Available online at www.sciencedirect.com
ScienceDirect
Intelligence 35 (2007) 305–318
INTELLIGENCE

The magnitude and components of change in the black–white IQ difference from 1920 to 1991: A birth cohort analysis of the Woodcock–Johnson standardizations

Charles Murray

American Enterprise Institute, United States

Received 27 September 2006; received in revised form 15 January 2007; accepted 5 February 2007
Available online 23 March 2007

Abstract

The black–white difference in test scores for the three standardizations of the Woodcock–Johnson battery of cognitive tests is analyzed in terms of birth cohorts covering the years from 1920 through 1991. Among persons tested at ages 6–65, a narrowing of the difference occurred in overall IQ and in the two most highly *g*-loaded clusters in the Woodcock–Johnson, Gc and Gf. After controlling for standardization and interaction effects, the magnitude of these reductions is on the order of half a standard deviation from the high point among those born in the 1920s to the low point among those born in the last half of the 1960s and early 1970s. These reductions do not appear for IQ or Gc if the results are restricted to persons born from the mid-1940s onward. The results consistently point to a B–W difference that has increased slightly on all three measures for persons born after the 1960s. The evidence for a high B–W IQ difference among those born in the early part of the 20th century and a subsequent reduction is at odds with other evidence that the B–W IQ difference has remained unchanged. The end to the narrowing of the B–W IQ difference for persons born after the 1960s is consistent with almost all other data that have been analyzed by birth cohort.

Keywords: Woodcock–Johnson; Black–white differences; IQ; Intelligence; Birth cohort

1. Introduction

The empirical record on changes in the black–white (B–W) difference in mental test scores over time informs many theoretical and policy issues, ranging from the causal roles of nature and nurture in ethnic differences in intelligence to policy arguments over the effects of affirmative action. But almost all of the reports of changes in the B–W difference have been framed in terms of the years in which the tests were administered (period effects) rather than the years in which the persons being given the tests were born (cohort effects). The choice can transform the interpretation of the data, as demonstrated by Huang and Hauser (2001) in their comparison of the B–W difference in the vocabulary test administered in the General Social Survey (GSS) from 1972 through 1998. Analyzed by survey year, a small reduction of the B–W difference is observed—statistically significant in Huang and Hauser's analysis, not significant in Lynn (1998). Analyzed by birth year, a substantial and significant narrowing of the difference emerged.

The contrast between period and cohort effects has special salience now because of the recent intensification of the debate about whether the B–W difference has

E-mail address: cmurray@aei.org.

0160-2896/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.intell.2007.02.001

narrowed in recent decades, and if so, on what kinds of measures. Rushton and Jensen (2005) and Gottfredson (2005) are two recent examples of one school of thought, arguing that the totality of the evidence indicates that the B–W difference in IQ has not significantly narrowed since IQ tests were first administered to large samples of African-Americans in World War I. They accept evidence of narrowing of the B–W difference in the National Assessment of Educational Progress (NAEP) during the 1970s and 1980s (Perie & Moran, 2005), but argue that the most highly *g*-loaded tests have not shown comparable narrowing. Nisbett (2005) argues that the correlations between some of the achievement tests and IQ tests are so high that narrowing in the B–W difference in the former is a strong evidence of narrowing in the latter. Dickens and Flynn (2006) present fresh direct evidence from IQ tests with period analyses of the standardization samples for the Wechsler scales for children and adults (WISC and WAIS), the Stanford–Binet, and the Armed Forces Qualification Test (AFQT), pointing to a reduction of 3–6 points in the B–W IQ difference between 1972 and 2002.

These findings are apparently inconsistent with Murray (2006), which employs a birth cohort analysis of test scores among children born to women in the 1979 sample of the National Longitudinal Survey of Youth (NLSY-79). These subjects were born from the mid-1970s through the mid-1990s, overlapping the same period covered by the standardizations examined by Dickens and Flynn, but showed no narrowing in the B–W difference in either achievement tests or a measure of verbal IQ. Murray's findings, combined with the reductions observed in the B–W difference gap observed in the Huang and Hauser cohort analysis and in the NAEP period analyses, suggest a hypothesis that reconciles apparently conflicting data: Narrowing in the B–W difference did occur, including narrowing in overall IQ, to persons born until the 1970s. Then the narrowing of the difference stopped. The present article uses the three standardizations of the Woodcock–Johnson cognitive test batteries as evidence for that hypothesis and examines the components of the changes that occurred.

2. Method

The Woodcock–Johnson test batteries include measures of both cognitive ability and academic achievement. The present article limits itself to the cognitive tests. The information about the tests that follows is taken from the technical manuals for the three standardizations (Woodcock, 1978; McGrew, Werder, & Woodcock, 1991; McGrew & Woodcock, 2001), supplemented by personal communication regarding cross-standardization comparability with Kevin McGrew, who directed the norming and psychometric analysis of the second and third editions of the Woodcock–Johnson batteries.

The initial version of the test, WJ1, was standardized using a sample of 4732 subjects aged 2 to 84, tested over the period from April 1976 to May 1977. WJ2 was standardized with a sample of 6359 subjects aged 2 to 95, tested from September 1986 to August 1988. WJ3 was standardized with a sample of 8818 aged 2 to 98, tested from September 1996 to August 1999. Samples for all three standardizations were stratified by region, community size, sex, race/ethnicity, and occupational status of adults, plus additional stratification criteria for WJ2 and WJ3. School-age subjects were randomly selected from class lists in the participating schools, and tested subjects randomly selected from among those who gave permission to be tested and met the stratification criteria. Compensation was provided for all participants, with monetary compensation for those above grade six. All scores are normed to a mean of 100 and a standard deviation (σ) of 15. All scores are age normed.

The conceptual basis for Woodcock–Johnson's approach to measuring intelligence is grounded in Gf–Gc theory, based on the distinction between fluid intelligence (Gf) and crystallized intelligence (Gc) originated by Raymond Catell (Cattell, 1941, 1943, 1950). As extended in Horn (1965), and then in Horn and Stankov (1982), the theory posits seven categories of broad cognitive abilities: Gc, comprehension-knowledge, corresponds to Cattell's crystallized intelligence. Gf, fluid reasoning, corresponds to Cattell's fluid intelligence. Gsm measures short-term memory. Glr measures long-term retrieval. Ga, auditory processing, measures the ability to comprehend patterns among auditory stimuli. Gs measures processing speed. Gv, visual processing, measures the ability to think using visual patterns. From these seven clusters, the Woodcock–Johnson test produces a measure it calls General Intellectual Ability (GIA) and that I will call IQ.

WJ2 and WJ3 both use a pair of subtests to measure each of the seven clusters. WJ1 provides comparable cluster measures for Gf, Gc, Gsm, and Gs, with single tests for measuring Glr and Ga. WJ1 had no measure of Gv.

Table 1 summarizes the subtests used to measure each of the clusters, the loadings of the subtests on a principal components analysis, and the rank order of the loadings. The sample consists of all subjects with complete test score data who were tested from ages 6 through 65.

