Psychological Bulletin
1987, Vol. 101, No. 2, 171-191

Copyright 1987 by the American Psychological Association, Inc.
0033-2909/87/$00.75

Massive IQ Gains in 14 Nations: What IQ Tests Really Measure

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

Data from 14 nations reveal IQ gains ranging from 5 to 25 points in a single generation. Some of the largest gains occur on culturally reduced tests and tests of fluid intelligence. The Norwegian data show that a nation can make significant gains on a culturally reduced test while suffering losses on other tests. The Dutch data prove the existence of unknown environmental factors so potent that they account for 15 of the 20 points gained. The hypothesis that best fits the results is that IQ tests do not measure intelligence but rather a correlate with a weak causal link to intelligence. This hypothesis can also explain differential trends on various mental tests, such as the combination of IQ gains and Scholastic Aptitude Test losses in the United States.

Over the last decade, scholars in many countries—largely unknown to one another—have been measuring IQ trends from one generation to the next. In this article their results are collected and analyzed so as to reap the benefits comparative data usually yield. It is also argued that this mass of data poses fundamental problems for developmental psychology, primarily concerning what factors have the most potent effect on IQ, what IQ tests measure, and how IQ tests should be used in making between-groups comparisons.

Data Collection

The method used to collect data can be simply put. Questionnaires, letters, or personal appeals—usually a combination of all three—were sent to all those researchers known to be interested in IQ trends on the basis of scholarly correspondence and the exchange of publications. One-hundred sixty-five scholars from 35 countries were contacted. They came from Europe—every nation except Albania, Denmark, Greece, and Portugal; Asia—Japan, India, and Israel; Latin America—Argentina, Brazil, Chile, Cuba, Mexico, and Venezuela; the Caribbean—Barbados and the U.S. Virgin Islands; and the Commonwealth—Australia, Canada, and New Zealand. American data were available from a previous study. Military authorities in charge of psychological testing were contacted in every European country, plus Australia, Canada, Greenland, Iceland, and New Zealand, as were 21 educational research institutes in Western Europe and the Commonwealth.

Data Presentation

The results are presented as they were in the original sources and are then converted into a uniform scoring convention, so that all trends over time can be expressed in terms of IQ points based on a mean and standard deviation set at 100 and 15, respectively. The mechanics of the conversion differ from nation to nation and are therefore detailed in that context.

The evaluation of the data involves four general criteria derived from suggestions made by A. R. Jensen (personal communications, January 12 and February 3, 1983): (a) The possibility of sample bias should be eliminated by comprehensive samples, such as mass testing of draft registrants; (b) tests should remain unaltered from one generation to another and estimates of trends should be based on raw score differences; (c) particular emphasis should be placed on culturally reduced tests like the Ravens Progressive Matrices Test, as distinct from tests with items that might easily be learned from one generation to another; and (d) particular emphasis should be placed on using mature subjects, subjects who have reached the peak of their raw score performance. This last criterion might eliminate the possibility that the current generation is merely reaching peak performance at an earlier age than the last generation and that at full maturity, the two generations would score about the same.

Data from the Netherlands, Belgium, and Norway meet all four of these criteria. However, the criteria of culturally reduced tests and mature subjects are relevant primarily to the significance of IQ gains. Thus, rather than being applied to each data set as it is presented they will be discussed in the section on implications. The criteria of sample quality and test continuity are relevant to assessing the status of each data set, that is, whether it provides reliable evidence that IQ gains have actually occurred at least within a given age group. Therefore, these two criteria, along with the obvious addition of clarity of results—whether the results can be converted into an estimate of IQ gains—are used as the data are presented. They are used to assess each data set and designate its status: Status 1 is evidence that verifies IQ gains beyond a reasonable doubt, Status 2 is probable evidence of gains, Status 3 is tentative evidence, and Status 4 is merely speculative evidence.

The data sets divide naturally into strong and weak data and so the former will not appear compromised by the limitations of the latter, they are kept separate and grouped under those

Correspondence concerning this article should be addressed to James R. Flynn, Department of Political Studies, University of Otago, P.O. Box 56, Dunedin, New Zealand.

Table 1
The Netherlands: IQ Gains on the Ravens Progressive Matrices Test, 18-Year-Old Men

Year	%[a]	*SD*s from mean	*M* IQ[b]
1952	31.2	−0.4914	100.00
1962	46.4	−0.0778	106.20
1972	63.2	+0.3374	112.43
1981/1982	82.2	+0.9154	121.10

Note. Data from de Leeuw and Meester (1984, p. 21), with data in the author's possession substituted for 1981/1982.
[a] Percentage of men who got more than 24 out of 40 items correct. [b] Calculated with 1952 mean and standard deviation set at 100 and 15, respectively, and assuming that the variance was constant over the whole period.

headings. There are a few exceptions, but they shall be clearly identified.

Strong Data

The Netherlands

Samples and test. The Dutch military examines virtually all 18-year-old men, the main exceptions being those with a grave mental or physical impairment and the prison population. Those men who pass a medical examination then take mental tests. In 1945, the military selected 40 of the most discriminating items of the 60 in the Ravens Progressive Matrices Test, and that test has remained unaltered to the present. Leeuw and Meester (1984, p. 21) used results from 1952, 1962, 1972, and 1982 to trace trends over time, but unfortunately, the 1982 performance was inflated because it was based on an elite group selected from the total population of recruits (Vroon, Leeuw, & Meester, 1984). P. A. Vroon (personal communication, November 5, 1984) has supplied the results of 57,897 men, all those tested in late 1981 and late 1982, so now accurate results are available for those years.

Results. Table 1 presents Dutch IQ gains in terms of the increasing percentage of men who got more than 24 out of the 40 Ravens test items correct; for example, the percentage rose from 31.2% in 1952 to 82.2% in 1981/1982. Loehlin, Lindzey, and Spuhler (1975, p. 143) describe the standard method of converting such comparisons into IQ comparisons. Given a cutting line (more than 24 items correct) and the distance between that line and the mean in percentages, a table of areas under a normal curve converts that distance into standard deviations. In 1952, the mean lay 0.4914 *SD*s below that line and by 1981/1982, it had risen to 0.9154 *SD*s above the line; thus Dutch men had gained a total of 1.4068 *SD*s. Multiplying 1.4068 times 15 (15 = 1 *SD*) yields a total gain of 21.10 IQ points over a period of 29.5 years. This method of converting results assumes that at both times, the IQ curve was roughly normal and variance did not significantly alter. Therefore, a more direct check on Dutch gains is desirable.

The data for the 1981/1982 examinees are complete and gives actual values for the mean, median, and standard deviation. The problem is to get similar values for an earlier year. Vroon took a sample of 2,847 men from the total population of 1981/1982 examinees and traced the scores of their fathers, whose median year of testing was 1954. He calculated actual values for both sons and fathers. The sons were elite compared to their 1981/1982 cohorts, by .64 Ravens items correct; because the correlation between sons and fathers was .33, the fathers would have been elite compared to their 1954 cohorts by .21 items (.64 × .33 = .21). This gives an estimated mean for 1954 of 21.39 Ravens items correct. The fathers' score distribution was normal, with their *SD* = 6.738.

The total population of 1981/1982 lacks a normal distribution because by that time, Dutch men were finding the Ravens items so easy that the test had an artificially low ceiling. This depressed the mean slightly below the median and through restriction of range diminished the standard deviation. To get the best estimate of performance in 1981/1982, the proper correction is to use the median rather than the mean and ignore the restricted standard deviation in favor of that from 1954. This gives the actual values needed for a comparison: the 1954 *M* = 21.39, the 1981/1982 *Mdn* = 29.50, and the 1954 *SD* = 6.738. Simple arithmetic converts the raw score gain over these years into IQ points: 29.50 − 21.39 = 8.11, 8.11 ÷ 6.738 = 1.204 *SD*s gained, and 1.204 × 15 = 18.054 IQ points gained (Leeuw & Meester, 1984, p. 21; P. A. Vroon, personal communications, September 24 to November 27, 1984).

To check these two estimates against one another, they must be projected over the same time span. The original estimate was a gain of 21.102 IQ points over 29.5 years, which would be 21.46 points over 30 years. The second estimate was 18.054 points over 27.5 years, which would be 19.70 points over 30 years.

Assessment. The Dutch drafted approximately 80,000 men in 1952, 65,363 of whom passed the medical examination and took the Ravens test (Leeuw & Meester, 1984, p. 11). Although census data do not allow an exact estimate, the draft included above 95%, and those tested about 79%, of all 18-year-old men (United Nations, 1979, p. 268). P. A. Vroon (personal communication, December 11, 1984) asserts that about 80% still take the Ravens at the present time. Thus the percentage who fail the medical exam and escape testing has remained constant from 1952 to 1981/1982 and should not be a significant source of bias.

The content of the 40-item version of the Ravens test has not been altered. The two conversions of raw score results into IQ scores tally well, but it is probably best to accept the second and more conservative estimate, based as it is on more detailed data. The second estimate evidences a gain of about 20 IQ points over a single generation, the 30 years between 1952 and 1982. It is also something unique in the literature, a measure of IQ gains over a real generation, a comparison of a generation of fathers with their own sons. In terms of sample quality, test continuity, and clarity of results, Dutch IQ gains should be taken as verified, or Status 1.

Belgium

Samples and tests. The Belgian military examines all 18-year-old men except those deferred to continue with their edu-

Table 2
Belgium: IQ Gains on Mental Tests, 18-Year-Old Men

Sample	Tests			
	Ravens[a]	Shapes	Arithmetic	Vocabulary[b]
French-speaking men				
1958 *M*	32.74	15.23	18.95	18.24
1967 *M*	37.88	18.42	21.51	20.70
Score gain	5.14	3.19	2.56	2.46
Average *SD*	11.92	7.74	12.14	13.92
SD unit gain	0.431	0.412	0.211	0.177
IQ gain	6.47	6.19	3.16	2.65
Dutch-speaking men				
1958 *M*	30.98	14.94	17.35	24.09
1967 *M*	36.99	18.29	20.80	28.30
Score gain	6.01	3.35	3.45	4.21
Average *SD*	11.53	7.50	11.88	14.04
SD unit gain	0.521	0.447	0.290	0.300
IQ gain	7.82	6.70	4.36	4.50

Note. Data from Bouvier (1969, pp. 2–3).
[a] The raw score for all tests is the total number of correct answers minus a penalty for wrong answers. Therefore, the Ravens Progressive Matrices Test means are not comparable to results obtained from the usual method of scoring. [b] The French and Dutch vocabulary tests are not comparable across the two linguistic groups; all other tests are.

cation and those classified as physically or mentally disabled (J. J. Deltour, personal communication, January 8, 1986). Most of those deferred are regained before the age of 25, but the Belgians are more liberal than the Dutch in exempting those classified as physically or mentally disabled. On the other hand, unlike the Dutch, those not exempt take both medical and mental tests. They take the full 60-item version of the Ravens test, but because wrong answers are penalized, raw score results must not be compared with those of other nations. Other tests include (a) "visualization of shapes," which consists of a series of geometrical designs; (b) elementary arithmetic problems; and (c) a vocabulary test, which divides into two different versions, one for French-speaking and one for Dutch-speaking examinees.

Results. Bouvier (1969) gives raw score means and standard deviations for each year from 1958 to 1967. Table 2 shows the means for 1958 and 1967 so as to measure gains over those 9 years. Variance was relatively constant, so the standard deviations from all years were averaged to give a representative value. The results were converted into IQ gains as follows: The raw score gain is divided by the raw score standard deviation, yielding the gain in standard deviation units that when multiplied by 15, yields the gain in IQ points. Because this sort of conversion is purely an arithmetic calculation, it is assumed for all remaining data sets that when raw score means and standard deviations are available, the method is understood. As Table 2 shows, both French- and Dutch-speaking examinees have made very similar gains over a period of 9 years: from 6.19 to 7.82 IQ points on the Ravens and Shapes tests, and from 2.65 to 4.50 points on arithmetic and vocabulary tests.

Assessment. Once again, the criteria adopted for assessing the reliability of evidence are sample quality, test continuity, and clarity of results. As regards sample quality, the 45,700 Belgians tested in 1958 were 83% of all the 18-year-old men and the 56,700 in 1967 were about 77% (Bouvier, 1969, pp. 2–3; United Nations, 1979, p. 260). Some of those sampled were older subjects whose testing had been deferred, but this was a constant factor. It should make no difference for the culturally reduced tests: Performance on these tests peaks in the late teens and remains stable throughout the early 20s (Jensen, 1980, p. 235). Test content remained unaltered and the availability of raw scores makes conversion into IQ gains automatic. Belgian gains should also be classed as Status 1, or verified.

France

Samples and tests. Girod and Allaume (1976) report results based on French military examinations in 1949 and a sample of those tested in 1974. Most examinees are 18-year-old men, but some are older, mainly 22-year-olds. In 1949, the mental tests included the Ravens test, a test of aptitude at manipulating verbal symbols, and a test of mathematical knowledge. The French Army's Laboratory of Applied Psychology administered these same tests to a representative sample of recruits from 1974. They selected men randomly from two regiments and compared their performance on the current battery of mental tests with that of the total population of 1974 recruits. Ten subjects were retained from each 20th of the total distribution, from the top 5% down to the bottom 5%, giving a sample of 200 subjects.

