Institute for Research on Poverty
Discussion Paper no. 1110-96

Trends in Black-White Test-Score Differentials

Robert M. Hauser
Min-Hsiung Huang

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

October 1996

Prepared for the conference, “Intelligence on the Rise: Secular Changes in IQ and Related Measures,” Emory University, Atlanta, 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, U.S. Department of Health and Human Services; the Vilas Estate Trust; and the Center for Demography and Ecology at the University of Wisconsin–Madison. The opinions expressed herein are those of the authors.

Abstract

Until the 1970s, there were few signs of change in the historic difference of one standard deviation between average ability or achievement test scores of blacks and whites in the United States. From about 1970 to the mid- to late 1980s, there was a substantial convergence of the average achievement test scores of black and white youth; however, from the mid- to late 1980s to 1992, test scores began to diverge again. Although we place the greatest weight on data from the National Assessment of Educational Progress (NAEP), the convergence also appeared in other test series. Herrnstein and Murray’s highly visible work, *The Bell Curve*, stands almost alone in minimizing the importance of the convergent trend. We also find a longer-term trend of convergence between the verbal abilities of blacks and whites in data from the General Social Survey (GSS), which covers adult cohorts born since 1909.

Trends in Black-White Test-Score Differentials

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 a test was normed with $\mu_W = 100$ and $\sigma_W = 15$ in the white population, then it was reasonable to expect that $\mu_B \approx 85$ in the black population. If test scores were normally distributed, then 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 change in test scores and the types of tests in which it has occurred. Our understanding of these events is incomplete, but, by the standards of evidence routinely accepted in psychological research on test performance, we believe that the available evidence of change is highly significant and incontrovertible.

TRENDS BEFORE 1970

Loehlin, Lindzey, and Spuhler (1975:140–41) 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, and Spuhler (p. 141) 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."[1]

Second, Loehlin, Lindzey, and Spuhler compared the test-score distributions of black and white military recruits from 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 [AGCT], and Armed Forces Qualifying Test [AFQT]), 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, these findings are of questionable quality. The population coverage is haphazard; scientific sampling is the exception; scores on a variety of tests are rendered nominally comparable, merely by assuming normality in the trait distribution and by expressing scores as deviations from the mean. In short, up to about 1980, data on trends in black-

[1] These conclusions would not change had we also considered Osborne and McGurk's (1982) continuation of Shuey's review, which covered the period from 1966 to 1979.

white differences in IQ were probably of even worse quality than those available for global assessments of trend in test performance (Flynn 1984, 1987). All the same, the Coleman-Campbell report of 1966, *Equality of Educational Opportunity*, should 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 United States (Coleman 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 dropout create coverage problems at age 17. Moreover, there is some nonresponse 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, we focus on the three tests 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 Aptitude 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 our schools are performing, as well as in comparisons among groups of students.

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 the less expensive tests of the American College Testing Program (ACT). As Wainer (1987:2) 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 the efforts to explain the long-term decline in SAT-V scores, but this work has yielded few definitive answers about academic performance (Wirtz et al. 1977; Flynn 1984; Zajonc 1976, 1986; Menard 1988; Alwin 1991; Morgan 1991; Murray and Herrnstein 1992).

Our favorite contribution to this literature is an elegant 1987 paper by Howard Wainer. 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. We know the average verbal and math scores of white, minority, and nonresponding test-takers. If, as Wainer posits, we assume that the scores of nonresponding test-takers are the same as those of respondents of the same race/ethnicity, then we can 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 nonrespondents. From 1980 to 1985, the estimated share of minorities among nonrespondents is never less than half and ranges as high as 70 percent, and the shares of minorities among nonrespondents are always estimated to be higher when using mathematical rather 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 test performance.

TEST PERFORMANCE DIFFERENTIALS SINCE 1970

We think the uncertainties of the SAT data are far greater than those of NAEP, and for that reason we focus mainly on trends and differentials in performance on NAEP.[2] 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, we ignore the official reports of performance in the NAEP and offer a brief

[2] 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 and Biddle 1995:23–24). We do not know whether it is possible to disaggregate these trend data by population groups.

review of their treatment in secondary sources. We also postpone our discussion of Herrnstein and Murray's 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 analyzed differentials in mathematics scores, and 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" (see Jones 1984: 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 1986:75–76). 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, and in nationally representative samples of high school seniors of 1971 and 1979, and in several state or local studies. Similarly, Humphreys (1988:240–41) 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 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 and Williams 1989:348–54).[3] Beyond finding signs of aggregate convergence, the NRC panel was also able to disaggregate the black-white

[3] The senior author of this paper was a member of the NRC Panel on the Status of Black Americans and contributed to chapter 7, "The Schooling of Black Americans," in Jaynes and Williams 1989.

differences by levels of proficiency and by region. With respect to levels of reading proficiency, they found that “The broad pattern is one of improvement over time at each level of reading proficiency.” Further, they found that the same broad pattern of improvement occurred in each of four geographic regions. At the same time, the NRC report 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 wrote a review critical of *A Common Destiny*, which appeared in *The Public Interest* (Herrnstein 1990a). 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 *A Common Destiny* 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). This criticism seemed misplaced, to say the least, for several of the contributors to *A Common Destiny* were demographers who had typically used distributional explanations throughout their work. Evidently, Herrnstein mainly intended to fault *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). Herrnstein went on to recap the data on black-white differences in functional literacy that were presented in *A Common Destiny*, without referring directly to the NAEP time series, and he concluded, “Other data, some of them mentioned in the book, lend support to the impression that there are intractable race differences in the performance of cognitively demanding tasks" (p. 10). The remainder of Herrnstein's review argued the centrality of intelligence in accounting for racial differentials in economic status, crime, health, and housing.

Together with the senior author of this paper, Jaynes and Williams responded at length to Herrnstein's misrepresentations of *A Common Destiny* (Hauser, Jaynes, and Williams 1990), and, among other things, they pointed out the report's documentation of large and consequential racial gaps in cognitive performance and to the findings (from NAEP and other sources) of decreasing differences in cognitive performance between blacks and whites. Moreover, two major themes of the NRC report were the heterogeneity within the black population and the report's finding that blacks were better off than whites in certain respects—findings that were not consistent with an ideological insistence on discrimination as the sole explanatory mechanism for black-white differences. Herrnstein (1990b) in turn responded and, in reference to black-white differences in cognitive performance, reiterated that the differences in reading comprehension and other cognitive skills "*are* intractable, for we do not know how to eliminate them" (Herrnstein 1990b:125). Subsequently, the next round of NAEP data appeared, and the senior author of this paper sent the extended trend series to Herrnstein, asking for his further comments on the intractability of black-white test-score differences. A copy of that letter was mailed to Herrnstein three times over a period of months, but he never replied.

After the hubbub about *A Common Destiny* subsided, other scholars continued to draw on the NAEP test series. Smith and O'Day (1991:72–77) 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 and 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, and Williamson (1994:11–17) reported decreasing gaps between blacks and whites in reading and mathematics achievement using the NAEP data for the mid- to late 1970s and for 1990 at ages 13 and 17, while Miller (1995:45–59) 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 also differed 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 large remaining 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 and Williams 1989:350–52). Smith and O'Day (1991:79–84) offered no specific analyses of changes in test-score gaps, but suggested that they might be explained by improved social background, 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: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, account for about one-third of the improvement in test scores among minority students from the 1970s to 1990.[4]

Achievement or Ability?

[4] See also Armor (1992) for similar findings.

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 about trends and differentials in academic achievement endlessly and in public. We spend a great deal of money to create and improve it in our public schools, and we blame the schools because we think they have not produced enough of it. We seem to think that if our levels of 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 and Garner 1982:3), 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 nonmembers of the psychometric profession, we are inclined to join those who elide or ignore the distinction between achievement and ability (see Jencks and Crouse 1982). We do not believe that ability can be assessed without reference to past learning and opportunity to learn. Moreover, we think it is difficult to maintain sharp distinctions in test content between aptitude and achievement. Thus, while we will not ignore the specific content of tests, we also think that any test performance partly indicates overall levels of realized ability.

For example, we think there is wide agreement that scores on the AFQT can justifiably be interpreted in much the same fashion as 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, and Spuhler (1975) and others, including Herrnstein and Murray (1994). At the same time, there is a great deal of evidence that schooling raises scores on IQ tests (Ceci 1991), and some of the strongest recent evidence refers specifically to the AFQT. Using data from school records for youth in the 1979 National Longitudinal Survey of Youth (NLSY), Herrnstein and Murray estimated that each year of schooling raised IQ, as measured by the AFQT, by about 1 point (1994:589–92). The Herrnstein and Murray estimates are seriously flawed, and corrected estimates of test-score gain, also based on the NLSY data, range from 2 to 3.5 points per year (Neal and Johnson 1994; Korenman and Winship 1995; Fischer 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 sixth, seventh, and eighth 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). The summer enrichment program specifies minimum scores on the SAT-M or SAT-V for each course and on the Test of Standard Written English (TSWE).

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 and 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 and 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, we consider *The Bell Curve*'s treatment of the NAEP findings in detail. After his experience with *A Common Destiny*, the senior author of this paper was curious to see how Herrnstein and Murray would treat trends in black and white test scores (Hauser 1995; Hauser and Carter 1995).[5] In our 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 "Ethnic Differences in Cognitive Ability." Most of the data on trends in test-score differences were put into one of the book's many appendices (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 and 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 National Center for Education Statistics from the early 1970s through 1990 (Mullis et al. 1991), and which we have reproduced here as Table 1. We were particularly interested in

[5] In personal correspondence, Stephen J. Ceci has also noted the discrepancy between Herrnstein and Murray's reading of the NAEP trends and that of other researchers.

this table and looked for something like it when first reading *The Bell Curve*. Across math, science, and reading examinations, and at ages 9, 13, and 17, Herrnstein and Murray report that the black-white difference declined by an average 0.28 standard deviations from 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 deviations in the verbal portion of the test and from 1.27 to .92 standard deviation in the mathematics portion of the test,” Herrnstein and 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 and 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 four 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 and 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 they added that it would be foolhardy to extrapolate the observed trends into the future. Indeed, whatever the specific estimate of test-score

TABLE 1

Herrnstein and Murray’s Computation of Test-Score Differences

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

	White-Black Difference, in Standard Deviations[a]		
	1969–1973	1990	Change
9-year-olds			
Science	1.14	.84	-.30
Math	.70	.54	-.16
Reading	.88	.70	-.18
Average	.91	.69	-.21
13-year-olds			
Science	.96	.76	-.20
Math	.92	.54	-.38
Reading	.78	.40	-.38
Average	.89	.57	-.32
17-year-olds			
Science	1.08	.96	-.12
Math	.80	.42	-.38
Reading	1.04	.60	-.44
Average	.97	.66	-.31
Overall average	*.92*	*.64*	-*.28*

Source: National Center for Education Statistics, 1991b.

[a] The computations assume a standard deviation of 50.

Note: Reproduced from Herrnstein and Murray 1994:291.

convergence between blacks and whites, one would be hard-pressed to find any acknowledgment of it once Herrnstein and Murray started drawing conclusions and making recommendations.

