Psychological Bulletin
1984, Vol. 95, No. 1, 29-51

The Mean IQ of Americans: Massive Gains 1932 to 1978

James R. Flynn
Department of Political Studies
University of Otago, Dunedin, New Zealand

This study shows that every Stanford-Binet and Wechsler standardization sample from 1932 to 1978 established norms of a higher standard than its predecessor. The obvious interpretation of this pattern is that representative samples of Americans did better and better on IQ tests over a period of 46 years, the total gain amounting to a rise in mean IQ of 13.8 points. The implications of this finding are developed: The combination of IQ gains and the decline in Scholastic Aptitude Test scores seems almost inexplicable; obsolete norms have acted as an unrecognized confounding variable in hundreds of studies; and IQ gains of this magnitude pose a serious problem of causal explanation.

Virtually since mental tests were introduced, there has been controversy over whether the mean IQ of Americans has been rising or falling. Two tests above all have dominated the American scene, the Stanford-Binet and various forms of the Wechsler, and insofar as these tests are accepted as measures of IQ, that controversy can be settled. Beginning at least as far back as 1932, when the old Stanford-Binet Form L was standardized, Americans have registered massive gains amounting to almost a full standard deviation. In this article I will undertake two tasks: (a) to use Stanford-Binet and Wechsler data as evidence for these gains; and (b) to argue that these gains have far-reaching implications, no matter whether they signal an increase in intelligence or a rise in test sophistication, or even if they are merely an artifact of the tests themselves.

IQ Gains and Standardization Samples

Method

The basic method of measuring IQ gains over time assumes that the standardization samples used to norm IQ tests are representative of Americans in general or more accurately, that each sample was representative of Americans as they were during the years when the sample was selected and tested. Therefore, if Americans made IQ gains over time, this would be reflected in the improved performance of standardization samples, which means that subjects should find early test norms easier to exceed than later ones. In other words, if the same group of subjects is given two IQ tests, one normed in 1932 and the other in 1947, they should score higher on the earlier test. The difference between their mean scores on the two tests serves as a measure of the magnitude of gains, that is, scoring 105 on the earlier test and 100 on the later would suggest a gain of 5 points in 15 years, or a rate of gain of .333 points per year. The reliability of this method depends on the quality of the tests, the degree to which the standardization samples were representative, and accurate measurement of the difference in performance as we go from one test to another.

Even though I have limited myself to Stanford-Binet and Wechsler tests, the problem of accurate measurement of performance on different tests is not a simple one, because of differing conventions of scoring, particularly in making the transition from standardization samples consisting of whites only to samples including minority groups. The results as we go from test to test are standard deviations for whites ranging from 14 to 16 and means for whites ranging from 100.00 to 102.81. This implies that we cannot take the scores reported at face value. Assume we find a study in which subjects took two IQ tests and averaged 88 on Test A and 84 on test B: The first test looks 4 points easier; but if the white mean and *SD* is 102 and 14 on one and 100 and 16 on the other, then appearances are deceiving. Both performances are exactly one standard deviation below the white mean (102 − 14 = 88; 100 − 16 = 84), and the tests are really equal in difficulty.

Therefore, accuracy requires a uniform convention of scoring, and I have chosen the traditional one of translating every score into so many white standard deviations above or below the white mean, assigning the mean and standard deviation values of 100 and 15, and then calculating a new score on that basis. For example, both of the scores mentioned above would translate into 85 (100 − 15 = 85). This uniform convention of scoring required certain standardization data for all tests used and these are presented in Table 1. The table includes both values for means and standard deviations, and the midpoint of the years during which the standardization sample for each test was actually selected and tested.

A few values in Table 1 cannot be used to translate scores reported for certain tests into our uniform scoring

Requests for reprints should be sent to James R. Flynn, Department of Political Studies, University of Otago, Private Bag Box 56, Dunedin, New Zealand.

Table 1
Stanford-Binet and Wechsler Standardization Data

Tests	Acronym	Date		M[a]	SD[a]
		Duration	Midpoint		
Stanford-Binet Form L	SB-L	1931–1933	1932	100.00	16.00[b]
Stanford-Binet Form M	SB-M	1931–1933	1932	100.00	16.00[b]
Stanford-Binet Form L-M	SB-LM	1931–1933	1932	98.00[b]	16.00
Stanford-Binet 1972 Norms	SB-72	1971–1972	1971½	102.81	15.00
Wechsler-Bellevue Form I	WB-I	1935–1938	1936½	100.00	14.83
Wechsler Intelligence Scale for Children	WISC	1947–1948	1947½	100.00	15.00
Wechsler Adult Intelligence Scale	WAIS	1953–1954	1953½	102.26	14.00
Wechsler Preschool and Primary Scale of Intelligence	WPPSI	1963–1966	1964½	102.26	14.00
Wechsler Intelligence Scale for Children—Revised	WISC-R	1971–1973	1972	102.26	14.00
Wechsler Adult Intelligence Scale—Revised	WAIS-R	1976–1980	1978	102.26	14.00

Note. Sources were Kaufman & Doppelt (1976, p. 167); Terman (1942, pp. 2–3); Terman & Merrill (1937, pp. 12–15); Terman & Merrill (1973, pp. 26–28, 64, 339, 353, 359–361); Thorndike (1975, p. 6); Seashore, Wesman, & Doppelt (1950, p. 102); Wechsler (1939, pp. 35–36, 41, 107–110); Wechsler (1949, pp. 4, 7); Wechsler (1955, pp. 3, 6, 10); Wechsler (1967, pp. 5, 13–15); Wechsler (1974, pp. iii, 17–19); Wechsler (1981, pp. 9, 16–19).
[a] Values given are for whites only.
[b] As discussed in the text, if these values are used to translate reported scores into our uniform convention, they will give only approximate results.

convention. To explain this requires a bit of history about the evolution of norms for Stanford-Binet tests. The old Stanford-Binet Form L (the same holds true for the alternate Form M) had norms carefully calculated to be representative of white Americans circa 1932, the midpoint of the actual testing period. These calculations made allowance for three known biases in the standardization sample: that it was elite in terms of parental occupational status, urban–rural balance, and lesser factors affecting a few age groups; for example, the sample did not include enough subjects out of school and over 15 years old (McNemar, 1942, pp. 20, 35–37; Terman, 1942, pp. 7, 12–13; Terman & Merrill, 1937, pp. 14–18). In 1960, Terman and Merrill brought out a new test, the Stanford-Binet LM, but because its content was essentially a selection from the old tests, they did not secure a new standardization sample. However, they did manipulate the norms as follows: They let the allowance for the elite bias of the 1932 sample in terms of occupation stand, and they discarded the allowances for the other elite biases with no reason given (Terman & Merrill, 1973, pp. 26–28). The effect of this was to toughen the norms by 0 to 5 points depending on the subject's age, the overall average being about 2 points for all ages. That is why Table 1 implies that a subject who scores 98 on the Stanford-Binet LM has really matched the 1932 norms, although in fact the necessary score would vary from 95 to 100 depending on age.

This is not to be critical of Terman and Merrill, for their avowed purpose was to update both the content and norms of the Stanford-Binet scale, and by 1960 they had much evidence that representative groups were scoring too high, a finding I would attribute to IQ gains over time (Terman & Merrill, 1973, pp. 21–23, 35–40). Our purpose, however, is to measure gains over time; therefore we want to score against representative norms circa 1932 without any updating. For the old SB-L and SB-M there is little problem: you merely accept the scores at face value with minor adjustments to convert the mental age IQs of those days into the deviation IQs of today. When we encounter scores for the new SB-LM, however, these have to be raised a few points to get back to the old norms. For specialists who wish to check the author's calculations, use the table Terman and Merrill (1973, pp. 339–341) have provided but use it in this way: SB-L and SB-M results, (Score − 100) K + 100; SB-LM results, Score + (Mean − 100). The adjusted scores will all refer to a convention in which the mean for whites circa 1932 was 100 and the standard deviation was 16.

Table 1 gives means for whites only, with not only blacks but other minority groups such as Hispanics, American Indians, and so forth, excluded as well. Again the rationale stems from the nature of the standardization sample used to norm the old Stanford-Binet Form L: As Thorndike (1973) has emphasized, "the black, the Mexican-American and the Puerto Rican-American were not included" (p. 360). Since this sample stands at the very beginning of our 46-year period, the values of all later tests must be comparable to its own, which is to say that they too must be based on whites only. Fortunately, thanks to Kaufman and Doppelt (1976, p. 167), we have exactly the values we need for those Wechsler tests that were normed on samples of all races: They give 102.26 (M for whites) and 14 (SD for whites) for the Wechsler Intelligence Scale for Children–Revised (WISC-R)—values that would be reasonable approximations for the other Wechsler tests because these were normed in precisely the same way.

Table 1 contains one test, the Stanford-Binet 1972 Norms (SB-72), whose mean and standard deviation for whites could not be obtained from any published source and had to be derived as follows. The Stanford-Binet standardization sample of 1971–1972 numbered 2,351. It was selected to be representative (in terms of levels of ability) of the larger sample used to standardize the Cognitive Abilities Test (CAT) the previous year, and therefore it is the CAT sample that is relevant for our purposes. Now the mean of the standardization sample was set at 100

and the standard deviation at 16, but it was a mixed-race sample, and therefore, we know that the mean for whites must have been somewhat higher and the standard deviation somewhat lower: The mean of an all white group will be higher than that of a group including both whites and low-scoring minority members; the standard deviation of an all white group will be lower than that of white and minority subjects together because the minority subjects tend to cluster at a point below the total population mean, thus adding some variance to the lower half of the curve. In fact, we can get a highly accurate estimate of the standard deviation for whites by analogy with other tests: military data gives a standard deviation for whites of 18.8 as compared with a mixed-race 20.0; the Wechsler above gave 14 as compared with 15; this dictates a standard deviation for whites for SB-72 of 15 as compared with the mixed-race value of 16.

In order to calculate the mean for whites, we need data specific to the CAT sample that was the parent sample for subjects used to norm the SB-72. E. C. Drahozal (personal communication, December 23, 1981) has been kind enough to supply estimates for the percentages of various racial and ethnic groups: whites 80.9%, blacks 15.0%, Hispanics 3.0%, American Indians .6%, and Orientals .5%. The Coleman Report as presented by Jensen (1980, p. 479) gives us differences between the means of whites and other groups, which can be expressed in IQ points: white = x; black = $x - 15.66$; Hispanic = $x - 12.79$; American Indian = $x - 10.58$; Oriental = $x - 3.38$. From this it follows algebraically that the mean for whites for the 1971–1972 Stanford-Binet standardization should be set at 102.81, the value given in Table 1.

This brief history of Stanford-Binet norms may seem tedious, but its importance emerges when it is applied to the data that shows IQ gains over time, the best example being the 2,351 subjects used to standardize the SB-72. We can think of this sample as having taken both the SB-72 and the earlier SB-LM, but actually these tests are identical in content; what really happened was that they took the one test and were scored against two sets of norms, the 1932 norms as toughened in 1960 and the 1971½ norms. The published results, when weighted so that each yearly age group from ages 2 to 18 counts the same, shows an average of 105.43 on the early norms and (by definition) 100.00 on the later norms, an apparent gain of 5.43 IQ points (Terman & Merrill, 1973, p. 359). However, we now know both that this is deceptive and how the true gain can be derived: Scoring against white 1932 norms, add 2.12 points as an allowance for the toughening of the 1932 norms by the SB-LM: 105.43 + 2.12 = 107.55; translating into our uniform scoring convention (where $SD = 15$ for whites, rather than 16) gives 107.08. Scoring against white norms, 1971½: subtract 2.81 points because the mean for whites on the SB-72 was really 102.81 rather than 100: 100 − 2.81 = 97.19; because the standard deviation for whites for this test is already 15, no further translation is needed, giving 97.19. We can now calculate the true gain as measured by these data, which comes to 107.08 minus 97.19, or 9.89 IQ points.

The Selection of Data

The method described dictated the following objective: application of our uniform convention of scoring to all data available for every Stanford-Binet and Wechsler test from the SB-L (or SB-M) normed in 1932 to the Wechsler Adult Intelligence Scale-Revised (WAIS-R) normed in 1978. An effort was made to locate every study in which two or more tests were administered to the same group of subjects, the tests having been normed at least 6 years apart. I cannot guarantee that I found every such study, but I did find more studies for each combination of tests than were listed in supposedly exhaustive surveys for each combination, and I would be surprised if more than a dozen studies have escaped the net.