Results. Table 3 shows that most of the 1974 sample matched the top half of the 1949 examinees. Consider a soldier who bettered 90% of his fellows in 1949; someone with the same raw score would better only 35.5% of the 1974 sample on the Ravens test, 75% on the mathematical test, and 77.5% on the verbal test. As has been discussed, a table of areas under a normal curve allows centile gains to be converted into standard deviation unit gains. Table 3 makes such comparisons at three cutting lines, averages them to get an overall estimate, and multiplies by 15 (1 *SD* = 15) to get the overall gain expressed in IQ points. Over 25 years, Ravens test gains were 25 points, and mathematics and verbal gains were just over 9 points.

Assessment. Girod and Allaume (1976, p. 121) claim to use

Table 3
France: IQ Gains on Mental Tests, 18-Year-Old Men

Test	1949 Centile	1974 Centile	*SD* unit gain	Average *SD* unit gain	IQ gain 1949–1974
Ravens	90.0	35.5	1.65	1.6745	25.12
	70.0	11.5	1.72		
	50.0	5.0	1.65		
Math	90.0	75.0	0.61	0.6429	9.64
	70.0	44.5	0.66		
	50.0	25.5	0.66		
Verbal	90.0	77.5	0.53	0.6037	9.06
	70.0	48.5	0.56		
	50.0	23.5	0.72		

Note. Data from Girod and Allaume (1976, p. 122). Ravens = Ravens Progressive Matrices Test.

data from virtually the total population of young men. This claim is reasonably accurate for 1974: they selected an excellent sample of all examinees and, inclusive of holdovers from previous years, the examinees were about 97% of all 18-year-old men. Pouvesle (personal communication, December 30, 1985) argues that it is not accurate for 1949. At that time, men were examined at the end of a period of premilitary training, and an entry in the archives asserts that only 30% of all young men were enrolled. He also notes that the military personnel administering mental tests had no specialized training in psychometric techniques. The last point may not be too important in that tests like the Ravens test are relatively easy to administer. Nonetheless, the estimates derived from these data must be labeled tentative, or Status 3. They appear among the strong data because of their comparative value, as will soon be apparent.

However, it would be wrong to dismiss Girod and Allaume completely. When 30% of the relevant age group have been recruited for military training, the sample size matches American military samples during periods of mass testing, for example, World War I and the principal years of the Vietnam war. These samples are regarded as reasonably reliable, usually substandard by about 3 IQ points, for the following reasons. Escape from military service has a class bias, but the correlation between class and IQ is of course not perfect. The lower 30% of the IQ curve's mean is about 17 points (1.159 *SDs*) below the population mean. The Dutch data yield a correlation between father's occupation and son's IQ for this age group of .33 (Leeuw & Meester, 1984, pp. 13 and 16), a value in accord with other data (Flynn, 1987, p. 231). Therefore, the bottom 30% of the total population in terms of class would be substandard by about 6 points ($17.4 \times .33 = 5.8$). Middle-class escape from military service never approaches total evasion, so it would be extraordinary for a 30% sample to be substandard by as much as 5 IQ points.

The estimate in Table 3 that the French gained 25 points on the Ravens test lacks precision. Yet it is highly probable they gained 20 points and almost certain they gained more than 15 points. France has a strong cultural affinity with both Belgium and the Netherlands, and comparative data confirm the plausibility of a Ravens gain of 20 points over 25 years. Projecting all Ravens gains over 30 years for the sake of comparability, the French gained 24 points, French-speaking Belgians gained 22 points, Dutch-speaking Belgians gained 26 points, and the Dutch gained 20 points. The French results also resemble those of neighboring Belgium in another respect: The Ravens gains are far greater than those on other kinds of tests, such as tests of verbal, mathematical, and vocabulary skills, the sort that make up the verbal subtests of IQ tests like the Wechsler group.

Norway

Sample and tests. Prior to 1978, the Norwegian military examined men between their 19th and 20th birthdays; from 1978 to the present, the ages have been between 18 years, 8 months and 19 years, 8 months. A few men tested are younger or older than the usual age; for example, serious illness or continuing education can mean a postponement of a year or two. Rist (1982, p. 34) notes there are few holdovers, for example, 3%–4% of all those tested in 1954. In addition to the usual exceptions, the Norwegians exempt resident aliens, seamen serving on foreign commercial vessels, and students studying in foreign countries. Rist used military data already available to analyze trends from 1954 to 1980, but compiled his own data for 1968 and 1977. Rather than using the total sample for those 2 years, he selected 788 subjects for 1968 and 718 for 1977 by a random method based on the 5-digit registration numbers assigned to each man (Rist, 1982, pp. 28–30, 34–38).

Table 4
Norway: Trends on Mental Tests, 19-Year-Old Men

Year	Test			Average[a]	General ability[b]
	Math	Verbal	Matrices		
1954	100.0	100.0	100.0	100.0	100.0
1961	—	—	—	—	105.8
1963	107.5	—	—	—	—
1965	—	—	—	—	107.1
1968	107.9	108.4	108.8	108.4	109.3
1974	106.1	110.7	110.1	109.0	110.0
1977	105.5	110.5	110.1	108.7	109.8
1980	104.8	109.9	111.4	108.7	—

Note. Data from Rist (1982, pp. 39–51).
[a] Average of scores on the mathematical, verbal, and matrices tests. [b] The general ability score is equivalent to an overall IQ score and is based on a stanine scale.

The mental tests are a mathematical test, with problems presented in prose assuming skills ranging from simple arithmetic to elementary algebra; a verbal test, which consists entirely of word-similarity items; and a matrices test. The matrices test has 36 items adapted from the Ravens but selected to provide a linear increase in difficulty, whereas the Ravens has five groups of items and each group begins with an easy item, so that there is an increase in difficulty only within each group.

Results. The Norwegian military used their 1954 population of examinees as a standardization sample with the mean set at 50 and the standard deviation set at 20. The results for most years are given as standard scores within that convention and can easily be converted into deviation IQ scores with the mean and standard deviation set at 100 and 15, respectively; for example, a standard score of 70 is one standard deviation above the mean (50 + 20), and converts into an IQ score of 115 (100 + 15). Rist gives only raw score results for some years, but there are enough data equating raw scores with standard scores to allow these to be converted as well. The major complication is that the mathematics test was restandardized in 1963 on an examinee population that, thanks to enhanced performance over time, set more demanding norms. Therefore, in order to make post-1963 math scores comparable to other results, they were raised by 7.5 IQ points, and any post-1963 scores that represented performance on the three tests collectively were raised by 2.5 points (Rist, 1982, p. 42).

Table 4 shows that Norwegian IQ trends divide into two periods. From 1954 to 1968, there were similar gains on all three tests, with matrices gains amounting to 8.8 points over 14 years, or 0.629 points per year. From 1968 to 1980, the matrices gains

continued at a diminished rate, 2.6 points over 12 years, or 0.217 points per year, the mathematics results began to show significant losses, and the verbal results were relatively stable.

The Norwegians also sum the scores recruits get on each of these three tests, yielding a measure of "general ability." They report these totals as a score on a stanine scale, a scale that allows one to make percentile comparisons between the current recruits and the 1954 standardization sample. As already discussed, such a comparison between two samples separated in time, plus a table of areas under a normal curve, means that the percentile differences can be converted into standard deviation unit differences and therefore into IQ gains. As a check on the accuracy of all calculations, Table 4 compares the average of the three separate tests with the overall general ability score. The former is always about 1 point below the latter, which makes sense: When recruits made gains over time on all three tests, this made them more of an elite overall, compared to 1954 recruits, than gains on each test taken separately would indicate.

Assessment. The fact that Norway enjoyed lower matrices test gains than other European countries after 1968 does not render the data suspect. Rist (1982, p. 52) emphasizes that enhanced performance over time created an artificially low ceiling and believes this may have been particularly potent in reducing matrices gains. More important, national differences probably come into play, for Scandinavia is not culturally identical with the rest of Europe. The tendency for greater gains on matrices than on verbal tests appears in Norway only after 1968; indeed, Norway shows that matrices gains can continue even when other kinds of mental test yield either stability or losses. Rist (1982, pp. 56–58, 63) notes that the mathematics test losses began when students trained in the new math began to reach military age.

Rist (1982, p. 36) shows that the 26,000 men tested in Norway each year, including holdovers, represent about 83% of all 19-year-old men, and if resident aliens are excluded, about 85%. There has been a slight tendency for the percentage to rise, mainly because of fewer youths serving as seamen. All mental tests were unaltered in content between 1954 and 1980. The results are detailed, and conversion into deviation IQ scores is straightforward, although the restandardization of the mathematics test in 1963 means that when math results after that date are scored against 1954 norms, the values are approximate. The restandardization had no effect on measuring the mathematics losses in that they occurred only after 1968. Norwegian IQ gains qualify as verified, or Status 1.

New Zealand

Samples and tests. Elley (1969) describes the standardization samples used to norm the Otis test in 1936 and to renorm it in 1968. In March 1936, the Otis Intermediate Test of Mental Ability, Form A, was administered to 26,000 schoolchildren, or one-fifth of the entire school population from 10 to 13 years old. Schools were chosen at random to represent all school districts, all kinds of communities, and all types and sizes of public and private schools; all children in each school chosen were tested. In March 1968, the same test was administered to 4,000 schoolchildren from the same age groups. The procedure was the same, except private schools were not included and rather than testing all children in each school chosen, subjects were selected at random from the total enrollment. Analysis of the 1936 results had shown that private schools were a negligible factor, affecting mean performance by less than 1/10th of 1 raw score point.

Table 5
New Zealand: IQ Gains on the Otis Test, Ages 10–13 Years

Age (years)	1936 Sample *M*	1936 Sample *SD*	1968 Sample *M*	1968 Sample *SD*	1936–1968 Gains, (IQ points)[a]
10	26.34	13.64	32.13	14.41	6.19
11	33.76	15.06	41.74	15.17	7.92
12	39.70	15.37	47.83	14.13	8.27
13	44.39	14.92	52.41	13.22	8.55
Average gain					7.73

Note. Data from Elley (1969, p. 145). The Otis test = Otis Intermediate Test of Mental Ability, Form A.
[a] Gains for each age were calculated by taking the difference between the means, dividing by the average of the standard deviations, and multiplying by 15 (the assigned standard deviation value).

Results. As Table 5 shows, the data provide both means and standard deviations and therefore, converting Otis test gains into deviation IQ gains is automatic (as already discussed). New Zealand subjects aged 10–13 years had an average gain of 7.73 IQ points over the 32 years from 1936 to 1968. There is a slight tendency for gains to increase with age. The old Otis Form A was a mix of vocabulary, verbal similarities and opposites, comprehension, information, reasoning, and arithmetic questions and gives a purely verbal IQ score, something analogous to the verbal side of Wechsler tests. The New Zealand rate of gain of 0.242 points per year yields 7.25 points over 30 years, and this clearly resembles the verbal and mathematical gains of other nations, which have ranged from 7 to 12 points per generation.

Assessment. Those who normed and renormed the Otis took advantage of New Zealand's small size to get samples approaching random samples of the entire school population. The test was unaltered and the results present no difficulties. New Zealand gains should be taken as verified, or Status 1.

Canada

Samples and tests. Clarke, Nyberg, and Worth (1978b) report test trends between 1956 and 1977 in Edmonton, Alberta's principal city. In May 1956, the 3,596 children in Grade 3 of the public schools all took the Ravens Coloured Matrices Test, the short form of the California Test of Mental Maturity (CTMM), and various achievement tests. By May 1977, there were 4,688 children in Grade 3, and they all took the CTMM and achievement tests unaltered; a 60% sample known to be representative based on CTMM scores (Clarke, Nyberg, & Worth, 1978a, p. 8) took the Ravens unaltered. Randhawa (1980) reports a similar study from Saskatchewan. In 1958, the Department of Education selected a 15% sample described as a stratified random representation of classrooms in the province. Grades 4–8 took the Otis Quick Scoring Beta, Form A, and

Table 6
Canada: IQ Gains in Two Locales, Ages 9–10 and 13 Years

Locale	Grade	Test	Period	IQ gains	
				By grade	By age
Edmonton	3	Ravens	1956–1977	6.74	8.44[a]
	3	CTMM	1956–1977	9.33	11.03[a]
Saskatchewan	4	Otis	1958–1978	6.98	12.55
	7	Otis	1958–1978	—	6.95

Note. Data from Clarke, Nyberg, and Worth (1978b, pp. 10, 108, and 130) and Randhawa, (1980, pp. 13–15). CTMM = California Test of Mental Maturity. Otis = Otis Quick Scoring Beta, Form A (administered in 1958) and Otis-Lennon, Form J, 1969 version (administered in 1978). Ravens = Ravens Coloured Progressive Matrices, Sets A, A_B, and B.

[a] These values are approximate.

the Iowa Test of Basic Skills, Form L. In 1978, they randomly selected 10% of the classrooms from each school jurisdiction. Grades 4 and 7 took the Otis-Lennon, Form J (1969 version), and the Canadian Test of Basic Skills, Form 4M. The Otis-Lennon revised the Otis Beta by converting it from mainly verbal items into a more conventional test, with about one-third nonverbal items.

Results. The Edmonton data give raw score means and standard deviations for Grade 3 students, and conversion into IQ gains was as follows: The difference between the means divided by the average of the standard deviations gave the gain over time in standard deviation units; multiplying by 15 converted this into IQ points. Deviation IQs normally compare subjects of the same age, but here the comparison is between subjects in the same grade. Because the average age for students in Grade 3 dropped over this period from 9.01 to 8.86 years, age-referenced gains should be somewhat larger. Every year of chronological age in this age group yields a difference of about 12.5 IQ points; that is, applying the standard formula, 9 years ÷ 8 years = 1.125, 1.125 × 100 = 112.5, and 112.5 − 100 = 12.5. The 0.15-year age difference here adds about 1.7 points: 9.01 years ÷ 8.86 years = 1.017, 1.017 × 100 = 101.7, and 101.7 − 100 = 1.7. The Saskatchewan data give both grade-related and age-related gains for Grade 4. The latter gains are greater by 5.57 IQ points, presumably because of a decline in age of about 6 months over the period. There are only age-related gains for Grade 7.