The more we thought about their findings, the more curious seemed Herrnstein and 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 (erroneous) 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 and Murray's table declares that they "assume a standard deviation of 50." We recalled some variation in the standard deviations of the NAEP test-scores across tests, from year to year, and between blacks and whites, so we 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 and 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 and Murray evidently confined their reading to a one-page summary of change in the test-score differences (Mullis et al. 1991:11), plus a footnote on page 1 of the source, which states 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 and 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 and 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. In Table 2, we have shown the change in test scores using the actual standard deviations of the total population of each age in 1990 as the unit of measure.[6] Using the revised standard deviations raises the overall average convergence from 0.29 to 0.39 standard deviations. Using Herrnstein and Murray's assumptions, this raises the implied convergence in IQ between blacks and whites to a range between 3.5 and 4.7 points. We wonder whether Herrnstein and 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 and 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 from 1986 to 1988 (Mullis 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.[7] 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,

[6] In science and mathematics, though not in reading, the variability of the tests declined across time. 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 we based the comparisons on the population standard deviations in 1990. 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 we had normed the changes on the standard deviations in 1970.

[7] Similar trends appear at age 17 and, to a lesser degree, at age 9.

TABLE 2

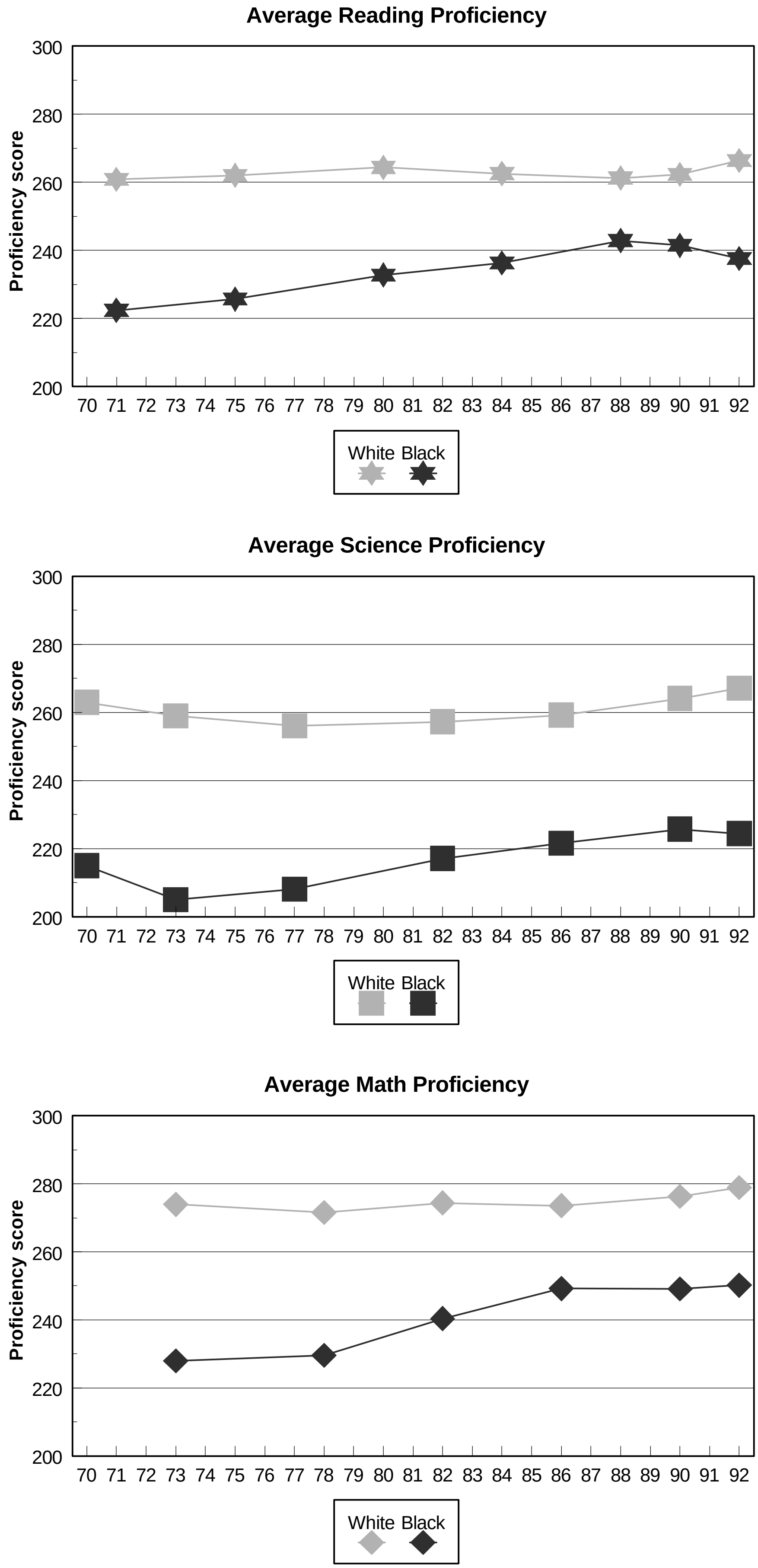
Revised Estimates of Reductions in the Black-White Difference on the NAEP

	Black-White Difference, in Standard Deviations		
	1969–1973	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 computation based on Mullis et al. (1991).

Note: The computations use standard deviations for each age group in 1990.

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



sometime in the mid- 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 et al. 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.

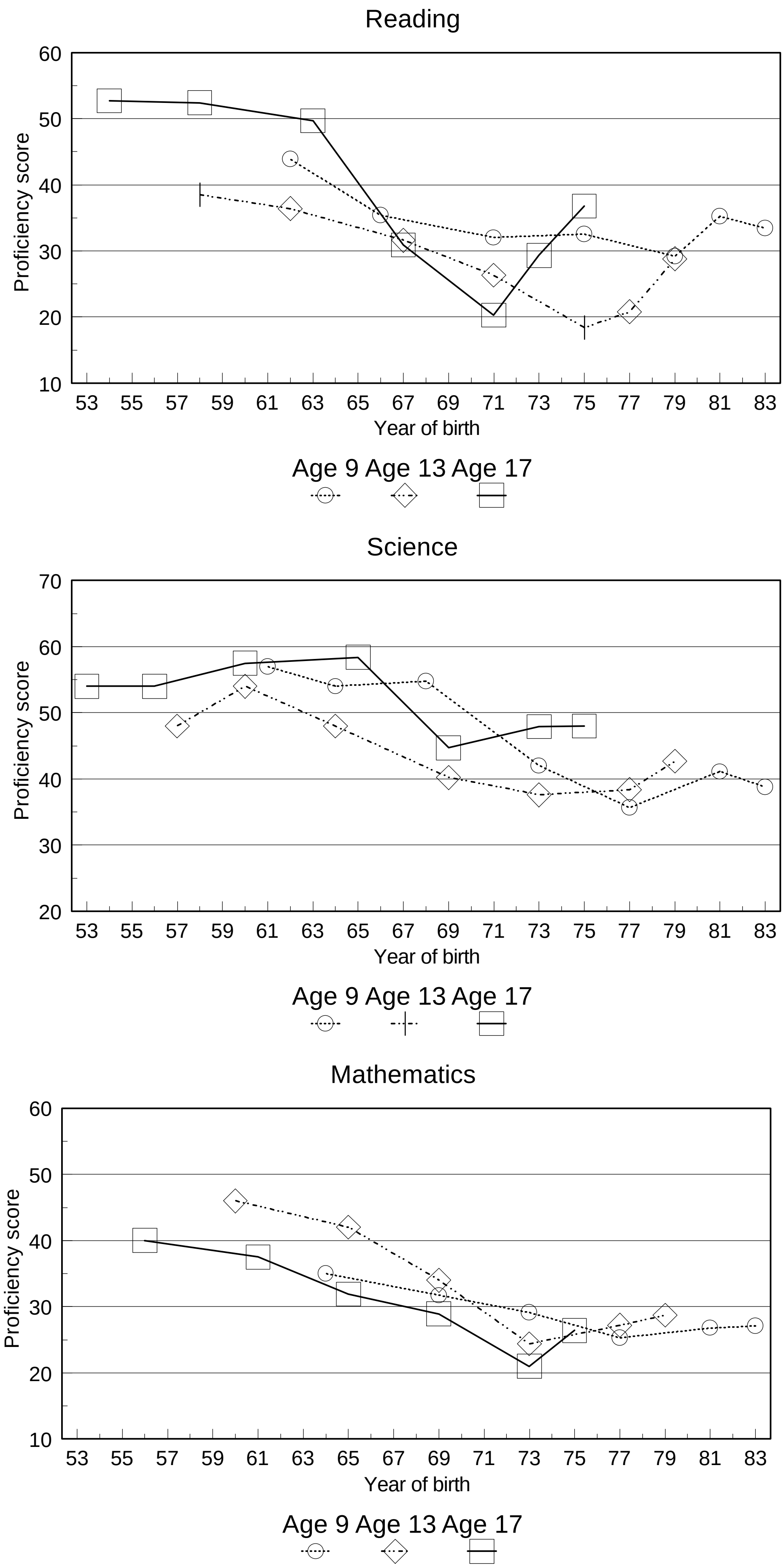
We thus looked further at the published NAEP series to learn whether our reading of the trend—and that appearing in the several reviews cited above—might have been mistaken.

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, we have used the same scale to show trend lines in mean test-score differences by age, but we 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, we 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.”

More important, Figure 2 clearly shows the major declines in white-black differences in reading, science, and mathematics achievement. This arrangement of the data suggests 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,