Table 1 arrays the clusters in rough order of their *g*-loadings. Gc and Gf are a clear first and second, with Gc and Gf together monopolizing five out of six of the most

Table 1
Cognitive clusters, subtests, and *g*-loadings by standardization

Cognitive cluster	Test	WJ1		WJ2		WJ3	
		g-loading	Rank	*g*-loading	Rank	*g*-loading	Rank
		(*n*=3973)		(*n*=5092)		(*n*=3450)	
Comprehension-knowledge (Gc)	Antonyms–synonyms (1)	0.825	1	0.766	1		
	Picture vocabulary	0.714	3	0.712	3		
	Verbal comprehension					0.781	1
	General information					0.729	3
Fluid reasoning (Gf)	Concept formation	0.666	4	0.697	4	0.734	2
	Analysis–synthesis	0.641	6	0.672	5	0.659	5
	Analogies	0.797	2				
Short-term memory (Gsm)	Memory for sentences	0.657	5	0.650	7		
	Numbers reversed	0.597	9	0.671	6	0.594	7
	Memory for words					0.557	9
Long-term retrieval (Glr)	Visual-auditory learning	0.635	7	0.715	2	0.684	4
	Memory for names			0.644	8		
	Retrieval fluency					0.513	12
Auditory processing (Ga)	Sound blending	0.624	8	0.624	9	0.611	6
	Incomplete words			0.561	11		
	Auditory attention					0.513	11
Processing speed (Gs)	Spatial relations (timed)	0.569	10				
	Visual matching	0.561	11	0.560	12	0.580	8
	Cross-out			0.609	10		
	Decision speed					0.518	10
Visual processing (Gv)	Picture recognition			0.543	13	0.374	14
	Visual closure			0.479	14		
	Spatial relations (untimed)					0.510	13

1. Called "oral vocabulary" in WJ2.
The *g*-loadings are the factor loadings for the first factor produced by a principal components analysis without rotation.
Sample consists of all subjects tested at ages 6–65.

highly *g*-loaded subtests in WJ1 and four out of the top five in WJ2 and WJ3. The rank of Gc and Gf is consistent with the typical results from IQ test batteries (Jensen, 1998). The proportion of variance explained by the first factor in the principal components analysis was 0.445 for WJ1, 0.410 for WJ2, and 0.368 for WJ3.

The Woodcock–Johnson technical manuals present detailed tables of reliabilities for both the individual subtests and the clusters. For the subtests shown in Table 1, test–retest reliabilities range from 0.80 to 0.95 for WJ1 (median=0.84); from 0.72 to 0.94 for WJ2 (median=0.87); and from 0.76 to 0.94 for WJ3 (median=0.87).

2.1. Sample for analysis

The sample for the subsequent analyses is limited to persons whose ethnicity/race was coded as black or as non-white, who took the test no younger than age 6 and no older than age 65, and whose birth year was no earlier than 1920 (only two blacks fitting the other criteria were born before 1920). The final criterion was that the subject has complete data on IQ and all of the cognitive clusters discussed above (seven clusters for WJ2 and WJ3; six clusters for WJ1). The sample selection criteria produced analysis samples of 3765, 4380, and 3018 for WJ1, WJ2, and WJ3 respectively.

The complete-data criterion had little effect on WJ1 and WJ2, resulting in the deletion of fewer than 3% of the cases that would otherwise have been eligible. But for WJ3, Woodcock–Johnson employed an advanced psychometric strategy of the type used in the National Assessment of Educational Progress, deriving its norms from partial test batteries administered to a large number of subjects. The requirement for complete data thus eliminated 2350 whites and 521 blacks in WJ3, constituting almost half of those who would have otherwise been eligible for the sample. Lacking means to assess adequately the randomness of the deleted cases, all of the analyses were replicated with a sample that included subjects with partial data from all three standardizations. The results were effectively identical to those produced by the more restricted sample. In Table 4, for example, which presents the fitted B–W difference for IQ, Gc, and Gf for various years, only one of the differences varied by as much as 0.06σ when the less restrictive sample was used (the 1921 baseline for Gc). None of the other differences varied by more than 0.03σ.

2.2. Sampling weights

The sample for WJ1 was designed to be nationally representative as it stood, and the norm sample is the same as the sample used for the following analyses. All of the WJ1 subjects are assigned a sampling weight of 1. The samples for WJ2 and WJ3 were also stratified geographically and demographically, but a sampling weight for each subject was computed as well, representing the ratio of the percentage of persons in the US population with the subject's socioeconomic and demographic profile to the percentage of persons with that profile in the sample. For all subsequent analyses in this article, these values were entered as weights in the computation of means, variances, and regression results.

2.3. Approach to the effects of test age

The Woodcock–Johnson scores are age normed, but test age nonetheless remains an issue in three respects.

The first test-age issue involves the young. The decision not to include subjects tested before age six was taken, first, because children younger than five were given a reduced battery of tests, and tracking changes in the full battery of tests is an objective of this analysis, and, second, because the B–W difference on mental tests among the very young is systematically different from the difference among older subjects. In infancy, the B–W difference can be close to zero (Fryer & Levitt, 2004). The difference rises through the preschool years, usually reaching about 0.70σ on full-scale IQ batteries by 5 to 6 years of age, then rising within a few years to about 1.0σ where it stabilizes for the rest of elementary school (Jensen, 1998). The results from the Woodcock–Johnson standardizations follow the common pattern for IQ tests. The B–W differences for children tested at age five were just 0.30σ, 0.12σ, and 0.69σ for WJ1, WJ2, and WJ3 respectively, with a combined difference of 0.57σ (white n=555, black n=83) rising immediately thereafter to 1.00σ, 0.81σ, and 1.13σ respectively at age 6, with a combined difference of 0.87σ (white n=733, black n=135).[1]

The second issue involves elderly adults. The effects of aging have been divided into "primary aging"—physiological changes that effect everyone in similar ways as they get older—and "secondary aging"—person-specific effects of environment and disease (Busse, 1969). Primary aging produces predictably lower scores on mental tests across the life span (Schaie & Willis, 1993), but the factor loadings of cognitive subtests are invariant over time, and most people maintain their relative ordering in intelligence (Taub & McGrew, 2004; Hertzog & Schaie, 1986; Schaie, 1996). The age-norming used in the Woodcock–Johnson data compensates for the standard deterioration patterns, making it reasonable to assume that an age-normed score of 100 obtained at age 60 is about the same as the score that the same subject would have achieved at 20. In contrast, secondary aging is associated with diseases that affect cognitive functioning differentially across individuals, and these become sufficiently prevalent as people reach their 70s and 80s that many people in that age range have scores that are markedly different than the scores they would have achieved when younger, despite age-norming. Following Huang and Hauser (2001), I use age 65 as the upper cutoff point for the sample.

The third issue involves the possibility of a relationship between the magnitude of the B–W difference and chronological age even among persons no older than 65. There is no direct evidence of such a relationship, either for or against, but the extensive literature on the overall effects of aging on cognitive ability provides indirect evidence pertinent to the findings presented here. Studies have found that IQ in old age is affected by variables such as social status (Arbuckle, Gold, & Andres, 1986), the intellectual demands of one's occupation, lifestyle variables such as reading habits and exercise (Clarkson-Smith & Hartley, 1990), and hypertension (Elias, Robbins, Elias, & Streeten, 1998). However, the magnitude of the effects of these independent variables on IQ is modest. Applying the results in Gold, Andres, Etezadi, Arbuckle, Schwartzman, & Chaikelson (1995), the implied effects on the B–W difference over a 40-year span on overall IQ are unlikely to be more than about 0.05σ, even assuming a 1σ B–W difference on all of the independent variables. Results for age-restricted groups permit an indirect test of the presence of an age-related artifact in the Woodcock–Johnson data (Table 5).

2.4. Treatment of data from different standardizations

Any comparison of means across test standardizations must assume that each standardization has yielded valid and reliable nationally representative norms for the period in which that standardization was conducted. The assumption is complicated in the present article because

[1] All B–W differences presented in the text are calculated on the basis of within-group standard deviations,

$$\frac{\left(\overline{X}_a - \overline{X}_b\right)}{\sqrt{\left(N_a\sigma_a^2 + N_b\sigma_b^2\right)/\left(N_a + N_b\right)}}$$

where N is the sample size, X is the sample mean, σ is the standard deviation, and the subscripts a and b denote each group.