Table 6 shows that Edmonton children gained from 6.74 (grade related) to 8.44 (age related) points on the Ravens test over the 21 years between 1956 and 1977. Projected over 30 years, this totals about 10–12 points, the lowest Ravens gain for any nation with strong data. Table 6 also gives results for the CTMM, which introduces a new kind of test. The CTMM gives essentially a Stanford-Binet IQ (Clarke et al., 1978b, p. 108), that is, a full-scale IQ based on a mix of verbal and performance items. It is broader than a purely verbal test like the old Otis and verbal subtests, with vocabulary or mathematical items like the typical military test. Yet it is not culturally reduced like the Ravens. Edmonton children gained from 9.33 (grade related) to 11.03 (age related) points of Stanford-Binet IQ between 1956 and 1977. Projected over 30 years, this is about 13–16 points. Wechsler-Binet data from other nations show that this is typical. Finally, Table 6 shows that children in Edmonton and Saskatchewan made roughly similar gains, although on different kinds of tests. Alberta and Saskatchewan are neighboring prairie provinces and similar, except that Alberta is oil-rich.

Canadian children have made gains on both IQ tests and academic achievement tests. In Edmonton, they gained 3.10 points (grade related) on five academic skills relevant to verbal IQ (Clarke et al., 1978b, p. 8). In Saskatchewan, the shift from one achievement test to another makes measuring trends difficult, but Randhawa (1980, p. 6) argues that indirect comparisons at least rule out a decline in academic skills.

Assessment. The sample quality, test continuity, and clarity of results are as near perfect as possible for Edmonton. When dealing with a city rather than a nation, selective migration over a period of 21 years becomes a significant possibility, and Edmonton grew from 226,000 to 471,000 during that time. However, Clarke et al. (1978b, pp. 17–18) compared schools that existed in both 1956 and 1977 with new schools and found no significant score differences. The Edmonton gains qualify as verified, or Status 1.

Saskatchewan is a much larger political entity and its population was relatively stable: It grew from 900,000 to 950,000 between 1958 and 1978. The samples for those years are described as being nearly random. However, replacing the original Otis with the Otis-Lennon test meant a transition from mainly verbal items to one-third nonverbal items. To compare results on the two tests, Randhawa (1980, p. 15) used the recommended table from the 1969 Otis-Lennon handbook. The table appears sound: It is based on 1,144 subjects (Grade 4) and 1,164 subjects (Grade 7) who took both tests in counterbalanced order (Otis & Lennon, 1969, pp. 37, 46). There is little doubt that children in Saskatchewan have enjoyed sizable IQ gains since 1958. Yet the actual estimates, when content is altered to this extent, must be put somewhere between probable and tentative, that is, Status 2/3.

The United States

Samples and tests. Between 1932 and 1978, the Stanford-Binet and Wechsler organizations selected seven nationwide samples to norm a variety of tests, four tests covering different age ranges—although always with some overlap—and revised versions of three of them. Flynn (1984c) found that when the same subjects took two Stanford-Binet or Wechsler tests, they almost invariably got lower scores on whichever test had been normed at the later date. This meant that going from past to present, standardization samples had set higher and higher standards of performance, standards a subject found more difficult to meet. After analyzing 73 studies containing almost 7,500 subjects with ages ranging from 2 to 48 years, Flynn concluded that white Americans had made IQ gains from 1932 to 1978 at a roughly uniform rate and at similar rates for all ages. He confined himself to studies with normal subjects, as gifted and retarded subjects give eccentric results, and translated all scores into a uniform convention based on the mean of white Americans set at 100 and the standard deviation set at 15.

Results. Table 7 traces the improved performance of Stan-

Table 7
White Americans: IQ Gains as Measured by Stanford-Binet and Wechsler Standardization Samples, Ages 2–75 Years

Year[a]	Sample	*M* IQ: Actual	*M* IQ: Smoothed[b]
1932	Stanford-Binet	100.00	100.00
1947/1948	WISC	106.62	104.65
1953/1954	WAIS	107.94	106.45
1964/1965	WPPSI	108.71	109.75
1971/1972	Stanford-Binet	109.89	111.85
1972	WISC-R	113.37	112.00
1978	WAIS-R	114.31	113.80

Note. From "The Mean IQ of Americans: Massive Gains 1932 to 1978" by J. R. Flynn, 1984, *Psychological Bulletin, 95,* p. 45. Copyright 1984 by the American Psychological Association, Inc. Adapted by permission. The ages covered, WISC mean IQ, and WAIS-R mean IQ have been revised as described in the text. WAIS = Wechsler Adult Intelligence Scale (WAIS-R = WAIS-Revised). WISC = Wechsler Intelligence Scale for Children (WISC-R = WISC-Revised). WPPSI = Wechsler Preschool and Primary Scale of Intelligence.

[a] The year represents the midpoint of the years during which the sample was actually tested. [b] The smoothed means give the mean IQ for each year, assuming a constant rate of gain of 0.300 IQ points per year.

ford-Binet and Wechsler standardization samples over time. The original version of this table (Flynn, 1984b, p. 45) has been revised to include a correction for the fact that the Wechsler Intelligence Scale for Children (WISC) sample was substandard by 0.86 IQ points because 2.5% of the sample were institutionalized mentally retarded children (Flynn, 1985, p. 238), and four additional studies of the Wechsler Adult Intelligence Scale (WAIS) and the WAIS-Revised (WAIS-R; Lippold & Claiborn, 1983; Mishra & Brown, 1983, p. 756; Smith, 1983, p. 416; Urbina, Golden, & Ariel, 1982). Table 7 is based on data using from 100 to 2,300 subjects to measure the differential performance of the standardization samples. This gives an estimated rate of gain of 0.311 IQ points per year, for a total gain of 14.31 points between 1932 and 1978. However, when data with 400 to 2,300 subjects are used, the rate falls to 0.295 points per year. This estimate rounded off to 0.300 produces a smoothed sequence of gains, with the total gain amounting to 13.80 points over a period of 46 years. Projected over 30 years, American gains would be 9 IQ points on this kind of test, lower than both the Edmonton Stanford-Binet gains and the Wechsler gains from other nations, albeit nations with weak data.

Assessment. The Stanford-Binet and Wechsler organizations take great pains to select representative samples of Americans, using census data to stratify for age, sex, geographic region, occupation, and urban–rural balance. Flynn (1984c, pp. 45–46) analyzed the test manuals for systematic bias and found only one: a geographic bias with a slight tendency to produce too low rather than too high an estimate of IQ gains. Nonetheless, these samples cannot match comprehensive military samples or samples of near-random quality. The content of the Binet test altered little over this period; Wechsler tests naturally vary their content for different age groups, and both the children's and adult tests were revised.

Despite reservations about each Wechsler and Binet sample taken individually, collectively they show that sampling error cannot account for the overall pattern of IQ gains over time. In Table 7, the rank order of standardization samples by quality of performance is a perfect match for the chronological order, and the odds against this arising by sampling error are 7 factorial or 5,040 to 1. As for content, the revised Wechsler tests inherited over 80% of their items from the earlier tests, and the balance between verbal and performance subtests was essentially unchanged. The correlation coefficients between various Wechsler tests are high, ranging from about .80 to .90 (Wechsler, 1974, pp. 11, 48–50; Wechsler, 1981b, pp. 11, 47). American IQ gains cannot qualify as verified, but they are probable, or Status 2.

Weak Data

East Germany

Samples and tests. G. Mehlhorn and H.-G. Mehlhorn (1981; H.-G. Mehlhorn, 1981) have begun to publish the results of studies done by the Central Institute for Youth Research in Leipzig. H.-G. Mehlhorn (personal communications, January 15 and February 24, 1985), who is responsible for the section on intelligence, has furnished valuable additional information by way of scholarly correspondence.

In 1968, the Institute selected 12 schools lying on a diagonal line from West to East across Leipzig, a method designed to obtain schools representative of the city as a whole. They administered the Ravens Progressive Matrices Test to all sixth-grade classes, children 11 to 12 years old. These same children were tested yearly for 4 more years as they progressed from the 6th to the 10th grade. The sample totaled 944 subjects, of whom 875 were from typical schools and 69 from an elite school for superior students. In 1978/1979, the Institute selected 18 schools, 11 from the original group (the elite school having been dropped), plus 7 from new communities surrounding Leipzig. They administered the Ravens to classes streamed as typical, omitting both elite and substandard classes. Rather than following a particular group of children as they progressed through the grades, all grades from 6 to 10 were tested during 1978/1979. Sample sizes for the various grades ranged from 700 to 1,100, except for the 10th grade, which had only 410 subjects.

Results. Table 8 presents the gains of Leipzig children aged 11–16 years on the Ravens test. The subjects tested from 1968 to 1972 as they progressed through the grades naturally benefited from practice effects after the first testing. Some of the subjects from 1978/1979 had previously taken the Ravens, the percentage varying from 21% to 33% depending on the school grade. The Institute found, using experimental groups, each with 120 subjects, that going from a first testing to a second produced a practice effect of 2 raw score points, with no further increment thereafter. Therefore, in Table 8, the raw score gains made at various grade or age levels have been adjusted accordingly. For example, 10th graders, aged 15–16 years, who were tested in 1972 had a raw score of 46.90 answers correct; because all students benefited from practice effects, the mean was lowered 2 points to 44.90. Those tested in 1978 had a raw score

Table 8
Leipzig: IQ Gains on the Ravens Progressive Matrices Test, Ages 11–16 Years

Grade	Age (years)	Period	Gains				
			Raw score unadjusted	Raw score adjusted[a]	IQ points[b]	Years	Rates[c]
6	11–12	1968 to 1978/1979	7.63	7.63	15.47	10.5	1.473
7	12–13	1969 to 1978/1979	3.31	4.81	12.15	9.5	1.279
8	13–14	1970 to 1978/1979	3.26	4.84	11.58	8.5	1.362
9	14–15	1971 to 1978/1979	2.98	4.49	8.00	7.5	1.067
10	15–16	1972 to 1978/1979	3.00	4.33	9.16	6.5	1.409

Note. Data from H.-G. Mehlhorn (1981) and G. Mehlhorn & H.-G. Mehlhorn (1981). These values are more exact than the published values thanks to H.-G. Mehlhorn (personal communications, January 15 and February 24, 1985).
[a] Raw score gains were adjusted to allow for practice effects. [b] IQ gains scored against 1979 Ravens test norms. [c] IQ points per year.

of 49.90; because one third benefited from practice effects, the mean was lowered 0.67 points (2 ÷ 3) to 49.23. The adjusted scores reveal a gain of 4.33 raw score points (49.23 − 44.90), and using the 1979 British norms, this converts into 9.16 IQ points. Because they excluded both elite and substandard subjects, the Leipzig samples have restricted variance and using local norms would inflate IQ gains.

The period in years during which different age groups made their gains varies. Therefore, the gain for each age has been divided by the period to get a rate in terms of IQ points per year. Table 8 shows that the rates are very similar except for age 14–15, which is a bit lower. When all rates are projected over a decade, they suggest that Leipzig children gained 10–15 IQ points on the Ravens test during the 1970s, an estimate identical to that made by the Institute.

Assessment. The Leipzig children were not selected for the purpose of measuring IQ gains over time, and the 1968–1972 and 1978/1979 samples do not match. There are two obvious sources of possible bias. First, focusing on the 11 schools common to the two samples, these schools may have been representative of Leipzig in 1968 but not a decade later because of unusual social changes in their catchment areas. Mehlhorn asserts that the social level of parents rose over the period, but that the rise was slight and matched nationwide trends. Second, the inclusion of the 12th elite school in the 1968–1972 sample would inflate performance; and the inclusion of 7 schools from surrounding communities in 1978/1979, schools slightly superior to the original 11, would also inflate performance. From data supplied by Mehlhorn, I have calculated that the earlier raw score means were raised by 0.80 points and the later by 0.66 points. Taken together, these biases would entail an underestimate of gains over time by 0.14 raw score points, or 0.28 IQ points.

It is clear that Leipzig has enjoyed substantial Ravens gains, but their magnitude is less certain. Despite the Institute's care to control for bias after the event, possible sources of bias are numerous enough to suggest caution. A conservative assessment of the status of these gains would be tentative over probable, that is, Status 3/2.

Great Britain

Samples and tests. The Ravens Progressive Matrices Test has remained essentially unaltered, so it is possible to compare the raw score performance of the original standardization samples, selected by J. C. Raven in 1938 and 1940, with that of the recent standardization sample, selected by J. Raven in 1979.

In 1938, J. C. Raven (1941) took a random sample of 919 from the birth registers of the County Borough of Ipswich. Of these, 660 were still attending school in Ipswich in 1938, now 8- to 14-years old, and they were tested individually. For each child tested individually, three of the same age were chosen randomly from the same classroom for group testing, and 1,407 of the 1,980 chosen were actually tested. In 1940, J. C. Raven also tested 3,665 subjects from 20 to 30 years old at a World War II training depot for militiamen, the men having already been passed as physically fit. J. Raven (1981, p. 56) says they were 22 years old, presumably their median age. Their occupational distribution matched that of the parents listed on the Ipswich birth registers and their educational distribution was typical of British men in general of their age.