Some studies were excluded on methodological grounds according to the following criteria: (a) Studies were rejected if there was insufficient data to calculate the means on the two tests, the tests had not in fact been given to the same subjects, or the subjects had already appeared in another study used. (b) If there was danger of practice effects the study was rejected; for example, there is a large carry-over of content from one Wechsler test to another, and if all subjects take one test first and then the other, their mean performance on the latter can be inflated by as much as 6 points. (c) Studies were rejected if there was too much time between administrations of the tests; if subjects took one test 2 or more years after the other, there was a danger their IQs had altered in the meantime. Also rejected were studies in which some obvious dramatic event had occurred, for example, if between tests the subjects had gone from an enriched school to a ghetto school. (d) All studies were eliminated whose subjects were limited to the highly gifted (mean above 130 on the more recent norms) or the mentally retarded (mean generally below 75).

The rationale behind the elimination of studies of the gifted and the retarded is that their performance represents only the top and bottom 2% of the population, that their results tend to be atypical, and that there are so many studies of deviant groups in the literature that to include them would both distort the normal distribution of IQ and bias the results. The studies that were accepted cover the full range of IQs and therefore naturally include many of the gifted and the retarded. If these groups are partially underrepresented due to this criterion, it would be reasonable to suppose that my data miss only about 1% at both the top and bottom of the curve. If such studies were included, the results should be given a weight of only 1% and therefore could not possibly affect my estimate of IQ gains for Americans in general.

I should add that my aim was to use as many studies as possible. Unless a study stated that tests were not given in counterbalanced order, it was assumed they were. Even for those not counterbalanced, I chose to ignore the possibility of practice effects when subjects took the Stanford-Binet as one test and some Wechsler test as another: Practice effects here would be small, and at any rate, they tended to cancel out, with the bias operating in some cases on behalf of the earlier test, in other cases on behalf of the later. I was much more likely to accept data when tests were given 1 year apart rather than close together or several years apart: A year seemed enough to avoid practice effects but not enough for the subjects to alter in intelligence. In summary, I doubt that anyone will object to the studies rejected: They are presented in an appendix to this article. If anything, others may wish to be stricter than myself and eliminate some studies used: A check of the more doubtful ones shows that their elimination would leave estimates of IQ gains over time essentially unchanged.

One test combination, studies in which subjects were given both the Wechsler Intelligence Scale for Children (WISC) and its revised version (WISC-R), has been challenged by the hypothesis that practice effects could be greater when the WISC-R is the first test taken than when the order is reversed (Davis, 1977; Tuma, Appelbaum, & Bee, 1978; Wheaton, Vandergriff, & Nelson, 1980). The evidence against this hypothesis is considerable. If going from the WISC-R to the WISC inflates the normal practice effect, then the average practice effect (the average of WISC-R to WISC plus WISC to WISC-R) should be unusually large, and in fact, it is not. Despite the fact that 76% of items on the WISC-R are inherited from the WISC with no real modification (Wechsler, 1974, p. 11), six studies reveal an average practice effect of 4.5 IQ points (Davis, 1977; Klinge, Rodziewicz, & Schwartz, 1976; Larrabee & Holroyd, 1976; Swerdlik, 1978; Tuma et al., 1978; Wheaton et al., 1980). This is small compared with the 6.2 points Karson, Pool, and Freud (1957) found for the Wechsler-Bellevue I and WAIS, two tests with a similar overlap of content, or compared with the 7.1 points Quereshi (1968b) found for WISC to WISC retests in his massive study. Moreover, the six studies listed above show a negative correlation between the magnitude of the practice effect and the magnitude of the IQ gain, a finding supported by the total body of data, which shows no correlation between size of IQ gain and increasing time between administration of the two tests. Increasing time between tests should produce diminishing practice effects, if only slightly. Moreover, the coding subtest, the only subtest that went totally unchanged both in content and administration from the WISC to WISC-R, shows the greatest IQ gains over time (Brooks, 1977; Catron & Catron, 1977; Schwarting, 1976; Solly, 1977; Solway, Fruge, Hays, Cody, & Gryll, 1976; Stokes, Brent, Huddleston, Rozier, & Marrero, 1978; Swerdlik, 1978; Weiner & Kaufman, 1979).

Finally and most significant of all, a number of studies whose research design was calculated to control for differential practice effects all show slightly greater gains than the other studies (Catron & Catron, 1977; Thomas, 1980; Solway et al., 1976). It should be noted that even if a differential practice effect exists, it would have to be very large to make much difference: WISC-R to WISC would have to have twice the practice effect of the reverse order to cut 1.5 points off our 8.4-point estimate of IQ gains from this test combination and three times the reverse effect to cut off 2.25 points. As for our overall estimate of IQ gains, which is of course based on many test combinations, it would be affected by only a miniscule amount.

It is arguable that the attempt to cover all Stanford-Binet and Wechsler tests led to one excess, namely, the inclusion of data based on the Wechsler-Bellevue I. This test was normed on the Wechsler standardization sample of 1935–1938 (1936½) and, as Anastasi (1961) pointed out, this sample was "drawn largely from New York City and its environs" (p. 304). Those familiar with regional differences in IQ would expect such a sample to be an elite, and five studies that compared the Wechsler-Bellevue I with the Stanford-Binet L, normed at much the same time, for subjects aged 8 to 18 years indicated that it was 3 or 4 points too hard for this age group (Anderson et al., 1942; Goldfarb, 1944; Halpern, 1942; Sartain, 1946; Weider, Levi, & Risch, 1943). The problem with a test that was too hard for its day is that it deflates gains over time when serving as the earlier test in a given combination, and inflates gains when serving as the later test. As for the Wechsler-Bellevue II, there is simply no satisfactory account of how or when it was normed, so it was omitted from this study.

Results

The results of applying our uniform convention of scoring to all relevant and reliable data are presented in Table 2, which includes 73 studies and almost 7,500 subjects with ages ranging from 2 to 48 years. The studies have been grouped in terms of 18 combinations of tests, and 17 of these show subjects scoring higher on earlier norms than later, the sole exception being a test combination including the Wechsler-Bellevue I. If we select out the eight combinations with the largest number of subjects, they evidence rates of gain whose consistency is quite remarkable, ranging from .250 points per year to .440 points, with a median of .332. An overall rate of this sort would entail an American IQ gain of over 15 points during the period 1932 to 1978. Given the magnitude of that result, some comments on Table 2 are in order.

As Jensen (1980, pp. 568–570) pointed out, any one combination of tests is suspect as evidence of IQ gains over time. But our 18 combinations of tests rest on eight standardization samples and these samples themselves constitute 15 distinct combinations. No doubt, one or two of these samples had a modest bias in its day, but with so many combinations, this would tend to be self-correcting. For example, assume that the Wechsler Preschool and Primary Intelligence Scale (WPPSI) sample of 1963–1966 was a bit substandard: Although this would mildly inflate gains between the WPPSI and WISC-R, it would also deflate gains between the WISC and WPPSI. It seems incredible that eight standardizations could make sampling errors so patterned as to mimic IQ gains. Note that this pattern of error could arise only if both the Stanford-Binet and Wechsler organizations, working quite independently, made mistakes in tandem.

Thorndike (1975) illustrated the disadvantage of working with a restricted body of data. Focusing on the single test combination of Stanford-Binet LM and Stanford-Binet 1972, he saw a pattern of disproportionate gains by young children (ages 2 to 6 years) extending all the way from 1932 to 1971–1972; this led

Table 2
White Americans: Evidence for IQ Gains 1932 to 1978

Test combination	Dates		Studies	*N*	Means		Gain	Years	Rate	Ages (years)
	Test 1	Test 2			Test 1	Test 2				
1. SB-L & WISC	1932	1947½	17	1,563	107.13	101.64	5.49	15½	.354	5–15
2. SB-M & WISC	1932	1947½	1	46	125.13	107.56	17.57	15½	1.134	5
3. SB-LM & WISC	1932	1947½	6	460	114.64	109.67	4.97	15½	.321	5–15
4. SB-L & WAIS	1932	1953½	3	271	113.02	105.48	7.54	21½	.351	16–32
5. SB-LM & WAIS	1932	1953½	2	79	109.08	101.75	7.33	21½	.341	16–48
6. SB-LM & WPPSI	1932	1964½	8	416	101.74	92.78	8.96	32½	.276	4–6
7. SB-LM & SB-72	1932	1971½	1	2,351	107.08	97.19	9.89	39½	.250	2–18
8. WB-I & WISC	1936½	1947½	2	110	103.51	105.54	−2.03	11	−.185	11–14
9. WB-I & WAIS	1936½	1953½	3	152	122.94	118.25	4.69	17	.276	16–39
10. WISC & WAIS	1947½	1953½	4	436	101.76	99.12	2.64	6	.440	14–17
11. WISC & WPPSI	1947½	1964½	2	108	93.56	90.86	2.70	17	.159	5–6
12. WISC & SB-72	1947½	1971½	1	30	96.40	84.42	11.98	24	.499	6–10
13. WISC & WISC–R	1947½	1972	17	1,042	97.19	88.78	8.41	24½	.343	6–15
14. WAIS & WISC–R	1953½	1972	1	40	102.94	96.29	6.65	18½	.359	16–17
15. WAIS & WAIS–R	1953½	1978	1	72	109.69	101.65	8.04	24½	.328	35–44
16. WPPSI & SB-72	1964½	1971½	1	35	93.06	88.65	4.41	7	.630	4–5
17. WPPSI & WISC–R	1964½	1972	2	140	112.84	108.58	4.26	7½	.568	5–6
18. WISC–R & WAIS–R	1972	1978	1	80	99.61	98.65	0.96	6	.161	16

Note. See Table 1 for full test names. Totals: 73 studies and 7,431 subjects. Age range in years: 2 to 48 (*Mdn* = 10.6). All means are weighted in terms of the number of subjects with the exception of Combinations 1, 3, and 7. These gave age-specific results and therefore were weighted so that each age counted equally. Sources: 1. Arnold & Wagner (1955); Barratt & Baumgarten (1957); Clarke (cited in Pastovic & Guthrie, 1951); Cohen & Collier (1952); Estes, Curtin, De Burger, & Denny (1961); Frandsen & Higginson (1951); French (cited in Zimmerman & Woo-Sam, 1972, p. 242); Holland (1953); Jones (1962, p. 121); Krugman, Justman, Wrightstone, & Krugman (1951, p. 476); Kureth, Muhr, & Weisgerber (1952, p. 282); Levinson (1959); McCoy (cited in Zimmerman & Woo-Sam, 1972, p. 242); Mussen, Dean, & Rosenberg (1952); Pastovic & Guthrie (1951); Rappaport (cited in Pastovic & Guthrie, 1951); Weider, Noller, & Schramm (1951). 2. Triggs & Cartee (1953). 3. Barclay & Carolan (cited in Zimmerman & Woo-Sam, 1972, p. 242); Brittain (1968); Estes (1965); Estes et al. (1961); Oakland, King, White, & Eckman (1971); Sonneman (cited in Zimmerman and Woo-Sam, 1972, p. 242). 4. Bradway & Thompson (1962, pp. 2–3, 13); Giannell & Freeburne (1963, p. 565); Wechsler (1955, p. 21). 5. Kangas & Bradway (1971); McKerracher & Scott (1966). 6. Barclay & Yater (1969); Fagan, Broughton, Allen, Clark, & Emerson (1969); Flynn (1980, pp. 184–185 plus Garber & Heber, 1977, pp. 122 & 125); Oakland et al. (1971); Pasewark, Rardin, & Grice (1971, p. 46); Prosser & Crawford (1971); Rellas (1969); Wechsler (1967, p. 34). 7. Thorndike (1973, p. 359). 8. Knopf, Murfett, & Milstein (1954); Price & Thorne (1955); 9. Karson et al. (1957); Neuringer (1963, p. 758); Rabourn (cited in Wechsler, 1958, p. 116). 10. Hannon & Kicklighter (1970); Quereshi (1968a, p. 77); Quereshi & Miller (1970, p. 108); Simpson (1970). 11. Oakland et al. (1971); Yater, Boyd, & Barclay (1975); 12. Brooks (1977). 13. Appelbaum & Tuma (1977, p. 142); Brooks (1977); Covin (1977); Davis (1977, p. 164); Hartlage & Boone (1977, p. 1285); Klinge et al. (1976, p. 74); Larrabee & Holroyd (1976); Reynolds & Hartlage (1979); Schwarting (1976); Solly (1977); Solway et al. (1976); Stokes et al. (1978); Swerdlik (1978, p. 119); Thomas (1980, p. 440); Tuma et al. (1978, p. 342); Weiner & Kaufman (1979); Wheaton et al. (1980). 14. Wechsler (1974, p. 50). 15. Wechsler (1981, p. 47). 16. Sewell (1977). 17. Rasbury, McCoy, & Perry (1977); Wechsler (1974, p. 49). 18. Wechsler (1981, p. 48).

