The Secular Increase in IQ and Longitudinal Changes in the Magnitude of the Black-White Difference: Evidence from the NLSY

Charles Murray
American Enterprise Institute

Behavior Genetics Association Meeting,
4 July 1999, Vancouver BC

Address correspondence to:
Charles Murray
American Enterprise Institute
1150 Seventeenth St., NW
Washington, DC 20036
E-mail: chasmurray@earthlink.net

Introduction

The secular, international rise in scores on mental tests, now commonly termed the Flynn effect, has been seen by Flynn and others as reason for optimism about the eventual convergence of black and white IQ scores. Part of the argument deals with specific issues in the ongoing debate about the sources of the black-white (BW) difference in test scores, such as those raised by the Spearman Hypothesis (Jensen 1985) or the relationship of subtest loadings on *g* and inbreeding-depression to the BW difference (Rushton 1989, Rushton 1999). In each case, Flynn has argued that the existence of the secular longitudinal rise in IQ raises provides evidence for an environmental explanation (Flynn 1987, Flynn 1998, Flynn 1999b).

The broader reason for the optimism derives from the certainty that the Flynn effect is overwhelmingly environmental, given the short time span over which it has occurred. If the temporal evolution of the environment can cause such a broad longitudinal drift in IQ, cross-sectional group differences are also plausibly due to an unknown environmental factor. If IQ scores can change so rapidly and so much, averaging almost a third of a point per year (Flynn 1999a), reasonably rapid convergence of group differences also seems within reach.

This aspect of the optimism is an appeal to analogy which can call on many situations in which different starting points converge on similar end states as the environment evens out (e.g., diffusion of heat into rooms that were initially at different temperatures). But analogies are not uniformly supportive. In some instances, a change in the environment increases the disparity between groups (e.g., economic growth tends to reduce poverty but increase inequality). In other cases, a change in the environment leads to a one-time narrowing of group differences which leaves a stable residual difference (e.g., a higher-protein diet increased average height among the Japanese, but a height difference between Japanese and Caucasians remains).

Which of these analogies applies to the BW difference in mental test scores? Until the causes of the BW difference are understood, it is a question that can be answered provisionally by asking whether the observed longitudinal patterns conform to the expectations of the Flynn argument. In taking this view, I draw from James Flynn's own common-sense test of whether the Flynn effect can be predominantly a change in real intelligence. Flynn asks if, as we look back over this century, we have any reason to think that people have gotten as much more intelligent as the gains in IQ scores would imply. His answer, with a high degree of face-validity, is "No." (Flynn 1999a). Similarly, the common-sense test of the optimism about the convergence of black and white test scores is whether the BW difference has in fact shown signs of converging during recent decades of change in the legal and economic status of American blacks. If the latent mean IQ is 100 for both whites and blacks, and if the environmental changes creating the Flynn effect are as powerful and ubiquitous as they seem to be, then it seems reasonable to expect that by this time the effects of environmental diffusion should be discernable.

History of the BW Difference

A mean difference in black and white mental test scores has been observed for as long as mental tests have existed. I divide the discussion into tests of cognitive ability and tests that, while they tap into cognitive ability, are intended primarily to assess academic achievement.

Mental tests measuring cognitive ability. Herrnstein and Murray (1994: 276–278) assembled 156 studies conducted from the early 1900s through the late 1980s that met basic standards of interpretability and for which the BW difference could be expressed as a standard deviation (SD). Overall, the mean BW difference was 1.08 SDs. For the 45 studies conducted after 1940, outside the South, with subjects older than 6, with full test batteries, the mean BW difference was 1.06 SDs. For the 24 studies meeting these same standards but conducted since the 1960s, the mean BW was 1.10 SDs.

The trend by decade is surprising. In the 1920s, when IQ tests were new and much more vulnerable to problems of bias and unreliability than later tests and at a time when black educational attainment was much lower than it is now, the mean BW difference in the 13 studies that used standardized tests was .86 SDs. The largest BW difference shows up in the 1960s, for which 37 studies are available (many of which include an admixture of tests of academic achievement), with a mean BW difference of 1.28 SDs. The other decades range from .82 SDs (1930s) to 1.12 SDs (1970s). There is no gross evidence in these data that the BW difference has narrowed over the century. On the other hand, none of the samples prior to the 1960s was nationally representative, and many of the samples were chosen in ways which would tend to have a selection bias toward higher-IQ black populations.

For the period from 1965 to the present, Hedges and Nowell (1998) examine every large, nationally representative survey of black and white academic test scores. Comparing five surveys from 1965–92 that used different instruments, they created a composite score from the vocabulary, reading, and math subtests. While no psychometric data about *g* loadings were presented, this composite might reasonably be interpreted as an approximation of IQ scores. Hedges and Nowell demonstrate a decrease from 1.18 SDs in the 1965 Equality of Educational Opportunity survey to .82 SDs in the 1992 wave of the National Education Longitudinal Study of 1988 and find that the trend among the five studies is statistically significant ($p<.05$).

The narrowing on the composite score occurred at the low end of the distribution, however. Hedges and Nowell found no evidence of diminishing racial disparities in the upper tail of the distribution. Herrnstein and Murray similarly found that the narrowing of the BW difference in SAT scores was almost exclusively the product of changes at the low end of the score range (Herrnstein and Murray 1994: 722–23). This pattern is consistent with a hypothesis that the convergence is primarily associated with improvements in basic skills, not increases in cognitive functioning across the range.

The only available time series for a cognitive test with a single instrument is the vocabulary test administered annually to a nationally representative population in the General Social Survey. Lynn (1998) has analyzed the BW difference for the period 1974–1996. The vocabulary test is very short (ten items) and thus represents a rough measure of cognitive ability for any

individual, but sample sizes are extremely large and the vocabulary subtest is highly correlated with full-scale IQ (r=.75 with the Wechsler). Lynn finds a small (0.004 SD per year) narrowing that does not reach statistical significance (p=0.29).

There have been two renormings of full-scale IQ tests during the last twenty years. The first was the renorming of the Wechsler Adult Intelligence scale in 1981 in which the BW difference was 1.0 SDs (Reynolds et al. 1987). The other was the renorming of the Stanford-Binet in 1986. The BW difference was .80 SDs for ages 2–11 and 1.10 SDs for ages 12–23 (Thorndike, Hagen, & Sattler 1986: 34–36). The rising BW difference with age is consistent with other IQ data showing a rising BW difference from infancy through post-pubescence. (Jensen 1998: 359).

Mental tests measuring academic achievement. The best longitudinal evidence on tests of academic achievement comes from the National Assessment of Educational Progress (NAEP), which uses identical instruments and sampling procedures from survey to survey. From 1971–94, Hedges and Nowell find statistically significant decreases in the BW difference for tests of reading and science and nonsignificant decreases for tests of mathematics and writing. The remaining gap as of 1994, expressed in standard deviations, stood at .66 for reading, .89 for mathematics, 1.08 for science, and .68 for writing (Hedges & Nowell 1998: 156–57). For the five nationally representative surveys since 1965 using different instruments, Hedges and Nowell conclude that "the racial gap appears to be getting slightly smaller over time for each measure except social science achievement," with a statistically significant decrease for reading comprehension, a nearly-significant decrease for mathematics, and nonsignificant decreases for vocabulary, science, and perceptual speed (Hedges & Nowell 1998: 154).

Herrnstein and Murray reviewed trends in the NAEP, Scholastic Aptitude Test (SAT), American College Testing program (ACT), Graduate Record Examination (GRE), and the national high school studies of 1972 and 1980. They found evidence of narrowing similar to that found by Hedges and Nowell, concluding that the gap on college entrance tests and national tests of educational proficiency had narrowed in the 1970s and 1980s. They associated this narrowing with a potential narrowing in IQ scores of two to three IQ points (Herrnstein and Murray 1994: 292, 637–642).