In 1979, J. Raven (1981, pp. 16–22) restandardized the test. He selected seven areas designed to give a representative sample of Great Britain as a whole, based on a 40-factor analysis. He then chose 169 schools that would maintain the representative character of the sample, 145 of which agreed to cooperate, to which 3 others were added; the subjects tested were chosen randomly from lists of all students 6 to 16 years old.

Aside from the Ravens test data, the literature for Great Britain provides two classic studies, both from an earlier period than any covered thus far. First, the Scottish Council for Research in Education (SCRE) study: On June 1, 1932, all Scottish children aged 11 years and attending school on that date took Moray House Test No. 12. It consisted primarily of a verbal section of 76 items, and only that section was scored for future reference. On June 4, 1947, the same sort of sample was given the same 76 items (SCRE, 1961, pp. 13–17). Second, Cattell's (1950) study of the city of Leicester: Cattell aimed at testing every 10-year-old child in the city in both 1936 and 1949 and appears to have largely succeeded, although the 1949 sample was probably more complete. Both samples took the Cattell Scale 1, Form A, a culturally reduced test of fluid intelligence.

Results. The Ravens's data (J. Raven, 1981, p. 27; J. C. Raven, 1941, p. 141) give percentile comparisons between the earlier and later standardization samples for all ages from 8 to 14 years. Raw score means and standard deviations were derived

Table 9
Great Britain: IQ Gains on Three Mental Tests

Test	Age, years	IQ gains	Period	Years	Rates[a]	Place[b]
Ravens	8–11	10.28	1938–1979	41	0.251	Ipswich/Great Britain
Ravens	12–14	4.37	1938–1979	41	0.107	Ipswich/Great Britain
Ravens	8–14	7.75	1938–1979	41	0.189	Ipswich/Great Britain
Ravens	20–30[c]	7.07	1940–1979	39	0.181	Great Britain
Cattell	10	0.91	1936–1949	13	0.070	Leicester
Moray House	11	2.16	1932–1947	15	0.144	Scotland

Note. Sources: Cattell (1950), Scottish Council for Research in Education (1961), J. Raven (1981), and J. C. Raven (1941).
[a] IQ points per year. [b] The gains for children on the Ravens Progressive Matrices Test are based on a comparison of a 1938 Ipswich sample with a 1979 sample drawn from seven areas of Great Britain. The adult gain on the Ravens test is based on a comparison between a 1940 sample of militiamen, representative of British men in general in terms of schooling, and the 1979 sample. [c] The gain for ages 20–30 on the Ravens test is based on a comparison of subjects of those ages in 1940 and subjects aged 15½ in 1979, to the disadvantage of the latter. The gain has been adjusted slightly to compensate.

for both samples, based on comparisons at the 10th, 50th, and 90th percentiles. The raw score gains were converted into IQ gains by the usual method. The gains for mature subjects are complicated by the fact that the 1940 sample had a median age of 22 years and, therefore, had reached peak performance. The 1979 sample stops at 15 1/2 years, at least 3 years short of their peak, and, therefore, to facilitate a comparison, one half of a raw score point was added to their mean. This gives quite a conservative estimate of their performance at maturity.

As Table 9 shows, between 1938 and 1979, British children aged 8–11 years had an average Ravens test gain of just over 10 IQ points, whereas those aged 12–14 years averaged over 4 points; together these two groups averaged 7.75 points. Adult gains were 7.07 points between 1940 and 1979. Projected over 30 years, the gain for mature subjects would be 5.43 points, far less than the Ravens gains of any other nation for which data exist. Gains on other tests were at an even lower rate: Leicester 10-year-olds gained about 1 point on the Cattell Scale between 1936 and 1949; Scottish 11-year-olds gained 2 points on the Moray House test between 1932 and 1947. These estimates differ very slightly from the reported estimates. On the assumption that the study samples were more representative than any other, their means and standard deviations were used to convert to IQ scores by the usual method. The difference between the means was divided by the average of the standard deviations, and then that was multiplied by 15.

Assessment. British Ravens test gains are so low as to suggest the possibility of some inhibiting factor, for example, an artificially low ceiling. Perhaps British subjects have been getting so many Ravens items correct as to leave little room for improvement. The Leipzig 1978/79 sample casts doubt on this hypothesis: They outscored their 1979 British counterparts by an extra three items correct, which amounts to 6 IQ points (H.-G. Mehlhorn, personal communication, February 24, 1985). The Ravens data are not suspect simply because the gains are so low. There is no reason why different nations should not show different rates of gain. Yet these data are suspect because of sampling problems. The 1979 sample covered all of Great Britain, whereas the 1938 sample was representative only of Ipswich, a city of about 100,000 located 60 miles northeast of London. If Ipswich was atypical by say 5 points, British Ravens gains could be anywhere from 2.75 to 12.75 IQ points over the 41 years. The 1940 adult sample may be a bit better in that although local, it was a war-time sample representative of the nation as a whole in terms of schooling. Nonetheless, all estimates of Ravens gains must be described as tentative, or Status 3.

The Cattell data apply only to the city of Leicester. Yet within that locale, the samples were excellent and the estimated gain should be taken as verified, or Status 1. The fact that Cattell gains were low may seem to add credibility to the Ravens data, reasoning that local rates of gain are less likely to be atypical of the nation as a whole than a local mean IQ. However, the Cattell data are entirely pre-1950, whereas the Ravens data apply mainly to the post-1950 generation. British gains in the current generation may well not be comparable to those of a previous generation. The same high status and the same limitation on its comparative value hold for the pre-1950 Scottish gains, this time on a purely verbal IQ test.

Lynn and Hampson (1986) have found eight studies that can be used to measure British IQ gains since 1950. Weighting the studies in terms of how much of the period from 1950 to 1980 they cover gives these results: four nonverbal tests (including the Ravens data already assessed) with a rate of gain of 0.227 IQ points per year, or 6.80 points over 30 years, and four verbal tests with a rate of gain of 0.177 IQ points per year, or 3.51 points over 30 years. The standardization samples involved cannot be linked into an overall pattern of enhanced performance over time, that is, each study must be assessed separately. Lynn and Hampson give no information about the quality of the samples. Still the odds against random sampling error giving gains in every one of eight separate cases are 2^8, or 256 to 1.

In sum, excellent data show that in the pre-1950 generation, British IQ gains were less than 5 points on both culturally reduced and purely verbal tests. A mass of data suggests that the post-1950 generation gained a bit more than 5 points on culturally reduced tests, a bit less on verbal tests. At present, Britain lacks post-1950 data for any particular kind of test that gives

better than a tentative estimate. However, more detailed information about sample quality could raise the estimates to a higher status.

Australia

Samples and tests. Marion de Lemos of the Australian Council for Educational Research (ACER) has found six tests that appear to exhibit IQ gains. Some of these are clouded by practice effects and testing at different times of the school year. I have chosen three cases without these handicaps and with tests that allow comparisons with other nations.

The Jenkins Non-Verbal Test is a test of fluid intelligence that is broader than the Ravens in that it has figural similarities and sequences as well as matrices. The first standardization was during late June and early July 1949 on a sample of over 4,000 schoolchildren aged 10–14 years. The ACER (1958a, pp. 23–24) designed the sample to include the correct proportion of children from each state, from private and public schools within each state, from classes of schools within each state, and from grades within schools; moreover, a random selection of schools within each type was made. The design aimed at 900 subjects at each of nine age levels, but failure to secure that target forced them to base the norms on groups of 300–660 instead. These subjects were chosen from those available to get an accurate representation of the state-by-state age–grade distribution. The second standardization was in April 1981. The ACER (1982, pp. 44–45) selected a sample of 1,299 schoolchildren aged 10, 12, and 14 years, who were drawn from 86 schools and were representative of the various states, kinds of school, grade level, and sex. The Jenkins test was administered to the 382 subjects of the Victorian section of the sample. The ACER's (1982, pp. 45–46) data show that on the ACER Intermediate Test F, a mental test about 60% verbal and 40% quantitative, the Victorian section was substandard by about 1 IQ point.

As for the Ravens test, the original Australian standardization was piecemeal. For 10- to 11-year-olds, 359 schoolchildren from Melbourne were used to equate the Ravens with the Junior Non-Verbal Test, a test similar to the Jenkins, with norms based on a 1949 standardization that is also similar to the Jenkins. For 12- to 14-year-olds, 496 schoolchildren from Melbourne were used to equate the Ravens with the Jenkins itself and its 1949 norms. For 18-year-olds, the Ravens was directly standardized on 783 national service trainees from Victoria, a group selected as representative in terms of educational level, occupational category, and geographical location of all 18-year-olds accepted for training between August 1951 and June 1953, the Korean War period (de Lemos, 1984). Norms for the missing ages, those between 14- and 18-years old, were based on interpolation between the Jenkins and the national service data. The Ravens has not yet been restandardized, but between 1971 and 1980, seven samples totaling 3,571 subjects, ages 10–16 years, were tested (M. M. de Lemos, personal communication, September 18, 1984). Most of those aged 10–14 years were schoolchildren from Melbourne. They were chosen to cover the different kinds of schools in the state of Victoria and a range of socioeconomic status, but these variables were not matched with the state as a whole. Those aged 14 to 16 years were randomly chosen from secondary students in the state of New South Wales.

Table 10
Australia: IQ Gains on Mental Tests, Ages 10 to 16 Years

Age (years)	Test	Period	IQ gains	Years	Rates*
10–14	Jenkins	1949–1981	15.67	32	0.490
10–16	Ravens	1950–1976	8.76	26	0.337
10–14	Otis	1936–1949	5.50	13	0.423

Note. Sources: Australian Council for Educational Research (1958a, p. 19) and de Lemos (personal communications, September 18, 1984 and May 15, 1985). Jenkins = Jenkins Non-Verbal Test. Otis = Otis Intermediate Test of Mental Ability, Form A (normed in 1936) and Form D (normed in 1949). Ravens = Ravens Progressive Matrices Test.
* IQ points per year.

The Otis Intermediate Test, Form A was standardized in March 1936 on a nationwide sample of 30,573 schoolchildren aged from 8 to 14 years, from public schools only. The selection of schools was not random, although school size was taken into account, and no data permit an evaluation of the effect of omitting church schools. Personal knowledge was used to secure a full range of socioeconomic status (McIntyre, 1938, pp. 23–27). The Otis Intermediate Test, Form D, a later revision, was standardized on the 1949 nationwide sample of 4,000 schoolchildren aged from 10 to 14 years, already described under the Jenkins. The ACER (1958b, p. 19) compared the later and earlier forms to determine whether the later standardization sample had set more demanding norms. Presumably, subjects were administered both tests, but no information is given. As for test content, Form A was 82% verbal items and Form D reduced this to 64%, with greater emphasis placed on quantitative reasoning (ACER, 1982, p. 4).

Results. As shown in Table 10, the 1981 Jenkins test sample outperformed the 1949 sample by almost 16 IQ points (M.M. de Lemos, personal communication, September 18, 1984). Projected over 30 years, this gives the post-1950 generation of Australian schoolchildren a gain of almost 15 points on a culturally reduced test of fluid intelligence, less than continental Europe, but still impressive. The same projection would put post-1950 Ravens test gains at 10 points. In Table 10, I have merged the Ravens data of de Lemos for simplicity's sake. The rate of gain of all ages was close to 0.337 points per year. Taking 1950 and 1976 as approximating the period between the standardization years and when later samples were tested, this rate of gain gives a total of 8.76 points over 26 years. This estimate corrects for an error in the published Ravens data. M. M. de Lemos (personal communication, May 15, 1985) has pointed out that de Lemos's (1984) Table 1 underestimates the performance of the later samples of 14- to 16-year-olds.

Table 10 also gives an estimate of Australian gains on the Otis Intermediate Test, based on the ACER's (1958b, p. 19) assertion that comparing the 1936 Form A norms and the 1949 Form D norms showed that the latter were more demanding by 5 or 6 points. Whatever the status of this estimate, it applies only to the pre-1950 generation.

Assessment. The Jenkins Non-Verbal Test was not altered between 1949 and 1981. The ACER stratified its 1981 sample for the major relevant variables and its quality was at least as good as the American Wechsler-Binet samples. The fact that only the Victorian section took the Jenkins means little given that this section is known to have not been elite. The 1949 sample is suspect because of school noncooperation in meeting the target figures for subjects at various age levels. There is no reason to believe schools furnished atypical subjects, and the fact that those tested were stratified for age and grade within each state should eliminate the possibility of too many children being ahead or behind for their age. Still, the estimated gain of 16 points over 32 years could easily be anything from 12 to 20 points and must be classed as tentative, or Status 3. The Ravens test was essentially unaltered, but all samples were only roughly representative and the estimated gain is speculative, or Status 4. The Otis test gains are speculative at best. They suffer from the defects of the 1936 rough sample, unknown studies as a measure of differential sample performance, and substantial changes in test content.

Japan

Samples and tests. Lynn and Hampson (1986) report a study in which 112 subjects took both the Japanese WISC and WISC-R. As discussed in the section on the United States, the difference between their WISC and WISC-R scores is a measure of the differential performance of the Japanese standardization samples used to norm these two tests. This study stands alone. However, the combined results of Lynn and Flynn can provide an indirect measure of how much the Japanese WISC-R standardization sample improved on the WISC sample, at least in terms of performance IQ.