Table 2
Summary data by standardization

Measure	WJ1 (1976–77)				WJ2 (1986–88)				WJ3 (1996–99)				B–W difference in SDs			
	White		Black		White		Black		White		Black		WJ1	WJ2	WJ3	Net, W3–WJ1
	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ	Mean	σ				
	(n=3329)		(n=436)		(n=3573)		(n=807)		(n=2592)		(n=426)					
General intellectual ability (IQ)	101.8	13.5	84.9	15.3	104.3	14.9	90.8	16.0	105.5	13.2	91.7	12.6	1.23	0.90	1.05	−0.18
Comprehension-knowledge (Gc)	101.7	13.5	86.1	16.0	104.4	14.6	90.5	14.4	106.0	12.3	89.8	12.4	1.13	0.95	1.31	0.17
Fluid reasoning (Gf)	100.8	14.5	91.6	15.0	103.1	14.6	92.8	15.6	104.3	13.5	92.9	13.9	0.63	0.70	0.84	0.21
Short-term memory (Gsm)	101.1	14.2	90.5	16.3	104.2	15.2	97.9	16.3	103.4	14.0	95.7	14.1	0.73	0.41	0.55	−0.18
Long-term retrieval (Glr)	100.8	14.4	91.2	15.5	103.7	15.5	95.4	17.1	105.4	13.5	97.7	13.1	0.66	0.53	0.57	−0.09
Auditory processing (Ga)	101.8	13.7	83.6	13.0	103.7	14.2	90.4	14.9	105.5	13.9	93.6	11.7	1.33	0.93	0.87	−0.47
Processing speed (Gs)	100.7	14.6	92.6	14.9	103.0	15.7	95.9	18.4	103.0	14.7	97.9	13.4	0.55	0.43	0.35	−0.20
Visual–spatial thinking (Gv)	–	–	–	–	101.8	14.9	95.8	16.6	103.4	14.3	97.0	13.4	–	0.40	0.45	

Note: Calculations are based on unrounded and unadjusted scores and, employing sampling weights.

the samples used here differ from those used to establish the norms. Thus the IQ means and standard deviations for the 6–65 samples used here are 100.0 and 14.7 for WJ1, 102.4 and 15.9 for WJ2, and 104.0 and 14.0 for WJ3, even though the norms for all three standardizations were set to a mean of 100 and a standard deviation of 15.

The use of sampling weights mitigates this problem in the multivariate analyses. For the presentation of the raw trendlines, I use two measures, "unadjusted" and "adjusted" scores. The unadjusted scores are just that—the scores in the database provided by Woodcock–Johnson. The adjusted scores renorms those scores to uniform means of 100 and standard deviations of 15 for the analysis sample within each standardization. The adjusted-score plot is thus the one that would be produced if each of the analysis samples were perfectly representative. As the graphs using the two measures will reveal, in practice the differences in results produced by these two measures are so small that they have no effect on the interpretation of the data. Scores reported in the text always refer to unadjusted scores.

3. Results

3.1. Period analysis

Table 2 shows the kind of period analysis of the black–white difference that has been most common in the literature, comparing test results according to the years when the test was administered. In this instance, Table 2 shows the means and B–W differences on overall IQ and the seven cognitive clusters for each of the three Woodcock–Johnson standardizations using the unadjusted scores.

A substantial reduction in the B–W difference in IQ occurred from WJ1 to WJ2, from 1.23σ to 0.90σ, a drop of one third of a standard deviation. The difference increased again to 1.05σ in WJ3, but a net reduction of 0.18σ from WJ1 to WJ3 remains, equivalent to about 3 IQ points.

Turning to the six clusters of cognitive functioning for which measures were available in all three standardizations, a substantial reduction in the B–W difference from WJ1 to WJ3 is observed for Ga (0.47σ), with small reductions for Gsm, Glr, and Gs. The two most highly *g*-loaded clusters, Gc and Gf, showed small increases.

In summary, the period analysis of the black–white difference in the Woodcock–Johnson shows reductions in the B–W difference in overall intelligence and most of its components. These improvements were apparently concentrated in the 1980s.

3.2. Birth cohort analysis of the B–W difference among persons born from 1920 to 1991

Fig. 1 recasts the numbers shown in Table 2 as a birth cohort analysis of overall IQ, using means by birth year.

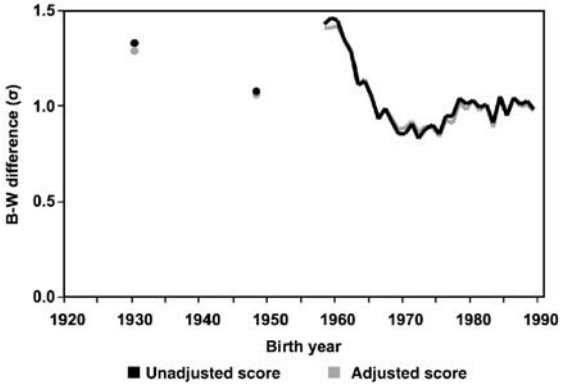

Fig. 1. The B–W difference in IQ by birth year. Note: Line represents results for moving 5-year aggregations. The two dots represent aggregations for birth years 1920–1939 and 1940–1955.

The line shows the B–W difference in standard deviations for 5-year aggregations of scores. The minimum black sample size for any 5-year aggregation is 114 and the median is 213; for whites, the minimum was 612 and the median was 1147. The opening point at 1958 represents the difference for persons born from 1956 through 1960. The most recent point, 1989, represents the difference for persons born from 1987 through 1991. The black line is based on the unadjusted scores and the gray line is based on the adjusted scores. For the birth years prior to 1956, black sample sizes were not large enough to produce interpretable means over short periods of time. The two dots show the B–W difference in standard deviations for persons born from 1920 to 1939 (white $n=429$, black $n=74$) and 1940–55 (white $n=693$, black $n=94$).

The plot is trivially affected by the choice of unadjusted or adjusted scores. The following observations apply to both.

The B–W difference among persons born from 1920 to 1939 was 1.33σ. The difference dropped to 1.08σ for those born from 1940 to 1955. When line begins in 1958, the difference was extremely large, reaching a high of 1.45σ in 1959. The difference dropped steeply throughout the 1960s, reaching its low in 1972, at 0.83σ. For those born most recently, 1987–1991, the difference was 0.98σ.

The corresponding plots are shown for the separate cognitive clusters in Fig. 2. The trends in the cognitive clusters have been widely divergent. The clusters with the highest *g*-loadings are Gc, comprehension-knowledge, and Gf, fluid reasoning. Both of them show a drop in the B–W