These same reviews have found that the convergence in academic achievement data has slowed or stopped since the late 1980s. For the NAEP, the BW gap has remained the same or increased on three of the four tests during the 1990s. For the SAT, the BW gap has been effectively unchanged since 1988.

It will be noted that the magnitude of the BW difference in the academic tests cited above tends to be smaller than the BW difference of 1 SD commonly found in IQ tests. Relevant to this finding is the extensive evidence accumulated by Jensen (Jensen 1985, Jensen 1992, Jensen 1998) for the Spearman Hypothesis. The Spearman Hypothesis asserts that variation in the size of the mean BW difference across tests is a positive function of variation in the tests' *g* loadings, and has by now been confirmed in sixteen independent studies. That tests of academic achievement, which often explicitly try to measure acquired knowledge rather than academic ability, have smaller BW differences than tests explicitly designed to measure *g*, is consistent with the Spearman Hypothesis.

Taken as a whole, the existing reviews of changes in the magnitude of the BW difference show a picture of reduction in most tests of academic achievement, occasionally substantial, and ambiguous evidence of a smaller reduction in the BW difference on tests of cognitive ability since the 1960s. In this context, the release of the 1996 interview wave of the National Longitudinal Survey of Youth (NLSY) offers an opportunity to add to the picture of trends in the BW difference. The nature of the data make them particularly relevant for testing expectations that the BW difference will converge via a gradual equalizing of the environment.

Measures and Methods

Data base. The NLSY, sponsored by the U.S. Department of Labor, began in 1979 with 12,686 participants, then ages 14–21, oversampling certain groups (blacks, Hispanics, low-income whites) but structured so that nationally representative estimates could be retrieved through the use of sample weights. Subsequently, the NLSY has been followed with annual interviews through 1992 and biannual interviews since then. These original NLSY subjects will subsequently be referred to as the 1st generation.

By the mid-1980s, the female subjects of the NLSY were well into their child-bearing years. In 1986, the project's sponsors added a biannual assessment of their children, subsequently to be called the 2nd generation. The data collection for the 2nd generation also included psychometric tests plus tests of educational achievement. As of the 1996 interview wave, the children born to the NLSY women represented approximately 90 percent of their eventual birth cohort's children.[1]

Racial identification. The NLSY reports both the screener's identification of race and the subject's self-identification of ethnic origin. To avoid the confounding influences of Hispanic ethnicity, these variables were used to select a population of non-Hispanic whites (screener's identification as white, ethnic self-identification as North American or European). A subject was classified as black if both the screener's identification was black and the ethnic self-identification was African. NLSY offspring are assigned the ethnic classification of the mother.

Cognitive measure for the 1st generation. The measure of cognitive ability for the original NLSY subjects is the AFQT, a highly *g*-loaded combination of four subtests of the Armed Services Vocational Aptitude Battery, using the revised scoring system established in 1989 (Ree & Earles 1991). Test-retest reliability for the AFQT is .9 (Earles & Ree 1992). Because the AFQT is age-sensitive, percentile scores were computed separately for each birth year, using sample weights. The age-equated percentile scores were then normalized and converted to a mean of 100 and SD of 15. The median correlation of these AFQT scores with other cognitive tests taken by members of the NLSY was .81 (Herrnstein and Murray 1994: 609). This is somewhat higher than the comparable correlations of the Wechsler Adult Intelligence Scale and the Stanford-Binet with other cognitive tests, .77 and .71 respectively (Jensen 1980: 314–15).

Cognitive and achievement measures for 2nd generation. The measure of cognitive ability for the children of the NLSY subjects is the Peabody Picture Vocabulary Test, revised version

[1] Based on current age distributions of women who give birth (National Center for Health Statistics 1998: Table 4).

(PPVT-R), a widely used test of verbal ability that was normed for a nationally representative sample in 1979.[2] The test is designed for administration to children ages $2^1/2$ through 18. The median split-half reliability is .81 (Robertson & Eisenberg 1981). The PPVT-R has been administered to the children of NLSY women in two-year intervals from 1986 through 1996. The scores reported by the NLSY are age-equated and standardized to a mean of 100 and an SD of 15, using the 1979 norms. To avoid exaggerating the effects of extremely low scores, the PPVT-R scores have been truncated to a range of ± 3 SDs, 55–145, assigning 55 and 145 to scores below and above those respective cut-off points. This truncation has increased the black mean by about 1 point and the white mean by about 0.2 point over the untruncated means. For children tested more than once, the mean PPVT-R score is used.

Expression of the BW difference. Point differences always refer to tests nationally normed to a mean of 100 and SD of 15. Differences expressed in standard deviations use the pooled variance weighted by sample size, via the equation

$$(1)\ \left(\overline{X}_a - \overline{X}_b\right) \Big/ \sqrt{\left[\left(N_a s_a^2 + N_b s_b^2\right) \Big/ \left(N_a + N_b\right)\right]},$$

where N is the sample size, $\overline{X}$ is the sample mean, s is the standard deviation, and the subscripts a and b designate the two groups (Jensen & Reynolds 1982).

Identification of full siblings in the 1st Generation. The original NLSY subjects were included as full siblings if (1) each identified the other as a full brother or sister and (2) each reported having lived with both biological parents at birth and in the year that the putative sibling was born.

Identification of full siblings in the 2nd generation. Among children of the NLSY women, a pair was included in the sibling sample if it was coded as either "full siblings" or "probable full siblings" in the classification by Charng & Baydar (1996). The PPVT-R correlation among siblings was .66 for the pairs identified as full siblings (not twins) and .60 among those identified as probable full siblings, indicating a high degree of accuracy in identification of the "probables."[3]

Construction of the paired samples. For the analyses of regression to the mean, two sets of matched samples were constructed, consisting of a pair of black siblings and a pair of white siblings.

The criteria for eligibility were that each pair be full siblings, classified either as black or non-Hispanic white, with valid mental test scores (AFQT for the 1st generation, PPVT-R for the 2nd generation). These conditions yielded a 1st generation sample of 1,592 white and 932 black sibling pairs, and a 2nd generation sample of 1,903 white and 1,178 black sibling pairs. One member of each sibling pair was randomly selected as the reference sibling and the other as the comparison sibling.

The first pair of samples, one for each generation, were matched exclusively for IQ. The black and white samples were sorted by the reference sibling's cognitive test score rounded to the

[2] Information on the Peabody test battery is taken from Baker, Keck, Mott, & Quinlan (1993: 133–151).
[3] The correlation of the mean PPVT-R score among the 44 twin pairs (mixed monozygotic and dizygotic) was a remarkable .89.

nearest point, and randomly within that point. Black and white reference siblings were matched on that rounded score. The first candidate subjects were used when the number of reference siblings with the same score was greater in one race than in another. This procedure yielded a 1st generation matched sample of 552 matched pairs and a 2nd generation matched sample of 616 matched pairs.

The second pair of samples, also one for each generation, were matched simultaneously for the reference sibling's IQ, parental income, and parental education. IQ was matched within ten-point ranges beginning with 55–64 and continuing through 125–134. Annual parental income (as of 1979–80 for the 1st generation sample, 1993–95 for the 2nd generation) was matched by categories of 0–$24,999, $25,000–$49,999, $50,000–$99,999, and $100,000+, expressed in 1995 dollars. Parental education was represented by the mother's completed years of education, grouped into 0–11 years, 12–15 years, and 16+ years.