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



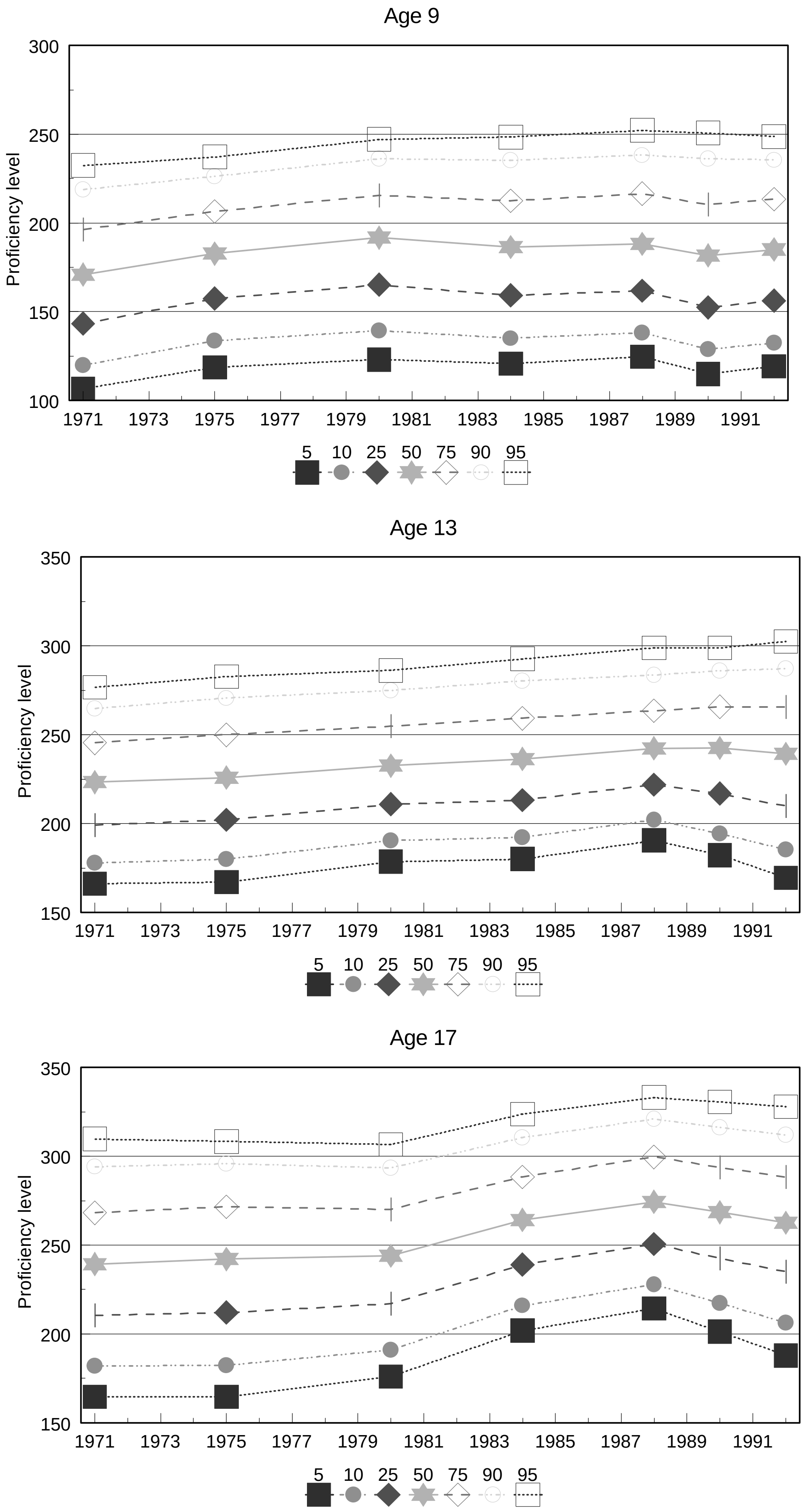
although not entirely consistently, for cohorts born after 1960.[8] In mathematics, the black gains appear for cohorts born from 1956 to 1973. We would hesitate to connect these test-score changes too closely to changes in IQ, though we note that Herrnstein and Murray showed no hesitation in doing so. However, in our 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 we have examined changes over time across the entire achievement distributions among blacks. Figure 3 shows selected percentile points of the NAEP reading 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. We have used line patterns and symbols for the data points to emphasize the rough symmetry in the distributions; the same lines and marker shapes are used for the three pairs of corresponding percentile points: 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.

Among 9-year-olds, there appears to have been growth in the reading scores of blacks across the entire distribution from 1971 to 1980, but little change thereafter. At age 13, growth appears across the whole distribution from 1971 to 1988, but performance in the bottom half of the distribution has declined since then. At age 17, rapid growth took place from 1980 to 1988 across the whole distribution, but performance declined thereafter.

[8] In science, but not in reading or mathematics, black gains were in part a consequence of declining test scores among whites.

Figure 3. NAEP Reading Trend Assessment: Percentiles of the Reading Distribution among Blacks by Age, 1971 to 1992



Age 9

Age 13

Age 17

Figure 4 gives percentile points of the distributions of the NAEP science assessment for black children from 1977 to 1992.[9] At age 9, science performance improved rapidly throughout the distribution from 1977 to 1984 and changed little thereafter. At age 13, science scores increased from 1977 to 1990. In this period, growth occurred throughout the distribution, but more rapidly at the bottom than at the top. After 1990, growth continued in the upper half of the distribution, but performance declined in the bottom half of the distribution. At age 17, performance declined between 1977 and 1982—a period during which performance also declined among 17-year-old whites—and the entire black distribution moved up between 1982 and 1986. Between 1986 and 1992, the distribution first spread and then narrowed, so performance levels were essentially the same at the beginning and end of that period.

Figure 5 gives percentile points of the distributions of the NAEP mathematics assessment for black children from 1978 to 1992.[10] 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. 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.

In our opinion, there is much more to these trends than the statement by Arvey et al. in "Mainstream Science" that "gaps in academic achievement have narrowed a bit for some races, ages,

[9] Unfortunately, the published series do not include percentile points from the assessments of 1970 and 1973. Except at age 13, the convergence in science performance of blacks and whites appears to have begun after 1973.

[10] 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:A63–A71).

Figure 4. NAEP Science Trend Assessment: Percentiles of the Science Distribution among Blacks by Age, 1977 to 1992

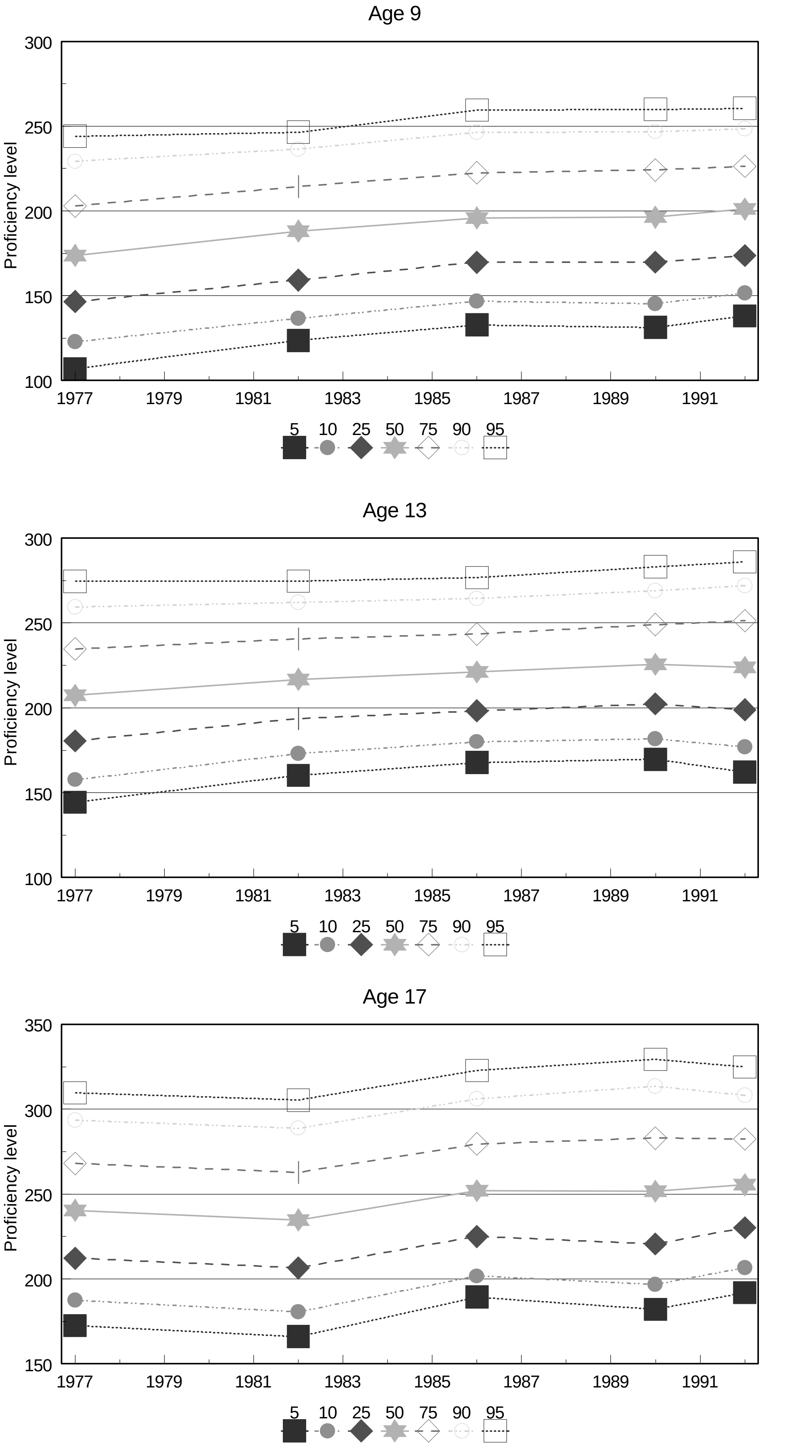
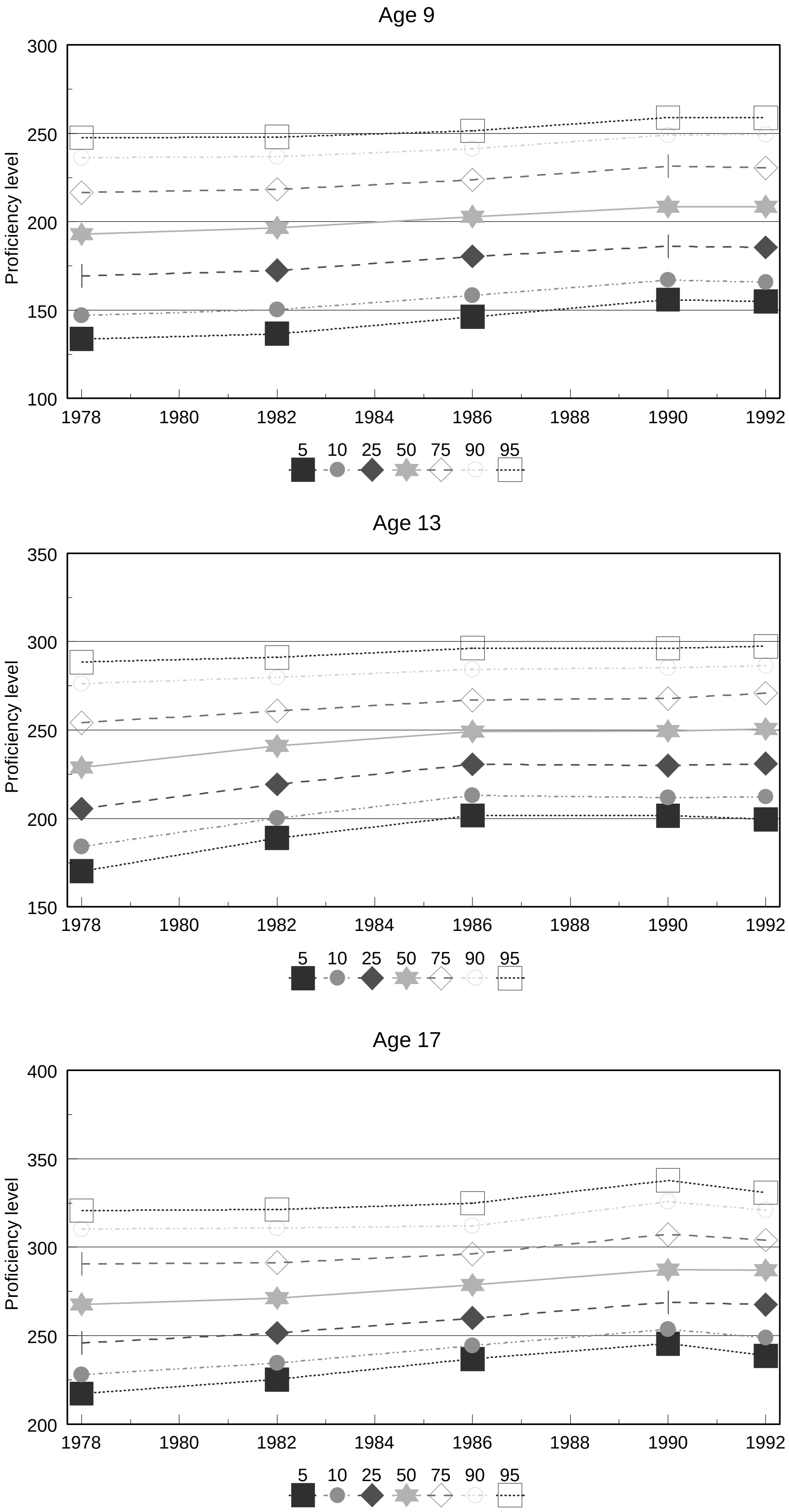