The five performance subtests of the American WISC were essentially unaltered in the Japanese version. Therefore, Lynn (1977) could score the Japanese WISC standardization sample of 1951 against the U.S. norms. The same American WISC-R subtests were unaltered and, therefore, Lynn (1982) also scored the Japanese WISC-R standardization sample of 1975 against the U.S. norms. Lynn's results show that the 1975 Japanese sample outscored the 1951 Japanese sample by 7.2 IQ points when both are scored against the white Americans who were their contemporaries. If Americans had made no gains during this period, that would represent the Japanese gain. Yet Flynn (1984a, p. 287) has shown that white Americans gained 8.3 IQ points during the period from the WISC to the WISC-R. Therefore, the Japanese results represent an extra gain, a gain over and above those that Americans were making, and the sum of the Japanese and American results yields an estimate of the total gain the Japanese truly made.

As stated previously, all of these results are merely ways of comparing the Japanese WISC and WISC-R standardization samples, so the quality of those samples is of direct relevance. Elsewhere it has been argued that the 1975 WISC-R sample was elite by probably 4 or 5 points because of a known urban bias and a suspected SES bias (Flynn, 1984b). The 1951 WISC sample was probably elite to much the same degree. The same organization selected 1,070 children with the same sort of bias toward schools associated with teaching colleges, and because all of these schools were from Tokyo, the urban bias should be at least as great (Kodama & Shinagawa, 1953; Schull & Neel, 1965, pp. 289–292).

Table 11
Japan: Wechsler IQ Gains, 1951–1975, Ages 6–15 Years

Gains	Verbal	Performance	Full scale
All ages			
Japanese data	16.02	17.36	20.03
Japanese and U.S. data	—	15.45	—
By age (years)	U.S. gain	Japanese: Extra gain[a]	Japanese: Total gain[a]
6	7.7	3.9	11.6
7	8.1	5.9	14.0
8	8.3	7.4	15.7
9	8.2	8.4	16.6
10	8.4	8.4	16.8
11	8.4	8.4	16.8
12	8.1	8.4	16.5
13	8.3	5.9	14.2
14	8.5	8.9	17.4
15	8.7	6.2	14.9
Averages	8.3	7.2	15.5

Note. Data from Flynn (1984a, p. 287), Lynn (1977, 1982), and Lynn and Hampson (1986).
[a] Performance gains only.

Results. When American IQ gains on Wechsler tests were put at about 9 points in one generation, they seemed massive. The Japanese are the first of a series of nations who may have gained at double that rate. As Table 11 shows, taking all ages from 6 to 15 years together, the Japanese gained 20 points of full-scale IQ in only 24 years, based on the one direct comparison of the WISC and WISC-R standardization samples. The verbal gain was 16 points and the performance gain was 17 points, a value that the Lynn-Flynn indirect comparison comes very close to matching. The latter data also give estimates for each age and reveal no tendency for gains to diminish from younger to older subjects.

Assessment. The Lynn-Flynn data pose a number of technical problems, such as allowing for differences between the American WISC and WISC-R samples. Solutions have been found (Flynn, 1984a, pp. 288–289) but are not detailed here. These indirect data are merely a stopgap until the Japanese accumulate more direct comparisons, which will no doubt occur shortly. A series of direct comparisons will give exact estimates of the extent to which the Japanese WISC-R sample outscored the Japanese WISC sample. However, they cannot solve the problem posed by the elite bias of those samples. The guess that they were elite to about the same degree will never be more than that. The chief hope is for better data based on better samples, and Lynn and Hampson (1986) have made a start. The Kyoto NX 9–15 intelligence test is similar to the Primary Mental Abilities Test used in the West and has both verbal and performance subtests. It was standardized in Kyoto in 1954, and in 1972, eight schools were retested. The sample of 711 children aged

from 10–11 years gained 16.56 IQ points over those 18 years; a smaller sample tested in 1963 suggests that gains were greater before that year versus after. Projecting the Kyoto gains over the same 24 years covered by the Wechsler data gives the Kyoto test 22 points, the Wechsler test 20 points. This sort of consistency pushes Japanese gains to the border between tentative and speculative, or Status 3/4.

West Germany and Switzerland

Samples and tests. Schallberger (1985) found three studies in which subjects took both the West German or Hamburg WISC (*Hamburg-Wechsler-Intelligenztest für Kinder* [HAWIK]), normed circa 1954, and the West German WISC-R (HAWIK-R), normed circa 1981. He added a fourth study of his own, bringing the total number of subjects to 257, from 7 to 15 years old. At least half of these subjects and perhaps as many as three-fourths were from 7 to 9 years old, precision being difficult because two studies are vague about their subjects' age. Once again, these studies are merely a device for comparing the West German WISC and WISC-R standardization samples.

The 1954 WISC sample consisted of 1,500 children from 6 to 15 years old, and care was taken to get an accurate stratification by kind of school (Priester, 1958, pp. 32–41). West German schools are part of a hierarchical system with divisions analogous to Britain's secondary modern (leave school at 15 to take up a trade) and grammar schools (college preparatory). The occupational profile of the parents of the children tested corresponds to the general population. Although they selected schools from varied locations, they do not claim a representative geographical distribution. About 83% of the sample were urban residents, as compared to 60% of the general population. They argue that urban–rural differences may not be too significant because many rural children attend urban schools. The 1981 WISC-R sample consisted of 1,898 subjects, and this time the geographical distribution was better, with 20 testing locations evenly distributed over West Germany, including West Berlin (Tewes, 1983, pp. 29–30). Half of the testing locations were in villages and towns with less than 50,000 people. They say that stratification for kind of school is not accurate for each separate German state; the reader is left to infer that it is accurate for the nation as a whole. The testing locations were given instructions about getting representative subjects in terms of SES, but they are forbidden to publish the data.

Schallberger (1985, pp. 6, 8–9, and 22) also presents results from German-speaking Switzerland. In 1977, a Zurich sample took the West German WISC: The sample was 150 nine-year-olds and 150 twelve-year-olds deemed representative based on test scores. In 1984, a sample from the canton of Solothurn took the same test: 120 subjects from 7 to 9 years old comprising all students from eight randomly selected classes. Both Swiss samples were scored against the norms set by the West German WISC sample of 1954.

Results. Table 12 presents results for the four samples discussed. As for West Germany, the 1981 WISC-R sample set much higher norms than the 1954 WISC sample, signaling a gain of 20 points of full-scale IQ over a period of 27 years; performance gains were still higher and ran at twice the rate of verbal gains. These values are based on a weighted average for all studies in which subjects took both tests. Schallberger (1985) adjusted some results for a variety of practice and regression effects, but makes a convincing case that these adjustments are appropriate. As for Switzerland, Schallberger believes that the fact the 1977 Zurich sample outscored the 1954 West German WISC sample argues for a massive Swiss gain of 15 points over that period. This assumes of course that Swiss and West German children were roughly equal in 1954. The samples internal to Switzerland show 1984 Solothurn only 1 point above 1977 Zurich.

Table 12
Four Samples from West Germany and Switzerland: Year Tested and Levels of Performance

Sample	Age (years)	Year	IQ		
			Verbal	Performance	Full scale
West German, WISC	7–15[a]	1954[b]	100	100	100
Zurich, WISC	9 and 12	1977	108	120	115
West German, WISC-R	7–15[a]	1981[b]	111[c]	125[c]	120[c]
Solothurn, WISC	8–9	1984	107	122	116

Note. Data from Schallberger (1985). WISC = Wechsler Intelligence Scale for Children. WISC-R = WISC-Revised.
[a] At least half of the 257 subjects used to compare the West German WISC and WISC-R samples were 7–9-years old. [b] Estimated as tested 2 years before publication. [c] Estimates of how the WISC-R sample would have performed against the West German WISC norms.

Assessment. The case for Swiss IQ gains rests on the assumption that Swiss and West German children were roughly equal in 1954. However, strong data from elsewhere imply the possibility of significant national differences: Dutch gains on the Ravens test were about 18 points between 1954 and 1980 (see Table 1) and Norwegian gains on their matrices test were only 11.4 points (see Table 4). Therefore, a national difference of 5 points was quite possible over that period. Admittedly, Swiss-Germans and West Germans are much more alike than are the Dutch and Norwegians. It seems best to grant the possibility that Swiss children have made sizable gains, while emphasizing the radical limitations of the data. This dictates an estimate of from 10 to 20 points between 1954 and 1977, but with only the most speculative status, that is, well down into Status 4. The estimate of little gain after 1977 is no better. To compare samples from Zurich and Solothurn is rather like comparing Boston with rural Ohio. Zurich canton houses the city of that name, the largest and most cosmopolitan city in the country, whereas Solothurn canton has no city of more than 20,000 and lost 3% of its population during the decade prior to 1984.

West German gains rest on the reliability of the West German WISC and WISC-R samples. The stratified samples of test publishers never attain the quality of near-exhaustive military samples: if American gains rested on only two such samples rather than seven, they would have to be called tentative. The West German WISC and WISC-R samples were done with reasonable care, but the stratification falls short of their American

Table 13
University of Vienna Clinic: IQ Gains on the West German WISC (HAWIK), Ages 6–15 Years

Year	IQ		
	Verbal	Performance	Full scale
1962	97.8	104.6	101.2
1965	103.7	106.8	105.8
1968	102.2	105.9	104.4
1971	105.6	107.6	107.3
1974	106.0	112.9	110.3
1977	100.5	106.0	104.0
1979	103.0	112.3	108.4

Rationale of estimates[a]	Estimates of full scale IQ gains			
	1962–1974	1974–1977	1977–1979	1962–1979
1974–1977, unadjusted	9.1	−6.3	4.4	7.2
1974–1977, adjusted	9.1	−1.4	4.4	12.1
1974–1977, discounted	9.1	2.9	4.4	16.4

Note. Data from Schubert and Berlach (1982, pp. 256–257). HAWIK = Hamburg-Wechsler-Intelligenztest für Kinder. WISC = Wechsler Intelligence Scale for Children.
[a] The various methods of estimating gains for 1974–1977 are described in the text.

counterparts. West German IQ gains should be put on the border of tentative and speculative, or Status 3/4.

The fact that East German children may have made massive gains on a culturally reduced test of fluid intelligence prompted a search for West German data on similar tests. The results are included for the sake of completeness, but their speculative nature is obvious. H.-G. Mehlhorn (personal communication, May 20, 1985) gave his subjects four subtests from the Horn Performance test, subtests that collectively constitute a culturally reduced test of fluid intelligence. They scored 9.13 IQ points above the West German standardization sample of 1961. He gave the same subjects the Ravens test, and this time they scored 0.63 IQ points below the West German standardization sample of 1978 (Kratzmeir & Horn, 1979, p. 25; H.-G. Mehlhorn, 1981, p. 163—Mehlhorn's values must be lowered slightly to account for practice effects). It looks as if West German children from 12½- to 15½-years old gained 10 points of fluid intelligence between 1961 and 1978.

Austria

Samples and tests. The University of Vienna maintains a children's clinic and between 1962 and 1979, children referred to it with scholastic or behavioral problems were given the West German WISC (HAWIK). There were 2,318 such subjects, from 6 to 15 years old, with an average sample size of 330 for ages 6–11 and 90 for ages 12–15.

Results. Table 13 shows large gains in terms of Wechsler verbal, performance, and full-scale IQ, the principal exception being a sharp drop between 1974 and 1977. Schubert and Berlach (1982, p. 257) discuss whether those years saw a real decline: They speculate that the four examiners became aware that the mean IQ had risen and, therefore, began to apply stricter standards in assessing responses. By 1974, thanks to a study by Wurst and Schubert (1975), the fact of massive IQ gains was known and the subsequent decline stands in sharp contrast to trends both before and after. In Table 13, three methods have been used to calculate the total gain from 1962 to 1979 for full-scale IQ. First, assuming the 1974–1977 drop is to be taken at face value, the gain was 7.2 points. Second, assuming the 1974–1977 drop was equivalent to the only other decline the data reveal, the 1.4 points lost between 1965 and 1968, the gain was 12.1 points. Third, assuming the 1974–1977 data are simply unreliable, the total gain should be calculated by taking the rate of gain for the other years. That rate was 0.964 points per year, and applied to the whole period of 17 years, it gives a gain of 16.4 points. Performance gains were significantly higher than verbal gains no matter what method of calculation is used.

Assessment. There is no easy solution to the problem of the suspect years. Unaltered data give an estimate of IQ gains that is almost certainly too low, whereas altered data give a very rough estimate of 12–16 points. However, the central question is whether the clinic's referrals can be taken as a measure of trends for either Austria or Vienna. Schubert and Berlach show appropriate caution about the status of the gains. On the other hand, they say that their subjects should be typical because scholastic and behavioral problems are not confined to those with either superior or inferior intelligence. This implies an unstated assumption: They know of no radical change, at least during these years, concerning the kind of children referred to the clinic for special help. Nonetheless, unless they have evidence for that assumption, it is all too easy to imagine sources of sample bias. Between 1962 and 1979, Viennese schools may have altered in their perception of bright children who underperform: Rather than calling underperformers lazy, they may have begun to see them as suffering from problems susceptible to diagnosis and treatment.

At present, all estimates of Austrian IQ gains must be classified as speculative, or Status 4. About all that can be said is that Austria may resemble the neighbor with which it has great cultural affinity—West Germany.

France

Samples and tests. Hanet (1984–1985) administered both the French WISC and WISC-R to a sample of French-speaking Belgian children. One-hundred fifty subjects were evenly divided among 6- to 7-year-olds, 8- to 9-year-olds, and 10- to 11-year-olds. He also reports data from a similar study with 70 subjects aged 11–15 years. The results do not measure Belgian gains, but rather French gains from the WISC standardization sample of 1954/1955 to the WISC-R standardization sample of 1978/1979.

