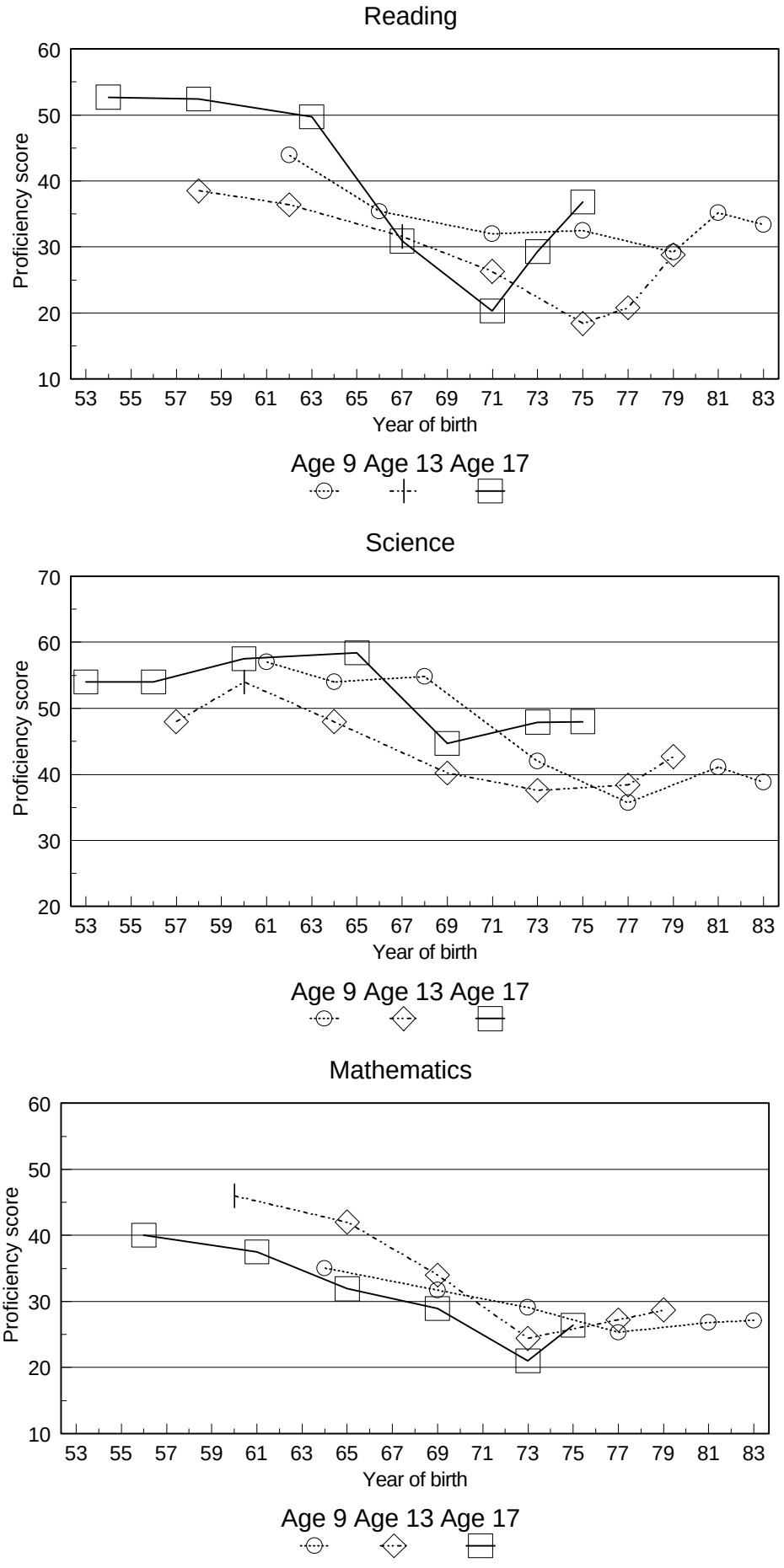


Figure 2. White-Black Differences in NAEP Proficiency Scores by Subject, Age, and Year of Birth



Center for Demography and Ecology
University of Wisconsin
1180 Observatory Drive Rm. 4412
Madison, WI 53706-1393
U.S.A.
608/262-2182
FAX 608/262-8400
comments to: hauser@ssc.wisc.edu
requests to: cdepubs@ssc.wisc.edu