Figure 5. NAEP Mathematics Trend Assessment: Percentiles of the Mathematics Distribution among Blacks by Age, 1978 to 1992

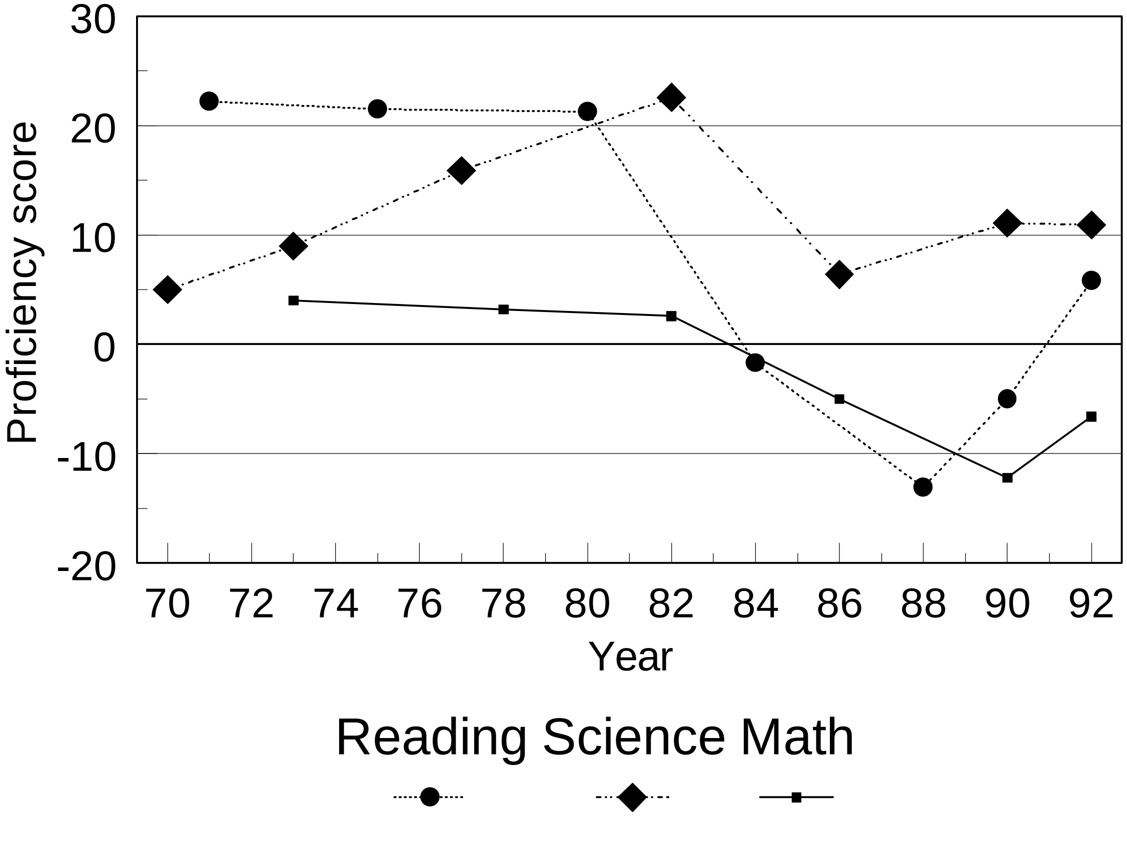


school subjects and skill levels.” Almost all of the growth in black scores led to convergence in the performance of blacks and whites. Although growth in black performance did not occur between every assessment from the 1970s to 1990, growth did occur throughout most of this period and throughout the entire distributions of performance levels. Thus, we think that the statement in “Mainstream Science” substantially understates the extent and pervasiveness of change in black achievement test scores.

In Figure 6, we show differences in mean NAEP performance differences between 13-year-old white students and 17-year-old black students by year and subject. We find it difficult to reconcile these data with the claim that, “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.” To be sure, in most assessments and in all three subjects, the younger whites outperform older blacks. However, in all three series, the white-black gap varies substantially across NAEP assessments. In reading, the younger whites had a 20-point advantage until 1984, but from 1984 to 1990, blacks had a modest advantage. In science, the younger whites had about a 10-point advantage in the early 1970s, which grew to 20 points by 1980, but declined again to about 10 points in the mid-1980s. Only in mathematics was there initially little difference between the performance levels of younger whites and older blacks, but a small advantage for whites in the 1970s turned into a small advantage for blacks in the 1980s. While the comparisons in Figure 6 are surely cautionary to anyone who would attempt to ignore or minimize black-white performance differentials, they, too, show some evidence of convergence. We doubt that many readers would independently describe these 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 mid-1980s and on the sources of the subsequent slowdown or reversal. One might think, for example, of the reduced enthusiasm for compensatory

Figure 6. Differences between NAEP Proficiency Scores of White 13-Year Olds and Black 17-Year Olds by Subject and Year



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. We are 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. We shall return later to consider the importance of these changes, whether or not they reflect changes in IQ.

Is There a Long-Term Convergence in Verbal Ability?

Unfortunately, the NAEP series on black-white differences in academic achievement begins only around 1970. For earlier periods, as we have seen, there are a great many studies of the test performance of blacks and whites, but few, if any, that can be used with confidence to measure differential performance trends. Contrary to Herrnstein’s claims (1990b:122), the fact that millions of people, black and white, have taken tests of cognitive performance does not turn their test scores into data worth analyzing.

One set of data on verbal ability provides clues about much longer trends in black-white differences. These data are from the General Social Survey (GSS) of the National Opinion Research Center, which has regularly administered a ten-item verbal ability test (WORDSUM) to cross-section samples of adults in U.S. households since the early 1970s (Davis and Smith 1994). Because there are repeated, usually annual, administrations of WORDSUM to cross-section samples of about 1500, it is possible to identify age and cohort effects on performance, on the assumption that there are no period effects. Using the GSS data, we can estimate trends in verbal ability among blacks and whites who were born from 1909 to 1974. These data show a consistent pattern of convergence between the test scores of blacks and whites throughout the period covered by the GSS cohorts. The convergence is not fully explained either by changes in social background or by changes in educational attainment across cohorts.

The GSS's ten vocabulary items were chosen from "Form A," one of two parallel, twenty-item vocabulary tests selected by Thorndike. Each form contained two vocabulary test items from each of the levels of the vocabulary section of the Institute for Educational Research (IER). Intelligence Scale Completions, Arithmetic Problems, Vocabulary, and Directions (CAVD) (Thorndike 1942). Form A was developed by Thorndike in response to the need for a very brief test of intelligence in a social survey (Thorndike and Gallup 1944), and it was also used in an attempt to study the feasibility of an aptitude census (Thorndike and Hagen 1952). Form A was later used by Miner (1957) in his monograph, *Intelligence in the United States*, which attempted to assess the intellectual ability of the U.S. population using a national household sample. Alwin (1991) used the GSS WORDSUM data from 1974 to 1990 in an analysis that asked whether changes in family configuration could account for the decline of verbal ability in the SAT-V.

For each of ten WORDSUM items, GSS respondents are asked to choose the one word out of five possible matches that comes closest in meaning to the word in capital letters. Figure 7 gives a set of sample items that are similar to those in WORDSUM.[11] The GSS obtains personal interviews, and each WORDSUM item is handed to the respondent on a preprinted card. Before 1988, WORDSUM was administered to the full GSS sample, but only every other year. Since 1988 it has been administered to two-thirds of the sample, using an alternate forms design. From 1974 to 1994, WORDSUM was completed by 11,160 whites and 1,418 blacks who were ages 20 to 65 at the survey date and who also provided valid data on years of schooling, number of adults in the household, number of siblings, and structure of the family of orientation.

[11] We have selected these items at random from the list in Miner (1957:53).

Figure 7

Illustrative Vocabulary Test Items

a. LIFT	1. sort out	2. raise	3. value	4. enjoy	5. fancy
b. CONCERN	1. see clearly	2. engage	3. furnish	4. disturb	5. have to do with
c. BROADEN	1. efface	2. make level	3. elapse	4. embroider	5. widen
d. BLUNT	1. dull	2. drowsy	3. deaf	4. doubtful	5. ugly
e. ACCUSTOM	1. disappoint	2. customary	3. encounter	4. get used	5. business
f. CHIRRUP	1. aspen	2. joyful	3. capsize	4. chirp	5. incite
g. EDIBLE	1. auspicious	2. eligible	3. fit to eat	4. sagacious	5. able to speak
h. CLOISTERED	1. miniature	2. bunched	3. arched	4. malady	5. secluded
i. TACTILITY	1. tangibility	2. grace	3. subtlety	4. extensibility	5. manageableness
j. SEDULOUS	1. muddled	2. sluggish	3. stupid	4. assiduous	5. corrupting

Source: Miner (1957: 53).

Miner (1957:28–30) argued that vocabulary tests are highly correlated with tests of general intelligence. He assembled some thirty-six studies in which a vocabulary measure had been correlated with a measure of general intelligence, and he found a median correlation of 0.83. Miner also noted that, whereas Wechsler had originally excluded a vocabulary test from the Wechsler-Bellevue scales because “he felt it might be unfair to illiterates and those with a foreign language background” (p. 30), Wechsler later decided that it was an excellent measure of general intelligence. Finally, Miner noted that the median correlation of verbal tests with full-scale IQ tests is larger than the median correlation Wechsler reported between the Wechsler-Bellevue Full Scale and fifteen other measures of general intelligence (Wechsler 1944).