Regressed True Scores. Following standard practice when matching subjects with dissimilar group means, regressed true scores were used for all analyses of the paired samples. Regressed true scores are computed as

$$(2)\quad \hat{T} = r_{XX'}(X - M_X) + M_X,$$

where $\hat{T}$ is the estimated true score, X is the observed test score, $r_{XX'}$ is the reliability of the test, and M_X is the mean of the group (Feldt & Brennan, 1989).

The Aggregate BW Difference on Cognitive Tests in the 1st and 2nd Generations

The original NLSY sample, the 1st generation, was born from 1957–64 and tested with the AFQT in 1980. A total of 6,502 non-Hispanic whites (hereafter, "white" should always be understood to mean non-Hispanic whites) and 3,022 blacks had valid AFQT scores. When sample weights are used to reach a nationally representative estimate, the white mean AFQT score was 103.3 with a standard deviation of 13.8 and a black mean was 86.7 with a standard deviation of 12.4. The difference between the two populations was thus 16.6 points or 1.24 SDs.[4]

The children of the NLSY women, the 2nd generation, included 5,072 white and 2,947 black subjects as of the 1996 interview wave. Of these, 3,697 white children and 2,467 black children had at least one valid score on the PPVT-R. The oldest was born in 1970 and the youngest in 1993, with 98.5 percent born from 1975–92.

[4] This result is obtained from the 1989 scoring version of the AFQT, age-equated and normalized to a mean of 100 and SD of 15. By way of comparison, the pre-1989 scoring system, not corrected for skew and not age-equated, yields a BW difference of 1.36 SDs.

Using the mean PPVT-R score for subjects with more than one test and applying sample weights, the white mean is 98.2 (SD=14.2) and the black mean 80.4 (SD=14.0), or a BW difference of 17.8 points and 1.26 SDs.[5]

These aggregate results for the two generations do not reveal any convergence of the BW difference. The difference measured in points goes from 16.6 points to 17.8 points; measured in SDs, from 1.24 to 1.26. This conclusion seems robust when subjected to more detailed examination regarding four questions: (1) What happens to the generational comparison when the 1st generation sample is limited to the mothers of the 2nd generation? (2) How might the observed results be affected if we had the test scores of the untested mothers and children? (3) How might the observed results be affected if we had the test scores of the ten percent of the NLSY generation's children yet to be born? (4) Is there any evidence of convergence for subsamples with maximum distance between the generations? I take up each question in turn.

The BW difference for mothers and offspring. Suppose we limit the sample to mother and children pairs in which both have valid cognitive test scores. For this subsample, the mean AFQT of the white 1st generation was 99.8 compared to a black mean of 84.6. The BW difference in the 1st generation for this limited sample was 15.2 points, or 1.26 SDs. The mean PPVT-R of white children was 98.3 compared to 80.3 for the black children, a BW difference of 18.0 points or 1.28 SDs. Limiting the sample to mother-child pairs with full test data has no appreciable effect on the magnitude of the BW difference in either generation.

Untested children. How might the observed results be affected if we had the test scores of the untested mothers and children? To answer this question, we may take a series of family background variables correlated with PPVT-R and examine the comparative means of families of the tested and untested children. The logic is straightforward: If the correlation between the background variable and PPVT-R is positive among the tested children (such as the correlation between the mother's AFQT score and the child's PPVT-R), and the mothers of the untested children have a higher AFQT mean than the mother's of the tested children, then we may expect that, *ceteris paribus*, the mean PPVT-R of the untested children would be higher (if test scores were to be obtained) than the PPVT-R of the tested children. This logic may be extended: Let us assume that the relationship of maternal AFQT to child's PPVT-R is the same for both the tested and the untested children. Using this assumption, we may use the regression of PPVT-R on AFQT among the tested children to calculate the expected mean PPVT-R among the untested children. Table 1 on the next page shows bivariate results for a variety of background variables with correlations of greater than .2 with PPVT-R. The continuous variables are mother's AFQT (r=.54), mother's year's of education (r=.30), and logged mean family income 1993–95 (r=.37). The binary variables, and their point-biserial correlations with PPVT-R, are married/unmarried as

[5] If instead the test is the unit of analysis, with multiple entries for children tested more than once, again using sample weights, the mean is 99.3 for whites (SD=15.2) and 80.8 for blacks (SD=15.6), or a BW difference of 18.5 points and 1.20 SDs.

Table 1. Tested and Untested Children in the 2nd Generation

	White tested children	*White untested children*	*Black tested children*	*Black untested children*	*Fitted BW diff. among the tested children (in points)**	*Expected BW diff. among the untested children (in points)***
Independent variable						
AFQT (mean)	98.7	102.2	84.4	86.9	16.5	16.8
Mother's years of education (mean)	13.0	14.0	12.4	12.9	18.4	19.1
Logged family income 1993–95 (mean)	10.60	10.90	9.80	9.95	18.5	19.0
Mother married as of 1996 (percentage)	78.5	88.0	33.4	39.4	18.5	18.6
Children born into poverty (percentage)	14.9	9.5	54.4	39.8	18.7	18.0
Mother ever on welfare (percentage)	28.8	12.5	76.5	61.1	18.8	18.8
Children born out of wedlock (percentage)	11.3	7.3	62.4	50.0	18.8	18.7

*For continuous variables, the mean for the tested children is applied to the beta coefficient when PPVT-R is regressed on the independent variable. For binary variables, the expected PPVT-R value associated with each state is applied to the proportion of children in each state and a weighted mean is calculated.
**Applies the sample value for the untested children to the parameters for the tested children.

of 1996 (r=.31), above/below the poverty line when the child was born (r=.37), never/ever on welfare (r=.38), and child born in/out of wedlock (r=.36). The first four columns shows the sample values for the tested and untested children. As an inspection of those sample values suggests, the white tested children appear to be at least as disadvantaged relative to the white untested children as the black tested children are disadvantaged relative to the black untested children. The last two columns show how this apparent relationship appears when fitted values are used. The fitted BW difference among the tested children is based on a solution of the regression results for PPVT-R regressed on the independent variable for that line, using the mean for the tested children (for continuous variables) or applying the observed proportions in each state to the expected PPVT-R values for each state (for binary variables). The fitted black PPVT-R is subtracted from the fitted white PPVT-R. The expected BW difference among the untested children represents the solution of the same regression results, but applies the sample values for the untested children.

Starting with AFQT as an example: The mean AFQT of the white mothers of tested children was 3.5 points lower than the mean AFQT of the white mothers of the untested white children, while the mean AFQT of black mothers of tested children was 2.5 points lower than the mean AFQT of the black mothers of the untested children. The downward bias in the estimate of child's IQ created by incomplete testing is fractionally greater than for the white than for the black sample. This is reflected in the results when the mean AFQT of the mothers of untested children is applied to the coefficient for maternal AFQT among the tested children: The expected net BW difference for the untested children is slightly greater (.3 points) than it is for the tested children.

As Table 1 indicates, the BW difference among the untested children could be expected to increase noticeably if we base our expectation on the differences in maternal AFQT, maternal years of education, and logged family income among the tested and untested children. For three variables (marital status in 1996, welfare recipiency, legitimacy status), the expected change is 0 or .1 point. For only one variable, children born into poverty, is there reason to expect that the BW difference might narrow nontrivially among the untested children.

As one may predict from the bivariate results shown in Table 1, multivariate analyses of these variables also lead to slightly increased estimates of the BW difference among the untested children. In short, it appears that testing 100 percent of the children in the 2nd generation would certainly not shrink the estimate of the BW difference taken from the observed samples of tested children, and probably would widen it.