Table 3
White Americans: Estimates of IQ Gains 1932 to 1978

Samples	*N*	IQ gain	Interval in years
SB, 1932 & W, 1947½	2,069	5.76	15½
SB, 1932 & W, 1953½	350	7.49	21½
SB, 1932 & W, 1964½	416	8.96	32½
SB, 1932 & SB, 1971½	2,351	9.89	39½
W, 1936½ & W, 1947½	110	−2.03	11
W, 1936½ & W, 1953½	152	4.69	17
W, 1947½ & W, 1953½	436	2.64	6
W, 1947½ & W, 1964½	108	2.70	17
W, 1947½ & SB, 1971½	30	11.98	24
W, 1947½ & W, 1972	1,042	8.41	24½
W, 1953½ & W, 1972	40	6.65	18½
W, 1953½ & W, 1978	72	8.04	24½
W, 1964½ & SB, 1971½	35	4.41	7
W, 1964½ & W, 1972	140	4.26	7½
W, 1972 & W, 1978	80	0.96	6
Total			
All samples		84.81	272
Samples with $N \geq 140$		52.10	164
Samples with $N \geq 400$		35.66	118

Note. SB = Stanford-Binet; W = Wechsler. Rates in IQ points per year (IQ gain divided by intervals) were, for all samples, .312; for samples with $N \geq 140$, .318; for samples with $N \geq 400$, .302. The samples amalgamate all test combinations normed on the same combination of standardization samples; see Table 2 for the test combinations, those having identical dates were the ones amalgamated.

him to suggest that IQ gains "are experienced primarily, perhaps even exclusively, in the preschool years" (p. 6). However, the full body of data in Table 2 presents a very different picture. Analysis of those test combinations that measure gains from the Stanford-Binet standardization sample of 1932 to the WISC sample of 1947–1948 shows that the disproportionate gains of young children were totally present by 1947. After that date they simply disappear, for example, the test combination of WISC to WISC-R, measuring gains for all ages of children from 1947–1948 to 1972, shows a pattern of almost complete uniformity. Indeed, all of the post-1947 data suggest uniformity; the reader need only turn to Table 2 and compare test combinations that include young children, older children, and adults to see that IQ gains have not been age-specific.

There are two possibilities: either age-specific gains occurred before 1947 and terminated abruptly at that date; or even those early gains were uniform and age-specific results are artifacts of irregularities in the Stanford-Binet sample of 1932, the earliest and most primitive of our samples. The latter conclusion is suggested by the fact that the only test combinations providing age-specific results have that sample as their starting point. At any rate, the dominant trend for our whole period 1932 to 1978 is one of IQ gains for all ages from 2 to 48 years.

Thus far we have used Table 2 to derive a very rough estimate of the rate of IQ gains prevalent in America since 1932, an estimate that suggested an overall rate of at least .300 points per year. In order to get a more accurate estimation, I have merged the data in a variety of ways, weighting for number of subjects, length of period covered, ages covered, and combinations of these. However, for the sake of economy, I wish to suggest a simple method based on three simplifying assumptions. The first assumption is that the rate of gain has been fairly uniform for all ages, which requires no further comment. The second is that rates have been fairly uniform year-by-year all the way from 1932 to 1978. Since I will examine this question in detail a few pages hence, for now the reader need merely divide Table 2 into test combinations before and after 1947, or before and after 1953, to be struck by the similarity of the rates. The last assumption is that rates have been fairly uniform for subjects whose IQs (on the more recent test) are above and below the population mean of 100, and once again, I invite the reader to divide the data along those lines.

This puts us in a position to understand Table 3, which generates estimates of the IQ gains of Americans in general since 1932. The 18 test combinations have been collapsed into 15 combinations of standardization samples because a number of tests have norms based on a common standardization sample, namely, the SB-L, SB-M, and SB-LM. Given our assumptions, each sample combination gives equally valid information about part of the 46 years we want to measure (1932 to 1978), periods ranging from only 6 years to as much as 39½ years. Therefore, the proper method of deriving an estimate for the whole 46 years is to weight the rates of various combinations in terms of the length of their periods. The mathematical method for doing this is to total all gains and divide by the total of all periods:

Table 3 does this and gives a rate of gain of .312 IQ points per year, very close to the rough estimate.

The number of subjects needed to measure the gap between the norms set by a given combination of standardization samples is important up to a point. Ideally, we would have at least 400 subjects who had taken each combination of tests so as to reduce measurement error to a minimum; beyond that numbers are of diminishing importance. At present, however, the only way to reduce the influence of combinations with small numbers, on the assumption that their measurement error might be great, is by selective elimination. That is why Table 3 offers estimates based on combinations with at least 140 subjects and combinations with at least 400. The fact that our various estimates differ so little from one another is gratifying, and because the estimate based on 400 subjects or more seems best, I will use it throughout this article, rounding off .302 points per year to .300 for ease of computation.

Our data suggests not only a rate of gain for the whole of 1932 to 1978 but also differential rates for shorter periods within those years. This has relevance for testing one variant of the "threshold" hypothesis, the notion that increased environmental quality beyond a certain point will yield diminishing IQ gains: It may be important to be raised in a home with some rather than no books, but less important that there are a thousand rather than a hundred books. Perhaps the United States during a period of prosperity has been approaching such a threshold, and if so, we would expect not a constant rate of IQ gain over our whole period but a falling off.

Table 4 provides estimates of rates of gain for both 1932–1948 and 1948–1972, and as the reader can see, the rates are very similar indeed. When the data are divided up among these shorter periods, each period possesses too few sample combinations to allow for our usual method of calculating rates of gain. Therefore, I adopted the expedient of taking all the combinations whose dates put them mainly into a certain period, merging the gains of their subjects, and thus treating them as if they were one large sample combination. This method is crude but it has the advantage of reducing the influence of combinations with small numbers of subjects. The usual method produces exactly the same pattern of results, but it has the disadvantage of generating less plausible values for the shorter periods. In calculating these rates, I tried to give the threshold hypothesis every possible chance by deriving a maximum rate of gain for the earlier years of 1932 to 1948. This meant excluding the Wechsler-Bellevue I data, which would have pulled the rate down and which are of dubious value, as we have seen.

Table 4
White Americans: IQ Gains Selected Periods

Period	Rate[a]	No. of studies	*N*	Data used[b]
1932–1948	.368	29	2,419	1–3, 4[c], 5[c]
1948–1972	.353	29	1,903	10–14, 15[c], 16, 17
1932–1948	.368	29	2,419	1–3, 4[c], 5[c]
1948–1960	.347	6	544	10, 11[c]
1960–1972	.359			Derived from 1948–1960 and 1948–1972

[a] IQ points per year.
[b] From Table 2. [c]Prorated.

Table 4 also attempts to divide our data among three short periods of approximately equal length. The rate of gain for 1932–1948 of course remains the same, and there were enough subjects whose test combinations matched the period 1948–1960 to get an estimate for those years. As for 1960–1972, there were plenty of subjects whose terminal date extended to 1972, but their initial date tended to be well before 1960. Therefore, the rate for that period was calculated from those for both 1948–1960 and 1948–1972 under the assumption that they were accurate. The end result offers differential rates for the early, middle, and late segments of the 40 years between 1932 and 1972, and once again the constancy of the rate of IQ gains over all those years is most striking.

As for the situation since 1972, we must look right back to Table 2, where we find two studies that extend to 1978, both of which indicate gains of some sort: Test Combination 15 shows a rate of .328 points per year, which is close to our overall rate, and Test Combination 18 gives .161 points, which is somewhat

lower. However, the numbers of subjects, 72 and 80, respectively, are too low to allow for a reliable estimate of IQ gains over the last decade. Even if the rate has begun to diminish, interpretation of the significance of such a phenomenon would be complicated by the fact that there is a serious case for a deteriorating environment during the 1970s, for example, those years witnessed the decline of the nuclear family and a growth in the number of children living in single-parent homes. The safest conclusion about the threshold hypothesis is a negative one: As yet, there is no evidence that further environmental enrichment for Americans in general will yield diminishing returns in regard to IQ.

Implications

IQ Gains and Levels of Reality

The magnitude of our estimate, that Americans gained 13.8 IQ points from 1932 to 1978 (.300 × 46 years), is sure to provoke a variety of opinions about the reality of such gains. I hope to convince others that the gains are real, but I wish to place equal emphasis on something else: convincing skeptics that they must not dismiss the data as unimportant because surprisingly enough, the implications are almost as great whether gains are real or simply an artifact of sampling error. As to levels of reality, thus far we have noted two possibilities: that Wechsler and Stanford-Binet standardization samples have been reasonably representative of Americans in general and therefore, the fact that each sample has set a higher standard than its predecessor signals real IQ gains; or that a series of mistakes produced standardization samples that just happened to err on the side of being more and more elite over time and therefore, these mistakes have mimicked IQ gains. Eventually, I will add a third possibility, that IQ gains are semireal, and the discussion of the implications of the data will attempt to cover all three of these possibilities.

IQ Gains and the SAT Score Decline

Between 1963 and 1981 (there was a slight upturn in 1982), American high school students who took the Scholastic Aptitude Test (SAT) showed a sharp decline in their average performance, particularly on the SAT-Verbal, the test most significant as a predictor of college grades. If American IQ gains are real, and if they extend into this period, then the SAT-V score decline becomes almost inexplicable and, insofar as we attempt an explanation, suggests societal trends of the most alarming sort. Let us explore the following propositions: that the years 1963 to 1981 saw either no net loss in IQ or steady gains of up to .300 points per year and that IQ accounts for about 64% of variance on the SAT-Verbal.

We can divide the years of the SAT score decline into two equal periods, 1963 to 1972 and 1972 to 1981. Our Stanford-Binet and Wechsler data show that IQ gains persisted until about 1972, but after that the evidence is fragmentary. As we have seen, Table 2 reveals one study of 72 adults with a high rate of gain and another study of eighty 16-year-olds with a lesser rate. Table 4 is more helpful in that it shows that the rate of gain was virtually constant for decades before 1972, with no tendency to diminish as that year was approached. Now it is logically possible that some environmental deterioration cuts through the year 1972 and we suddenly go from gains of .300 points per year to losses at that rate, but such a turnabout seems unlikely. For example, SAT scores did not suddenly go from gains to losses but rather after a long period of stability, began a gradual decline that then gained momentum. It seems much more likely that the IQ gains that persisted for so many years before 1972 have continued either at the same rate or at a diminishing rate. Nevertheless, I will offer two possibilities as limiting estimates of IQ trends from 1963 to 1981. One is the possibility of 9 years of gain being balanced by 9 years of loss, which would result in nil gain overall and will be called our *safe* estimate. The other is the possibility of gains at .300 points per year throughout, which would result in an overall gain of 5.4 IQ points (.300 × 18 years), or .360 standard deviation units and will be called our *speculative* estimate.

That IQ accounts for 64% of SAT-V variance assumes that the Stanford-Binet and the various Wechsler tests correlate with the SAT-V at about .80 (correlation squared equals percentage of variance explained). Incredible as it seems, there appear to be no correlational studies in the literature, but a good case can be made for a value of .80 (assuming correction for restriction of range) on the following

grounds: Jensen (1981, pp. 29–30) points out that IQ tests correlate from .50 to .80 with scholastic achievement tests; scholastic aptitude tests are closer to IQ than achievement tests, and indeed, Jensen argues that IQ and aptitude tests measure general intelligence to about the same degree and are functionally more or less equivalent.

This brings us to the SAT-V score decline: Between 1963 and 1981 the average score fell from 478 to 424 for an apparent loss of 54 points. We cannot, however, assume that such a loss among the candidate sample indicates a similar loss among the larger population. The 1.5 million high school students who take the SAT represent about one-third of the total number of 18-year-olds in America, but of course they are a scholastic elite. Moreover, the examinees went from a group in which low-scoring minorities were underrepresented in 1963 to the representative group of today. When the College Entrance Examination Board established an advisory panel to analyze the score decline, the latter estimated that about half of the score drop had to do with the broadening of the candidate sample, the other half reflecting a downward trend in the general population itself. Although the panel did not disclose its evidence, they said their estimate rests on a "relatively firm statistical basis" and they do cite two compelling facts: from 1963 to 1970 the score decline was accompanied by a broadening candidate sample, but from 1970 to 1977 the score decline was even worse despite the fact that changes in the candidate sample were insignificant. They also note that since 1970, the score decline has shown up within all categories of SAT takers, within students with good high school grades as well as bad, within offspring of high-income families as well as low-income, and within whites as well as blacks (Wirtz, 1977, pp. 18–24, 46).