More recently, Wolfle (1980:110) reported that the correlation between the GSS vocabulary test and the AGCT is 0.71. His account is worth quoting in full:

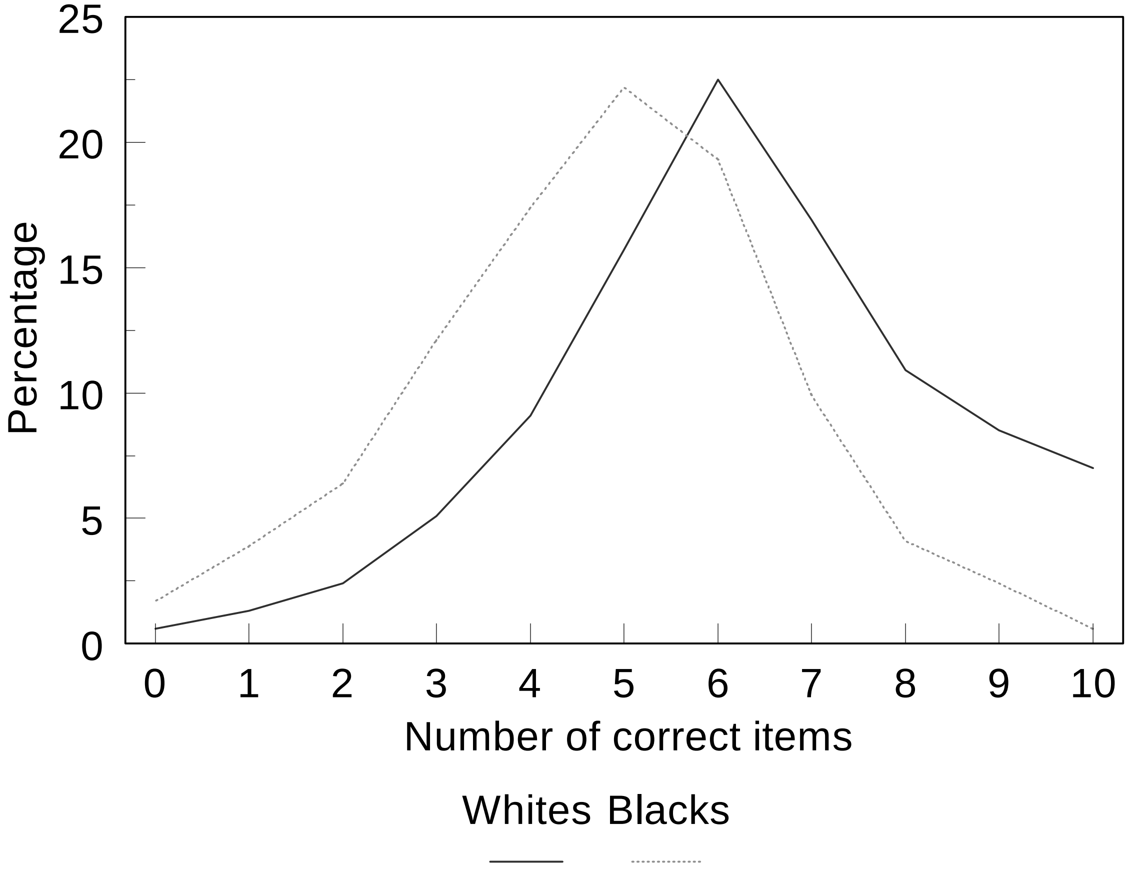
> The correlation between adult intelligence and the vocabulary score was calculated from information kindly provided by Robert L. Thorndike (1967). He reported on two normative studies by Irving Lorge conducted among Army enlisted men who were administered the AGCT and the Vocabulary-GT. Correlations of 0.70 and 0.77 were reported for Forms 1 and 2, respectively. These two values were averaged (0.74), and to this value the Spearman-Brown correction of test length was applied (because NORC used only half the items) using Thorndike’s revised estimate of the test’s reliability (Miner 1957:50). These computations resulted in a revised correlation of 0.71 between adult intelligence and the NORC vocabulary score.

Despite these indications of validity, we would urge caution in the use of WORDSUM. First, the test is very short, so its reliability is low. The internal consistency reliabilities are 0.712 among whites and 0.628 among blacks. In the 1994 GSS, in addition to WORDSUM, half the sample was administered eight of the fourteen similarity (abstract reasoning) items from the Wechsler Adult Intelligence Scale-Revised (Wechsler 1981).[12] The correlations between WORDSUM and the WAIS-R similarity scores were 0.394 for whites and 0.409 for blacks, without correction for unreliability. After correction for internal consistency reliability, the correlations were 0.589 among whites and 0.608 among blacks. Second, as we will discuss in more detail, despite the common use of vocabulary tests in IQ instruments and the high correlations between tests of vocabulary and of general intelligence, there is evidence of divergent trends between IQ and verbal ability in the past several decades. Thus, even if verbal ability is a valid indicator of general intelligence, neither the overall trend in WORDSUM nor the specific trends in WORDSUM among blacks and whites need follow those in more general tests of ability.

Figure 8 shows the vocabulary test-score distributions of blacks and whites in the GSS, 1974 to 1994. The mean numbers of correct items are $\overline{x}_w = 6.23$ ($s_w = 2.09$) for whites and $\overline{x}_b = 4.78$ ($s_b = 1.93$) for blacks. From the display of relative frequencies, it is obvious that there is a ceiling effect on the scores for whites in the general population. Similarly, there are ceiling or floor effects on the number of correct answers in other subpopulations, e.g., persons with many or few years of schooling. For this reason, our analyses of the WORDSUM data are based on a two-sided Tobit specification, which compensates for censoring at both ends of the distribution on the assumption that the true distribution of test scores is Gaussian (Maddala 1983).

[12] Hauser and Mare (1993) added the similarity items in order to calibrate them against WORDSUM for use in supplementary telephone interviews with brothers and sisters of the GSS respondents.

Figure 8. WORDSUM Distributions of Blacks and Whites: General Social Survey, 1974-94



METHODOLOGICAL ISSUES

Before attempting a trend analysis of the GSS data, we carried out several methodological analyses. Only 460 GSS respondents (3.7 percent) refused to answer any of the WORDSUM items, and we ignored them throughout the analysis. The refusals had completed slightly fewer years of schooling ($\overline{x}$ = 10.5 vs. 12.9 years), but were also more variable in schooling (s = 3.6 vs. 2.8 years). We considered whether item nonresponses—other than complete refusal—should be treated as incorrect answers, or whether total scores should be adjusted for the number of items answered. We compared the internal consistency reliabilities among complete responses and among incomplete responses, treating nonresponses in the latter group as errors. These were virtually the same, 0.62 and 0.64. Further, the correlations between educational attainment and WORDSUM among the complete and incomplete responses were identical. Thus, we assumed that item-specific nonresponses were erroneous responses. We experimented with a range of corrections for guessing and examined the correlations of WORDSUM with educational attainment, occupational status (Duncan's Socioeconomic Index), and—for 1994 GSS respondents only—the WAIS-R similarities for each correction scheme. We found that the correlations with external criteria were lowered by corrections for guessing, so we did not introduce such a correction.

Finally, we carried out logistic regression analyses of each item, by race, in order to look for differential item functioning. In the absence of differential item functioning, the odds of a correct response to an item would be the same for blacks and whites at every ability level, as indicated by the total test score. Uniform differential item functioning exists when the odds of answering an item are equally greater or equally less for blacks than whites at every ability level. Nonuniform differential item functioning exists when the difference between blacks and whites in the odds of a correct answer are not the same across ability levels (Swaminathan and Rogers 1990).

Let u = 1 if an item is answered correctly and u = 0, otherwise. The logistic regression model for predicting the probability of a correct response to an item is

$$p\ (u = 1) = e^{z} / (1 + e^{z}), \tag{1}$$

where

$$z = \tau_0 + \tau_1\theta + \tau_2 g + \tau_3(\theta \times g), \tag{2}$$

g = 1 for whites, otherwise g = 0; θ is the observed ability of an individual; and the τ_i are parameters. In this case, θ is estimated by the respondent's total vocabulary test score. An item shows uniform differential item functioning if $\tau_2 \neq 0$ and $\tau_3 = 0$, and it shows nonuniform differential item functioning if $\tau_3 \neq 0$, whether or not $\tau_2 = 0$. Since verbal ability generally increases with age, we also include the respondent's age in the regression. Thus, we contrast three versions of the model in equation 2. The first model includes total test scores and age as independent variables. The second model adds the black-white contrast. The third model adds the interaction of the black-white contrast and total test score.

Several of the tests of differential item functioning are nominally significant, but this means little because the total sample size is so large (N = 12,578). For that reason, we have used the Bayesian Information Criterion (*bic*) as a guide to model selection (Raftery 1995). Using that criterion, we find four cases in which there is differential item functioning and two in which there is nonuniform differential item functioning. Among relatively easy items, two are easier for blacks than for whites, while two others are relatively easier for whites than for blacks with the same total score. Two relatively difficult items are easier for blacks than for whites with low total scores, but easier for whites than for blacks with high total scores. Given these equivocal findings, it would probably be a good idea to test the sensitivity of our regression analyses to the elimination of selected items, but we have not done so here.

Trend Analyses of WORDSUM

We estimate three models of intercohort trends in verbal ability in the 1974–1994 GSS data. In each case, we analyze the total number of correct answers to the ten WORDSUM items using a two-sided Tobit specification. The estimated coefficients and their standard errors are shown in Table 3. The baseline model includes sex, race, age, birth cohort, and interactions between race and sex, and between race and birth cohort.[13] This model is intended to describe trends in verbal ability among blacks and whites, free of the confounding influences of temporal changes in the age composition of the population and of the association of age with test scores. If we had observations from a single cross-section sample, we could not conceivably separate the effects of chronological ages from those of birth cohorts. Because the GSS provides repeated cross-section measures in samples of the same cohorts at different ages, we can estimate distinct effects of chronological age and year of birth. However, this requires a strong identifying assumption, namely, that there are no period effects on test scores. That is, there are no effects on test scores, specific to the year of the survey, above and beyond the combination of age and birth cohort effects that pertain to persons in each survey year.

The social background model adds eight social background characteristics to the baseline model: father's educational attainment, mother's educational attainment, number of siblings, nonintact family (at age 16), father's occupational status, foreign residence (at age 16), farm background, and Southern residence (at age 16). In addition, the social background model includes three dummy variables that flag missing values on father's education, mother's education, and father's occupation. The education model adds respondent's years of schooling to the social background model. Years of education is coded into a series of twenty-one dummy variables with twelve years of education as the

[13] The number of cases for Whites born in 1909–19, 1920–29, 1930–39, 1940–49, 1950–59, 1960–69, and 1970–74 are 543, 1304, 1660, 2542, 2860, 1555, and 187, respectively. The counts for blacks are 60, 186, 298, 426, 587, 342, and 28, respectively. The small number of cases in the oldest and youngest cohorts of blacks implies that the findings about them should be interpreted with great caution.