Unborn children. How might the observed results be affected if we had the test scores of the ten percent of the NLSY generation's children yet to be born? As of 1996, all of the NLSY women were 29 or older. Among them, 22.7 percent of the white women and 18.6 percent of the black women were childless. The white and black age distributions for these women were equivalent. The mean AFQT of the white women without children was 105.8, compared to 90.8 for the black women without children. These figures compare to white and black AFQT scores of 101.3 and 85.6 respectively for 1st generation women who had at least one child as of 1996. The difference between the mothers and nonmothers is thus 4.5 points for white women and 5.2

points for black women. If the remaining children were born equally to black and white women, we could expect a slight convergence of test scores. However, birth rates tail off much more quickly after age 30 for blacks than for whites. Through 1996, 1st generation white women over 30 had produced 35 babies per 100 women while black women over 30 had produced 21 babies per 100 women. The birth rate for 1st generation white women over 30 with at least a BA was twice that of black women over 30 with at least a BA. Both of these tendencies in the NLSY correspond to national statistics (National Center for Health Statistics 1998: Tables 4 & 18). Combining these figures, the expectation is that the unborn babies will be disproportionately white, of high-IQ mothers, suggesting that the net downward bias from this source in the observed PPVT-R scores is also modestly greater for whites than for blacks. To estimate the degree of this bias more precisely would require speculative estimates of how the remaining babies will be split between women who are already mothers and those who have not yet had any children, which I will not attempt.

In summary, the minimal expectation must be that universal testing of the entire 2nd generation, untested and unborn children alike, will produce a BW difference in the children no smaller than has been observed in the available samples. The more likely possibility is that universal testing of the entire generation will produce a slightly larger BW difference than was actually observed.

Maximizing generational distance. The youngest members of the 1st generation were born in 1964; the oldest member of the 2nd generation was born only six years later. What happens if the 2nd generation sample is limited to those born from 1985 onward, thus separating the two generations by a minimum of 20 years? In the 2nd generation, the black mean for those children born from 1985 onward is 79.5, compared to 81.0 for black children born before 1985. More broadly, there are no trends toward convergence over the 1975–1992 period in which 98 percent of the 2nd generation was born.[6] On the contrary, the uncorrected BW difference tends to increase with time for the 2nd generation, but this is an artifact of differences in age-at-testing and trends in the AFQT scores of mothers over time. Once these factors are taken into account, the magnitude of the BW difference appears to be flat throughout the two decades in which the tested children were born.

Black and White Sibling Pairs

Knowing merely that the BW difference in means has persisted unchanged from the 1st to the 2nd generation is of limited value. A richer way of comparing the BW difference over the two generations is to take advantage of the large number of siblings that are present in both the 1st and 2nd generation NLSY samples and to compare patterns of sibling regression to the mean.

[6] Because of the nature of the AFQT, estimating trends within the 1957–64 period would require complex analyses that have not been attempted here. The AFQT scores used here, being age-equated by birth year, cannot be used for this purpose.

Some Basic Considerations about Regression to the Mean and Sibling Comparisons

Regression to the mean is a commonly observed statistical phenomenon, with IQ scores providing one of innumerable examples. The standard linear regression equation

$$(3)\quad \hat{Y} = r_{xy}\frac{s_y}{s_x}(X - \bar{X}) + \bar{Y}$$

predicts the magnitude of regression to the mean, independently of whatever causal mechanism may be involved (Humphreys 1978). It may be applied to any set of paired observations in which the correlation on the variable of interest is importantly different from zero. In the case of IQ, the most common type of analysis involves pairs of blood relatives. For example, given a correlation of .5 between the child's IQ and the mid-point of parental IQ, and equal variances for parental and child IQ, the IQs of the offspring are expected to regress by half of the difference between the parental IQ and the parental population mean. One may also apply sample data to equation (3) to calculate the mean to which comparison sample is regressing, by determining the value at which $X = \hat{Y}$.

The first important point to remember about regression to the mean in the following discussion is that differential regression to the mean in two groups does not imply any particular cause. It implies only that, for whatever reasons, the pairs in the two groups are drawn from different underlying populations. The second and equally important point is that two groups with different means will nonetheless regress to the same mean if they are drawn from the same underlying population. This mathematical characteristic of regression to the mean can easily be verified with simulated data, as described in the appendix. Thus when blacks and whites with different group means also regress to different means, the result is not a mathematical tautology.

With these considerations in mind, I present the results from matched sibling samples without interpretation, then turn to the question of how these results correspond to the logic of the Flynn effect. All scores throughout this section of the paper refer to regressed true scores. None of the analyses use sample weights.

Results from Matched Sibling Samples

Jensen (1973) first discovered that siblings in a large sample of black and white California elementary school students matched for IQ regressed to their respective group means. He also found that the higher the IQ score of the matched children, the greater the difference in regression. As examples of the magnitude, Jensen reported that white children with IQs of 120 had siblings with a mean of about 110, while black children with IQs of 120 had siblings with a mean of about 100. At the other end of the scale, white and black children with IQs of 70 had siblings with means of about 85 and 78 respectively. In other words, at all points along the IQ scale, the white siblings appeared to be regressing to a population mean of 100 and the black students to a mean of about 85 (Jensen 1973: 118).[7]

[7] Osborne (1980) using data collected and analyzed for twins ages 12–20 as of 1972, found similar results using samples of white twins (n=133) and black twins (n=47), but these samples were not matched for IQ.

1st generation of the NLSY. The sibling pairs in the 1st generation NLSY sample were born in roughly the same time period (1957–64) as the sample with which Jensen was working. The mean AFQT score of the sample of 552 matched reference siblings was identical to the nearest tenth of a point, 89.9 for both the black and white samples, and the distributions were correspondingly identical.

The means for the white and black comparison siblings were 94.9 and 87.2 respectively. Figure 1 shows the regression lines for each race and the observed means of comparison siblings grouped by the reference siblings' IQ. The groupings were: scores of less than 75, 75–84, 85–94, 95–104, and 105+. Sample sizes for these groupings were 36, 109, 268, 100, and 39 respectively. The coordinates for the subgroups are based on the actual mean of the reference subgroup, not the midpoint of the subgroup range.

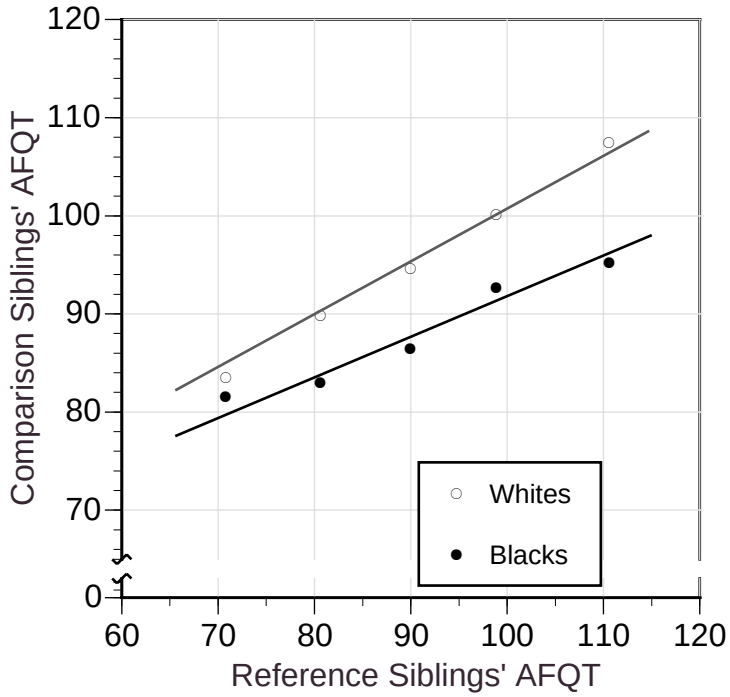

Figure 1. Regression to the Mean among Siblings in the 1st Generation, Matched Sample

Regarding the regression lines: If the reference sibling in the 1st generation has an AFQT score of 80, for example, the white comparison sibling is expected to have an AFQT ten points higher, while the black comparison sibling is expected to be only three points higher. For reference siblings with AFQT scores of 110, the white comparison sibling is expected to be five points lower while the black comparison sibling is expected to be 16 points lower. The difference in the slopes of the regression lines does not quite reach significance at the .05 level. For the 1st generation, the white comparison siblings regressed to a mean of 100.8, while the black comparison siblings regressed to a mean of 85.2, a difference of 15.6 points.