If the advisory panel is correct, the candidate sample loss of 54 points translates into a general population loss of 27 points. However, both of these must be raised thanks to the work of Modu and Stern (1977, p. 1), who found that despite efforts to equate all versions of the SAT for difficulty over the years, there has been an "upward scale drift." The 1973 test was easier than the 1963 test by 8 to 13 points, and thus the SAT-V score decline is actually greater than it seems. I have chosen the mid-point of their estimates and have added 10.5 points on to the apparent losses to get a real decline from 1963 to 1981 of 64.5 points for the candidate sample and 37.5 points for the general population.

Table 5
Scholastic Aptitude Test–Verbal: Data and Trends

Date	*M*	*SD*	No. of candidates
1941	500	100	19,247
1952	476		81,646
1963	478	109	924,833
1970	460	110	1,610,800
1977	429	110	1,425,000
1981	424	110	1,600,000

Note. Candidate loss 1963–1981: points, 54; real points, 64.5; *SDU*s, .645. Population loss 1963–1981: points, 27; real points, 37.5; *SDU*s, .288. Real points are the apparent losses adjusted for the declining difficulty of the SAT-Verbal which amounted to 10.5 points 1963–1973. The candidate loss in standard deviation units is based on the candidate *SD* of 1941, i.e. 100; the population loss is based on the estimated *SD* for the general population of 130. Sources: Educational Testing Service (ETS, 1977, p. 14; 1981, p. 12); Jackson (1976, p. 2); *Memorandum for Mrs. Sharp* (undated, pp. 1–2); Modu & Stern (1977, p. 1); Wirtz (1977, p. 6).

To state the drop as so many points means little; it must be put in terms of standard deviation units. The standard deviation of the SAT-V was set at 100 back in 1941 when the test was standardized, which gives a drop of .645 standard deviation units (*SDU*; 64.5 ÷ 100) for the candidate sample. The standard deviation for the general population would be considerably greater: the 1941 standardization sample was an elite group, and as for all elite groups the variance would suffer from restriction of range, for example, note that during recent years when the candidate sample has become more representative, the standard deviation has increased to 110. I decided to get at least a rough estimate of variance for the general population by using WISC-R data: several studies of elite samples (mean IQ = 120) revealed a ratio of 10 to 13 between their standard deviations and that of the general population. This suggests a population standard deviation for scholastic aptitude of 130, and that value gives a drop of .288 *SDU* (37.5 ÷ 130) as applicable to American 18-year-olds in general. All of the relevant SAT-V data including the author's adjustments are summarized in Table 5.

The advisory panel went beyond the facts of the SAT-V score decline to discuss possible causes, which meant listing the personal traits that contribute to scholastic aptitude: intellectual ability, motivation, study habits, self-discipline, and acquired verbal and writing skills. They recognized that these personal traits could only be the proximate cause of the test performance decline, with the ultimate causes being such things as less demanding textbooks, less demanding school standards in general, rates of student absenteeism commonly running above 15%, the erosion of the nuclear family, and the advent of television. I am going to take the liberty of identifying the main ability component that goes into scholastic aptitude as IQ and will describe as non-IQ factors the rest of the personal traits the panel lists.

We now have everything we need to analyze the combination of IQ trends 1963 to 1981 and SAT-V trends during the same period. To recapitulate, there were IQ gains somewhere between nil (safe estimate) and .360 *SDU* (speculative estimate); SAT-V losses were about .288 *SDU*; and SAT-V variance was 64% due to IQ and 36% due to non-IQ personal traits. To anticipate: these values entail a decline in non-IQ personal traits, motivation, self-discipline, and so forth, from 1963 to 1981 of such magnitude as to constitute a national disaster. The first step is to calculate how much a one-standard-deviation drop in non-IQ traits would lower SAT-V scores. Assuming these traits have a roughly normal distribution, that means taking the square root of .36, the non-IQ portion of SAT-V variance, which gives .600 *SDU*. The next step is to calculate what drop in non-IQ traits has occurred assuming nil IQ gain during this period, which is simply .288 (the SAT loss) ÷ .600 (the SAT loss per *SD* of non-IQ drop) and gives .480 standard deviation units as the non-IQ decline. The final step is to take the possibility of steady IQ gains during this period into account, for naturally such gains would tend to boost SAT-V test performance and thus have to be overwhelmed by an even greater deterioration in non-IQ personal traits. Just to make this clear: an IQ gain of .360 *SDU* would tend to boost SAT performance by, coincidentally, .288 *SDU* (the square root of .64 × .360 = .288); yet what we have is an SAT loss of .288 *SDU*; thus the total SAT-V loss to be explained by non-IQ factors amounts to .576 *SDU*. When we divide this by .600, we get .960 standard deviation units as the non-IQ decline.

Our analysis dictates the conclusion that American 18-year-olds have deteriorated .480 to .960 *SDU* in terms of a total package consisting of motivation, study habits, self-discipline, and acquired verbal and writing skills. Which is to say that only the upper 17% to 32% of today's 18-year-olds can match the upper half of young people as recently as 1963! The calculations above should not be taken literally of course: They are merely meant to show that if both IQ gains and SAT losses are taken to be real, rather than artifacts of sampling error, then the deterioration of non-IQ personal traits among young Americans must have been very great.

I say this because the calculations above can be fairly described as oversimplified; for example, they assume that IQ gains and non-IQ losses affect scholastic aptitude in a simple additive fashion. When we go beyond the personal traits—intelligence, motivation, and so on—that are the proximate causes of SAT test performance and look at the ultimate causes, surely these will interact in a complex and nonadditive fashion. But it is precisely at this point that one's head begins to spin: do less demanding textbooks and low-level TV programs raise intelligence while lowering verbal skills; do declining standards in schools sharpen the mind while undermining study habits; does student absenteeism mean students are engaged in mentally demanding tasks while missing out on knowledge; does a demoralized family environment boost IQ while lowering motivation? Going beyond simple models to speculate about ultimate causes makes no sense whatsoever of the trends in question.

In sum, the combination of IQ gains and SAT-V losses carries an unpalatable implication and problems of causal explanation of a baffling nature. Given this, we must take the possibility that IQ gains are not real more seriously: after all, if IQ gains are a mere artifact of sampling error, they tell us nothing about intelligence; and if the intelligence of young Americans began to decline in 1963, there is no mystery as to why SAT-V performance declined. Let us then grant at least the possibility

that IQ gains are not real and see whether or not this possibility robs our mass of Stanford-Binet and Wechsler data of all of its significance.

IQ Gains and Confounding Variables

IQ gains produce obsolete norms and obsolete norms have acted as unrecognized confounding variables in literally hundreds of studies, misleading researchers about the nature and significance of their results. This is true no matter whether IQ gains are real or not. After all, standardization samples have set higher and higher norms over time: This fact remains whether the cause is a series of representative samples reflecting genuine IQ gains or a series of unrepresentative samples growing steadily more elite because of sampling errors. Thus, for whatever reason, if two Stanford-Binet or Wechsler tests were normed at different times, the later test can easily be 5 or 10 points more difficult than the earlier, and any researcher who has assumed the tests were of equivalent difficulty will have gone astray.

In analyzing the results of such research, begging the question of the reality of IQ gains merely means using certain terms with care: Obsolete norms are simply ones that are earlier and easier than later norms; current norms are simply those attached to a Stanford-Binet or Wechsler test that happened to be standardized at about the time the researcher in question actually tested subjects. In adjusting the scores of subjects to compensate for their having taken tests of unequal difficulty, I will use current norms as the point of reference. Assuming total unreliability of standardization samples, they will do as well as any other, and if we assume some sort of progress, the more recent norms would have at least some advantage in terms of reliability. Table 6 details how many points we should subtract from an earlier (and easier) test to make its scores equivalent with those from a later (and more difficult) test.

This table requires several words of caution. First the allowances to be subtracted have been calculated on the basis of our uniform scoring convention (white $M = 100$, $SD = 15$), which means that researchers must first use Table 1 to translate all test scores into that convention and only then use Table 6. Second, simple allowances of this sort are sometimes reliable only for scores in the normal range of 90 to 110: I know of no pair of tests in this table where high and low scores require special adjustments, but such pairs do exist (Hannon & Kicklighter, 1970). The WISC and WISC-R may appear to be such a pair but once again, if scores are first translated into our uniform scoring convention, the 8-point deduction from the WISC seems to hold throughout the scale except perhaps at the very extremes. Third, if a researcher's subjects are all of a certain age, say 10 years old, the allowance required may be different than the average for all ages. Finally, the number of subjects who took both tests and thus provided the basis for the allowance ranges from about 100 to 400 for most combinations, a situation that must be remedied by further studies.

Table 6
Allowing for Obsolete Norms

Tests	Allowance
SB-LM & WPPSI	Subtract 9 points from SB-LM
WISC & WPPSI	Subtract 3 points from WISC
WISC & WISC-R	Subtract 8 points from WISC
WPPSI & WISC-R	Subtract 4 points from WPPSI
WPPSI & SB-72	Subtract 4 points from WPPSI
SB-72 & WISC-R	None

Note. See Table 1 for full test names. Use this table only after using Table 1 to translate all test scores into our uniform scoring convention. For SB-LM results, ignore the value 98 for white mean in Table 1: rather take them as scored against a white mean and *SD* of 100 and 16 and translate into our convention of 100 and 15. See text for other cautionary notes. Sources: Table 2 and Wechsler (1974, pp. 51–52).

Despite all of these qualifications, using both Table 1 (for translation into uniform scoring convention) and Table 6 (for its allowances) to compare performances on different tests is far better than present practice: accepting scores at face value as if the tests were comparable and equivalent in difficulty. To demonstrate this, from many available candidates, I have selected a few studies for analysis: the Milwaukee Project because of its great notoriety and because of growing frustration about replicating its results; and three other studies that show how scholars who thought they were measuring the predictive value of tests, or the effects of modes of administering tests, or cul-

Table 7
Heber's Experimental Group: Mean IQ Performances with Increasing Age

	Age (years)				
Test	2–3	4–6	7–9	10–11	12–14
	Results as presented				
SB-LM	122	121			
WPPSI		111			
WISC			103	104	100
	Results adjusted				
SB-LM	108	107			
WPPSI		105			
WISC			95	96	92

Note. See Table 1 for full test names. Adjusted results scored against WISC-R norms, sample tested 1972, translated into $M = 100$ and $SD = 15$ for whites. Sources: Garber (1982, May, pp. 12–13, 18, 33a; personal communication, December 14, 1982), Heber & Garber (1975, p. 40).

tural bias, were really measuring the rate of obsolescence.

Heber and the Milwaukee Project. Heber and Garber selected 20 children whose risk of mental retardation, based on all known indices, was 16 times greater than the average, indeed, children who promised to have an eventual mean IQ of about 70. Who can forget the great days of the early 1970s when the first reports emerged: that from ages 2 to 4, the experimental children were not merely normal but superior, that their mean IQ on the Stanford-Binet was above 120. The news spread beyond America to the whole English-speaking world: When A. D. Clarke delivered the third Fred Esher lecture, he gave 25% of his time to Heber and, despite words of caution, enthused over "the remarkable acceleration of development" of the children of the experimental group (Clarke, 1973, p. 15). Heber's results quickly found their way into the textbooks: Mussen, Conger, and Kagan (1974) was typical with its references to this "exciting" study, its "most encouraging" results, its "impressive" findings—not the usual language of a text.

Table 7 shows why the early excitement soon became mixed with bewilderment and some consternation. If readers look at the Results as presented data, which represent the scores actually released, they will see the early Stanford-Binet scores that ranged 120 and above; but soon there came WPPSI scores that seemed to show the children had fallen to a mean IQ of about 111 at ages 4 to 6. Or had they fallen? Their Stanford-Binet performance remained high; that is, the children maintained a mean of 121 on the SB-LM while at the same time performing 10 points lower on the WPPSI. Then the WISC results began to arrive and at age 7, the experimental group fell from 111 to 103, a drop of 8 points in a single year. Oddly enough, this new loss caused less concern, perhaps because it seemed less spectacular than the ground lost going from the early Stanford-Binet scores to the WPPSI. However, it occurred just when the children left the enrichment program and entered public schools, all but one entering innercity Milwaukee schools that have low levels of academic performance.