The standardization of the French WISC aimed at 1,000 children stratified for age and sex, urban–rural locale, public and private schools, kind of school, and geographical region (Wechsler, 1965, pp. 13–15). Because of administrative difficulties, they only tested subjects at the five ages of 6, 8, 10, 12, and 14 years, the norms for other ages from 5 to 15 years being derived

Table 14
France: Wechsler IQ Gains, From 1954/1955 to 1978/1979, Ages 6–15 Years

Age (years)	IQ		
	Verbal	Performance	Full scale
6–7	4.82	17.78	11.98
8–9	−0.22	13.46	6.98
10–11	0.56	14.96	8.66
11–15	2.57	15.00	9.00
6–15[a]	2.06	15.24	9.12

Note. Data from Hanet (1984–1985, pp. 75–77).
[a] The average gain for all ages calculated by giving each value the weight of the number of ages it covers, with values for ages 11–15 taken as covering ages 12–15.

by extension. The target for age 14 was not met because insufficient numbers for each variable, such as "out of school," were found. They divided the country into three geographical zones, essentially the Paris region; a collection of regions surrounding Amiens, Lille, Lyon, and Marseille; and the rest of France. There is no mention of parental occupation or SES. The standardization of the French WISC-R aimed at 1,100 children aged 6–16 years, but small deficits above age 11 put the actual number at 1,066 (Wechsler, 1981a, pp. 30–32). The sample was stratified for age and sex, urban–rural locale, kind of school, geographical region, and father's occupational status. This time they drew a balanced sample from every region of France. Unlike the WISC, they did not include those attending private schools.

Results. Table 14 shows that during the 24 years from 1954/1955 to 1978/1979, French children gained only 2 points on the Wechsler verbal scale; however, they gained 15 points of performance IQ, and as a consequence, 9 points of full-scale IQ. The fact that the only real gains are performance gains is unique. Military data for France already analyzed suggested that Wechsler performance gains would be larger than verbal gains but not that the latter would be nonexistent. There is no tendency for gains to diminish with age, that is, the values for older subjects match those for all ages.

Assessment. The French WISC sample was not stratified for parental occupation and the norms for 13- to 15-year-olds were based on projections or inadequate sampling. Therefore, the estimates of Wechsler gains for older subjects can never be better than speculative or Status 4. The estimates for younger subjects are at present Status 4 but could rise to Status 3 if future studies confirm Hanet's results.

Summary of Data

Table 15 summarizes results from 14 nations or locales. It includes only those results relevant to the current generation, that is, results for periods either entirely or mainly post-1950. The IQ trends by place have been grouped in terms of kind of test, that is, culturally reduced, Wechsler-Binet, or purely verbal; and within these groups, they have been ranked by rate of gain. Status numbers show the reliability of each data set, Status 1 and 2 for strong data, Status 3 and 4 for weak data.

The rates of gain are expressed in IQ points per year, so multiplied by 30 years, they give an estimated gain over one generation. Setting aside the most extreme cases, and discounting France a bit, the generational gains range from 5 to 25 points and the median is 15 points, or a full standard deviation. These data do not show that all nations have made massive IQ gains; they cover only the developed world, and even there, only 14 nations out of more than 30. However, these data do show that no psychologist in any part of the developed world can discount the possibility of massive gains. Table 15 covers all the major areas: eight nations from Western Europe, one from Eastern Europe, and three from the English-speaking Commonwealth, as well as the United States, and Japan as representative of the industrialized nations of Asia. Some of these areas have only weak data, but that puts their psychologists in the worst possible position: They have reason to believe massive gains have occurred but no reliable estimates of their magnitude. Flynn (1984c, 1985) has described the penalty of ignoring the existence of IQ gains. Their effects are mistakenly assigned to compensatory education, modes of test administration, and cultural bias; and they call into question the use of IQ tests to classify subjects as mentally retarded.

The chance that reliable data exist for the current generation in every developed nation is as slight as that of observing every swan. The best that can be said is that nobody thus far has discovered even one black swan. This is not for want of trying: The phrasing of questionnaires was neutral, that is, correspondents were asked for evidence of either gains, stability, or losses; all data received were used. The fact that massive IQ gains dominate the current generation does not, of course, mean anything for the next.

Implications

In this section I state conclusions, describe their derivation, and assess their status. The assessment criterion is the status of the evidence on which the conclusion is based or the status of the propositions from which it is derived as a logical corollary.

IQ Gains and Learned Content

Conclusion. Learned content has inhibited rather than promoted IQ gains.

Derivation. Wechsler and verbal IQ tests attempt to test for decontextualized problem-solving ability, but they do so through a vehicle with many items that also test for something more specific. For example, they include a general information subtest, an arithmetic subtest, and so forth. If IQ gains from one generation to another merely reflect that the later generation has mastered more of these items, they would raise no interesting theoretical questions. The average person today would outscore Aristotle or Archimedes on general information, but this hardly shows greater intelligence. On the other hand, if generational IQ gains reflect massive gains in decontextualized problem-solving ability, they do pose interesting theoretical questions, for example: Those who wish to identify the problem-solving ability IQ

Table 15
Recent IQ Gains: Locations Grouped by Test and Ranked by Rate of Gain

Location	Test	Rate[b]	Age (years)	Period	IQ gain (points)	Status[c]
Leipzig	Ravens	1.250	11–16	1968–1978	10–15	3/2
France	Ravens	1.005	18	1949–1974	25.12	3
Belgium	Ravens	0.794	18	1958–1967	7.15	1
Belgium	Shapes	0.716	18	1958–1967	6.45	1
Netherlands	Ravens	0.667	18	1952–1982	20.00	1
Norway	Matrices	0.629	19	1954–1968	8.80	1
West Germany	Horn-Ravens[a]	0.588	12–16	1961–1978	10.00	4
Australia	Jenkins	0.490	10–14	1949–1981	15.67	3
Edmonton	Ravens	0.402	9	1956–1977	8.44	1
Australia	Ravens	0.337	10–16	1950–1976	8.76	4
Norway	Matrices	0.217	19	1968–1980	2.60	1
Great Britain	Ravens	0.189	8–14	1938–1979	7.75	3
Great Britain	Ravens	0.181	20–30	1940–1979	7.07	3
Japan	Wechsler[a]	0.835	6–15	1951–1975	20.03	3/4
Vienna	Wechsler	0.824	6–15	1962–1979	12–16	4
West Germany	Wechsler[a]	0.741	7–15	1954–1981	20.00	3/4
Zurich	Wechsler	0.652	9 and 12	1954–1977	10–20	4
Edmonton	CTMM	0.525	9	1956–1977	11.03	1
France	Wechsler[a]	0.380	6–15	1955–1979	9.12	4
United States	Wechsler-Binet[a]	0.300	2–18	1932–1972	12.00	2
United States	Wechsler[a]	0.243	16–75	1954–1978	5.95	3
Solothurn	Wechsler	0.186	8–9	1977–1984	1.30	4
Saskatchewan	Otis[a]	0.628	10	1958–1978	12.55	2/3
Norway	Verbal-Math	0.582	19	1954–1968	8.15	1
Belgium	Verbal-Math	0.408	18	1958–1967	3.67	1
France	Verbal-Math	0.374	18	1949–1974	9.35	3
Saskatchewan	Otis[a]	0.348	13	1958–1978	6.95	2/3
New Zealand	Otis	0.242	10–13	1936–1968	7.73	1
Norway	Verbal-Math	−0.133	19	1968–1980	−1.60	1

Note. Data from Tables 1–14; see these tables for full test names.
[a] The content of these tests was substantially altered. [b] IQ points per year. [c] Key: 1 = verified, 2 = probable, 3 = tentative, and 4= speculative.

tests measure with intelligence must argue that the current generation is radically more intelligent than the last. A clarification: The very fact that massive IQ gains have occurred means that something has been learned at least in the broadest sense of the word. Yet that something could either be decontextualized problem solving or something more specific, such as information, the meaning of a word, or arithmetic. The learned content hypothesis does not cover both of these possibilities, but opts for the latter, the notion that massive IQ gains do not represent the sort of skill gains that might be identified with increased intelligence.

A consensus about the significance of generational IQ gains depends, therefore, on whether they manifest themselves on culturally reduced tests like the Ravens. These tests maximize problem solving and minimize the need for more specific skills and familiarity with words and symbols. Table 15 contains strong data for massive gains on culturally reduced tests: Belgium, the Netherlands, Norway, and Edmonton show gains ranging from 7 to 20 points over periods of from 9 to 30 years; when the rates of gain are multiplied by 30 years, they suggest that the current generation has gained 12–24 points on this kind of test. Tentative data from other nations are in full agreement. This settles the question at issue: IQ gains since 1950 reflect a massive increase in problem-solving ability and not merely an increasing body of learned content.

The next question is whether problem-solving gains have been greater or lesser than learned-content gains. Table 15 gives separate results for culturally reduced and verbal-only tests, the latter having the highest loading of learned content. The median rate of gain for all nations is 0.588 for culturally reduced tests and 0.374 for verbal tests, a pattern that holds both for all data and for strong data only. Table 16 selects out nations that have results for both kinds of tests: All three nations show gains on culturally reduced tests at twice the size of verbal gains.

At least for the current generation, the more a test uses learned content as a vehicle for measuring problem-solving ability, the more it prevents gains in that ability from becoming visible. Norway since 1968 emerges as a case in which gains in problem-solving ability, as captured by the matrices test, were turned into a loss, thanks to the learned content of verbal and math subtests. The only exception to the rule is Edmonton, where 9-year-olds made slightly higher gains on the CTMM than on the Ravens test. There is not one instance of the phenomenon the learned-content hypothesis posits: learned-content gains but no gains in terms of decontextualized problem-solving ability.

Table 16
Recent IQ Gains: Comparison Between Culturally Reduced and Verbal Tests

Nation	Gains in IQ points: Culturally reduced	Gains in IQ points: Verbal only	Ratio[a]	Status[b]
France	25.12	9.35	2.69:1	3
Belgium	6.80	3.67	1.85:1	1
Norway	11.40	6.55	1.74:1	1

Note. Data from Table 15.
[a] Culturally reduced gain divided by verbal gain. [b] The status of the data on which the estimates of IQ gains are based. Key: 1 = verified, 3 = tentative.

Wechsler verbal subtests rely more heavily on learned content than do Wechsler performance subtests, and in Table 17, five nations offer a comparison between verbal and performance gains. The periods for all nations fit the post-1950 generation except that of America, and that problem was solved by using only data from the WISC and WISC-R standardization samples. All gains were projected over 30 years for the sake of comparability. The results: Performance gains were greater than verbal gains in every nation, sometimes by as much as 16 points.

The Wechsler data are weak for each nation taken in isolation but despite this, the collective pattern allows for strong inference. The major defect of these data is sampling problems. If one assumes that sampling errors had an even chance of favoring either performance or verbal gains, the odds against a sequence of five higher performance gains are 2^5, or 32 to 1. Moreover, the greater magnitude of performance gains is huge in three of the five cases. Plotting the results as random events on a normal distribution, the odds against this kind of difference between performance and verbal scores are 840 to 1. Anyone who wishes to explain away the data must look for this: a systematic bias that favors performance gains and that had some impact in America and Japan, a large impact in Austria, and a huge impact in West Germany and France.

Assessment. The conclusion rests on strong data and is, therefore, also strong. It applies only to the current generation and admits of exceptions.

IQ Gains and Early Maturation

Conclusion. IQ gains persist to full maturity in Belgium, the Netherlands, and Norway, and may persist in France, Great Britain, and the United States.

Derivation. If IQ gains merely showed that this generation had reached its peak ability at an earlier age than the last, but that the mature level was actually no higher, the significance of IQ gains would be diminished. The present generation might be expected to show radically altered behavior at school but not in the adult world of cultural, scientific, and technological achievement. Once again, a distinction between different kinds of tests is relevant, this time between tests of fluid intelligence and tests of crystalized intelligence. The culturally reduced tests mentioned herein, the Ravens, the Norwegian matrices, the Belgian Shapes test, the Jenkins and the Horn tests, are tests of fluid intelligence. They make the subject demonstrate problem-solving ability on the spot. The other tests, the Wechsler-Binet and purely verbal tests, measure both fluid and crystalized intelligence. Some items are direct measures of problem-solving ability, but others test for the sort of things able people accumulate throughout life—general information, vocabulary, and mathematical skills.

The growth curve for fluid intelligence parallels certain physical growth curves, like lung capacity and brain weight, and peaks sometime in the late teens or early 20s (Jensen, 1980, p. 235). Table 15 shows that the age of subjects administered tests of fluid intelligence in Belgium, the Netherlands, and Norway varies from 18 to 19 years of age. Because all of this is Status 1 data, the persistence of IQ gains to full maturity in these nations is verified. France and Great Britain also show Ravens test gains for mature subjects, but these are Status 3 data, or tentative.

The growth curve for crystalized intelligence is quite different: Because people can expand their general information and vocabularies throughout their lives, it may peak as late as age 60. As far as the Wechsler-Binet and verbal tests are concerned, Table 15 shows the United States as the sole nation with data for ages that span all the years of adulthood. Collectively these data reveal the following pattern: (a) Schoolchildren gained 9.00 points during the current generation, (b) adults aged 35 years and less gained 5.55 points, and (c) adults aged 35–75 years gained 10.19 points. American gains on Wechsler tests appear to persist into late adulthood, but note that the adult estimates are based on one pair of standardization samples (WAIS and WAIS-R) and, therefore, can be only Status 3, or tentative. Parker (1986) offers an additional piece of evidence. He analyzed all Wechsler standardization samples for adults beginning with the old Wechsler-Bellevue Form I and found that the peak age for raw score performance actually rose from age 22 to age 30 between 1936/1937 and 1978.