Regarding the grouped means: Two of the grouped means fall conspicuously above the regression line: blacks in the under 75 group and blacks in the 95–104 group. The statistical

significance of the deviation of the grouped means from the regression line was tested by comparing the grouped mean with the regression line that results when the cases in question are deleted from the sample. The deviations of the black groupings for 55–75 and 95–105 reached the .05 significance level. No other groupings in either race significantly deviated from the regression line.

2nd generation of the NLSY. Turning to the 616 matched pairs in the 2nd generation, means of the true regressed scores for the matched white and black reference siblings were 85.0 and 85.1 respectively. The means of the white and black comparison siblings in the 616 matched cases for the 2nd generation were 89.3 and 82.2 respectively.

Figure 2 uses the same format as Figure 1 did for the 1st generation, showing both the regression line for the entire sample and grouped means for the comparison siblings based on groups of reference siblings. Once again, the groupings were less than 75, 75–84, 85–94, 95–104, and 105+. Sample sizes were 92, 213, 217, 71, and 23 respectively.

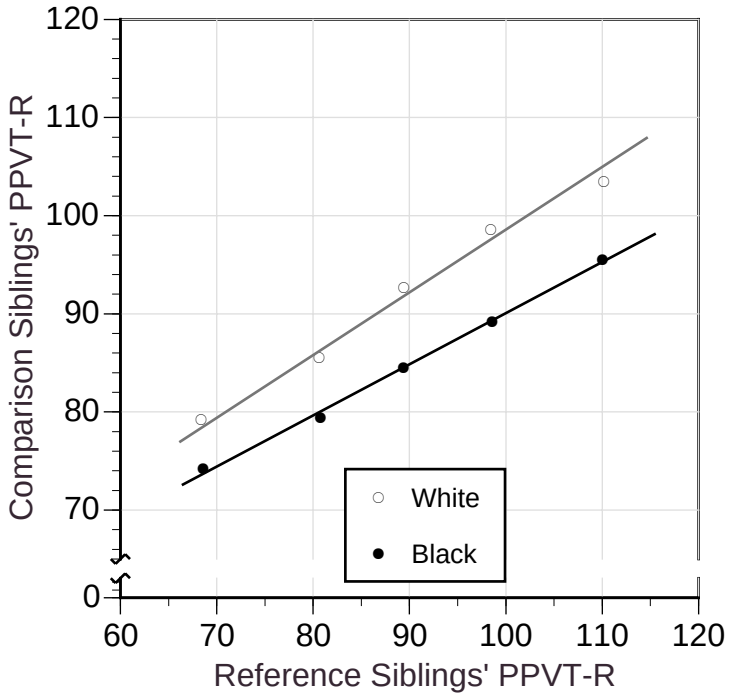

Figure 2. Regression to the Mean among Siblings in the 2nd Generation, Matched Sample

Regarding the regression lines: The general shape of the regression lines for the two generations is very similar. The divergence in the slopes of the regression lines in the 2nd generation reaches the .05 level of statistical significance. It is worth noting some differences at the lower end of the distribution. In the 1st generation, if the reference sibling has a score of 80, the score of the white comparison sibling is expected to be ten points higher; in the 2nd generation, only six points higher. The black comparison sibling of someone with a score of 80 in the 1st generation is expected to have a score of 83 but a score of only 80 in the 2nd generation. The differences were small at the higher levels of IQ, however. For reference siblings with a test score of 110, the

average white comparison sibling was five points lower in both the 1st and 2nd generations; the black comparison sibling was 17 and 16 points lower respectively. Overall, the white comparison siblings in the 2nd generation regressed to a mean of 96.7 and the black comparison siblings regressed to a mean of 79.0, a difference of 17.7 points.[8]

Regarding the grouped means: None of the groupings for either race departed significantly from the regression line.

It should be emphasized that the results for high-IQ pairs are sparse. Only 16 reference siblings in the 1st generation data and 10 in the 2nd generation data involved regressed true scores of 110 or higher. The possibility cannot be excluded that the regression patterns were different for the upper-IQ cases than for the rest of the range.

Regression to the Mean by Social and Economic Category

The above analysis was replicated with samples matched for parental education and income in addition to the reference siblings' cognitive test score. Figure 3 on the following page shows the trendlines from these matched samples. The regression lines from the groups matched only for IQ are shown as broken lines.

Regarding the regression lines: As the figure makes apparent, matching for parental income and education had virtually no effect on the regression lines in either generation.

Regarding the grouped means: The only anomalies are in the 1st generation and both come from the white sample, with white comparison siblings far underperforming their predicted AFQT in the under-75 group and exceeding their predicted AFQT the 75–84 group ($p<.01$ in both cases). The under-75 group constitutes only 13 pairs in each race, so not much should be made of the anomaly. It is worth noting, however, that only two out of the 13 white siblings grew up in a household where either parent had gone beyond the ninth grade (the exceptions had reached 12th grade). Such a low level of parental education is highly exceptional among the white 1st generation and suggests not only a poor environment for nurturing cognitive development but exceptionally low parental IQ. The anomaly in the 75–85 group has no obvious explanation.

[8] The use of the mean PPVT-R score makes the use of regressed true scores redundant for those who have been tested more than once: In effect, they have accomplished in practice what the regressed true score tries to accomplish using statistical theory. The analyses reported in the test were therefore replicated using the actual mean score for such subjects and using regressed true scores for subjects who had taken the test only once. The results were substantively indistinguishable from the ones presented in the text.