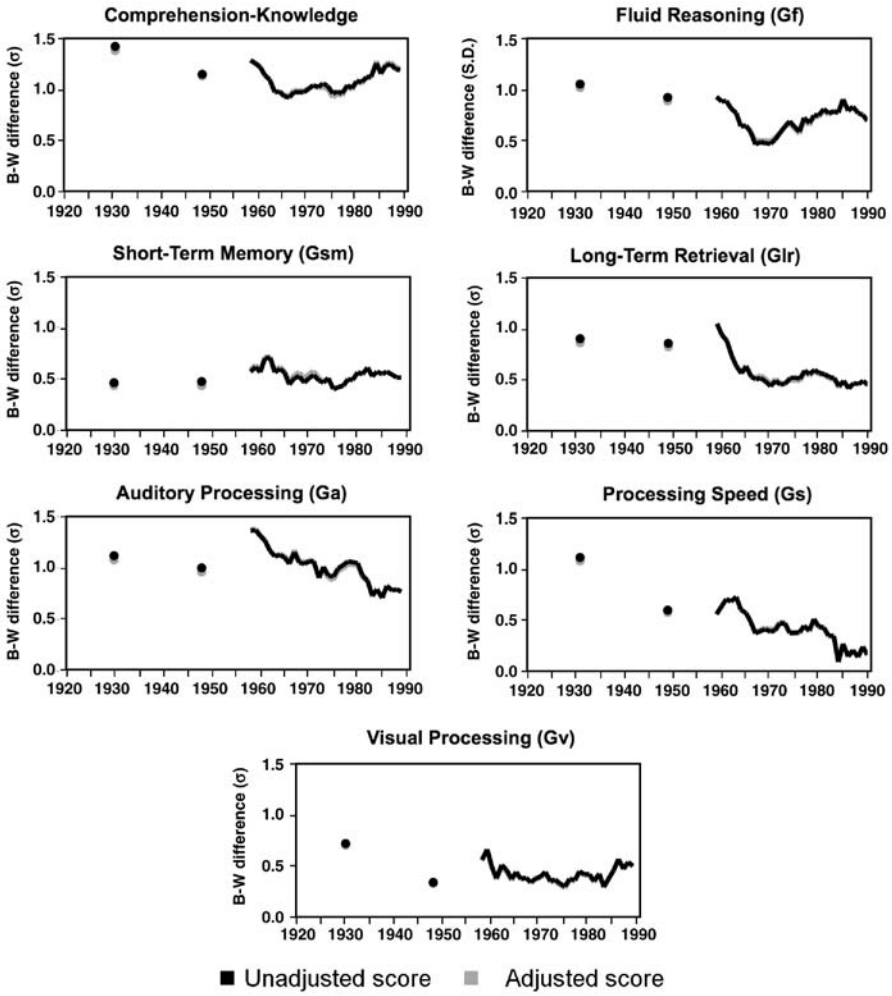

Fig. 2. The B–W difference in the cognitive cluster. Note: Line represents results for moving 5-year aggregations. The two dots represent aggregations for birth years 1920–1939 and 1940–1955.

difference, reaching lows in 1966 of 0.94σ for Gc and 0.57σ for Gf. For those born most recently, 1987–1991, the difference had risen to 1.19σ and 0.71σ respectively.

The largest and most consistent reduction in the B–W difference occurred for processing speed, from 1.10σ for subjects born in 1920–1939 to less than 0.2σ for subjects born in the 1980s. The B–W difference in long-term retrieval and auditory processing has also shown substantial drops since the 1960s.

The B–W difference in short-term memory has been effectively level throughout the entire period of observations. The difference in visual processing has been about the same since those born in 1940–1955.

The trendlines in Figs. 1 and 2 are consistent with the propositions that (1) the B–W difference IQ difference narrowed over the course of the 20th century and that (2) the B–W difference in IQ has done no better than remain unchanged for persons born since the first half of the 1970s. These statements apply not only to overall IQ but to the two most highly *g*-loaded components of IQ, Gc and Gf. A subsidiary issue raised by Figs. 1 and 2 involves the 1950s, when the B–W difference on IQ and all but two of the clusters were higher than they had been from 1940 to 1955.

Controls for the effects of birth date, race, and standardization (which also roughly captures the date of

Table 3
Regressions of IQ, Gc, and Gf on birth date, race, standardization, and the interactions

Independent variable	Model 1			Model 2		
	IQ	Gc	Gf	IQ	Gc	Gf
Birth date, segment 1	0.08 **	−0.01	0.01	0.11	0.01	−0.25
	(0.03)	(0.03)	(0.03)	(0.14)	(0.14)	(0.15)
Birth date, segment 2	0.01	0.05	0.16 *	−0.01	0.07	0.32 **
	(0.05)	(0.08)	(0.08)	(0.07)	(0.11)	(0.11)
Birth date, segment 3	−0.08	0.04	−0.14 **	−0.11	0.01	−0.16 **
	(0.09)	(0.05)	(0.05)	(0.10)	(0.05)	(0.06)
Race (black = 1)	−23.47 ***	−22.07 ***	−18.41 ***	0.46	−11.59	−10.43
	(4.06)	(4.06)	(4.21)	(21.70)	(22.10)	(22.89)
WJ1 dummy	6.31 **	2.80	3.96	5.13	4.71	15.09 **
	(2.22)	(2.21)	(2.30)	(5.11)	(5.51)	(5.71)
WJ3 dummy	12.07 ***	13.28 ***	10.40 ***	3.13	4.28	9.76
	(2.28)	(2.23)	(2.32)	(5.69)	(5.59)	(5.79)
Birth date × WJ1	−0.14 ***	−0.08 *	−0.11 **	−0.13	−0.11	−0.27 **
	(0.03)	(0.04)	(0.04)	(0.08)	(0.08)	(0.09)
Birth date × WJ3	−0.15 ***	−0.16	−0.13 ***	−0.03	−0.04	−0.12
	(0.03)	(0.03)	(0.03)	(0.08)	(0.08)	(0.08)
Race × WJ1	−3.67 ***	−2.48	1.06	−3.86 ***	−2.73 *	1.13
	(1.10)	(1.08)	(1.12)	(1.20)	(1.17)	(1.21)
Race × WJ3	2.46	−0.28	1.92	2.33	−0.47	1.27
	(1.55)	(1.47)	(1.53)	(1.67)	(1.60)	(1.66)
Birth date × race, segment 1	**0.17** *	**0.15**	**0.09**	**−0.25**	**−0.03**	**−0.07**
	(0.08)	(0.08)	(0.09)	(0.38)	(0.39)	(0.40)
Birth date × race, segment 2	**0.13**	**0.04**	**0.57** **	**0.20**	**0.10**	**0.77** ***
	(0.13)	(0.20)	(0.21)	(0.15)	(0.23)	(0.24)
Birth date × race, segment 3	**−0.43** *	**−0.15**	**−0.20**	**−0.47** *	**−0.16**	**−0.20**
	(0.21)	(0.12)	(0.12)	(0.21)	(0.12)	(0.13)
Constant	99.67	104.20	102.41	98.32	103.11	116.68
Observations	9616	9616	9616	8610	8610	8610
R-squared	0.12	0.13	0.06	0.12	0.12	0.06

Notes:
Birth date segment 1 is 1920–1958 for Model 1, 1947–1958 for Model 2.
Birth date segment 2 is 1959–1972 for IQ, 1959–1966 for Gc and Gf in both models.
Birth date segment 3 is 1973–1981 for IQ, 1967–1981 for Gc and Gf in both models.
Standard errors in parentheses.
Birth date is expressed as a fractional value, omitting century (e.g., a birth date of July 1, 1960, is expressed as 60.5).
* Significant at 0.05.
** Significant at 0.01.
*** Significant at 0.001.

testing) clarify these findings. Table 3 shows two models of a regression of test scores on birth date, a dummy variable for race (black = 1), dummy variables for WJ1 and WJ3 (using WJ2 as the reference group), and interaction terms for race × standardization, birth date × standardization, and birth date × race. The variables of chief interest are the interaction terms between birth date and race after the other main effects and interactions have been taken into account. A positive coefficient indicates that black scores rose with birth date, narrowing the B–W difference, while a negative coefficient indicates that the black score dropped with birth date, widening the B–W difference.

Both models employs a linear spline regression (Greene, 2000), also known as a piecewise regression, that permits separate estimates of the slopes for different segments of the range.[2] The knots, corresponding to inflection points, are based on the post-World War II high point of the B–W difference for IQ, Gc, and Gf during the late 1950s (set at 1958 for all three measures) and its low point (1972 for IQ and 1966 for Gc and Gf). Model 1 uses the full range of years in which at least two standardizations are represented, 1921 through the end of 1981. Model 2 is restricted to persons who were under 40 years old when they were tested and to the range of birth years in which more than one standardization is represented (1947–1981), providing an estimate when potential interactions between the B–W difference and advanced test age have effectively been eliminated. Additional analyses were limited to persons tested no younger than age 20, thereby eliminating age effects on the B–W difference through adolescence. They produced results that were substantively indistinguishable from those in models 1 and 2.