We will never make sense of Heber's experimental group until we have an accurate picture of their IQ test performance over the last 14 years. I ask readers to return to Table 7, assume for the moment that the Results as presented are deceptive and the Results adjusted are reliable, and consider the difference between the two. Heber's reported results have dominated the minds of scholars: a total loss of over 20 points (Jensen, 1981, p. 187), half lost while the children were still in the enrichment program, about a third when they left, a few points during their seven years of public school. The adjusted results tell a very different story: These children never did have a mean IQ above the normal range and their performance was essentially stable just so long as they remained in Heber's program; indeed, the only loss we can be certain about occurred when they left the program, and it was even larger than we thought, amounting to a full 10 IQ points! In other words, we have no good reason to believe that the experimental group ever had a mean IQ above 105, we have excellent reason to believe they suffered a massive 10-point loss down to 95 the year they entered public schools, and since then they have lost little ground if any. With the illusion of the lost high IQs dispelled, with the conflict between the SB-LM and WPPSI results reconciled, we can focus on the only real problem: How could the transition from the program to public schools have been so devastating?

Heber and Garber made certain observations at the time and now that the children are 14, it may be too late to add much. They emphasize both problems of adjustment to innercity schools, one little girl of high intellectual ability refused to speak at all during the first 2 months, and factors signaling a deteriorating home environment, increased father absence, increased economic distress, and so forth (Garber & Heber, 1977, pp. 125–126; Heber, Garber, Hoffman, and Harrington, 1977). The effects of deterioration at home must have been gradual, which leaves the shock of entry into innercity schools as the obvious dramatic event. Thanks to the adjusted scores, we now know that the 10-point IQ loss was real, rather than an artifact of tests of unequal difficulty, and can hope that those conducting similar experiments will prepare their subjects for such a transition.

As to why Heber and Garber's results as reported were deceptive, we already know the answer: The high scores on the SB-LM were inflated by about 30 years of obsolescence because although the 1932 norms were updated a bit for the SB-LM, this did not amount to much. The WPPSI scores were inflated by 8 years of obsolescence, what with norms dating from 1963 to 1966; and the WISC scores were inflated by some 25 years of obsolescence because the test was normed in 1947–1948. The adjusted scores, on the other hand, are the result of using Table 1 to translate Heber's scores into our uniform scoring convention and then using Table 6 to allow for obsolete norms. As for a test that would serve as a measure of current norms at the time the testing was done, the WISC-R was the obvious choice: It was normed in 1972 and the subjects were tested from 1969 to 1982.

The most important reduction of Heber's scores was from the SB-LM results: It dispelled the myth of superior IQs lost. Since it amounts to a full 14 points, some detail is in order, particularly since we have no studies that yield a direct comparison between the SB-LM and the WISC-R. First, translating SB-LM scores into our uniform scoring convention reduces them by just over 1 point, which leaves almost 13 points to be explained. Second, I used Table 6 to equate the SB-LM with WPPSI scores and then equated these in turn with the WISC-R: This gave 9 plus 4 equals 13 points. As a check on the second step, I used data that compares the SB-LM with the WISC and that in turn with the WISC-R, which gave 12 points, very close to our estimate of 13. The high scores of the experimental group were no real index of what their performance would have been in terms of current norms.

The adjusted scores for the mean IQ of Heber's children, 105 at ages 4 to 6 years, 95 at age 7 and thereafter, must be taken into account in assessing attempts to replicate his results. Naturally these studies will use tests with relatively current norms, and if the myth of IQs of 120 plus persists, they will all appear unsuccessful. Ann Clarke (1981) refers to the failure "of a quite close attempt to replicate Heber's Milwaukee study" (p. 324); we can only hope that the definitions of success and failure were realistic. Jensen (1981, p. 188) tells us of a North Carolina program similar to Heber's in which the experimental children have attained a Stanford-Binet mean (presumably SB-72) of 95 with a control group at 81; given my values for the Milwaukee Project this comes close to successful replication. As a help to the scholars in question, remember that my values represent scores translated into a uniform scoring convention. The corresponding scores before translation would be SB-72, 97.8–107.8; WISC-R, 97.6–106.9; WPPSI, 101.3–110.7, these last being a bit higher due to the need to allow for obsolete norms as well.

Crockett, Rardin, and Pasewark and predicting IQ. Crockett, Rardin, and Pasewark (1975) tested 42 Headstart children on both the WPPSI and Stanford-Binet LM and then in 1972, 4 years later, retested them on the WISC. Taking scores at face value, the subjects gained 7 points over 4 years, going from WPPSI to WISC, whereas there was no statistically significant gain going from SB-LM to WISC. These scholars concluded that "the WPPSI IQs tend to underestimate subsequent IQs of lower-class children as measured by the WISC" and that "in comparison to the WPPSI, the Stanford-Binet seems better able to predict WISC scores" (p. 922). As we have seen, the recommendation of the SB-LM as a reliable predictor of eventual IQ could hardly have been more mischievous: It was the inflated scores of Heber's experimental children on this test that led to unrealistic expectations about their

Table 8
Crockett, Rardin, and Pasewark's Headstart Group: IQ Stability or IQ Gains?

Results	WPPSI (age 5½)	SB-LM (age 5½)	WISC (age 9½)
As presented	87.46	91.91	94.43
Translated[a]	84.14	92.42	94.43
Adjusted[b]	80.14	79.42	86.43

Crockett et al.'s (1975) conclusion: SB-LM at age 5½ predicted WISC mean IQ at age 9½, that is, stability over 4 years.
Author's conclusion: all results at age 5½ lower than mean IQ at age 9½, that is, 6 to 7 points gain over 4 years.

Note. WPPSI = Wechsler Preschool and Primary Scale of Intelligence; SB-LM = Stanford-Binet Form L-M; WISC = Wechsler Intelligence Scale for Children.
[a] Results translated into uniform scoring convention, M = 100, SD = 15 for whites. [b]Results further adjusted to allow for obsolete norms, that is, scored against WISC–R norms, sample tested 1972.

eventual IQ. The only reason the unreliability of the SB-LM escaped notice was that the WISC, with norms dating from 1947–1948, went a long way toward matching its radical obsolescence.

Table 8 under Results as presented shows how things appeared to Crockett et al.: looking at the test results from age 5½ years, the SB-LM scores do look a good match for the mean IQ of these children 4 years later, whereas the WPPSI scores seem rather low. This gave them the following options: They could choose to emphasize the gap between the WPPSI and WISC results, which implied that the children had made IQ gains with increasing age, or they could choose to emphasize the apparent similarity of the SB-LM and WISC results, which implied that the children had maintained much the same IQ with increasing age. On the face of the results, there was no way to choose between these two options, although for reasons not stated they gave preference to the latter and compared the SB-LM favorably with the WPPSI.

Table 8 also shows the results translated into our uniform scoring convention and then adjusted for obsolete norms, that is, adjusted to compensate for the fact that the tests used were simply not comparable in terms of difficulty. As a consequence, the conflict between the WPPSI and SB-LM scores at age 5½ has disappeared in favor of a remarkably consistent performance signaling a mean IQ of about 80; and at age 9½, the mean IQ stands at almost 86.5. There is no doubt that these disadvantaged children made gains at school rather than exhibiting the usual pattern of decline, a decline that leaves such children actually labeled as retardates as they enter high school. However, since the phenomenon was overlooked, it could not be investigated.

Fagan and colleagues and administering tests. When lower-class or black children do better on a particular test because scores are inflated by obsolescence, that test is likely to be hailed as a fairer or more objective measure of their intelligence. For example, Fagan, Broughton, Allen, Clark, and Emerson (1969) administered both the SB-LM and the WPPSI to thirty-two 5-year-old lower-class children, half black and half white, and found that the Stanford-Binet mean was 8 points higher than the WPPSI mean. They explain these results by arguing that the content and administration of the WPPSI do not suit the temperament of the lower-class child as well as the SB-LM. The lower-class child is "frequently shy, nonverbal, activity-oriented, and sensitive to failure." The WPPSI requires frequent changes in instructions and materials in midtest, asks children to give additional reasons for their answers, and takes too long to administer. The Stanford-Binet "enabled the examiners to achieve better rapport and led to their willingness to accept the Binet IQ score as the more accurate" (p. 609).

If readers turn back to Table 2, they will see that whether children find the WPPSI harder than another test has nothing to do with content or administration but is essentially a function of obsolescence. It is harder than every test normed at an earlier time and easier than every test normed at a later time. Fortunately, the question in hand is subject to a direct empirical test: The SB-LM with norms from about 30 years before the WPPSI, and the SB-72, normed 7 years after the WPPSI, are virtually identical in content and administrative procedure. If content or administration are operative factors, both should be easier than the WPPSI. If obsolescence is at work, the earlier SB should be about 9 points easier than the WPPSI, and it is, whereas the later SB should be 2 or 3 points harder. Sewell (1977) has given us the first study of the WPPSI and the SB-72

and, almost exactly as I would predict, the latter was 4.41 IQ points harder (using uniform scoring convention). As to the moral Sewell himself drew from his study, he sees it as a useful supplement to Fagan et al.: Fagan et al. showed that the SB-LM was better than the WPPSI as a test for lower-class blacks because the Stanford-Binet gave higher scores; now Sewell has shown that the WPPSI is better than the SB-72 because the WPPSI gives higher scores!

Table 9 manipulates Fagan et al.'s results in the usual way. Note that when translated into our uniform scoring convention, these lower-class children had a greater advantage on the SB-LM over the WPPSI than Fagan et al. thought: almost 12 points rather than 8 points. When we allow for obsolescence, however, the advantage is reduced to less than 3 points. Still, that was enough to make the author wonder if there was at least something to the hypothesis that lower-class children found the SB-LM more to their taste than did middle-class subjects. Therefore, the total data on the SB-LM and WPPSI was divided in terms of class: As Table 9 shows, the score advantage the SB-LM conferred on lower-class children was actually less than that enjoyed by middle-class children; both groups were of course profitting from obsolescence, and their "unearned increments" are so close that no real distinction should be made.

Table 9
Social Class and Differential Performance on the Stanford-Binet Form L-M (SB-LM) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI)

Results	SB-LM	WPPSI	Difference
Fagan et al.'s lower-class children (N = 32)			
As presented	95.20	87.10	8.10
Translated[a]	95.50	83.76	11.74
Adjusted[b]	86.50	83.76	2.74
Lower-class vs. middle-class children			
Middle class (N = 100)[c]	119.18	109.76	9.42
Lower class (N = 296)[c]	94.12	85.89	8.23

Note. Sources: Barclay & Yater (1969); Fagan et al. (1969); Flynn (1980, pp. 184–185); Garber & Heber (1977, pp. 122, 125); Oakland, King, White, & Eckman (1971); Pasewark et al. (1971); Prosser & Crawford (1971); Rellas (1969); Wechsler (1967, p. 34).
[a] Results translated into uniform scoring convention, M = 100, SD = 15 for whites.
[b] Results further adjusted to allow for obsolete norms, that is, scored against WPPSI norms, sample tested 1963–1966.
[c] Results translated into uniform scoring convention but no allowance made for obsolete norms.

Yater and standardizing tests. This brings us to one of the most peculiar hypotheses ever put forward in the history of psychology: the notion that a test normed on a standardization sample including all races will somehow be fairer to blacks, or a better measure of their IQ, than a test normed on whites only. Assume you have a test on which the average black gets the same number of correct answers as the 16th percentile of the white population. So long as that is the case, nothing you do about *norming* the test will alter the performance gap between the races. If you norm on a standardization sample of whites only and define the white mean as 100, blacks will get 85. On the other hand, if you norm on a standardization sample of all races, mainly whites plus blacks, and put the mean at 100, blacks will get about 88. Naturally, blacks are closer to whites plus blacks than to whites only, but this is simply playing with numbers. If you want to give blacks a higher score while in no way affecting the real performance gap between the races in terms of correct answers, it would be simpler to just norm blacks on themselves and assign them a score of 100.