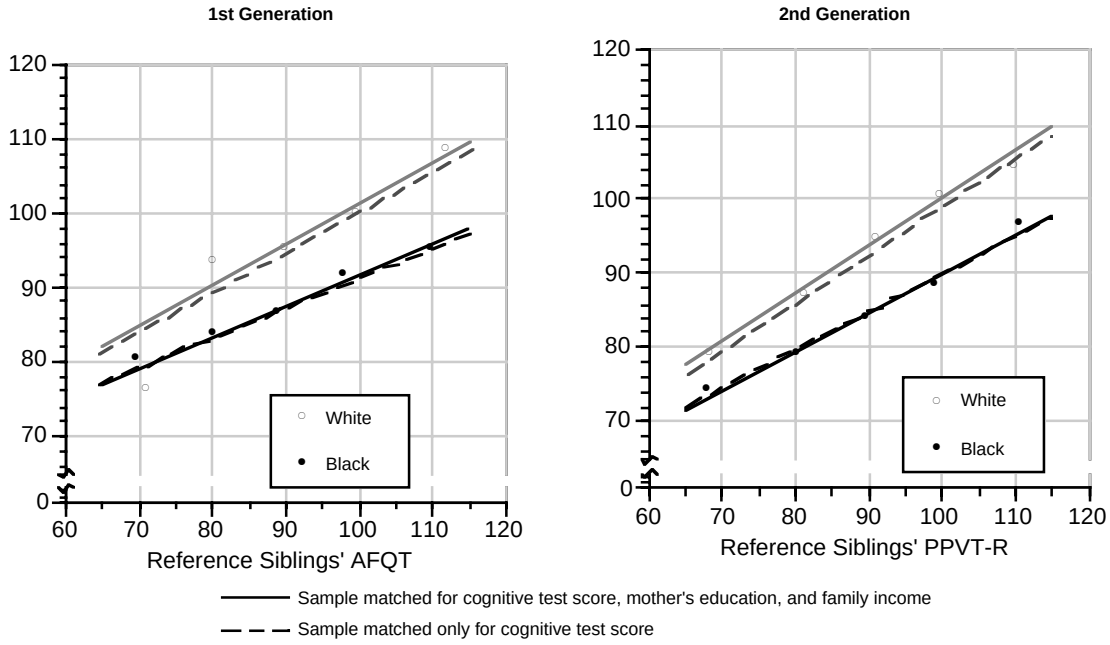

Figure 3.
Regression to the Mean in a Sample Matched for Cognitive Test Score, Mother's Education, and Family Income

These two anomalies aside, the noteworthy feature of Figure 3 is how little is noteworthy. The patterns of the regression lines are remarkably similar to the patterns in the sample matched only for IQ and, for that matter, similar to the unmatched samples of siblings. Table 2 summarizes the results that have been presented, adding data on the unmatched full samples of siblings.

Table 2. Comparison of Means to Which the Comparison Siblings Regress

Sample	*n*	*Observed mean of the reference siblings*			*Mean to which the comparison siblings regressed*		
		White	*Black*	*Difference*	*White*	*Black*	*Difference*
1st Generation							
Total sibling sample	1,147 (w) 932 (b)	103.0	84.9	18.1	104.2	85.8	18.4
Sample matched for AFQT	552	89.9	89.9	0.0	100.8	85.2	15.6
Sample matched for AFQT plus maternal education and family income	355	93.2	92.0	1.2	102.6	85.8	16.8
2nd Generation							
Total sibling sample	1,903 (w) 1,178 (b)	95.7	78.7	17.1	96.1	79.0	17.1
Sample matched for PPVT-R	616	85.1	85.0	0.1	96.7	79.0	17.7
Sample matched for PPVT-R plus maternal education and family income	379	88.1	87.2	0.9	100.4	78.8	21.6

It is usually assumed that the more closely black and white samples are matched, the smaller the BW gap becomes, as has been the case when using SES variables to control for the BW difference in regression equations. But matching has very little effect on the BW difference if the outcome in question is the mean to which comparison siblings regress. In both generations, the BW difference was larger for the sample matched on cognitive test score, maternal education and family income, than for the sample matched only for cognitive test score.

It is not possible with the sample at hand to examine fine-grain comparisons of many different socioeconomic configurations, because cell sizes usually become too small when more than one variable is included. But to get a sense of the differential regression to the mean for various socioeconomic groups, a few examples, drawing from the full sibling samples (including cases not part of the matched samples) will serve to illustrate what seems to be a more general pattern.

The first example consists of households in the 1st generation of the NLSY, earning a working-class income of $25,000–$50,000 per year. Consider the children of such families with AFQTs of 100–114 (continuing to use regressed true scores), bright enough to consider college even if they are not prime college material. Seventy-four white reference siblings in the 1st genera-

tion fit this category, with a mean AFQT score of 106.2 Their comparison siblings had a mean of 105.2, only one point lower. Seventy-five percent of the comparison siblings had scores of 100 or more, making college a realistic possibility. The 16 black reference siblings who fit the same category had a mean AFQT of 104.9, but their comparison siblings had a mean of only 94.4, more than ten points lower. Only 38 percent of the comparison siblings had scores of 100 or more.

Turning to the 2nd generation, consider the families in which the mother had completed a bachelor's degree or higher but the reference sibling had a PPVT-R of only 85–99, indicating a level of cognitive ability not ordinarily sufficient to complete a genuine college program. The 87 white reference siblings in the 2nd generation who fit this description had a mean PPVT-R of 94.1. Their comparison siblings had a mean of 99.8, about six points higher. Half of the white comparison siblings had scores of 100 or higher. The 21 black reference siblings who fit this description had a mean PPVT-R of 91.5. Their comparison siblings had a mean of 87.6, about four points lower. None had a PPVT-R of 100 or higher.

The final example consists of households near the bottom of the socioeconomic ladder, with total income of less than $25,000 and a mother with no more than 12 years of education. I focus on the reference siblings with a cognitive test score in the 75–89 range, conspicuously below average in measured intelligence and almost never able to get through college. In the 1st generation, 22 white reference siblings met this description, with a mean AFQT of 83.7. Their comparison siblings a mean almost 12 points higher, 95.3. More than a quarter of the white comparison siblings had scores of 100 or higher. The 194 black reference siblings in this category had a mean AFQT of 82.3. Their siblings had a mean of 83.3, one point higher. Five percent had scores of 100 or higher.

The above examples illustrate a general pattern of large between-race differences within socioeconomic configurations. There are no countering examples of substantial between-race convergence within the combinations of IQ, maternal education, and income. Once again, however, the number of cases at the upper end of the IQ distribution is too small to permit strong conclusions about that part of the range.

Implications of the Sibling Results for Convergence Via the Flynn Effect

At issue is the argument that dissipation of the environmental disadvantages facing blacks will lead to the eventual convergence of black and white test scores. In ordinary samples of sibling pairs, examining regression to the mean is not informative. That black and white siblings regress to their respective means rather than their combined population mean merely expresses, in another form, that something is creating a large group difference between blacks and whites on mental tests. Nor does matching blood relatives for IQ necessarily help. When, for example, black and white mothers are matched for IQ, little analytic leverage is gained when examining regression to the mean among their offspring, because mothers and children are not matched for shared environment within race.

But matching sibling samples for IQ does offer important analytic leverage. Between races, the mean, variance, and even skew of the distribution of the reference siblings' IQ scores are identical for the black and white samples. Within race, full siblings are matched on both their biological heritage and their shared environment. On what theoretical basis might we predict that the comparison siblings of each race will regress to different means? Matching siblings for IQ across races puts constraints on the logically acceptable answers to that question.

The answers that are also consistent with the logic of the Flynn effect demand that the BW difference be exclusively environmental. Flynn himself has discussed the general problem of environmental explanations in terms of what Jensen has called a "Factor *X*" (Jensen, 1973). It is known that the standard SES variables explain about 30–40 percent of the BW difference (Herrnstein and Murray 1994; Jensen 1998; Flynn 1999a). The residual difference needs to be explained by some more general, diffuse environmental disadvantage (hence the label "Factor *X*") that that has two characteristics: It affects blacks and whites differentially, and its depressing effect on black test scores is uniform across the IQ range.

In thinking about what Factor *X* might be, a candidate immediately comes to mind: racism. Racism obviously affects blacks and whites differentially, and it is diffuse and omnipresent. The difficulty is to explain how racism can have uniform effects across the range. As Flynn has pointed out,

> Racism looks like a potent environmental factor that affects all Blacks both negatively and with considerable uniformity…. [But] racism is not some magic force that harms Blacks without a chain of causality. Racism harms Blacks because of certain effects, such as lack of self-confidence, low self-image, emasculation of men, the welfare-mother home, poverty. Who could argue that these same factors do not vary significantly within the Black population?… If these factors both are potent and vary among Blacks, why do they explain so little IQ variance within the Black population? (Flynn 1999: 13).