Table 3 shows the regression results, with the interactions between race and birth date in boldface. Table 4 summarizes these results in terms of the B–W difference produced by fitted values and compares them with the corresponding scores used to produce Figs. 1 and 2.

Models 1 and 2 imply very different interpretations of changes in the B–W difference on the Woodcock–Johnson. In Model 1, when persons tested at all ages are included and the observations begin in 1921, the B–W difference in IQ, Gc, and Gf all drop through both of the first two segments. The decline in the B–W difference in IQ occurred at a similar slope through both segments. Almost all of Gc's decline occurred in the first segment, 1921–1958, while Gf saw its steepest reduction in the B–W difference in segment 2, from 1959 through 1966. The reduction from the 1920s baseline to the low points as shown in Table 4 were 0.54σ for IQ, 0.42σ for Gc, and 0.54σ for Gf. All three measures saw an increase in the B–W difference from the low point to the end of observations in 1981, but fitted differences as of the end of 1981 were all smaller than the fitted differences as of the beginning of 1921, with net reductions of 0.28σ for IQ, 0.27σ for Gc, and 0.34σ for Gf.

If the Woodcock–Johnson had tested only people under the age of 40 from mid-century (Model 2), the same regression equation provides evidence that the B–W difference in IQ not only failed to narrow, but widened from a fitted values of 0.75σ for those born in 1947 to 1.05σ for those born in 1981. The fitted values show a smaller increase for Gc, from 0.86σ to 0.99σ, and a decrease in the B–W difference for Gf, from 0.90σ to 0.74σ.

Table 4
Changes in the B–W difference on IQ, Gc, and Gf

	B–W difference in standard deviations[a]		
	Unadjusted scores[b]	Fitted Values	
		Model 1	Model 2
IQ			
1921 baseline	1.59	1.33	
1947 baseline	1.15		0.75
Post WW-II high point (1958)	1.42	0.91	0.94
Low point (1972)	0.83	0.79	0.77
End point (1981)	0.97	1.05	1.05
Gc			
1921 baseline	1.47	1.26	
1947 baseline	1.15		0.86
Post WW-II high point (1958)	1.27	0.86	0.88
Low point (1966)	0.94	0.84	0.83
End point (1981)	1.04	0.99	0.99
Gf			
1921 baseline	1.17	1.09	
1947 baseline	0.89		0.90
Post WW-II high point (1958)	0.93	0.86	0.95
Low point (1966)	0.57	0.55	0.54
End point (1981)	0.77	0.75	0.74

Values were fitted the reference group, WJ2, for 1/1/1921, 1/1/1947, 12/31/1958, 12/31/66, 12/31/72, and 12/31/81.

[a] The point difference produced by the fitted values were divided by 15 to produce the estimated difference in standard deviations.

[b] Birth years 1920–1929 for the 1921 baseline, 1940–49 for the 1945 baseline, five-year aggregations for 1958 (1956–1960), 1966 (1964–68), 1972 (1970–74), and 1981 (1979–83).

[2] A linear spline regression allows the slope to change for specified ranges of one of the independent variables. In the present instance, using IQ as the example, it integrates the results that would be produced by running separate regressions for subjects born within each segment. Unlike separate regressions, a linear spline regression thus produces coefficients for the other independent variables that are the same for subjects born across the entire range. Separate regressions grouped by birth date were also examined, but these variations did not add to nor contradict the information conveyed in Table 3. A fully specified model with two knots proved to be oversaturated. The model reported in the text omits the set of birth date × race × standardization interactions.

The choice of interpretation depends on the weight one places on the scores of persons tested at older ages in the 1920s and 1930s, when the B–W difference was extremely large. For the period 1920–39 represented in Fig. 1, the B–W difference for the combined samples was 1.33σ (n=74). But the B–W difference was even larger for those born in the 1920s, standing at 1.59σ (black n=31).

The large difference among those born in the 1920s takes on interest despite the small samples because of the additional data can be brought to bear from the 444 whites and 56 blacks in the Woodcock–Johnson data who were born before 1930 but were older than 65 when tested and thus not included in the sample used for the analyses presented so far. The mean B–W difference in IQ in this group was 1.55σ.

Since the correlation between birth date and test age is −1.0 within standardizations, the question is whether these large differences among persons born in the 1920s and earlier reflect a widening B–W difference among those tested at older ages rather than a difference that was greater in the early 20th century. Given the sample sizes and overlap across standardizations in the Woodcock–Johnson data, it is not possible to model the interaction effect of race × test age × birth date with any of the accepted techniques. Table 5 presents the results of an indirect test for such an effect, showing the birth date × race interaction term for each 10-year test-age group from ages 45–54 through 56–65. Restricting the test-age range to 10 years means that the average absolute test-age difference of subjects within a group is too small (ranging from 2.3 to 2.7 years for the groups in Table 5) for test-age effects to be a plausible explanation for the observed coefficient.

For IQ and all of the components except Gf, the median coefficient for the 10-year age groups is consistent with the trend for the entire 45–65 age range that produced the pre-1950s results shown in Figs. 1 and 2. The results for Gf are not. When the entire 45–65 age range is included in the regression, the interaction of birth date and race is +0.37, corresponding to a 3.7-point increase in the black Gf score per decade, net of the main effects of birth date, race, and standardization. None of the coefficients for the individual age groups is that large, and the median coefficient is slightly negative, at −0.035. A candidate for explaining this discrepancy is an interaction between test age and race.

Collateral evidence casts doubt on this explanation. In the literature, Gs is the component of IQ that has shown the greatest deterioration with age, and the deterioration in Gs has also been shown to explain age-related losses in Gf that are not paralleled in Gc (Zimprich & Martin, 2002; Buehner, Krumm, Ziegler, & Pluecken, 2004). These findings take on significance for a comparison of blacks and whites because high blood pressure, historically more prevalent among blacks than among whites, has been found to be a significant predictor of decline in performance in both Gs and Gf, but not in Gc (Elias et al., 1998). But the median change in Gs for the older Woodcock–Johnson subjects was both small and about the same for the restricted age groups (+0.04) as it was for the entire 45–65 age range (+0.07).

Table 5
The birth date × race interaction term by test age

Test age	Black n	Birth years covered	b for birth date × race interaction							
			IQ	Gc	Gf	Gsm	Glr	Ga	Gs	Gv
Test-age group										
45–54	50	1922–1953	0.309	0.140	0.209	0.328	0.206	−0.064	0.223	−0.064
46–55	49	1921–1952	−0.011	−0.011	−0.205	0.201	−0.040	−0.304	−0.181	−0.304
47–56	52	1920–1950	0.319	0.356	−0.038	0.311	0.237	−0.069	0.144	−0.069
48–57	48	1920–1950	0.008	0.041	−0.326	0.250	0.026	−0.198	−0.149	−0.198
49–58	45	1920–1949	0.239	0.211	−0.178	0.575	0.230	−0.078	−0.032	−0.078
50–59	44	1920–1948	0.363	0.306	−0.031	0.593	0.261	0.001	0.025	0.001
51–60	40	1920–1946	0.135	0.075	−0.128	0.275	0.132	−0.162	−0.134	−0.162
52–61	39	1920–1945	0.217	0.055	0.024	0.241	0.021	0.035	0.061	0.035
53–62	40	1920–1944	0.729	0.438	0.202	0.815	0.388	−0.044	0.465	−0.044
54–63	37	1920–1943	0.588	0.403	−0.065	0.639	0.438	−0.159	0.334	−0.159
55–64	32	1920–1943	0.701	0.314	0.113	0.693	0.495	0.144	0.222	0.144
56–65	28	1920–1941	0.009	−0.293	0.186	−0.025	−0.281	−0.695	−0.265	−0.695
Median			0.274	0.176	−0.035	0.320	0.218	−0.074	0.043	−0.074
All tested at 45–65	85	1920–1953	0.354	0.139	0.366	0.353	0.260	−0.014	0.067	−0.014

Independent variables: birth date, race, standardization, birth date × race. Interaction terms with standardizations were omitted because of sample size considerations.