TABLE 3
Baseline, Social background, and Education Models of Vocabulary Test Scores:
General Social Survey

	Baseline		Social Background		Education	
Variable	Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Intercept	3.86	0.32	3.86	0.30	5.33	0.28
Sex (men = 1)	0.02	0.12	-0.02	0.11	0.03	0.10
Race (white = 1)	2.46	0.31	1.69	0.28	1.36	0.26
Age 20–29	----	----	----	----	----	----
Age 30–39	0.47	0.06	0.47	0.06	0.28	0.05
Age 40–49	0.51	0.08	0.53	0.07	0.29	0.07
Age 50–59	0.29	0.10	0.30	0.09	0.01	0.08
Age 60–65	0.18	0.13	0.22	0.11	-0.07	0.11
Birth year (1909–1919)	----	----	----	----	----	----
Birth year (1920–1929)	0.21	0.35	-0.12	0.32	-0.41	0.29
Birth year (1930–1939)	0.61	0.33	0.27	0.30	-0.34	0.28
Birth year (1940–1949)	0.85	0.33	0.14	0.30	-0.77	0.28
Birth year (1950–1959)	0.63	0.33	-0.19	0.30	-1.13	0.28
Birth year (1960–1969)	0.65	0.34	-0.56	0.31	-1.41	0.29
Birth year (1970–1974)	1.37	0.49	-0.13	0.45	-0.87	0.41
Interaction of sex and race	-0.23	0.13	-0.18	0.11	-0.36	0.10
Interaction of birth year 1909–1919 and race	----	----	----	----	----	----
Interaction of birth year 1920–1929 and race	-0.38	0.37	-0.22	0.33	-0.05	0.30
Interaction of birth year 1930–1939 and race	-0.91	0.35	-0.92	0.31	-0.54	0.29
Interaction of birth year 1940–1949 and race	-0.89	0.34	-0.84	0.30	-0.31	0.28
Interaction of birth year 1950–1959 and race	-0.87	0.33	-0.93	0.30	-0.35	0.28
Interaction of birth year 1960–1969 and race	-1.08	0.34	-0.89	0.31	-0.35	0.28
Interaction of birth year 1970–1974 and race	-2.05	0.51	-1.78	0.46	-1.21	0.42
Father's education			0.05	0.01	0.01	0.01
Mother's education			0.10	0.01	0.04	0.01
Number of siblings			-0.09	0.01	-0.05	0.01
Not living with both own parents at age 16			-0.00	0.06	0.09	0.05
Father's occupational status			0.02	0.00	0.01	0.00
Lived in foreign country at age 16			-0.98	0.10	-0.96	0.09
Farm background			-0.44	0.04	-0.30	0.04
Lived in the South at age 16			-0.45	0.04	-0.34	0.04
Missing father's education			-0.43	0.06	-0.09	0.06
Missing mother's education			-0.64	0.06	-0.25	0.06
Missing father's occupational status			0.19	0.08	0.04	0.07
Respondent's years of education is 0					0.39	0.66
Respondent's years of education is 1					-3.19	1.01
Respondent's years of education is 2					-2.16	0.73
Respondent's years of education is 3					-2.39	0.39
Respondent's years of education is 4					-2.50	0.30
Respondent's years of education is 5					-2.39	0.25

(table continues)

TABLE 3, continued

Variable		Baseline		Social Background		Education	
		Coefficient	S.E.	Coefficient	S.E.	Coefficient	S.E.
Respondent's years of education is	6					-1.50	0.20
Respondent's years of education is	7					-1.57	0.15
Respondent's years of education is	8					-1.24	0.09
Respondent's years of education is	9					-1.00	0.10
Respondent's years of education is	10					-0.75	0.08
Respondent's years of education is	11					-0.68	0.08
Respondent's years of education is	12					----	----
Respondent's years of education is	13					0.40	0.06
Respondent's years of education is	14					0.75	0.06
Respondent's years of education is	15					1.06	0.08
Respondent's years of education is	16					1.67	0.06
Respondent's years of education is	17					1.89	0.10
Respondent's years of education is	18					2.07	0.10
Respondent's years of education is	19					2.57	0.15
Respondent's years of education is	20+					2.32	0.14

Note: Variable values marked by “----” are reference groups in dummy variables with more than two classes.

reference group. Comparisons among estimates from these models may help us understand the effects of social background and respondent’s years of education on black-white differences and their trends over time.

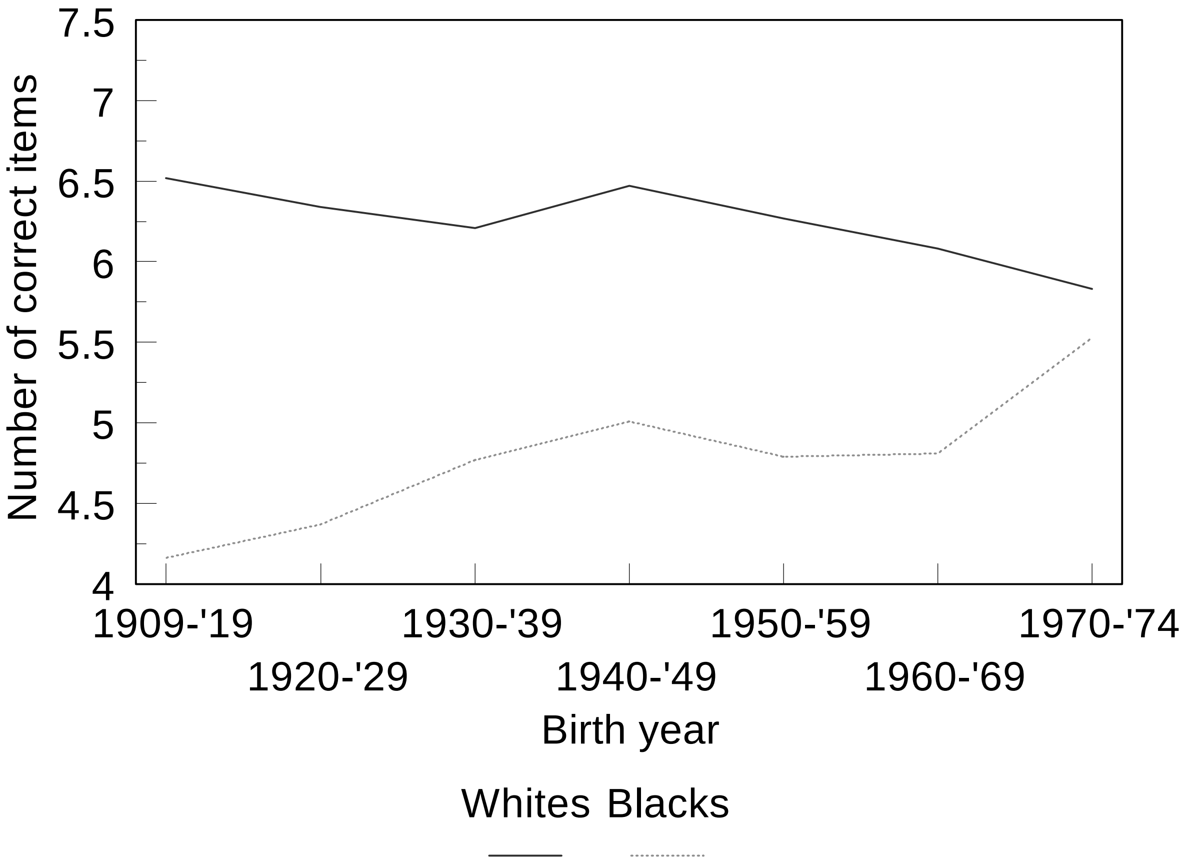
The effects of age in the three models are of interest in their own right. They have been estimated with 20 to 29-year-olds as the reference group. In the baseline model and, again, when social background has been controlled, test scores increase by about half a point at ages 30 to 39 and 40 to 49. Then, they fall by about 0.2 and 0.3 at ages 50 to 59 and 60 to 65. However, this age pattern is partly a consequence of the variation in schooling by age in the GSS data. When years of school are also controlled, there remains a smaller increase in test scores at ages 30 to 39 and 40 to 49, about 0.3 years relative to persons ages 20 to 29. However, the two older age groups no longer have any advantage relative to the youngest age group. This age pattern would appear to contradict some suggestions in the research literature that a relatively crystallized ability, like vocabulary knowledge, increases indefinitely with age.

Figure 9 shows the intercohort trends in WORDSUM scores of blacks and whites born from 1909 to 1974, as estimated in the baseline model. Black-white differences in test scores have been decreasing.[14] In the most recent birth cohort, 1970 to 1974, the black-white difference is 13 percent as large as in the earliest birth cohort, 1909 to 1919.[15] The differences decline irregularly across birth cohorts, but never increase substantially from one cohort to the next. Relative to the oldest cohort, they are 84 percent, 61 percent, 62 percent, 62 percent, 54 percent, and 13 percent as large in successive cohorts. About two-thirds of the convergence can be attributed to an upward trend in test scores among blacks, from 4.2 in the oldest cohort to 5.5 in the youngest. About one-third of the

[14] In the baseline model, the likelihood-ratio test statistics for intercohort differences in test scores and for the race by cohort interactions are both highly significant.

[15] Again, because of the small number of cases in the youngest cohort, this observation should not be taken too seriously.

Figure 9. Baseline Model of WORDSUM Trends: General Social Survey, 1974-94



convergence can be attributed to a downward trend among whites, from 6.5 in the oldest cohort to 5.8 in the youngest. Most of the decline in white scores took place in cohorts born after 1950, while the growth in test scores among blacks took place between the cohorts of 1909 and 1949 or after 1970.

Figure 10 shows the trends in test scores for blacks and whites when social background variables are added to the baseline model. Because there have been persistent but declining differences in social background between black and white cohorts, one might expect the initial differences between black and white test scores to be smaller and the convergence to be more rapid when social background is controlled. This appears to be the case.[16] The initial difference between blacks and whites in the cohort of 1909 to 1919 is 1.6 points in the social background model, while it was 2.4 points in the baseline model. Moreover, convergence is greater than in the baseline model. Black-white differences in WORDSUM disappear in the 1970 to 1974 birth cohort when social background variables are controlled. The convergence appears to have been driven largely by a continuous decline in the verbal ability of whites, while the test scores of blacks varied irregularly from one cohort to the next. That is, the estimates suggest that verbal ability was relatively stable among blacks with the same social background throughout most of this century, while it declined steadily among whites of constant social background throughout the century. In any event, black-white differences in social background account for part of the initial difference between average test scores, and improvements in the social background of blacks relative to whites account for part of the convergence.

Figure 11 shows the intercohort trends in test scores when years of schooling as well as social background are added to the baseline model.[17] Net of schooling, as one would expect, average black

[16] In the social background model, the likelihood-ratio test statistics for intercohort differences in test scores and for the race by cohort interactions are both highly significant.

[17] In the schooling model, the likelihood-ratio test statistic for intercohort differences in test scores are highly significant. However, the race by cohort interactions are only marginally significant ($p < 0.03$); thus, the cohort trend lines for blacks and whites are almost parallel.