Despite this, Yater, Boyd, and Barclay (1975) believed that the WPPSI might be a more appropriate test of disadvantaged black children than the WISC. The WISC was normed on whites only, whereas the WPPSI included black children in its standardization sample and therefore "cultural bias effects with the WPPSI would not be expected to be operative" (p. 80). They administered both tests to 60 disadvantaged black children and found, not surprisingly, no statistically significant difference. The reader may wonder why scores were not a few points higher on the WPPSI, thanks to a scoring convention that should elevate a WISC score of 85 into a WPPSI score of 88. The answer of course is that the WISC was standardized some years earlier than the WPPSI. As usual, the earlier test is the easier test and the inflation of WISC scores by obsolescence just about matches the inflation of WPPSI scores

Table 10
Race, Class, and Differential Performance on the Wechsler Intelligence Scale for Children (WISC) and the Wechsler Preschool and Primary Scale of Intelligence (WPPSI)

Results	WISC	WPPSI	Difference[a]
Lower-class black (N = 84)			
As presented	89.92	89.36	0.56
Translated[b]	89.92	86.18	3.74
Middle-class white (N = 24)			
As presented	106.30	109.00	−2.70
Translated[b]	106.30	107.22	−0.92

[a] Difference equals WISC score minus WPPSI.
[b] Results translated into uniform scoring convention, M = 100, SD = 15 for whites.

by way of the latter's scoring convention. Table 10 makes this clear by combining Yater et al.'s results with those of Oakland et al. (1971), the only other study in which children were given both the WISC and WPPSI. The Results as presented show black children getting virtually identical scores on both tests. The Results translated show what really happened: When all scores are translated into our uniform scoring convention, the WPPSI mean drops about 4 points below the WISC, and the advantage the WISC gains from obsolescence is highly visible. The scores of the 24 white children do not show the effect of obsolescence, but the small sample size is probably the cause of their atypical results.

It might be thought that because all current tests have been normed on all races, this issue would die a natural death. But Evans and Richmond (1976) have reservations about the SB-72 because of the racial mix of its standardization sample: The sample included minorities, but in the absence of detailed data, they wonder if minorities were properly represented. When Sewell (1977) found that black children scored lower on the SB-72 than on the WPPSI, he brought the doubts of Evans and Richmond to the readers' attention. Actually, minorities were slightly overrepresented in the CAT sample on which the SB-72 norms were based, but the real point is the irrelevance of whether minorities are included in standardization samples at all. Indeed, obsolescence has done us at least one good turn: Those comparing a test with mixed-race norms against a test with whites-only norms have been comparing a later test against an earlier one; obsolescence has worked to guarantee that blacks would get lower scores on the all-races test precisely because it was later, thus falsifying the hypothesis of its greater fairness. However, we have had a close call: Imagine that the all-races tests had been the earlier ones; then we would have been told there was overwhelming "evidence" in favor of the hypothesis in question.

Obsolescence and the future. Earlier Stanford-Binet and Wechsler IQ tests suffer from obsolescence in the sense that their norms are easier to meet than those of later tests. No matter whether this reflects real IQ gains or whether it is an artifact of cumulative errors in standardization samples, obsolescence is a fact that must be taken into account in reinterpreting dozens of studies done in recent decades. There is also much in the general literature that must be called into question, for example, the claim that regression from parent to child is less for high-IQ parents than for others. In fact, children may have been tested on an earlier and easier test (say the SB-LM or WISC), whereas parents were tested on a later and harder test (say the WAIS). The effect would be to deflate regression from high-IQ parents down to the mean and inflate regression from low-IQ parents up toward the mean.

The number of studies that must be reinterpreted rises into the hundreds if one takes into account the likelihood of norms of unequal difficulty for IQ tests other than the Stanford-Binet and Wechsler: if IQ gains are real, these other tests will have been affected; if the Stanford-Binet and Wechsler organizations have made persistent and large sampling errors, it is unlikely that all others have escaped unscathed. The tables presented here for the SB and Wechsler must be improved; but then someone must do a similar job for the Peabody, first comparing earlier and later versions, then comparing all versions with Stanford-Binet and Wechsler tests, and then do a similar job for Ravens, and so on. We have no choice: Allowing for obsolescence in intelligence testing is just as essential as allowing for inflation in economic analysis.

Still, when dealing with tests normed from 1932 to 1972, at least we have a body of studies

which show that obsolescence was at work and suggest a rough estimate of rates. What of scholars using the WAIS–R normed in 1978 or those who will use the WPPSI–R when it is released later this decade? They will have no notion of where they stand: Even if IQ gains up to 1972 were real, the trend may have stopped or even reversed thereafter; even if sampling errors made every test harder than its predecessor up through 1972, they may begin to make for easier tests at any time. At present, we are in the intolerable situation of knowing the extent and direction of obsolescence only years after the event, after lots of little studies have accumulated and a tell-tale pattern has begun to show through their results. Test publishers should assume responsibility for giving advance warning of the existence of obsolescence as a matter of professional ethics.

When the Psychological Corporation publishes the WPPSI–R, the manual should contain not only data on a standardization sample of 1,200, there must also be results from a matching sample of 400 given both the WPPSI–R and the old WPPSI, another sample given both the WPPSI–R and the WISC–R, and another given both the WPPSI–R and the SB-72. Otherwise, it may not be long before the SB-72 begins to play the destructive role that the SB-LM played with Heber and others, which is to say that the 1990s will yield the same confusion as the 1960s and 1970s.

The Reality of IQ Gains Revisited

Having explored the possibility that IQ gains are not real, I now wish to reverse direction and address the hypothesis that a series of sampling errors have mimicked IQ gains. Assume that sampling errors for Stanford-Binet and Wechsler tests from 1932 to 1978 were likely to run as high as plus or minus 7 IQ points (a generous assumption). Assume that it is unlikely that the errors for any two standardization samples would be identical. Then sampling errors could mimic IQ gains and produce what we find in Table 11: a perfect correspondence between the chronological order of our seven standardization samples and their rank order when listed in terms of ascending levels of performance. However, the odds against having this occur by chance are 7 factorial, or 5,040 to 1. Only seven standardization samples are listed because the eighth, the Wechsler-Bellevue I, was omitted for reasons already given.

Table 11
Stanford-Binet and Wechsler Standardization Samples: Comparison of Chronological Order and Order of Performance

Samples	IQ differentials		
	Actual[a]	Predicted[b]	Period[c]
SB, 1932	.00	.00	0
W, 1947½	5.76	4.65	15.5
W, 1953½	7.94	6.45	21.5
W, 1964½	8.71	9.75	32.5
SB, 1971½	9.89	11.85	39.5
W, 1972	13.37	12.00	40.0
W, 1978	14.33	13.80	46.0

Note. SB = Stanford-Binet; W = Wechsler. Source: Table 3 with differentials calculated by averaging all sample sequences that terminate in a given sample and contain no more than four samples. Sample combinations with number less than 100 omitted except for WAIS–R (1978) for here, the sole study offering comparison with WISC–R (1972) has number = 80.

[a] Standardization samples ranked in terms of number of IQ points by which their performance bettered that of 1932.

[b] Standardization samples ranked in terms of number of IQ points by which their performance would have bettered that of 1932, assuming a rate of .300 points per year.

[c] Standardization samples ranked in terms of number of years elapsed since 1932.

The exact differentials in Table 11 that rank our standardization samples for quality of performance are approximations based on Table 3. Nonetheless, the magnitudes of these differentials are an extraordinary match for what we would predict, that is, the enhanced performance differentials predicated under the assumption of real IQ gains by Americans at a constant rate of .300 from 1932 to 1978. The discrepancies cluster at about 1 IQ point, and the largest is less than 2 points: The odds against sampling errors producing this kind of match between actual and predicted values cannot be calculated exactly but would be little short of astronomical.

This naturally suggests the possibility that systematic rather than random bias has been at work to mimic IQ gains. Test manuals do reveal one persistent trend: Early standardization samples tended to have a geographical bias in favor of more progressive locales. For

example, both Stanford-Binet 1932 (Terman & Merrill, 1937, p. 12) and Wechsler 1947–1948 (Wechsler, 1949, p. 16) virtually omit the Deep South with its low IQs and represent the South by states such as Texas, Kentucky, North Carolina, and Virginia. As Terman and Merrill (1973, p. 36) have pointed out, efforts to counteract such a bias by using other stratification variables are probably never entirely successful. By contrast, later samples, the CAT sample on which the SB-72 was based and the Wechsler 1972 sample, have excellent geographical coverage of the nation as a whole (Anastasi, 1976, pp. 258, 310). Now the effect of this trend would be to give earlier samples an elite bias lacking in later ones, that is, it would actually work against enhanced performance by later standardization samples. Another possible source of systematic bias would be the fact that subjects taking both an early and later test would encounter obsolete items on the early test, items whose outmoded content lends them artificial difficulty. However, this would make it more difficult for subjects to meet early norms than later and once again, the bias would operate against estimates of enhanced performance over time.

Recall that nothing in the IQ data itself evidences gross sampling error, rather we were driven to this hypothesis by the contrast between IQ trends and SAT-Verbal trends. It seemed impossible that tests correlated at the .80 level and measuring much in common would permit the following: that over a period of 18 years, performance on the two kinds of tests had diverged by something between .288 *SDU* (safe estimate) and .648 *SDU* (speculative estimate). However, acting on a hunch, the author undertook a more detailed analysis of the IQ data, studies in which subjects had taken both the WISC (1947–1948) and the WISC–R (1972), and these revealed a fact of great interest. Eight studies involving a total of 623 subjects provide a measure of not only full-scale IQ gains during this period but also gains on each of the Wechsler subtests (Brooks, 1977; Catron & Catron, 1977; Schwarting, 1976; Solly, 1977; Solway et al., 1976; Stokes et al., 1978; Swerdlik, 1978, pp. 120–121; Weiner & Kaufman, 1979).

These studies show a full-scale IQ gain of 9.27 points, about 1 point higher than the total body of WISC to WISC–R data, and a gain of only 3.09 points on the Vocabulary subtest. The Vocabulary subtest correlates with full-scale IQ at .80 (Wechsler, 1974, p. 47), yet these two trends diverge by fully 6.18 points or .412 *SDU*. Indeed, if we compare the Vocabulary subtest with the rest of the WISC (the whole test minus Vocabulary), the divergence is .458 *SDU*. These values fall virtually in the middle of the gap posited between IQ trends and SAT-V trends, that is, about halfway between .288 and .648 *SDU*. So what seemed impossible is possible: IQ and at least one central verbal skill can diverge radically despite a high correlation coefficient; the full-scale IQ versus Vocabulary contrast shows that the IQ versus SAT-Verbal contrast really may have occurred. On the other hand, these new data do nothing to resolve the problem of explaining the IQ versus SAT-V contrast. It merely poses that problem with renewed force: how can school children gain so much in overall intelligence and make so little progress in terms of enhanced vocabulary? What causal factors could boost intelligence and yet somehow withhold their potency from the world of words?

IQ Gains and Causal Explanation

Setting aside the combination of trends discussed above, can we even explain the phenomenon of IQ gains taken in isolation? In my opinion, the massive gains Americans appear to have made from one generation to another, about 14 IQ points over 46 years, pose a serious problem of causal explanation, given the present state of our knowledge. I will argue for three propositions: that the familiar causal factors we use to explain IQ variance within each generation, such as socioeconomic status (SES), have very limited promise; that the causal factors that are plausible candidates for operating primarily between generations have greater explanatory potential; but that even these may fall short and force us to conclude that our knowledge of environmental determinants of IQ is more limited than we suspected.

Recent research indicates that IQ gains over the last two generations must be due to environmental progress rather than to improved genes (Loehlin, Lindzey, & Spuhler, 1975, pp. 306–307). Among environmental factors potent within each generation, Jensen (1973, p. 357) designates SES as the most important

variable, with additional variables adding very small increments; he also argues that an SES gap of a full standard deviation makes a difference of about 2 IQ points in the American setting, assuming that confounding genetic factors have been eliminated (Jensen, 1981, p. 216). Clearly, measured in terms of present-day differences, the SES gap between 1932 and 1978 would have to be too enormous for credence in order to play much of a role in explaining a 14-point IQ gain. As for the total of environmental factors active at a particular time, the latest estimates attribute a maximum of .25 of IQ variance to between-families factors (I say maximum because the value declines sharply after adolescence; Plomin & DeFries, 1980, p. 19). Assuming that systematic environmental effects have a roughly normal distribution, one standard deviation of between-families difference in environmental quality makes a difference of 7.5 IQ points (the square root of .25 × 15 as value for *SD* for whites). Therefore, if we tried to explain our between-generations IQ gap in terms of within-generations environmental factors, we would have to say that the average environment in 1932 was 1.84 standard deviations (13.8 ÷ 7.5) below the average in 1978. This would put the Americans of 1932 at the 3rd percentile of environmental quality as measured today, which again taxes our credence.