Thus we have indirect but equivocal evidence that the decline in the B–W difference in Gf from those born in 1920–1950 may have been exaggerated by an underlying interaction between test age and race. This possibility does not affect the validity of the much larger decline in the B–W difference in Gf in the 1960s. Taking the results from all the components into account, any corresponding exaggeration in the decline in the B–W difference in overall IQ from 1920 to 1950 must be extremely small.

4. Discussion

This analysis has used data from the Woodcock–Johnson standardizations to explore a hypothesis for explaining for the disparate findings in the literature on the B–W difference over time: Narrowing in the B–W difference on highly *g*-loaded measures did occur during the 20th century, but the difference stopped narrowing for persons born in the 1970s and thereafter.

The evidence for the first half of the hypothesis is inconclusive. If the evidence is restricted to persons tested under the age of 40 (Model 2), the multivariate analysis provides no support for a narrowing B–W difference in IQ and Gc for persons born from the late 1940s onward, even if the low points of the B–W difference (early 1970s for IQ and the mid-1960s for Gc) are used as the end point for the comparison. A case can be made for a substantial reduction in the B–W difference in Gf if the mid-1960s is used as the end point for the comparison, but not if the end point is extended to the 1980s or early 1990s.

If the evidence includes persons tested through age 65 (Model 1), the interpretation of the evidence rests on the meaning of the B–W difference among persons born in the 1920s and 1930s and tested at older ages. The literature on the effects of adult socioeconomic status and lifestyle variables does not indicate that B–W differences on those variables would produce much widening of the IQ difference among persons tested through age 65. The analysis of restricted age groups in the present article indicates that such effects cannot explain the increases in black scores in IQ and Gc among persons born in the 1920s, 1930s, and 1940s, although they may explain the corresponding increases in Gf during those birth years. I interpret the balance of the evidence as substantiating a B–W difference in IQ and *Gc* that was probably not less than 1.3σ and perhaps higher for people born in the 1920s. If this is correct, the B–W difference in IQ and Gc decreased from the 1920s to the low points in the mid-1960s and early 1970s by magnitudes on the order of half a standard deviation if the fitted results in Table 4 are applied. Even if we discount the evidence on Gf for persons tested at older ages, the B–W difference in Gf decreased substantially from 1948 to 1966 for persons tested at age 40 or younger.

The conclusion that the B–W difference narrowed is countered by the earliest measures of racial differences in IQ, which consist of a large number of studies catalogued in Shuey (1966) showing an average B–W difference of no more than 1σ (Loehlin, Lindzey, & Spuhler, 1975; Gottfredson, 2005) and the Army Alpha and Army Beta tests used during World War I, representing men born around 1900, which showed a B–W difference of 1.16σ (Loehlin, Lindzey, & Spuhler, 1975, based on Yerkes, 1921). If those results are taken at face value, they overwhelm the evidence for a higher B–W difference during that era obtained from the Woodcock–Johnson standardizations.

They cannot be taken at face value, however. At the time of World War I, almost 70% of all blacks still lived in the rural South (Myrdal, 1944), unschooled or very poorly schooled. This population, presumptively with the lowest mean black IQ, is effectively unrepresented in the Shuey studies, and there is reason to believe that it was radically underrepresented among those draftees who reached the point of being administered the Army Alpha and Army Beta tests (Keith, 2004).

The caution with which one must approach the World War I data is accentuated by the data from World War II. The B–W difference on the Army General Classification Test for inductions in 1944–1945 has been put at 1.52σ (Loehlin, Lindzey, & Spuhler, 1975). This represents the scores of men born from 1925 to 1927, and is very close to the 1.59σ difference observed among the Woodcock–Johnson subjects born in the 1920s. How could the B–W difference in IQ have risen from 1.16σ to 1.52σ in 20 years? The simplest explanation is that the World War II testing produced a more accurate nationally representative estimate of the B–W difference than did the World War I testing.

The argument for an unchanging B–W difference in IQ over the course of the 20th century is not refuted. The defenders of the Shuey studies and of the Army Alpha and Beta estimates can mount counter-arguments. They can also call upon the WAIS standardization data, which show a B–W difference of only 0.99σ for the 41 blacks born from 1904 to 1923.[3] The evidence from the Woodcock–Johnson standardizations should be seen as another piece of the puzzle, adding weight to the argument that the B–W difference in the early part of the 20th century was

[3] For this and subsequent statements about the standardizations of the WAIS, the Stanford–Binet and WISC, I am indebted to William Dickens for providing me with the data used in Dickens and Flynn (2006).

substantially larger than has been observed in recent decades, but without settling the issue.

Regarding the second half of the hypothesis explored here, that narrowing in the B–W difference stopped for persons born in the 1970s, the evidence from the Woodcock–Johnson is not only internally consistent across standardizations but consistent with a substantial body of other longitudinal data.

The NAEP, the WAIS standardizations, and the Stanford–Binet adult scale standardizations all show a narrowing of the B–W difference prior to the 1970s and an end to narrowing sometime during the 1970s. Specifically:

The NAEP. The B–W difference in the NAEP was first measured for the cohort born in 1954. On reading, math, and science, the difference at baseline was at least 1.03σ and as high as 1.29σ. Subsequent rounds of the NAEP showed a narrowing B–W difference for cohorts born throughout the 1960s. The narrowest point in the B–W difference through the 2004 administration of the NAEP occurred among cohorts born from 1969 to 1979, varying by test from 0.38σ to 1.04σ (Perie & Moran, 2005).[4]

The WAIS standardizations. The B–W difference in the WAIS was at its widest among cohorts born in the 1920s, at about 1.14σ. It reached its narrowest level, 0.73σ, among cohorts born from 1971 to 1975.

Stanford–Binet adult scale standardizations. The narrowing in the standardizations of the Stanford–Binet adult scale, from 1.11σ to 0.98σ, occurred in the period between cohorts born from 1962 to 1973 and cohorts born from 1978 to 1989.

One source of data, the AFQT standardizations conducted in 1980 and 1997, shows a narrowing of the B–W difference from 1.21σ to 0.97σ among persons born in the 1960s and/or 1970s (author's analysis of NLSY-79 and NLSY-97). The narrowing in the B–W difference occurred sometime after 1964, the last birth year for NLSY-79, and before 1980, the earliest birth year for NLSY-97. There is no way to know what years from 1965 to 1979 account for the bulk of the reduction.

Two sources show an unchanging B–W difference for persons born since the mid-1970s with no information on when the stable difference began:

Children of the NLSY-79 Women. The B–W difference on the Peabody reading recognition, reading, mathematics, and picture vocabulary tests increased slightly for the children of women in the NSLY-79 cohort born from the mid-1970s through the mid-1990s (Murray, 2006).

The Stanford–Binet children's scale standardizations. The B–W difference was effectively flat between cohorts born from 1974 to 1978 and cohorts born from 1990 to 1994 (0.65σ and 0.62σ respectively).

Only two data sources that I have been able to identify are seemingly inconsistent with the Woodcock–Johnson results:

The GSS vocabulary test. GSS data are now available through the 2004 survey, 6 years longer than the observation period available to Huang and Hauser (2001), and they show a continuing decline in the B–W difference for persons born into the early 1980s (author's analysis of the GSS). But if the question is whether black performance on the vocabulary test has improved, there is no inconsistency with the Woodcock–Johnson results. The GSS has an absolute scale of correct answers, from 0 to 10, and the vocabulary items have remained unchanged since the advent of the GSS. The highest black mean score, whether measured in a single birth year or in five-year aggregations, occurred among blacks born in 1945–1949. The decline in the B–W difference in the GSS vocabulary test for persons born since mid-century is entirely attributable to a decline in white performance, not improvement in black performance.