Figure 10. Social Background Model of WORDSUM Trends: General Social Survey, 1974-94

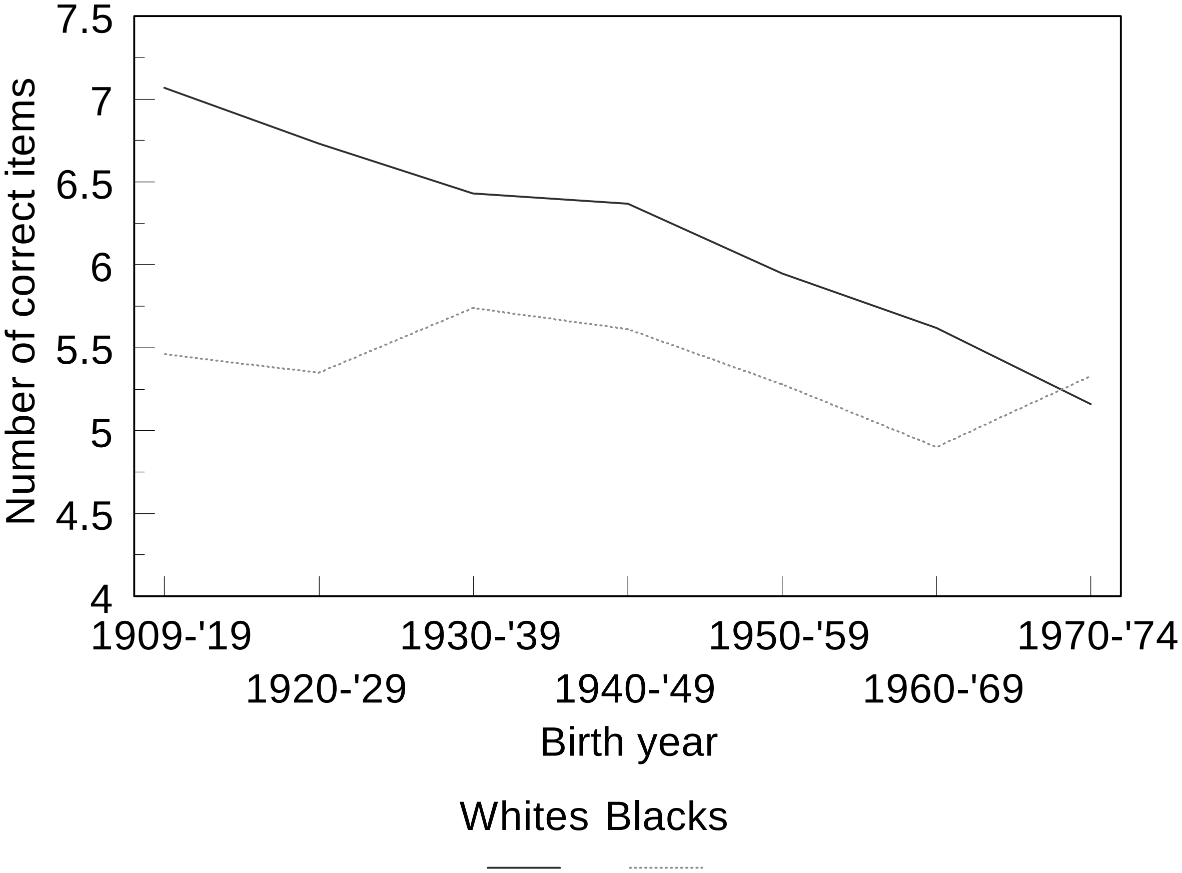
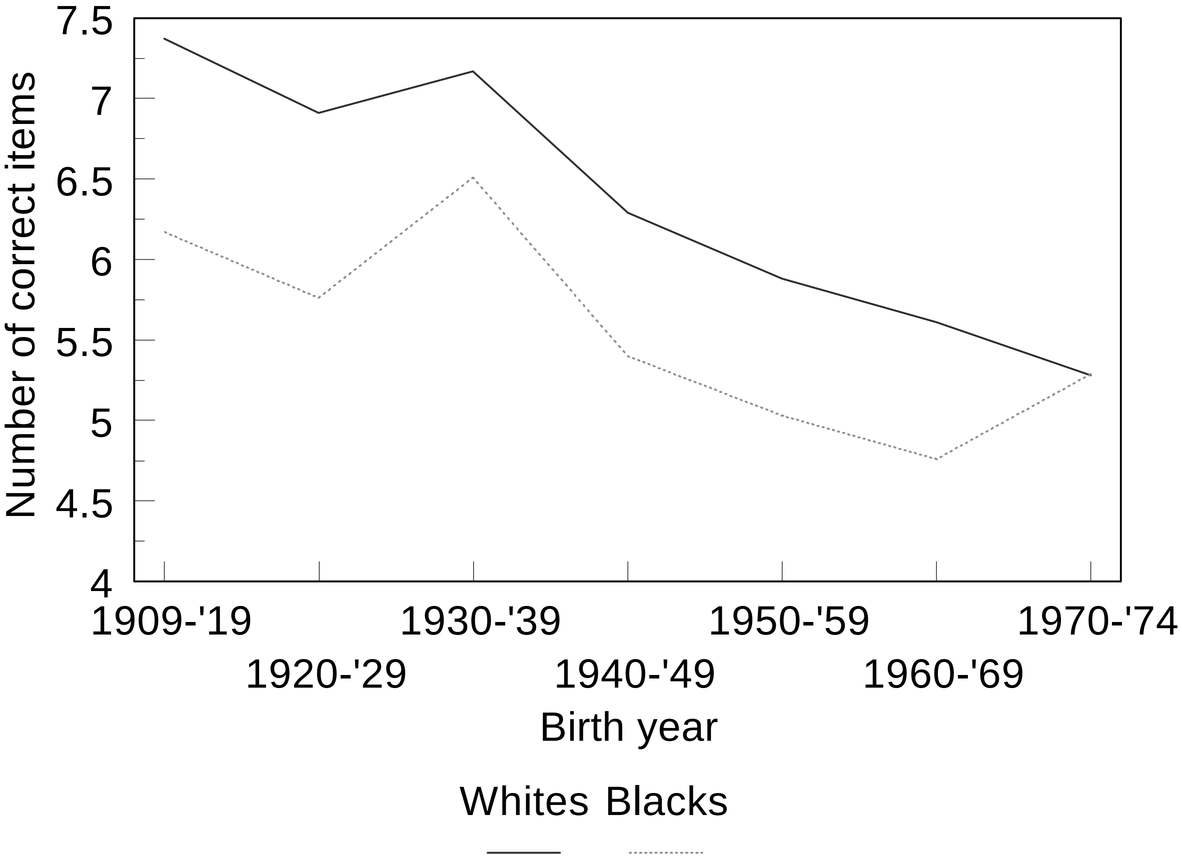


Figure 11. Education Model of WORDSUM Trends: General Social Survey, 1974-94



and white test scores are less far apart. In the oldest cohorts, the differential is slightly more than one correct answer on the WORDSUM test (half a standard deviation), and in all but the youngest cohort, the differential is slightly less than one correct answer. Thus, much of the black-white difference in vocabulary scores is associated with differences in social background and schooling. They account for more than a 1-point test score difference in the oldest cohort and for about a half-point difference in the next-to-youngest cohort.

The data also suggest that changes in social background and schooling contributed to the convergence in test scores. For example, in the baseline model, the difference in scores between the oldest and youngest cohorts was 2 points; in the social background model, it was 1.9 points; and in the schooling model, it was 1.2 points. Moreover, there are also signs of convergence in test scores, above and beyond the salutary effects of social background and schooling. That is, in the schooling model, the black and white test score series are almost parallel, but there remain some signs of convergence, especially in the post-1970 period. Finally, among blacks and whites, the net trends in Figure 6 are no longer linear. Both groups show an overall decline in vocabulary knowledge, but with a relative peak in the Depression cohorts. Thus, as successive cohorts completed higher nominal levels of schooling, those levels of schooling appear to have become less selective with respect to a basic level of verbal proficiency. Some might argue that this finding indicates a decline in school quality, but we prefer to reserve such terminology for direct evidence that the effects of schools have declined.

SUMMARY AND DISCUSSION

An increasing array of evidence suggests that black-white differences in cognitive tests have been reduced for cohorts born after the mid-1960s. Although several test-score series show some signs of convergence, we are especially impressed by the NAEP series in reading, science, and mathematics, which cover ages 9, 13, and 17, and are more nearly representative of the general population than other testing programs. As Smith and O'Day summarized the findings, between 1970 and the mid- 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.

There is new evidence of a longer-term convergence between black and white test scores in verbal ability. Under reasonable assumptions with respect to the effects of age and sex, data collected by the NORC GSS since the mid-1970s suggest a sixty-five-year trend toward convergence between adult blacks and whites in verbal ability, as indicated by a short test of vocabulary knowledge. Differences between blacks and whites in social background and schooling account for a large share of the observed differences in test scores. The convergence in verbal ability appears to have been close to complete for persons of similar background and schooling who were born after 1970. Improvements in the social background and schooling of blacks account in part for the convergence in test scores, but other sources also appear to have reduced the test-score difference, especially in recent cohorts.

One singular aspect of the GSS series is that they show a long-term decline in verbal ability in the white population. A similar trend appears for blacks and whites when social background and schooling have been controlled, though the decline among blacks is less than that among Whites. Are the test-score series from the GSS plausible in light of Flynn's findings of rising IQ scores (Flynn 1984, 1987), both in the United States and nineteen other countries? We believe that the GSS series are not necessarily inconsistent with Flynn's findings, though we certainly hope that it will be possible to test those findings directly or indirectly with data from other sources. For example, Loehlin, Lindzey, and Spuhler's (1975:135–39) review of global trends in intelligence took note of two studies in which verbal scores had declined across time, while other ability measures had increased. It should also be recalled that, in Flynn's initial report of temporal gains in test scores, much of the discussion concerned the possible inconsistency between IQ gains up through 1978 and the post-1963 declines in the SAT-V (1984:36–39). In “IQ Gains over Time,” Flynn (1994:617) notes that tests may be ordered from those with reduced cultural content, “many of which are indicators of fluid intelligence,” to those that measure crystallized intelligence, “less emphasis on on-the-spot problem solving and more on whether someone has acquired the skills, or general knowledge, or vocabulary we would expect an intelligent person to gain in a normal life.” Thus, he places “pure vocabulary tests” at the extreme of tests of crystallized intelligence.

Flynn’s reading of the available evidence, worldwide, is that

> IQ gains over time diminish as tests get farther and farther from measuring fluid intelligence. . . . Verbal IQ gains vary from almost nil to 20 points per generation, with 9 as a rough median, and some of this is adult data from military testing. Among the eleven countries that allow a comparison, there is not one in which verbal gains match the gains on culture-reduced, or performance, or nonverbal tests and often the ratios run against verbal gains by two or three to one. Where vocabulary gains can be distinguished from verbal gains in general, they rarely match them.

If we accept this reading, we should not be entirely surprised to find a valid time series of vocabulary measures with a pronounced downward trend. At the same time, the clear distinction between the behavior of vocabulary tests and other tests of ability across time is a valuable reminder of the limits of the present findings.

In this context, leaving aside our specific findings in the GSS, we might want to consider other applications of similar research designs. There is no hope of going as far back in time in the measurement of fluid intelligence as the GSS series permits us in the assessment of verbal ability. But we could cover a considerable span of cohorts by administering a variety of tests to a single large cross-section sample, provided only that we truly understood the intrinsic relationship between age and performance on each test. Alternatively, a new series of cross-section samples, following the GSS model, would permit estimation both of age and cohort effects and could provide both retrospective and prospective assessments of trends and differentials in abilities.

Of what importance is convergence in achievement test scores or verbal ability scores of blacks and whites? Herrnstein and 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 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" (1986:318).[18] If that were the case, we should be most concerned about the strength of the link of IQ with test series like the NAEP assessments and GSS WORDSUM, and we 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 is neither the sole nor necessarily the most important cognitive factor in adult success. For example, the Armed Services Vocational Aptitude Battery (ASVAB), which includes ten distinct tests, was normed on the NLSY, a large, national sample of young men and women who were 14 to 21 years old in 1979. The NLSY data were extensively analyzed by Herrnstein and Murray to demonstrate the ubiquitous influences of IQ. They estimated IQ with a composite of four of the ten ASVAB components, which together make up the AFQT: Word Knowledge (WK), Paragraph Comprehension (PC), Arithmetic Reasoning (AR), and Mathematics Knowledge (MK) (1994:580–83). Among the remaining six subtests—General Science (GS), Electronics Information (EI), Coding Speed (CS), Automobile/Shop Information (AS), Numerical

[18] The emptiness of this statement is suggested by the fact that it would be equally true if one substituted "educational attainment" for "*g*" both times it appears.