This suggests turning to environmental factors that differ greatly between two generations while perhaps explaining a small amount of IQ variance within a generation. Scholarly correspondents of high competence (H. J. Eysenck, personal communication, December 14, 1982; J. C. Loehlin, personal communication, January 3, 1983; D. Zeaman, personal communication, January 13, 1983) have offered two possible causes of IQ gains over time, namely, increased test sophistication and a rising level of educational achievement. In passing, it is worth noting that even if these can explain IQ gains in isolation, neither does anything to solve the puzzle posed by the combination of IQ gains (or even IQ stability) and SAT-V losses. Young Americans who enjoy increased test sophistication or better education ought to show gains on both of these tests, indeed, it seems inconceivable that one could be affected and not the other.

Test sophistication is unique among environmental explanations of IQ gains in affecting the status of those gains. Assume that a massive gain from one generation to another merely shows that Americans were getting more practice in taking standardized tests: the gains would be real in the sense that enhanced performance on IQ tests really exists rather than being an artifact of sampling error; but they would not be real in the sense that they would signal no increase in intelligence, the trait IQ tests attempt to measure. Given this assumption, we should call IQ gains semi-real: they are like the improved times of athletes benefitting from a lighter running shoe; although performances on the clock really do improve, no one would claim that the athletes are better as such.

However, test sophistication is going to require a great deal of evidence to render it plausible as a dominant cause. Jensen (1980, pp. 590–591) emphasizes that even working with entirely naive subjects, repeated testing with parallel forms of IQ tests gives gains that total only 5 or 6 points. Further, he argues that by the 1950s, we reached a point of virtual saturation in exposure to standardized tests, and if that is so, it should have paid diminishing returns after that date. Test sophistication above all factors exhibits a threshold effect and no such effect is revealed in our data. It is interesting to note that many years ago when R. D. Tuddenham (1948) provided evidence from Army mental tests of massive IQ gains beginning in 1918, the immediate reaction was to suggest test sophistication as a major factor (Fulk, 1949, p. 17). If it is true that Americans have gained 6 IQ points every 20 years from 1918 to 1978, we cannot keep explaining this away by calling upon test sophistication time after time.

Enhanced educational achievement is a much more impressive candidate for explaining IQ gains over time. For example, Tuddenham (1948) analyzed his army data and found that weighting the 1918 mental test performance in terms of years of school completed, so as to match the 1943 educational distribution, caused about 55% of the mental test gains to disappear. On the other hand, selecting an elite in terms of schooling from 1918 to match 1943 inflates the influence of education as an environmental variable: Such an elite will be to some degree a genetic elite as well, and the influence of superior genes for IQ will be confounded with better education. Also

note that an educational elite will be superior in terms of a whole range of other environmental variables, such as SES, nutrition, child care, and test sophistication in particular. My own guess is that our current notions of the nature and potency of environmental variables put us about halfway, maybe a bit further, toward a plausible explanation of massive IQ gains.

The difficulty of our task, given current notions of the potency of environmental factors, can be illustrated by some comparative data. In order to explain a Japanese IQ gain on Wechsler tests of 7 points over 23 years (note that the rate of .304 points per year is familiar), A. M. Anderson (1982) had to call attention to environmental changes of the most radical sort. As he pointed out, since 1930 Japan has experienced massive urbanization, a cultural revolution from feudal to western attitudes, the decline of inbreeding and consanguineous marriages, and huge advances in nutrition, life expectancy, and education. The magnitude of gain Anderson attempted to explain in the Japanese context, we have to explain in the American context for the period from 1948 to 1972. And yet, to find analogous environmental changes in the United States, we would have to go back to the turn of the century.

Summary of Implications

Assuming that American IQ gains 1932 to 1978 are not real, but an artifact of sampling error, they have acted as a confounding variable in hundreds of studies and require altered practices from testing organizations. Assuming that these gains are semi-real, due primarily to test sophistication, they imply all of the above and also reveal the inexplicable combination of IQ gains and SAT-V losses. Assuming that these gains are real, they imply all of the above and pose a serious problem of causal explanation. Moreover, all of these implications hold even if real gains ceased in 1978 or even 1972: The period in question shows the radical malleability of IQ during a time of normal environmental change; other times and other trends cannot erase that fact.

References

Anastasi, A. (1961). *Psychological testing* (2nd ed.). New York: Macmillan.

Anastasi, A. (1976). *Psychological testing* (4th ed.). New York: Macmillan.

Anderson, A. M. (1982). The great Japanese IQ increase. *Nature, 297,* 180–181.

Anderson, E. E., Anderson, S. F., Ferguson, C., Gray, J., Hittinger, J., McKinstry, E., Motter, M. E., & Vick, G. (1942). Wilson College studies in psychology: I. A comparison of the Wechsler-Bellevue, Revised Stanford-Binet, and American Council of Education tests at the college level. *Journal of Psychology, 14,* 317–326.

Appelbaum, A. S., & Tuma, J. M. (1977). Social class and test performance: Comparative validity of the Peabody with the WISC and WISC-R for two socioeconomic groups. *Psychological Reports, 40,* 139–145.

Arnold, F. C., & Wagner, W. K. (1955). A comparison of Wechsler children's scale and Stanford-Binet scores for eight- and nine-year-olds. *Journal of Experimental Education, 24,* 91–94.

Austin, J. J., & Carpenter, P. (1970). The use of the WPPSI in early identification of mental retardation and preschool special education. *Proceedings of the 78th Annual Convention of the American Psychological Association, 5,* 913.

Barclay, A., & Yater, A. (1969). Comparative study of the Wechsler Preschool and Primary Scale of Intelligence and the Stanford-Binet Intelligence Scale, Form L-M among culturally deprived children. *Journal of Consulting and Clinical Psychology, 33,* 257.

Barratt, E. S., & Baumgarten, D. L. (1957). The relationship of the WISC and Stanford-Binet to school achievement. *Journal of Consulting Psychology, 21,* 144.

Bradway, K. P., & Thompson, C. W. (1962). Intelligence at adulthood: A twenty-five year follow-up. *Educational Psychology, 53,* 1–14.

Brittain, M. (1968). A comparative study of the Wechsler Intelligence Scale for Children and the Stanford-Binet Intelligence Scale (Form L-M) with eight-year-old children. *British Journal of Educational Psychology, 38,* 103–104.

Brooks, C. R. (1977). WISC, WISC-R, S-B L & M, WRAT: Relationships and trends among children ages six to ten referred for psychological evaluation. *Psychology in the Schools, 14,* 30–33.

Catron, D. W., & Catron, S. S. (1977). WISC vs WISC-R: A comparison with educable mentally retarded children. *Journal of School Psychology, 15,* 264–266.

Clarke, A. (1981). Sir Cyril Burt and Rick Heber. *Bulletin of the British Psychological Society, 34,* 324.

Clarke, A. D. (1973). The prevention of subcultural subnormality: Problems and prospects. *British Journal of Mental Subnormality, 19,* 7–20.

Cohen, B. D., & Collier, J. N. (1952). A note on the WISC and other tests of children six to eight years old. *Journal of Consulting Psychology, 16,* 226–227.

Cole, P., & Weleba, L. (1956). Comparison data on the Wechsler-Bellevue and the WAIS. *Journal of Clinical Psychology, 12,* 198–200.

Covin, T. M. (1977). Comparability of WISC and WISC-R scores for 30 8- and 9-year-old institutionalized Caucasian children. *Psychological Reports, 40,* 382.

Crockett, B. K., Rardin, M. W., & Pasewark, R. A. (1975). Relationship between WPPSI and Stanford-Binet IQs and subsequent WISC IQs in headstart children. *Journal of Consulting and Clinical Psychology, 43,* 922.

Davis, E. E. (1977). Matched pair comparison of WISC and WISC-R scores. *Psychology in the Schools, 14,* 161–166.

Delattre, L., & Cole, D. (1952). A comparison of the WISC and the Wechsler-Bellevue. *Journal of Consulting Psychology, 16,* 228–230.

Educational Testing Service. (1977). *National report on college bound seniors, 1977.* Princeton, NJ: College Entrance Examination Board.

Educational Testing Service. (1981). *National report on college bound seniors, 1981.* Princeton, NJ: College Entrance Examination Board.

Estes, B. W. (1965). Relationship between the Otis, 1960 Stanford-Binet, and WISC. *Journal of Clinical Psychology, 21,* 296–297.

Estes, B. W., Curtin, M. E., De Burger, R. A., & Denny, C. (1961). Relationship between the 1960 Stanford-Binet, 1937 Stanford-Binet, WISC, Raven, and Draw-a-Man. *Journal of Consulting Psychology, 25,* 388–391.

Evans, P. L., & Richmond, B. O. (1976). A practitioner's comparison: The 1972 Stanford-Binet and the WISC-R. *Psychology in the Schools, 13,* 9–14.

Fagan, J., Broughton, E., Allen, M., Clark, B., & Emerson, P. (1969). Comparison of the Binet and WPPSI with lower class 5-year-olds. *Journal of Consulting and Clinical Psychology, 33,* 607–609.

Flynn, J. R. (1980). *Race, IQ and Jensen.* London: Routledge & Kegan Paul.

Frandsen, A. N., & Higginson, J. B. (1951). The Stanford-Binet and the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology, 15,* 236–238.

Fulk, B. E. (1949). *A comparison of Negro and white Army General Classification Test scores.* Unpublished master's thesis, University of Illinois, Urbana.

Garber, H. L. (1982, May). *The Milwaukee Project: Preventing mental retardation in children of families at risk.* Paper presented at the CIBA-GEIGy Conference on Mental Retardation from a Neurobiological and Sociocultural Point of View, Lund, Sweden.

Garber, H., & Heber, F. R. (initials incorrect—correct: R. F.). (1977). The Milwaukee Project: Indications of the effectiveness of early intervention in preventing mental retardation. In P. Mittle (Ed.), *Research to practice in mental retardation: Vol. 1. Care and intervention* (pp. 119–127). Baltimore, MD: University Park Press.

Gehman, I. H., & Matyas, R. P. (1956). Stability of the WISC and Binet tests. *Journal of Consulting Psychology, 20,* 150–152.

Gerboth, R. (1950). A study of two forms of the Wechsler-Bellevue Intelligence Scale. *Journal of Consulting Psychology, 14,* 365–370.

Giannell, A. S., & Freeburne, C. M. (1963). The comparative validity of the WAIS and the Stanford-Binet with college freshmen. *Educational and Psychological Measurement, 23,* 557–567.

Gibby, R. G. (1949). A preliminary survey of certain aspects of Form II of the Wechsler-Bellevue Scale as compared to Form I. *Journal of Clinical Psychology, 5,* 165–169.

Goldfarb, W. (1944). Adolescent performance on the Wechsler-Bellevue Intelligence Scales and the Revised Stanford-Binet Examination, Form L. *Journal of Educational Psychology, 35,* 503–507.

Goolishian, H. A., & Ramsay, R. (1956). The Wechsler-Bellevue Form I and the WAIS: A comparison. *Journal of Clinical Psychology, 12,* 147–151.

Halpern, F. (1942). A comparison of the Revised Stanford L and the Bellevue Adult Intelligence Test as clinical instruments. *Psychiatric Quarterly Supplement, 16,* 206–211.

Hannon, J. E., & Kicklighter, R. (1970). WAIS vs WISC in adolescents. *Journal of Consulting and Clinical Psychology, 35,* 179–182.

Harlow, J. E., Price, A. C., Tatham, L. J., & Davidson, J. R. (1957). Preliminary study of comparison between Wechsler Intelligence Scale for Children and Form L of the Revised Stanford-Binet Scale at three age levels. *Journal of Clinical Psychology, 13,* 72–73.

Hartlage, L. C., & Boone, K. E. (1977). Achievement test correlates of Wechsler Intelligence Scale for Children and Wechsler Intelligence Scale for Children—Revised. *Perceptual and Motor Skills, 45,* 1283–1286.

Heber, R., & Garber, H. (1975). Progress report II: An experiment in the prevention of cultural-familial retardation. In D. A. Primrose (Ed.), *Proceedings of the Third Congress of the International Association for the Scientific Study of Mental Deficiency, Vol. 1,* (pp. 34–43). Warsaw: Polish Medical Publishers.

Heber, R. F., Garber, H. L., Hoffman, C., & Harrington, S. (circ. 1977). *Establishment of the High Risk Population Laboratory* (Research project report). Unpublished manuscript, University of Wisconsin, Madison.