The WISC standardizations. The only known clear contradiction between the Woodcock–Johnson results and another longitudinal data set involves the WISC standardizations. For children tested at ages 6–16, the B–W difference on the WISC narrowed from 1.08σ among cohorts born from 1973 to 1983 to 0.78σ among cohorts born from 1986 to 1996. With that exception, all the available data on changes in the B–W difference across birth cohorts is consistent with the proposition that narrowing in the difference ended no later than the close of the 1970s, with the bulk of the evidence pointing to the first half of the 1970s.

Many other analyses might be applied to the Woodcock–Johnson standardizations. The results presented here generally support arguments (e.g., Rushton & Jensen 2005) that the B–W difference has narrowed least on the most highly *g*-loaded and most highly heritable cognitive clusters, but I have not attempted to explore that proposition using Jensen's method of correlated vectors (Jensen, 1998). The data also lend themselves to Multigroup Confirmatory Factor Analysis, a method used by Wicherts, Dolan, Hessen, Oosterveld, van Baal, Boomsma, & Span (2004) to explore the nature of the Flynn effect and of the nature of B–W differences across test batteries. I leave those analyses to specialists in the relevant techniques. The limited purpose of the present paper has been to describe when, how, and for what kinds of cognitive ability the B–W difference has changed across birth cohorts on the Woodcock–Johnson standardizations.

[4] Information for converting the point differences reported in Perie and Moran (2005) to standard deviations was provided by the National Center for Education Statistics.

If the results include those tested from ages 6 to 65, they indicate a narrowing of the difference by about half a standard deviation in IQ, Gc and Gf, after controlling for standardization and interaction effects, from persons born in the 1920s to persons born through the last half of the 1960s. If the results are limited to those tested before the age of 40, there is evidence of a decline in the difference in Gf of about a third of a standard deviation for those born from the last half of the 1940s through the last half of the 1960s, but no decline in the difference in overall IQ or Gc during the same period. The results consistently point to a B–W difference that has increased slightly on all three measures for persons born after the 1960s.

Acknowledgments

The author wishes to thank the Woodcock–Muñoz Foundation for providing the data and technical manuals for the Woodcock–Johnson standardizations, with special thanks to Kevin McGrew for his detailed responses to many questions. Thanks go as well to William Dickens, Greg Duncan, Kevin Hassett, and reviewers Nathan Brody, Conor Dolan, Earl Hunt, and Jan te Nijenhuis.

Appendix

Three alternative explanations for the results presented in the main text were explored but proved to be invalid. The three are: (1) the secular rise in IQ distorts the results of a birth cohort analysis, (2) the standardizations of the Woodcock–Johnson were too disparate to permit the interpretation of combined data, and (3) racial differences in longitudinal trends in school dropout and incarceration created artifacts in the samples.

The secular rise in IQ during the 20th century

Perhaps the most obvious objection to combining standardizations arises from the secular rise in IQ scores dubbed the Flynn effect (Herrnstein & Murray, 1994), initially identified for Japan in Lynn (1982) and documented as an international phenomenon since at least the first half of the 20th century by James Flynn (Flynn, 1983, 1985, 1987). During the period covered by the three standardizations of the Woodcock–Johnson test, the expected value of the Flynn effect has been about 0.3 points per year. Since the standardizations of the Woodcock–Johnson were about 10 years apart, subjects tested with WJ3 in the last half of the 1990s could be expected to have produced a mean about 3 points higher if they had been tested with the WJ2 battery and about 6 points higher if they had been given the WJ1 battery. Put more formally, a Flynn-corrected score for subjects in WJ2 and WJ3 would be

$$F = S + A(T - 76.75)$$

where F is the test score corrected for the Flynn effect, S is the unadjusted score, A is the annual magnitude of the Flynn effect expressed in points, T is the test date, and 76.75 is the imputed test date for all WJ1 subjects.

If the objective of the present analysis was to trace the true mean IQ of blacks and whites over several decades, the Flynn effect makes the interpretation of trends in IQ across standardizations problematic, but it poses a much simpler problem for the interpretation of group differences. If the Flynn effect has the same impact on two groups, it is a constant added to the scores of both groups and has no effect on the observed difference. The only way that the Flynn effect can mask or artificially create between-group changes is if the distribution of birth dates differs systematically between the groups being compared. This is not an issue with the standardizations of the Woodcock–Johnson, where the differences in those distributions were minor.

If the impact of the Flynn effect is not the same for two groups, the most plausible difference is that the effect is larger at the low end of the IQ range than at the high end. Since blacks are disproportionately represented at the low end, the Flynn effect would have greater impact on black scores than on white scores. There is no evidence for such a differential effect in the United States, but, for purposes of testing the sensitivity of the results, I simulated a large differential effect by calculating a set of Flynn-corrected scores based on $A = 0.6$ for IQs of less than 85, $A = 0.3$ for IQs of 85–114, and $A = 0$ for IQs of 115 and higher. Fig. A1 below shows the results. Even a large disproportionate effect on the lower end of the IQ range does not translate into changes in the B–W difference that affect the interpretations presented here.

The comparability of the Woodcock–Johnson test batteries across standardizations

Despite the consistent theoretical approach brought to the creation of the three Woodcock–Johnson standardizations, it may be argued that the use of different subtests across standardizations creates a presumptive case against trying to compare results across standardizations.

This argument may be tested for overall IQ by restricting the computation of an overall score to the subtests that were identical in rationale and design for all three standardizations. Seven subtests qualify: visual-auditory processing, sound blending, visual matching,

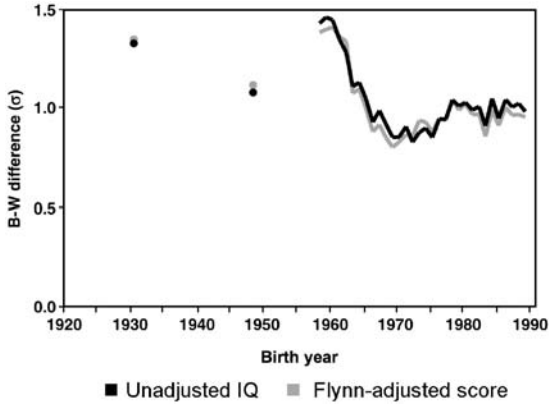

Fig. A1. The B–W difference in IQ incorporating a simulated Flynn effect. Note: Line represents results for moving 5-year aggregations. The two dots represent aggregations for birth years 1920–1939 and 1940–1955.

analysis–synthesis, numbers reversed, picture vocabulary, and concept formation.[5] The correlations of these summed common subtests with the overall IQ measure are 0.95, 0.94, and 0.96 for WJ1, WJ2, and WJ3 respectively. The tests that were unique to each standardization did not affect the comparability of the measure of general intellectual ability. The interpretations of trends in IQ presented above would have been identical if the analysis had been restricted to subtests common to all three standardizations.

For the clusters, an analogous test can be applied: instead of using cluster scores reported by Woodcock–Johnson, proxy cluster scores restricted to subtests that were common across standardizations can be substituted for them. As in the case of overall IQ, these proxies are inferior conceptually the cluster scores discussed in the text—the proxy cluster scores consist of a single subtest instead of at least two—but substituting them for the cluster scores does not change any of the interpretations that have been presented.