Lacking adult results, other nations provide only indirect evidence. When generational gains of schoolchildren can be traced from younger to older subjects, New Zealand (see Table 5), Leipzig (see Table 8), Japan (see Table 11), and France (see Ta-

Table 17
Wechsler Data From Five Nations: Comparison of Full Scale, Verbal, and Performance IQ Gains

Nation	IQ gains projected over 30 years: Full scale	Verbal	Performance	P – V[a]
United States	9	8	10	2
Japan	25	20	22	2
Austria	25	20	28	8
West Germany	22	12	28	16
France	11	3	19	16

Note. Data from Tables 11–14; see the comment in the text on this table and Table 13 for the basis of the United States and Austrian values.
[a] Performance gains minus verbal gains.

ble 14) show no tendency for gains to diminish with age. Saskatchewan (see Table 6) is the sole exception, with 13-year-olds showing lower gains than 10-year-olds on the Otis test, although note that the content was substantially altered.

Assessment. The conclusion is supported by strong data for three nations, tentative data for three nations, and almost all of the indirect evidence existent.

IQ Gains and Aging

Conclusion. Cross-sectional data, as a measure of the effects of aging on IQ, are suspect.

Derivation. Cross-sectional data compare, for example, 80-year-old subjects with a group of 20-year-old subjects, with both groups being tested at the same time. This makes sense only if current 20-year-olds have the same IQ as 20-year-olds did two generations ago, that is, when today's 80-year-olds were 20. However, generational gains have occurred throughout the developed world and they may have contaminated adult samples of all ages. Horn and Donaldson (1976, 1977) defend cross-sectional data against Baltes and Schaie (1976, 1977) on the grounds that such data show a more rapid decline with age for fluid intelligence than crystalized intelligence. They argue that if generational gains had occurred, they should almost certainly have been greater for crystalized intelligence. Hence they conclude that cross-sectional data do not show the pattern to be expected if they were contaminated by generational gains. In fact, generational IQ gains have been greater for fluid intelligence; therefore, cross-sectional data show exactly the pattern to be expected if they were contaminated.

Assessment. The conclusion is a logical corollary of the two strong conclusions already established and is itself strong.

What IQ Tests Really Measure

Conclusion. The Ravens Progressive Matrices Test does not measure intelligence but rather a correlate with a weak causal link to intelligence; the same may apply to all IQ tests.

Derivation. The literature makes clear what real-world behavior we have a right to expect from those at various high IQ levels: Above 130 they find school easy and can succeed in virtually any occupation; above 140 their adult achievements are so clear that they fill the pages of *American Men of Science* and *Who's Who;* above 150 they begin to duplicate the life histories of the famous geniuses who have made creative contributions to our civilization (Jensen, 1980, pp. 111–114). Just as an elite with a massive IQ advantage should radically outperform the rest of its generation, so a generation with a massive IQ gain should radically outperform its predecessors. There are a few escape clauses where generational IQ gains are concerned: Perhaps they do not reflect gains in the problem-solving ability IQ tests attempt to measure, or perhaps they do not persist to maturity and should not affect the world of adult achievement. These escape clauses have been shown to be nonapplicable.

The current generation in the Netherlands must radically outperform the last or the Ravens test does not measure intelligence. Table 18 shows the effect of a 20-point gain on high IQ levels as measured in terms of 1952 norms. The same effect has probably occurred in France in that French gains, although tentative, are almost certainly as high as 15 or 20 points. These effects should be highly visible: 25% of the children teachers face qualify as gifted; those with IQs of 150 and above have increased by a factor of almost 60, which means that the Netherlands alone has over 300,000 people who qualify as potential geniuses. The result should be a cultural renaissance too great to be overlooked.

Table 18
The Netherlands: The Effect of IQ Gains on High IQ Levels

	Percentages		
1952 IQ level	1952	1982	Ratio[a]
130 and above	2.27	25.25	11:1
140 and above	0.38	9.12	24:1
150 and above	0.04	2.27	57:1

[a] 1982 percentage divided by 1952 percentage.

The Centre Parisien d'Etudes et de Documentation pour l'Enseignement et le Rayonnement du Français (CPEDERF) was asked to survey the French and Dutch scene. *L'Express* (1967 to the present), *Le Point* (1973 to the present), and *Le Monde de l'Education* (1974 to the present) contained not a single reference to a dramatic increase in genius or mathematical and scientific discovery during the present generation; no one has remarked on the superiority of contemporary schoolchildren. The *Handelsblad, Volkskrant,* and *Elseviers Weekblad,* all from 1974 to the present, gave the same result. The Dutch do feel they have achieved a marked superiority in sport. As for inventions, the number of patents granted has actually diminished: the 1980s show 60%–65% of the yearly rate for the 1960s (CPEDERF, personal communications, November 18, 1985 to June 19, 1986).

The only hypothesis that can cover the facts is that the Ravens test does not measure intelligence but rather a correlate with a weak causal link to intelligence. Imagine that we could not directly measure the population of cities but had to take aerial photographs, which gave a pretty good estimate of area. In 1952, ranking the major cities of New Zealand by area correlated almost perfectly with ranking them by population, and in 1982, the same was true. Yet anyone who found that the area of cities had doubled between 1952 and 1982 would go far astray by assuming that the population had doubled. The causal link between population and its correlate is too weak, thanks to other factors that intervene, such as central city decay, affluent creation of suburbs, and more private transport, all of which can expand the city's area without the help of increased population.

Clearly much the same is true of the Ravens and intelligence. The Ravens test measures a correlate of intelligence that ranks people sensibly for both 1952 and 1982, but whose causal link is too weak to rank generations over time. This poses an important question: If a test cannot rank generations because of the cultural distance they travel over a few years, can it rank races or groups separated by a similar cultural distance? The problem is not that the Ravens measures a correlate rather than intelli-

gence itself, rather it is their weak causal link. When measuring the real-world phenomenon we call a hot day, we use the height of a column of mercury in a thermometer as a correlate, but note that this correlate has a strong causal link that allows it to give sensible readings over time. A thermometer not only tells us the hottest day of 1952 and the hottest day of 1982, it also gives a sensible measure of whether the summer of 1952 was hotter than the summer of 1982.

As to whether other IQ tests measure intelligence, the best path to clarity is to go from an ideal evidential situation to the actual one. Imagine the following situation in at least one nation: For every culturally reduced test in existence, strong data show massive IQ gains, say at least 10 points in a single generation; for every other IQ test in existence, a strong collective pattern of data shows that gains on them were less than massive only because of the inhibiting presence of learned content; that nation enjoyed no cultural renaissance in that generation. Then the conclusion that all IQ tests measure only a correlate of intelligence would be verified. The next best situation would be this: Strong data are not available for every test existent but are available for a fair selection of the various kinds of IQ tests, that is, culturally reduced, Wechsler-Binet, and purely verbal tests; that is, for Ravens, Shapes, Jenkins, Cattell, Wechsler, Binet, Otis, verbal and math tests. Then the conclusion would be probable. Now, to describe the actual situation: The Ravens test matches the evidential ideal; the Wechsler, CTMM (a stand-in for the Stanford-Binet), and Otis tests, as well as various verbal and math subtests, all fall short primarily because of a lack of data from mature subjects; the Shapes data are from a limited time period, the Jenkins data are tentative, and there are no data for the Cattell, that is, none from this generation.

Therefore, the conclusion as applied to all IQ tests is tentative. However, until the matter is settled, psychologists should stop saying that IQ tests measure intelligence. They should say that IQ tests measure abstract problem-solving ability (APSA), a term that accurately conveys our ignorance. We know people solve problems on IQ tests; we suspect those problems are so detached, or so abstracted from reality, that the ability to solve them can diverge over time from the real-world problem-solving ability called intelligence; thus far we know little else. A disclaimer: Language of this sort should not be taken as a commitment to a unitary theory of intelligence. When speaking about mental ability, the singular is used purely as a matter of convenience. No conclusion would be substantially altered by using the plural, such as saying that IQ tests measure a variety of abilities, or that what they measure are correlates of a variety of real-world abilities, or that what they really measure are APSAs.

Assessment. The conclusion is strong for the Ravens test and tentative for all other IQ tests.

Causes of Massive IQ Gains

Conclusion. Most of the Dutch IQ gains have been caused by unidentified environmental variables; these variables may account for as much as 15 points in the current generation.

Derivation. Massive IQ gains cannot be due to genetic factors. Reproductive differentials between social classes would have to be impossibly large to raise the mean IQ even 1 point in a single generation (Flynn, 1986; Vining, 1986). The Dutch data are the only strong data that give information about the role of various environmental variables. The impact of higher levels of education can be estimated from Leeuw and Meester (1984, pp. 18, 21, Figures 8 and 10). As already discussed, their data are inaccurate for 1982, so education levels from 1972 were compared with their 1952 counterparts: five groups ranging from those with primary school educations to those at gymnasium. Collectively, these groups outscored their earlier counterparts by 11.62 points, which was 93.5% of the Dutch IQ gains between 1952 and 1972 (see Table 1). This means that matching subjects to hold education level constant should eliminate just over 1 point of the 20-point Dutch gain.

Leeuw and Meester (1984, pp. 14, 16, Figures 5 and 7) allow an estimate of SES gains from 1952 to 1962, as measured by father's occupation. When projected over 30 years, this amounts to 1.18 *SD*s. The correlation between father's occupation and son's IQ is .33 (Leeuw & Meester, 1984, pp. 13, 16), and, therefore, SES gains might appear to account for 5.84 of the 20-point Dutch gain ($1.18 \times .33 = 0.3894$ *SD* units; $0.3894 \times 15 = 5.84$ points). However, the correlation between father's occupation and son's IQ may not represent a causal link: When Vroon controlled for father's IQ and father's education level, variables with a high genetic loading, the path correlation between father's occupation and son's IQ was .02, or virtually zero (P. A. Vroon, personal communication, October 9, 1984). A generous estimate for SES, as an environmental variable, would be that it caused a 3-point IQ gain in the current generation.

There is indirect evidence that test sophistication is not a major factor. It has its greatest impact on naive subjects, that is, repeated testing with parallel forms gives gains that total 5 or 6 points. It seems unlikely that a people exposed to comprehensive military testing from 1925 onward were totally naive in 1952. Moreover, test sophistication pays diminishing returns over time as saturation is approached, and as Table 1 shows, Dutch gains have actually accelerated, with the decade from 1972 to 1982 showing the greatest gains of all. Reviewing the factors discussed, higher levels of education contribute 1 point, SES may contribute 3 points, and what for test sophistication, perhaps 2 points? These estimates cannot simply be summed because the factors are confounded; for example, higher SES encourages staying in school longer, which raises test sophistication. Together they appear to account for about 5 points.

The Dutch data can only tell us so much. They reveal an irony about the Ravens test: Culturally reduced content has maximized culturally influenced performance. That is, the effort to avoid culturally loaded words or symbols has made the test the purest measure of those problem-solving skills that are most subject to cultural determination. They do not reveal whether these skills can be taught. Additional years of schooling have not been a significant causal factor, but these data provide no assessment of qualitative changes in school experience. A number of scholars have argued that decontextualized problem-solving ability is the most pervasive of formal school skills (Cole & Means, 1981; Scribner & Cole, 1981; Sharp, Cole, & Lave, 1979). If schools in the developed world are producing massive gains in APSA, they appear to be doing this better than

enhancing traditional academic skills. Perhaps some self-examination is in order: APSA gains pay few dividends in terms of creativity, inventions, and scientific or mathematical discovery.

At present, the Dutch data leave unknown the environmental factors responsible for about 15 points of a 20-point gain. The fact that the factors are unknown does not mean that when identified, they will prove exotic or unfamiliar: Television and greater exposure to information stimuli of all sorts join formal schooling as possibilities. Nonetheless the factors at work must be identified and their great potency poses the real problem. When Archimedes wanted to impress Hiero with the power of the lever, he took a ship in drydock, heavy laden with many passengers and freight, and clasping the end of a compound pulley, drew her along smoothly as if moving under full sail at sea. It would be uninformative to say that Archimedes was using something familiar, his muscles, because without his knowledge of the principle of the lever, what he could do with his muscles was quite inexplicable. If environmental factors have an unanticipated potency, at least in advanced societies like the Netherlands, it does no good to stress their familiarity. We must look at the environment with new eyes and try to identify the unknown forces that have done so much to transform the factors we know.

Assessment. The conclusion is based on strong data, but should not be taken as dismissing the possibility that familiar environmental variables play an important role.

IQ Gains and Between-Groups Differences

Conclusion. Between-groups score differences on IQ tests may not be equivalent to intelligence differences.

Derivation. As a tentative conclusion for IQ tests in general, it has been shown that between-generations score differences cannot be equated with intelligence differences. The magnitude of these score differences matches the size of all major between-groups differences in the literature, whether these refer to races, classes, or nations. Between-generations IQ differences occur because cultural distance separates the generations—after all, the mere passage of time counts for nothing. Therefore, unless two groups are separated by a lesser or dissimilar cultural distance, their score differences cannot be equated with intelligence differences. Until the causal problem of what factors engender between-generations IQ differences is solved, no one knows what cultural variables are relevant. Therefore, no one can show that between-groups cultural distance is dissimilar to between-generations cultural distance.

Assessment. This conclusion is a corollary of two previous conclusions, one of which is tentative and, therefore, this conclusion is also tentative. It does not claim that no between-groups intelligence differences exist, merely that IQ differences cannot, at present, be used as evidence for them.