Holland, G. A. (1953). A comparison of the WISC and Stanford-Binet IQs of normal children. *Journal of Consulting Psychology, 17,* 147–152.

Jackson, R. (1976). *A summary of SAT score statistics for College Board candidates.* Princeton, NJ: College Entrance Examination Board.

Jensen, A. R. (1973). *Educability and group differences.* New York: Harper & Row.

Jensen, A. R. (1980). *Bias in mental testing.* London: Methuen.

Jensen, A. R. (1981). *Straight talk about mental tests.* New York: The Free Press.

Jones, S. (1962). The Wechsler Intelligence Scale for Children applied to a sample of London primary school children. *British Journal of Educational Psychology, 32,* 119–132.

Kangas, J., & Bradway, K. (1971). Intelligence at middle age: A thirty-eight-year follow-up. *Developmental Psychology, 5,* 333–337.

Karson, S., Pool, K. B., & Freud, S. L. (1957). The effects of scale and practice on WAIS and W-B I test scores. *Journal of Consulting Psychology, 21,* 241–245.

Kaufman, A. S., & Doppelt, J. E. (1976). Analysis of WISC-R standardization data in terms of the stratification variables. *Child Development, 47,* 165–171.

Klinge, V., Rodziewicz, T., & Schwartz, L. (1976). Comparison of the WISC and WISC-R on a psychiatric adolescent inpatient sample. *Journal of Abnormal Child Psychology, 4,* 73–81.

Knopf, I. J., Murfett, B. J., & Milstein, V. (1954). Relationships between the Wechsler-Bellevue Form I and the WISC. *Journal of Clinical Psychology, 10,* 261–263.

Krugman, J. I., Justman, J., Wrightstone, J. W., & Krugman, M. (1951), Pupil functioning on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology, 15,* 475–483.

Kureth, G., Muhr, J. P., & Weisgerber, C. A. (1952). Some data on the validity of the Wechsler Intelligence Scale for Children. *Child Development, 23,* 281–287.

Larrabee, G. J., & Holroyd, R. G. (1976). Comparison of WISC and WISC-R using a sample of highly intelligent students. *Psychological Reports, 38,* 1071–1074.

Levinson, B. M. (1959). A comparison of the performance of bilingual and monolingual native born Jewish preschool children of traditional parents on four intelligence tests. *Journal of Clinical Psychology, 15,* 74–76.

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

McKerracher, D. W., & Scott, J. (1966). IQ scores and the problem of classification: A comparison of the WAIS and S-B Form L-M in a group of subnormal and psychopathic patients. *British Journal of Psychiatry, 112,* 537–541.

McNemar, Q. (1942). *The revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin.

Memorandum for Mrs. Sharp: Scholastic Aptitude Test candidate volumes. (Undated). (Available from Eleanor V. Horne, Educational Testing Service, Princeton, NJ)

Modu, C. C., & Stern, J. (1977). *The stability of the SAT-Verbal score scale.* Princeton, NJ: College Entrance Examination Board.

Mussen, P. H., Conger, J. J., & Kagan, J. (1974). *Child development and personality* (4th ed.). New York: Harper & Row.

Mussen, P. H., Dean, S., & Rosenberg, M. (1952). Some further evidence of the validity of the WISC. *Journal of Consulting Psychology, 16,* 410–411.

Neuringer, C. (1963). The form equivalence between the Wechsler-Bellevue Intelligence Scale, Form I and the Wechsler Adult Intelligence Scale. *Educational and Psychological Measurement, 23,* 755–763.

Oakland, T. D., King, J. D., White, L. A., & Eckman, R. (1971). *A comparison of performance on the WPPSI, WISC, and SB with preschool children: Companion studies. Journal of School Psychology, 9,* 144–149.

Pasewark, R. A., Rardin, M. W., & Grice, J. E. (1971). Relationship of the Wechsler Pre-school and Primary Scale of Intelligence and the Stanford-Binet (L-M) in lower class children. *Journal of School Psychology, 9,* 43–50.

Pastovic, J. J., & Guthrie, G. M. (1951). Some evidence on the validity of the WISC. *Journal of Consulting Psychology, 15,* 385–386.

Plomin, R., & DeFries, J. C. (1980). Genetics and intelligence: Recent data. *Intelligence, 4,* 15–24.

Price, J. R., & Thorne, G. D. (1955). A statistical comparison of the WISC and Wechsler-Bellevue, Form I. *Journal of Consulting Psychology, 19,* 479–482.

Prosser, N. S., & Crawford, V. B. (1971). Relationship of scores on the Wechsler Preschool and Primary Scale of Intelligence and the Stanford-Binet Intelligence Scale Form LM. *Journal of School Psychology, 9,* 278–283.

Quereshi, M. Y. (1968a). The comparability of WAIS and WISC subtest scores and IQ estimates. *Journal of Psychology, 68,* 73–82.

Quereshi, M. Y. (1968b). Practice effects on the WISC subtest scores and IQ estimates. *Journal of Clinical Psychology, 24,* 79–85.

Quereshi, M. Y., & Miller, J. M. (1970). The comparability of the WAIS, WISC, and WB II. *Journal of Educational Measurement, 7,* 105–111.

Rasbury, W., McCoy, J. G., & Perry, N. W. (1977). Relations of scores on WPPSI and WISC-R at a one-year interval. *Perceptual and Motor Skills, 44,* 695–698.

Rellas, A. J. (1969). The use of the Wechsler Preschool and Primary Scale (WPPSI) in the early identification of gifted students. *Journal of Educational Research, 20,* 117–119.

Reynolds, C. R., & Hartlage, L. (1979). Comparison of WISC and WISC-R regression lines for academic prediction with black and with white referred children. *Journal of Consulting and Clinical Psychology, 47,* 589–591.

Ross, R. T., & Morledge, J. (1967). A comparison of the WISC and WAIS at chronological age 16. *Journal of Consulting Psychology, 31,* 331–332.

Sartain, A. Q. (1946). Comparison of the new Revised Stanford-Binet, the Bellevue Scale, and certain group tests of intelligence. *Journal of Social Psychology, 23,* 237–239.

Schachter, F. F., & Apgar, V. (1958). Comparison of preschool Stanford-Binet and school-age WISC IQs. *Journal of Educational Psychology, 49,* 320–323.

Schwarting, F. G. (1976). A comparison of the WISC and WISC-R. *Psychology in the Schools, 13,* 139–141.

Seashore, H., Wesman, A., & Doppelt, J. (1950). The standardization of the Wechsler Intelligence Scale for Children. *Journal of Consulting Psychology, 14,* 99–110.

Sewell, T. E. (1977). A comparison of the WPPSI and Stanford-Binet Intelligence Scale (1972) among lower-SES black children. *Psychology in the Schools, 14,* 158–161.

Simpson, R. L. (1970). Study of the comparability of WISC and the WAIS. *Journal of Consulting and Clinical Psychology, 34,* 156–58.

Solly, D. C. (1977). Comparison of WISC and WISC-R scores of mentally retarded and gifted children. *Journal of School Psychology, 15,* 255–258.

Solway, K. S., Fruge, E., Hays, J. R., Cody, J., & Gryll, S. (1976). A comparison of the WISC and WISC-R in a juvenile delinquent population. *Journal of Psychology, 94,* 101–106.

Stokes, E. H., Brent, D., Huddleston, N. J., Rozier, J. S., & Marrero, B. (1978). A comparison of WISC and WISC-R scores of sixth grade students: Implications for validity. *Educational and Psychological Measurement, 38,* 469–473.

Stroud, J. B., Blommers, P., & Lauber, M. (1957). Correlation analysis of WISC and achievement tests. *Journal of Educational Psychology, 48,* 18–26.

Swerdlik, M. E. (1978). Comparison of WISC and WISC-R scores of referred black, white and Latino children. *Journal of School Psychology, 16,* 110–125.

Terman, L. M. (1942). The revision procedures. In Q. McNemar, *The revision of the Stanford-Binet Scale* (pp. 1–14). Boston: Houghton Mifflin.

Terman, L. M., & Merrill, M. A. (1937). *Measuring intelligence.* London: Harrap.

Terman, L. M., & Merrill, M. A. (1973). *Stanford-Binet Intelligence Scale: 1973 norms edition.* Boston: Houghton Mifflin.

Thomas, P. J. (1980). A longitudinal comparison of the WISC and WISC-R with special education pupils. *Psychology in the Schools, 17,* 437–441.

Thorndike, R. L. (1973). *Stanford-Binet Intelligence Scale 1972 norms tables.* Boston: Houghton Mifflin.

Thorndike, R. L. (1975). Mr. Binet's test 70 years later. *Educational Researcher, 4,* 3–7.

Triggs, F. O., & Cartee, J. K. (1953). Pre-school performance on the Stanford-Binet and the Wechsler Intelligence Scale for Children. *Journal of Clinical Psychology, 9,* 27–29.

Tuddenham, R. D. (1948). Soldier Intelligence in World Wars I and II. *American Psychologist, 3,* 54–56.

Tuma, J. M., Appelbaum, A. S., & Bee, D. E. (1978). Comparability of the WISC and the WISC-R in normal children of divergent socioeconomic backgrounds. *Psychology in the Schools,* 15, 339–346.

Wechsler, D. (1939). *The measurement of adult intelligence.* Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1949). *WISC manual.* New York: The Psychological Corporation.

Wechsler, D. (1955). *WAIS manual.* New York: The Psychological Corporation.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore, MD: Williams & Wilkins.

Wechsler, D. (1967). *WPPSI manual.* New York: The Psychological Corporation.

Wechsler, D. (1974). *WISC-R manual.* New York: The Psychological Corporation.

Wechsler, D. (1981). *WAIS-R manual.* New York: The Psychological Corporation.

Weider, A., Levi, J., & Risch, F. (1943). Performance of problem children on the Wechsler-Bellevue Intelligence Scales and the Revised Stanford-Binet. *Psychiatric Quarterly, 17,* 695–701.

Weider, A., Noller, P. A., & Schramm, T. A. (1951). The Wechsler Intelligence Scale for Children and the Revised Stanford-Binet. *Journal of Consulting Psychology, 15,* 330–333.

Weiner, S. G., & Kaufman, A. S. (1979). WISC-R versus WISC for black children suspected of learning or behavioural disorders. *Journal of Learning Disabilities, 12,* 100–107.

Wheaton, P. J., Vandergriff, A. F., & Nelson, W. H. (1980). Comparability of the WISC and WISC-R with bright elementary school students. *Journal of School Psychology, 18,* 271–275.

Wirtz, W. (Chairperson). (1977). *On further examination: Report of the Advisory Panel on the Scholastic Aptitude Test Score Decline.* Princeton, NJ: College Entrance Examination Board.

Yater, A. C., Boyd, M., & Barclay, A. (1975). A comparative study of WPPSI and WISC performance of disadvantaged children. *Journal of Clinical Psychology, 31,* 78–80.

Zimmerman, I. L., & Woo-Sam, J. (1972). Research with the Wechsler Intelligence Scale for Children: 1960–1970. *Psychology in the Schools, 9,* 232–271.

Appendix

Studies Rejected for Computing IQ Gains Over Time

Reason	Study
Same subjects as another study	Crockett, Rardin, & Pasewark, 1975 Gehman & Matyas, 1956
Insufficient data about means	Cole & Weleba, 1956 Harlow, Price, Tatham, & Davidson, 1957
Tests not given to same subjects	Goolishian & Ramsay, 1956
Two or more years between tests	Austin & Carpenter, 1970 Schachter & Apgar, 1958 Stroud, Blommers, & Lauber, 1957
Altered environment between tests	Garber & Heber, 1977[a]
Practice effects	Delattre & Cole, 1952 Ross & Morledge, 1967
No date for norms of Wechsler-Bellevue II	Gerboth, 1950 Gibby, 1949 Quereshi & Miller, 1970[b]

Note. Exceptions: As explained in the text, I rejected all studies whose subjects had a mean IQ above 130 on the test with the more recent norms and all studies of the mentally retarded, unless the subjects had a mean IQ above 75. A few studies of the combination SB-LM and SB-72 were omitted as superfluous given the large numbers (N = 2,351) provided by the Stanford-Binet 1972 standardization sample. Two studies of the WISC and WISC-R (Thomas, 1980; Solway et al., 1976) were included despite not meeting the usual criteria because their research design was calculated to overcome methodological problems peculiar to that test combination.

[a] Rejected data for the Wechsler Intelligence Scale for Children only. [b] Rejected Wechsler-Bellevue II data only.

Received March 9, 1983
Revision received May 20, 1983 ■