Jensen has recently presented a fully elaborated statement of the mathematical implications of a strict environmentalist explanation of the BW difference (Jensen 1998: 447–62), focusing on the constraints posed by evidence that the within-race heritability of IQ is similar (though perhaps not identical) for blacks and whites (Jensen 1998: 446–47), and the within-race developmental processes are similar (Rowe & Cleveland 1996, Rowe, Vazsonyi, & Flannery, 1994).

The sibling results presented here add another constraint to the burden on those who take a strict environmentalist position. Proponents of the convergence hypothesis must not only posit an environmental Factor *X* that affects blacks but not whites and that is relatively uniform in its effects across the range of IQ (possibly having a greater effect as IQ goes up). Because the shared environment is the same for both siblings by definition, they must also posit a causal mechanism that is expressed through the nonshared environment. This requirement is accentuated by the results presented in Figure 3, showing that differential regression to the mean is virtually unaffected by matching for parental income and education along with subject IQ. As a final consideration, a Factor *X* that satisfies the rest of the conditions must also be quite powerful if it is

to produce BW differences in regression to the mean commensurate with those observed in the sibling samples.

Combining these constraints—a Factor *X* that is powerful, pervasive, uniform across the range of IQ, located in the nonshared environment, and consistent with equivalence of within-race heritability and within-race developmental processes—poses difficult problems, beginning with a basic problem of description. A defining feature of the nonshared environment is that it is random across siblings. How does one conceptualize a nonshared environment that is random with respect to siblings and both powerful and systematic with respect to race?

I am unable to describe a Factor *X* in the nonshared environment that meets these requirements. The possibilities are complex, however. The restricted claim here is that the challenge is real and must be met if the convergence logic of the Flynn effect is to be sustained.

The sibling results from the NLSY offer a second kind of challenge to the logic of convergence. Let us assume that the first challenge has been met, and that a Factor *X* having all the required properties can be defined theoretically. Having stipulated that, it remains to examine the nature of the BW difference over time, for the Flynn-effect logic also says that over time this Factor *X* will dissipate and the scores of whites and blacks in similar environments will tend to become more similar.

Progress in the dissipation of Factor *X* when comparing two different generations could be reflected in the regression patterns in one of three ways. The simplest indication would be that the regression lines move closer together, accompanied by reduced differences in IQ groups across the range. This effect is the least interesting, since we would not need the sibling samples to know it—we would already have observed a reduction in the observed black and white population means.

The second possible effect is that the relationship of the black and white regression slopes change. Suppose, for example, that the observed means from the 1st to 2nd generations remained unchanged, but the slope of the black regression line became steeper relative to the white slope. This result would be consistent with a changing environment in which the effects of Factor *X* had diminished for high-IQ blacks while increasing for low-IQ blacks. Much in the history of the last 30 years, with the black middle class and black underclass growing contemporaneously, makes such a possibility plausible.

The third possibility is that one or more subgroups would move conspicuously off the regression line. In this scenario, the Factor *X* is unchanged for most blacks, but shifts importantly for some subgroups but not others. In many ways, this is the most plausible of all scenarios, with the increase in opportunities for high-IQ blacks once again being the area in which positive change might be expected to occur without necessarily being accompanied by dissipation of Factor *X* for low-IQ blacks.

None of the three possible changes in the pattern of sibling scores occurred in the 1st and 2nd generations of the NLSY. The regression lines did not get closer. The relative slopes of the black and white regression lines did not change. None of the black subgroups moved off the regression line.

In considering how to interpret this unchanging pattern, it is important to remember the period covered by these two generations. Almost 24 years separate average members of the two generations. The 1st generation was born in 1957–64. Ninety percent of the 2nd generation of the NLSY was born after 1978. No period since the end of the Civil War has seen more extensive changes in the environment in which black children grow up than the period from the 1960s to 1980s, which makes the stability of the patterns of regression to the mean in the 1st and 2nd generations of the NLSY the more striking.

Conclusion

The purpose of this paper has been to examine whatever light the NLSY may shed on the question of convergence in BW test scores, with particular attention to the expectations raised by the logic of the Flynn effect.

In the two generations of the NLSY, no convergence has occurred. The BW difference on a highly *g*-loaded cognitive test for the 1st generation of the NLSY, born from 1957–64, was 16.6 points, amounting to 1.24 SDs relative to the black and white distributions. For the 2nd generation, born primarily in the 1980s, the difference on a widely used test of verbal cognitive ability was 17.8 points, or 1.26 SDs. The estimated magnitude of the BW difference in the 2nd generation is robust, surviving a variety of hypotheses about possible sources of attenuation.

When sibling samples from the 1st and 2nd generations are matched for cognitive ability, it is found that the comparison siblings regress to different means about as widely separated as the BW difference found in unmatched samples. This finding persists when the samples are matched not only for cognitive ability but for maternal education and family income as well. These findings pose a challenge to the strict environmentalist interpretation of the BW difference. The matched sibling pairs in both the 1st and 2nd generation of the NLSY effectively rule out the traditional types of environmental analysis (parental socioeconomic status, family background variables, quality of education) as explanations for the differences in comparison sibling IQ. An adequate theory of the environmental source of the difference must include causal mechanisms that are consistent with the logical constraints posed by the sibling results and also lend themselves to empirical test.

These findings from the NLSY will, like all findings about race and IQ, become part of a hotly contested debate. Three recent reviews of the American BW difference published as recently as 1998 reach apparently opposing conclusions. Hedges and Nowell (1998: 167) write that "The data provide convincing evidence that racial differences have decreased over time." Lynn (1998: 1001) writes that "the best reading of the data as a whole is that there is no conclusive evidence that the black-white difference in intelligence has been narrowing over time." Jensen (1998: 357) writes that "The mean W-B IQ difference has remained fairly constant at about 1σ for at least eighty years, with no clear trend upward or downward since the first large-scale testing of representative samples of blacks and whites in the United States."

But the conflict in these conclusions may be more apparent than real. Consensus seems to be broadening on several key issues that could lead to a reconciliation of the opposing view-

points: (1) The tests that show the clearest convergence are tests designed to measure academic achievement, not full-scale IQ tests. (2) Variation in the magnitude of the BW difference is a positive function of variation in the *g* loading of the tests. (3) The convergence that has occurred has been produced predominantly by improvements at the low end of the range. (4) Convergence has effectively stopped since the late 1980s.

Drawing these strands together, I suggest the following hypothesis: the BW difference on tests of academic achievement narrowed during this century while the BW difference in tests of cognitive ability did not. An elaboration of this hypothesis is that the narrowing on tests of academic achievement was largely confined to the post-war period from 1945 to 1985 and was associated with broad-spectrum improvements during that period in elementary and secondary education for the average and below-average student in general (Herrnstein and Murray 1994: 419–27), which had the greatest effect on black students, the group that had been most broadly deprived of good education.

It is unlikely that the hypothesis is strictly correct, insofar as it seems *prima facie* unlikely that there has been no change whatsoever in the BW difference on cognitive tests over the course of the century. But a clear-cut hypothesis is useful for sharpening the investigation, which leads to this important point: A comprehensive, systematic investigation is badly needed. Despite the many books and articles that have reviewed the BW difference, a full-scale meta-analysis of all the accumulated data has yet to be done. The task is eminently feasible. Technical descriptions of all the extant studies from the earliest tests through the mid-1970s are available in just two volumes (Shuey 1966 and Osborne & McGurk 1982). Much of the work of assembling comprehensive information on subsequent studies has been done by Jensen (Jensen 1985, Jensen & Naglieri 1987, and Jensen 1992). Most of the post-1970 national surveys are publicly available on-line or on CD-ROM. Meta-analytic methodology is by now well developed, and the IQ literature, involving many disparate instruments and samples, seems especially appropriate to the strengths of meta-analysis.