IQ Gains and the Scholastic Aptitude Test (SAT) Score Decline

Conclusion. Wechsler-Binet IQ gains over time are compatible with SAT score losses over the same period.

Derivation. These divergent American trends are a special case of an international phenomenon. France, Belgium, and Norway all show different trends on different kinds of mental tests. The problem of reconciling the American trends separates into two aspects: explaining what goes on in the examination room, and explaining what is happening in the larger society.

When subjects take a mental test, it measures their problem-solving ability through a vehicle. If it does so through the vehicle of academic skills, and those skills are in decline, problem-solving gains must overcome academic skill losses. Wechsler IQ tests require no more than elementary academic skills, and some performance subtests minimize the need even for these. Therefore, the problem-solving gains win and the result is score gains, albeit diminished ones. However, as D. R. Vining (personal communication, February 29, 1984) has pointed out, a subject simply cannot perform well on the SAT-Verbal (SAT-V) subscale without the advanced academic skills taught in high school English courses. Therefore, the problem-solving gains lose and the result is score losses. In addition, a new piece of the puzzle has fallen into place. Korb (1982) found no gains on armed forces mental tests between 1944 and 1980, at least by white Americans from 18 to 23 years old. The content of the armed forces tests is illuminating: simple arithmetic, both presented as such and verbally, word knowledge, and paragraph comprehension, all of which are on an elementary school level (Korb, 1982, pp. 4–6; U.S. Department of Defense, undated but circa 1984, pp. 4–7).

In sum, between 1963 and 1981, trends on various mental tests were as follows. The WAIS to WAIS–R data for young adults give a gain of 3.33 IQ points ($4.53 \div 24.5$ years = 0.185, 0.185×18 years = 3.33). The armed forces data suggest no gain during the same period. The SAT-V score decline amounted to 0.288 *SD*s, or 4.32 IQ points (Flynn, 1984c, p. 37). The pattern is now complete: As young adults went from a test of problem-solving ability with a moderate reliance on elementary academic skills, to one with heavy reliance on elementary academic skills, and then to one with heavy reliance on advanced academic skills, trends changed from gains to no gains and from no gains to losses.

This leaves us with the deeper problem of the dynamics of American society. Thanks to gains on Wechsler-Binet tests, it seemed that those entering American high schools were getting more and more intelligent, and yet they were leaving high school with worse and worse academic skills. Unless nonintellectual traits, such as motivation, study habits, and self-discipline were deteriorating at an incredible rate, how could more intelligent students be getting so much less education? Now the solution is apparent: High school students in 1981 did not necessarily have higher intelligence than their counterparts in 1963, they merely had higher APSA. The hypothesis that Wechsler-Binet tests are like the Ravens test, that they measure APSA rather than intelligence, makes everything fall into place.

Assessment. The conclusion is a corollary of the conclusion that Wechsler-Binet IQ gains over time cannot be equated with intelligence gains. Because the latter is tentative, the former is also tentative.

Summary of Implications

The current generation has made massive gains on all kinds of IQ tests. These gains persist to maturity and, therefore, cross-

sectional data are suspect as a measure of the effects of aging on IQ. These gains suggest that IQ tests do not measure intelligence but rather a correlate with a weak causal link to intelligence. Therefore, between-groups IQ differences cannot, at present, be equated with intelligence differences. Environmental factors with a large impact on IQ have not been identified. The hypothesis that Wechsler-Binet IQ gains may not represent intelligence gains shows how they can be reconciled with the SAT score decline.

References

Australian Council for Educational Research. (1958a). *Manual for ACER Intermediate Test D.* Hawthorn, Australia: Author.

Australian Council for Educational Research. (1958b). *Otis Intermediate Test manual.* Melbourne, Australia: Author.

Australian Council for Educational Research. (1982). *Manual for ACER Intermediate Test F.* Hawthorn, Australia: Author.

Baltes, P. B., & Schaie, K. W. (1976). On the plasticity of intelligence in adulthood and old age. *American Psychologist, 31,* 720–725.

Baltes, P. B., & Schaie, K. W. (1977). Some faith helps to see the forest. *American Psychologist, 32,* 1118–1120.

Bouvier, U. (1969). *Evolution des cotes a quelques tests* [Evolution of scores from several tests]. Brussels: Belgian Armed Forces, Center for Research Into Human Traits.

Cattell, R. B. (1950). The fate of national intelligence: Test of a thirteen-year prediction. *Eugenics Review, 42,* 136–148.

Clarke, S. C. T., Nyberg, V., & Worth, W. H. (1978a). *Alberta Grade III achievement study.* Edmonton, Canada: University of Alberta.

Clarke, S. C. T., Nyberg, V., & Worth, W. H. (1978b). *Technical report on Edmonton Grade III achievement: 1956–1977 comparisons.* Edmonton, Canada: University of Alberta.

Cole, M., & Means, B. (1981). *Comparative studies of how people think: An introduction.* Cambridge, MA: Harvard University Press.

de Lemos, M. M. (1984). A note on the Australian norms of the Standard Progressive Matrices. *ACER Bulletin for Psychologists, 36,* 9–12.

Elley, W. B. (1969). Changes in mental ability in New Zealand schoolchildren. *New Zealand Journal of Educational Studies, 4,* 140–155.

Flynn, J. R. (1984a). IQ gains and the Binet decrements. *Journal of Educational Measurement, 21,* 283–290.

Flynn, J. R. (1984b). Japanese IQ. *Nature, 308,* 222.

Flynn, J. R. (1984c). The mean IQ of Americans: Massive gains 1932 to 1978. *Psychological Bulletin, 95,* 29–51.

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

Flynn, J. R. (1986). Sociobiology and IQ trends over time. *Behavioral and Brain Sciences, 9,* 192.

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

Girod, M., & Allaume, G. (1976). L'évolution du niveau intellectuel de la population française pendent le dernier quart de siecle [The evolution of the intellectual level of the French population during the last quarter century]. *International Review of Applied Psychology, 25,* 121–123.

Hanet, J.-L. (1984–1985). *Etude comparative du WISC-R et des autres echelles de D. Wechsler pour enfants sur un echantillon de 150 sujets normaux ages de 6 a 10 ans* [Comparative study of the WISC-R and the other scales of D. Wechsler for children on a sample of 150 normal subjects ages from 6 to 10 years]. Unpublished mémoire de licence, Universite de Leige, Leige, Belgium.

Horn, J. L., & Donaldson, G. (1976). On the myth of intellectual decline in adulthood. *American Psychologist, 31,* 701–719, 878.

Horn, J. L. & Donaldson, G. (1977). Faith is not enough. *American Psychologist, 32,* 369–373.

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

Kodama, H., & Shinagawa, F. (1953). *Wechsler Intelligence Scale for Children* (Japanese ed.). Tokyo: Nihon Bunka Kagakusha.

Korb, L. J. (1982). *Profile of American youth.* Washington, DC: Office of the Assistant Secretary of Defense.

Kratzmeier, H., & Horn, R. (1979). *Manual: Raven-Matrizen-Test* [Manual: Raven's Standard Progressive Matrices]. Weinheim, West Germany: Beltz Test.

Leeuw, J. de, & Meester, A. C. (1984). Over het intelligente—onderzoek bij de militaire keuringen vanaf 1925 tot heden [Intelligence—as tested at selections for the military service from 1925 to the present]. *Mens en Maatschappij, 59,* 5–26.

Lippold, S., & Claiborn, J. M. (1983). Comparison of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale-Revised. *Journal of Consulting and Clinical Psychology, 51,* 315.

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

Lynn, R. (1977). The intelligence of the Japanese. *Bulletin of the British Psychological Society, 30,* 69–72.

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

Lynn, R., & Hampson, S. (1986). The rise of national intelligence: Evidence from Britain, Japan and the U.S.A. *Personality and Individual Differences, 7,* 23–32.

McIntyre, G. A. (1938). *The standardization of intelligence tests in Australia.* Melbourne, Australia: Melbourne University Press.

Mehlhorn, G., & Mehlhorn, H.-G. (1981). Intelligenz-Tests und Leistung [Intelligence tests and achievement]. *Wissenschaft und Fortschritt, 31-9,* 346–351.

Mehlhorn, H.-G. (1981). Intellektuelles Potential der Jungend yur Nutzung für kreative Leistungen [Intellectual potential of youth: On its use for creative achievement]. In *Internationales Wissenschaftliches Kolloquium,* Technische Hochschule Ilmenau, Ilmenau, East Germany.

Mishra, S. P., & Brown, K. H. (1983). The comparability of WAIS and WAIS-R IQs and subtest scores. *Journal of Clinical Psychology, 39,* 754–757.

Otis, A. S., & Lennon, R. T. (1969). *Otis-Lennon Mental Ability Test: Technical handbook.* New York: Harcourt, Brace & World.

Parker, K. C. H. (1986). Changes with age, year-of-birth cohort, age by year-of-birth cohort interaction, and standardization of the Wechsler adult intelligence tests. *Human Development, 29,* 209–222.

Priester, H.-J. (1958). *Die Standardisierung des Hamburg-Wechsler-Intelligenztests für Kinder (HAWIK)* [The standardization of the Hamburg Wechsler Intelligence Test for Children (HAWIK)]. Bern, Switzerland: Huber.

Randhawa, B. S. (1980, July). *Change in intelligence and academic skills of grades four and seven pupils over a twenty-year period.* Paper presented at the 22nd International Congress of Psychology, Leipzig, East Germany. (ERIC Document Reproduction Service No. ED 194 580)

Raven, J. (1981). *Manual for Raven's Progressive Matricies and Mill Hill Vocabulary Scales* (Res. Suppl. No. 1). London: H. K. Lewis.

Raven, J. C. (1941). Standardization of progressive matricies. *British Journal of Medical Psychology, 19,* 137–150.

Rist, T. (1982). *Det intellektuelle prestasjonsnivaet I befolkningen sett I lys av den samfunns-messige utviklinga* [The level of the intellectual

performance of the population seen in the light of developments in the community]. Oslo: Norwegian Armed Forces Psychology Service.

Schallberger, U. (1985). *HAWIK und HAWIK-R: Ein empirischer Vergleich* [HAWIK and HAWIK-R: An empirical comparison] (Tech. Rep.). Zurich: Psychologisches Institut der Universität.

Schubert, M. T., & Berlach, G. (1982). Neue Richtlinien zur Interpretation des Hamburg Wechsler-Intelligenztests für Kinder (HAWIK) [New guidelines for the interpretation of the Hamburg Wechsler Intelligence Tests for Children (HAWIK)]. *Zeitschrift für Klinische Psychologie, 11,* 253–279.

Schull, W. J., & Neel, J. V. (1965). *The effects of inbreeding on Japanese children.* New York: Harper & Row.

Scottish Council for Research in Education (1961). *The level and trend of national intelligence.* London: University of London Press.

Scribner, S., & Cole, M. (1981). *The psychology of literacy.* Cambridge, MA: Harvard University Press.

Sharp, D., Cole, M., & Lave, C. (1979). Education and cognitive development: The evidence from experimental research. *Monographs of the Society for Research in Child Development, 44,* 1–92.

Smith, R. S. (1983). A comparison study of the Wechsler Adult Intelligence Scale and the Wechsler Adult Intelligence Scale-Revised in a college population. *Journal of Consulting and Clinical Psychology, 51,* 414–419.

Tewes, U. (1983). *HAWIK-R: Hamburg-Wechsler-Intelligenztest für Kinder, revision 1983* [HAWIK-R: Hamburg Wechsler Intelligence Test for Children, 1983 revision]. Bern, Switzerland: Huber.

United Nations Department of International Economic and Social Affairs, Statistical Office. (1979). *Demographic yearbook: Historical supplement, special issue.* New York: United Nations.

Urbina, S. P., Golden, C. J., & Ariel, R. N. (1982). WAIS/WAIS-R: Initial comparisons. *Clinical Neuropsychology, 4,* 145–146.

U.S. Department of Defense. (circa 1984). *Armed Services Vocational Aptitude Battery (ASVAB) information pamphlet* (DoD Pub. No. 1304.12Z). Washington, DC: Author.

Vining, D. R. (1986). Social versus reproductive success: The central theoretical problem of human sociobiology. *Behavioral and Brain Sciences, 9,* 167–187.

Vroon, P. A., Leeuw, J. de, & Meester, A. C. (1984). *Correlations between the intelligence levels of fathers and sons.* Unpublished manuscript, Department of Theoretical Psychology and History of Psychology, University of Utrecht, The Netherlands.

Wechsler, D. (1965). *Manuel: Echelle d'intelligence de Wechsler pour enfants (Etalonnages Francais), WISC* [Manual: Wechsler Intelligence Scale for Children (French standardization), WISC]. Paris: Les Editions du Centre de Psychologie Appliquée.

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

Wechsler, D. (1981a). *Manuel: Echelle d'intelligence de Wechsler pour enfants, forme révisée WISC-R* [Wechsler Intelligence Scale for Children, Revised Form WISC-R]. Paris: Les Editions du Centre de Psychologie Appliquée.

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

Wurst, E., & Schubert, M. T. (1975). Statistischer Beitrag zur Interpretation des Hamburg-Wechsler-Intelligenztests für Kinder (II) [A statistical contribution to the interpretation of the Hamburg Wechsler Intelligence Tests for Children (II)]. In F. Poustka & W. Spiel (Eds.), *Therapien in des Kinder- und Jugend-Psychiatrie, Band 2* (pp. 899–903). Vienna: Engermann.

Received January 30, 1986
Revision received April 29, 1986 ■