Artifacts involving changes in school dropout rates and incarceration

The Woodcock–Johnson standardizations did not seek out school dropouts or test in correctional institutions. Could contemporaneous racial trends in school dropout or incarceration explain the observed changes in scores? The answer is no. As the period analysis (Table 2) showed, the reductions in the B–W difference occurred from WJ1 to WJ2, administered in 1976–1977 and 1986–1988 respectively. During that same decade, black dropout from secondary school dropped from approximately 26% to 17% while white dropout remained unchanged at about 13%. The racial differential, if it had any effect, would have tended to understate the underlying reduction in the B–W differences. From WJ2 to WJ3, administered in 1996–1999, dropout rates remained nearly constant for both whites and blacks (U.S. Bureau of the Census, 2006: Table 268). Regarding imprisonment, the effects of omitting prisoners on any of the standardizations were necessarily trivial. Additions of nationally representative proportions of prisoners would have augmented the Woodcock–Johnson samples by less than 2% for either blacks or whites for any of the standardizations (author's calculations from Bureau of the Census and Department of Justice statistics on general and prison populations).

[5] Gc in WJ3 was measured by the verbal comprehension and general information subtests, but a stand-alone picture vocabulary score was also obtained.

References

Arbuckle, T. Y., Gold, D., & Andres, D. (1986). Cognitive functioning of older people in relation to social and personality variables. *Psychology and Aging*, *1*, 55–62.

Buehner, M., Krumm, S., Ziegler, M., & Pluecken, T. (2004). Cognitive abilities and their interplay: Reasoning, crystallized intelligence, working memory components, and sustained attention. *Journal of Individual Differences*, *27*(2), 57–72.

Busse, E. W. (1969). Theories of aging. In E. W. Busse & W. Schaie (Eds.), *Behaviour and adaptation in later life* (pp. 11–32). Boston: Little, Brown.

Cattell, R. B. (1941). Some theoretical issues in adult intelligence testing. *Psychological Bulletin*, *38*, 592.

Cattell, R. B. (1943). The measurement of adult intelligence. *Psychological Bulletin*, *40*, 153–193.

Cattell, R. B. (1950). Personality: A systematic theoretical and factoral study. New York: McGraw-Hill.

Clarkson-Smith, L., & Hartley, A. A. (1990). Structural equation models of relationships between exercise and cognitive abilities. *Psychology and Aging*, *5*, 437–446.

Dickens, W. T., & Flynn, J. R. (2006). Black Americans reduce the racial IQ gap: Evidence from standardization samples. *Psychological Science*, *17*(10), 913–920.

Elias, M. F., Robbins, M. A., Elias, P. K., & Streeten, D. H. P. (1998). A longitudinal study of blood pressure in relation to performance on the Wechsler Adult Intelligence Scale. *Health Psychology*, *17*(6), 486–493.

Flynn, J. R. (1983). Japanese intelligence: Now the great augmentation of the American IQ. *Nature*, *301*, 655.

Flynn, J. R. (1985). Wechsler intelligence tests: Do we really have a criterion of mental retardation? *American Journal of Mental Deficiency*, *90*, 236–244.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin*, *101*(2), 171–191.

Fryer, R. G., & Levitt, S. D. (2004). Understanding the black–white test score gap in the first two years of school. *Review of Economics and Statistics*, *86*(2), 447–464.

Gold, D. P., Andres, D., Etezadi, J., Arbuckle, T. Y., Schwartzman, A., & Chaikelson, J. (1995). Structural equation model of intellectual

change and continuity in predictors of intelligence in older men. *Psychology and Aging*, *10*(2), 294−303.

Gottfredson, L. S. (2005). Implications of cognitive differences for schooling within diverse societies. In C. L. Frisby & C.R. Reynolds (Eds.), *Comprehensive handbook of multicultural school psychology* (pp. 517−554). New York: Wiley.

Greene, W. H. (2000). Econometric analysis, 4th ed. Upper Saddle River, NJ: Prentice Hall.

Herrnstein, R. J., & Murray, C. (1994). The bell curve: Intelligence and class structure in American life. New York: Free Press.

Hertzog, C., & Schaie, K. W. (1986). Stability and change in adult intelligence: 1. Analysis of longitudinal covariance structures. *Psychology and Aging*, *1*(2), 159−171.

Horn, J. L. (1965). Fluid and crystallized intelligence. Urbana-Champaign: University of Illinois.

Horn, J. L., & Stankov, L. (1982). Auditory and visual factors of intelligence. *Intelligence*, *6*, 165−185.

Huang, M. -H., & Hauser, R. M. (2001). Convergent trends in black–white verbal test score differentials in the U.S.: Period and cohort perspectives. *EurAmerica*, *31*(2), 185−230.

Jensen, A. R. (1998). The *g* factor: The science of mental ability. Westport, CT: Praeger.

Keith, J. (2004). Rich man's war, poor man's fight: Race, class, and power in the rural south during the first World War. Chapel Hill, NC: University of North Carolina Press.

Loehlin, J. C., Lindzey, G., & Spuhler, J. N. (1975). Race differences in intelligence. San Francisco: W.H. Freeman and Company.

Lynn, R. (1982). IQ in Japan and the United States shows a growing disparity. *Nature*, *297*, 222−223.

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

McGrew, K. S., & Woodcock, R. W. (2001). Technical manual. *Woodcock–Johnson III.* Itasca, IL: Riverside Publishing.

McGrew, K. S., Werder, J. K., & Woodcock, R. W. (1991). WJ-R technical manual. Allen, TX: DLM.

Murray, C. (2006). Changes over time in the black–white difference on mental tests: Evidence from the children of the 1979 Cohort of the National Longitudinal Survey of Youth. *Intelligence*, *34*(6), 527−540.

Myrdal, G. (1944). An American dilemma. *The Negro problem and modern democracy.* New York: Harpers.

Nisbett, R. E. (2005). Heredity, environment, and race differences in IQ: A commentary on Rushton and Jensen (2005). *Psychology, Public Policy, and Law*, *11*(2), 302−310.

Perie, M., & Moran, R. (2005). NAEP 2004 trends in academic progress: Three decades of student performance in reading and mathematics. Washington: National Center for Education Statistics.

Rushton, J. P., & Jensen, A. R. (2005). Thirty years of research on race differences in cognitive ability. *Psychology, Public Policy, and Law*, *11*(2), 235−294.

Schaie, K. W. (1996). Intellectual development in adulthood. *The Seattle longitudinal study.* New York: Cambridge University Press.

Schaie, K. W., & Willis, S. L. (1993). Age difference patterns of psychometric intelligence in adulthood: Generalizability within and across ability domains. *Psychology and Aging*, *8*(1), 44−55.

Shuey, A. M. (1966). The testing of Negro intelligence, 2nd ed. New York: Social Science Press.

Taub, G. E., & McGrew, K. S. (2004). Confirmatory analysis of the factor structure and cross-age invariance of the Wechsler Adult Intelligence Scale—Third Edition. *Psychological Assessment*, *16*(1), 85−89.

U.S. Bureau of the Census. (2006). Statistical abstract of the United States 2006. Washington: Government Printing Office.

Wicherts, J. M., Dolan, C. V., Hessen, D. J., Oosterveld, P., van Baal, G. C. M., Boomsma, D. I., et al. (2004). Are intelligence tests measurement invariant over time? Investigating the nature of the Flynn effect. *Intelligence*, *32*, 509−537.

Woodcock, R. W. (1978). Development and standardization of the Woodcock–Johnson psycho-educational battery. Hingham, MA: Teaching Resources Corp.

Yerkes, R. M. (1921). Psychological examining in the United States Army. *Memoirs of the National Academy of Sciences*, 15.

Zimprich, D., & Martin, M. (2002). Can longitudinal changes in processing speed explain longitudinal changes in fluid intelligence? *Psychology and Aging*, *17*, 690−695.