Operations (NO), and Mechanical Comprehension (MC)—Herrnstein and Murray point out that Numerical Operations (NO) and Coding Speed (CS) are both speeded tests and have relatively low correlations with the general intelligence factor, *g*. However, following a long and deeply flawed tradition of psychometric analysis, Herrnstein and Murray looked only at relationships among the ASVAB components—not at their effects on adult outcomes—when they chose their test-score composite.

Goldberger (1995) has shown the importance of numerical operations (NO) and computational speed (CS) in predicting the earnings of workers in the NLSY.[19] For example, Goldberger points out that 84 minutes of testing time were devoted to the components of the AFQT, while only 10 minutes were given to NO and CS. Thus, one would expect the AFQT to be substantially more reliable than a composite of NO and CS, as well as more closely related to *g*. Yet among 4,165 men in the NLSY, a composite of NO and CS was more highly predictive of total income in 1989 than was the AFQT. Among 3,737 women, the regression of income on the AFQT was 0.414, while the regression of income on the composite of NO and CS was slightly less, 0.367. Among men and women, the regressions of income on the combination of NO and CS were greater than its regressions on the AFQT, once educational attainment was controlled.

In a broad-ranging reanalysis of Herrnstein and Murray's work, Corcoran (1996) regrouped the ASVAB into five components, Math (AR, MK), Verbal (PC, WK), Fluent Production (NO, CS), General Science, and Technical Knowledge (MC, AS, EI).[20] Then she ran separate regression analyses of the effects of all five components, plus age, socioeconomic status, and—for the post-school outcomes—educational attainment, on the several outcomes treated in *The Bell Curve*. These are educational outcomes (attainment, high school dropout, completion of a GED, completion of a BA),

[19] See also Heckman (1995).

[20] Note that Corcoran's "Fluent Production" is the same as Goldberger's combination of NO and CS.

economic outcomes (income, wages, poverty), labor force outcomes (entry into a high-IQ occupation, being out of the labor force, being unemployed), and social outcomes (having been in jail, having "middle-class values" [Herrnstein and Murray 1994:263], and being married by age 30). Among white men and women in the NLSY, no one component or combination of them consistently dominated the regressions. For example, in the case of schooling outcomes, Math and Verbal scores were important, but so was General Science and—in the case of high school dropout—Fluent Production. We have already mentioned the importance of Fluent Production in economic outcomes, and Corcoran found that it tended to be more important than either Math or Verbal scores. Net of schooling, General Science and Technical Knowledge determined entry into a high-IQ occupation among men, while Math scores determined occupational entry among women. Among men and women, Fluent Production was more important than Math in determining labor force participation, which was unaffected by the other three components. Experience in jail was affected by Math scores and Fluent Production among men and by Verbal scores and Fluent Production among women.

In short, 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, we 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 we do not have 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, Duane F. 1991. “Family of Origin and Cohort Differences in Verbal Ability.” *American Sociological Review* 56 (5, October):625–38.

Armor, David. 1992. “Why Is Black Educational Achievement Rising?” *The Public Interest* (108, Summer):65–80.

Arvey, Richard D., et al. 1994. “Mainstream Science on Intelligence.” *Wall Street Journal*, December 13.

Berliner, David C., and Bruce J. Biddle. 1995. *The Manufactured Crisis: Myths, Fraud, and the Attack on America’s Public Schools*. Reading, Mass.: Addison-Wesley.

Ceci, Stephen J. 1991. “How Much Does Schooling Influence General Intelligence and Its Cognitive Components? A Reassessment of the Evidence.” *Developmental Psychology* 27 (5):703–22.

Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York. 1966. *Equality of Educational Opportunity*. Office of Education, U.S. Department of Health, Education, and Welfare. Washington, D.C.: U.S. Government Printing Office.

Corcoran, Jill. 1996. “Beyond *The Bell Curve* and *g:* Rethinking Ability and Its Correlates.” Thesis. Department of Sociology, Harvard University, Cambridge, Mass.

Davis, James Allan, and Tom W. Smith. 1994. *General Social Surveys, 1972–1994: Cumulative Codebook*. Chicago: National Opinion Research Center.

Fischer, Claude S., Michael Hout, Martin Sanchez Jankowski, Samuel R. Lucas, Ann Swidler, and Kim and Voss. 1996. *Understanding Inequality in America: Beyond the Bell Curve*. Princeton: Princeton University Press.

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

———. 1987. “Massive IQ Gains in Fourteen Nations: What IQ Tests Really Measure.” *Psychological Bulletin* 101 (2):171–91.

———. 1994. “IQ Gains over Time.” In *Encyclopedia of Human Intelligence*, ed. Robert J. Sternberg, Stephen J. Ceci, J. Horn, Joseph Matarazzo, and Sandra Scarr. New York: Macmillan.

Goldberger, Arthur S. 1995. “Abilities, Tests, and Earnings.” MacArthur Foundation Conference on Meritocracy and Inequality. Madison, Wis., December.

Grissmer, David W., Sheila N. Kirby, Mark Berends, and Stephanie Williamson. 1994. *Student Achievement and the Changing American Family*. Santa Monica, Calif.: RAND Institute on Education and Training.

Hauser, Robert M. 1995. “*The Bell Curve*.” *Contemporary Sociology* 24 (2, March):149–53.

———, and Wendy Y. Carter. 1995. “The Bell Curve as Social Science.” Presented at the 1995 Meetings of the American Sociological Association. Washington, D.C., August.

———, Gerald D. Jaynes, and Robin M. Williams, Jr. 1990. “Explaining Black-White Differences.” *The Public Interest* (99, Spring):110–19.

———, and Robert D. Mare. 1993. “A General Social Survey Mini-Module on Trends and Differentials in Cognition.” A Proposal to the Governing Board of the General Social Survey.

Heckman, James J. 1995. “Lessons from *The Bell Curve*.” *Journal of Political Economy* 103 (5, October):1091–120.

Herrnstein, R. J. 1990a. “Still an American Dilemma.” *The Public Interest* (98, Winter):3–17.

———. 1990b. “On Responsible Scholarship: A Rejoinder.” *The Public Interest* (99, Spring):120–27.

Herrnstein, Richard J., and Charles Murray. 1994. *The Bell Curve: Intelligence and Class Structure in American Life*. New York: The Free Press.

Humphreys, Lloyd G. 1988. “Trends in Levels of Academic Achievement of Blacks and Other Minorities.” *Intelligence* 12:231–60.

Jaynes, Gerald David, and Robin M. Williams, 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, Christopher S., and James Crouse. 1982. “Aptitude vs. Achievement: Should We Replace the SAT?” In *New Directions for Testing and Measurement, Guidance, and Program Improvement*, ed. W. B. Schrader. San Francisco: Jossey-Bass.

Jensen, Arthur R. 1986. “g: Artifact or Reality?” *Journal of Vocational Behavior* 29 (3, December):301–31.

Jones, Lyle V. 1984. “White-Black Achievement Differences: The Narrowing Gap.” *American Psychologist* 39 (11, November):1207–13.

Korenman, Sanders, and Christopher Winship. 1995. “A Reanalysis of *The Bell Curve*.” Unpublished paper, October.

Koretz, Daniel. 1986. *Trends in Educational Achievement*. Washington, D.C.: Congressional Budget Office.

Lemann, Nicholas. 1995. “The Structure of Success in America: The Untold Story of How Educational Testing Became Ambition’s Gateway—and a National Obsession.” *The Atlantic Monthly*, August, 41–60.

Loehlin, John C., Gardner Lindzey, and J. N. Spuhler. 1975. *Race Differences in Intelligence*. San Francisco: W. H. Freeman.

Maddala, G. S. 1983. *Limited-Dependent and Qualitative Variables in Econometrics*. New York: Cambridge University Press.

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

Miller, L. Scott. 1995. *An American Imperative: Accelerating Minority Educational Advancement*. New Haven: Yale University Press.

Miner, John B. 1957. *Intelligence in the United States: A Survey*. New York: Springer.

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

Mullis, Ina V. S., John A. Dossey, Jay R. Campbell, Claudia A. Gentile, Christine O’Sullivan, and Andres Latham. 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.: U.S. Government Printing Office.

———, John A. Dossey, Mary A. Foertsch, Lee R. Jones, and Claudia A. Gentile. 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.: U.S. Government Printing Office.

Murray, Charles, and R. J. Herrnstein. 1992. “What’s Really Behind the SAT-Score Decline?” *The Public Interest* (106, Winter):32–56.

Neal, Derek A., and William R. Johnson. 1994. “The Role of Pre-Market Factors in Black-White Wage Differences.” Unpublished paper, University of Chicago, November 21, 1994.

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

Osborne, Travis R., and Frank C. J. McGurk. 1982. *The Testing of Negro Intelligence, volume 2*. Athens, Ga.: The Foundation for Human Understanding.

Raftery, Adrian E. 1995. “Bayesian Model Selection in Social Research.” In *Sociological Methodology 1995*, vol. 25, ed. Peter V. Marsden, 111–63. Cambridge, Mass.: Basil Blackwell.

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

Smith, Marshall S., and Jennifer O'Day. 1991. "Educational Equality: 1966 and Now." In *Spheres of Justice in Education: The 1990 American Education Finance Association Yearbook,* ed. D. Verstegen and J. G. Ward, 53–100. New York: Harper Collins.

Solomon, Robert 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, N.J.: Educational Testing Service.

Swaminathan, H., and H. J. Rogers. 1990. "Detecting Differential Item Functioning Using Logistic Regression Procedures." *Journal of Educational Measurement* 27:361–70.

Thorndike, Robert L. 1942. "Two Screening Tests of Verbal Intelligence." *Journal of Applied Psychology* 26:128–35.

———. 1967. "Vocabulary Test G-T: Directions and Norms." Mimeograph. New York, Institute of Psychological Research.

———, and George H. Gallup. 1944. "Verbal Intelligence of the American Adult." *Journal of General Psychology* 30:75–85.

———, and Elizabeth P. Hagen. 1952. *Analysis of Results of Field Trials to Determine the Feasibility of an Aptitude Census*. Human Resources Research Center Research Bulletin, vol. 52-22.

Wainer, Howard. 1985. "Some Pitfalls Encountered While Trying to Compare States on Their SAT Scores: Page and Fiefs as an Example" 85-6285-62. Princeton, N.J.: Educational Testing Service.

———. 1987. "Can We Accurately Assess Changes in Minority Performance on the SAT?" Princeton, N.J.: Educational Testing Service, January.

Wechsler, David. 1944. *The Measurement of Adult Intelligence*, 3rd ed. Baltimore: Williams and Wilkins.

———. 1981. *WAIS-R Manual: Wechsler Adult Intelligence Scale-Revised*. San Antonio: The Psychological Corporation.

Wigdor, Alexandra K., and Wendell R. Garner, 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, Willard, Harold Howe, 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.

Wolfle, Lee M. 1980. “The Enduring Effects of Education on Verbal Skills.” *Sociology of Education* 53 (April):104–14.

Zajonc, Robert B. 1976. “Family Configuration and Intelligence.” *Science* 192:227–36.

———. 1986. “The Decline and Rise of Scholastic Aptitude Scores: A Prediction Derived from the Confluence Model.” *American Psychologist* 41:862–67.

Zwick, Rebecca. 1992. “Special Issue on the National Assessment of Educational Progress.” *Journal of Educational Statistics* 17 (2, Summer):93–232.