The key to conducting the analysis is to discriminate between measurement of academic proficiency and measurement of cognitive ability. Since all mental tests show intercorrelations, the distinction cannot be a simple one, but *g* loadings provide a natural basis for discriminating among tests. In effect, the analysis can follow the lead of Jensen's exploration of the Spearman Hypothesis. The study I am proposing would test the Spearman hypothesis longitudinally, hypothesizing that the estimated slope of the BW difference over time flattens as the *g* loadings of the tests used to estimate the slope rise.

A rigorous test of the BW difference over time, embracing every interpretable study ever conducted, decomposed according to the *g* loading of the test, could inform several outstanding questions about the BW difference. With specific regard to the Flynn effect, a study that establishes that the slope for highly *g*-loaded tests throughout the century is effectively flat would be important evidence against the Flynn-effect logic for convergence of black and white test scores. If the study establishes instead that the BW difference on highly *g*-loaded tests has changed over time, knowing the magnitude of that change and its timing could inform many ongoing debates not only about the nature of the Flynn effect but also about the prospects for eventual convergence.

Appendix

When presenting the evidence of differential regression to the mean as in Figures 1 and 2 of this paper, a common reaction is that the results are tautological—two groups with different means must necessarily regress to different means as well. Apart from the settled statistical theory that says that two groups drawn from a common underlying population will regress to a common mean (Humphreys 1978), the question may easily be settled in a few minutes using any statistical package with a random number function and standard capabilities for creating variables.

The hypothesis to be falsified is that two groups with different means will necessarily regress to different means. The method of falsification is two create two groups that have different means but are known to be drawn from the same underlying population, and then demonstrate that they regress to a common mean.

Step 1. Create a large sample of simulated sibling data comparable to IQ data; i.e., two normally distributed variables correlated at the .4–.6 level, with one variable designated as the reference score *R* and the other as the comparison score *C*.

Step 2. Separate the cases into two groups using a selection algorithm that favors high *R* scores over low *R* scores, thereby producing two groups with different means.

Step 3. For each group, regress *C* on *R.*

Given large samples, it will be found that the means to which the two groups are regressing (when $\hat{C} = R$) are the same within a few tenths of a point.

To preserve the parallelism with the IQ example, the group difference in means should approach the 1 SD magnitude ordinarily observed in the BW difference and the algorithm should preserve a more-or-less normal distribution for the two groups of reference scores, but neither the magnitude of the difference in means nor the normality of the distributions affects the demonstration.

References

Baker, P. C., Keck, C. K., Mott, F. L., & Quinlan, S. V. (1993). *NLSY Child Handbook, Rev. Edition*. Columbus, Ohio: Center for Human Resource Research, Ohio State University.

Charng, H.-W., & Baydar, N. (1996). *1992 Children of the NLSY Kinship Links*. Seattle, WA: Battelle Centers for Public Health.

Earles, J. A., & Ree, M. J. (1992). The predictive validity of the ASVAB for training grades. *Educational and Psychological Measurement, 52*, 721–725.

Eckland, B. K. (1979). Genetic Variance in the SES-IQ Correlation. *Sociology of Education, 52* (3), 191-196.

Feldt, L. S., & Brennan, R. L. (1989). Reliability. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed.,). New York: MacMillan.

Flynn, J. R. (1987). Race and IQ: Jensen's case refuted. In S. Modgil & C. Modgil (Eds.), *Arthur Jensen: Consensus and controversy* (pp. 221–232). Lewes, England: Falmer Press.

Flynn, J. R. (1998). Evidence against Rushton: The genetic loading of WISC-R subtests and the causes of between-group IQ differences. *Personality and Individual Differences, 26,* 373–379.

Flynn, J. R. (1999a). Searching for justice: The discovery of IQ gains over time. *American Psychologist, 54* (1), 5–20.

Flynn, J. R. (1999b). Reply to Rushton: A gang of *g*s overpowers factor analysis. *Personality and Individual Differences, 26*, 391-393.

Hedges, L. V., & Nowell, A. (1998). Black-White Test Score Convergence since 1965. In C. Jencks & M. Phillips (Eds.), *The Black-White Test Score Gap* (pp. 149–181). Washington: Brookings Institution Press.

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

Humphreys, L. G. (1978). To understand regression from parent to offspring, think statistically. *Psychological Bulletin, 85* (6), 1317-1322.

Jensen, A. R. (1973*). Educability and Group Differences*. London: Methuen & Co.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Free Press.

Jensen, A. R. (1985). The nature of the Black-White difference on various psychometric tests: Spearman's hypothesis, *The Behavioral and Brain Sciences* (1985 ed., Vol. 8, pp. 193-258). Cambridge: Cambridge University Press.

Jensen, A. R. (1992). Spearman's hypothesis: Methodology and evidence. *Multivariate Behavioral Research*, 27, 225-233.

Jensen, A. R. (1998). *The* g *factor: The Science of Mental Ability*. Westport, CT: Praeger.

Jensen, A. R., & Naglieri, J. A. (1987). Comparison of Black-White differences on the WISC-R and the K-ABC: Spearman's hypothesis. *Intelligence, 11* (1), 21.

Jensen, A. R., & Reynolds, C. R. (1982). Race, social class and ability patterns on the WISC-R. *Personality and Individual Differences, 3*, 423-438.

Lynn, R. (1998). Has the black-white intelligence difference in the United States been narrowing over time? *Personality and Individual Differences, 25*, 999-1002.

National Center for Health Statistics. (1998). *Advance Report of Final Natality Statistics, 1996* (Monthly Vital Statistics Report 46-11). Washington: National Center for Health Statistics.

Osborne, R. T. (1980). *Twins: Black and White.* Athens, GA: Foundation for Human Understanding.

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

Ree, M. J., & Earles, J. A. (1991). The stability of convergent estimates of *g*. *Intelligence, 15*, 271-278.

Reynolds, C. R., Chastain, R. L., Kaufman, A. S., & McLean, J. E. (1987). Demographic characteristics and IQ among adults: Analysis of the WAIS-R standardization sample as a function of the stratification variables. *Journal of School Psychology, 25*, 323-342.

Robertson, G. J., & Eisenberg, J. L. (1981). *Technical Supplement, Forms L and M, Peabody Picture Vocabulary Test-Revised*. Circle Pines, MN: American Guidance Service.

Rowe, D. C. (1994). *The Limits of Family Influence: Genes, Experience, and Behavior*. New York: Guilford Press.

Rowe, D., Vazsonyi, A., & Flannery, D. (1994). No more than skin deep: Ethnic and racial similarity in developmental process. *Psychological Review, 101* (3), 396-413.

Rowe, D. C., & Cleveland, H. H. (1996). Academic achievement in African Americans and whites: Are the developmental processes similar? *Intelligence, 23*, 205-228.

Rushton, J. P. (1989). Japanese Inbreeding Depression Scores: Predictors of Cognitive Differences between Blacks and Whites. *Intelligence, 13*(1), 43-51.

Rushton, J. P. (1997). Race, intelligence, and the brain: the errors and omissions of the 'revised' edition of S.J. Gould's *The Mismeasure of Man* (1996). *Personality and Individual Differences, 23*(1), 169-180.

Rushton, J. P. (1999). Secular gains in IQ not related to the *g* factor and inbreeding depression—unlike Black-White differences: A reply to Flynn. *Personality and Individual Differences, 26*, 381–389.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Technical Manual for the Stanford-Binet Intelligence Scale: Fourth Edition*. Chicago: Riverside.