WAIS-III and WISC-III IQ gains in the United States from 1972 to 1995: how to compensate for obsolete norms

Article *in* Perceptual and Motor Skills · June 1998

DOI: 10.2466/pms.1998.86.3c.1231

1 author:

James Flynn
University of Otago
82 PUBLICATIONS **6,979** CITATIONS

Perceptual and Motor Skills, 1998, 86, 1231-1239.

WAIS–III AND WISC–III IQ GAINS IN THE UNITED STATES FROM 1972 TO 1995: HOW TO COMPENSATE FOR OBSOLETE NORMS[1]

JAMES R. FLYNN

University of Otago

Summary.—Flynn used data covering 1932 to 1972 to put U.S. gains at about 0.300 IQ points per year. For post-1972, comparison of WISC–R versus WISC–III, particularly the data of Zimmerman and Woo-Sam, gives a rate of 0.312. However, comparison of WAIS–R versus WAIS–III shows that the current rate may be as low as 0.171. It is hypothesized that the discrepancy may be due to sampling error and it is suggested that post-1972 US gains be put at about 0.25 points per year.

Periodically, the Psychological Corporation selects standardization samples to renorm Wechsler IQ tests. These samples are probably the best available and if reliable, they can be used to estimate IQ gains over time. If Americans are performing better and better on IQ tests, later norms will be harder to beat than earlier ones. Therefore, subtracting the (lower) scores achieved on the later test from the (higher) scores achieved on the earlier test gives the number of IQ points gained in the period between the selection of the two samples. Using a wealth of data from tests normed between 1932 and 1972, Flynn (1984) calculated a rate of gain during that period of about 0.300 IQ points per year. Since then, the norming of the WISC–III and the WAIS–III provide the relevant data with which to calculate post-1972 gains. However, they appear to yield quite different rates, which poses a problem for those of us who want an estimate of American IQ gains from 1972 to 1995.

Confirming The WISC–III Estimate

The WISC–III manual contains data from 206 subjects who took both the WISC–R, normed in 1972, and the WISC–III, normed in 1989, in a counterbalanced design (Wechsler, 1992, Table 6.8). The first band of Table 1 subtracts their mean IQ on the WISC–III from their mean IQ on the WISC–R, which yields the following estimates for IQ gains over the 17 years that separate the two tests: 5.30 points for the Full Scale, 2.40 points for the Verbal Scale, and 7.40 points for the Performance Scale.

The manual's 206 subjects are not drawn from the standardization samples, rather they are a vehicle for comparing the norms set by the two samples. So we have two possible sources of error: the standardization sam-

[1]Address enquiries to J. R. Flynn, Department of Political Studies, University of Otago, Box 56, Dunedin, New Zealand.

TABLE 1
U.S. GAINS 1972 TO 1989 MEASURED BY WISC–III VS WISC–R

Source		Full Scale	Verbal	Performance
WISC–III manual ($N = 206$)	WISC–R	108.20	103.90	111.60
	WISC–III	102.90	101.50	104.20
	Difference	5.30	2.40	7.40
Zimmerman and Woo-Sam (1997)	WISC–R	92.94	92.09	95.47
All 26 studies ($N = 2266$)	WISC–III	87.75	88.24	89.77
	Difference	5.19	3.85	5.70
Zimmerman and Woo-Sam (1997)	WISC–R	103.28	101.04	105.12
Selection of 13 studies ($N = 1033$)*	WISC–III	98.16	97.51	99.21
	Difference	5.12	3.53	5.91
Zimmerman and Woo-Sam (1997)	WISC–R	103.28	101.04	105.12
13-study WISC–III scores adjusted †	WISC–III	97.98	97.33	99.03
	Difference	5.30	3.71	6.09
Rate of Gain 1972 to 1989 (IQ points per year)		0.312	0.218	0.358

*The 13 studies with the highest WISC–R Full Scale mean IQs selected out to approximate a normal IQ distribution (see text). †All WISC–III scores reduced by 0.18 points to compensate for practice effects (see text).

ples themselves may not be fully representative of Americans in general; the subjects used to compare the scores may be either too few or atypical in the sense of not covering the full range of IQs in a normal curve. Fortunately, Zimmerman and Woo-Sam (1997, pp. 532-533) can be used to eliminate the second source of error. They collated 26 studies, which collectively contained 2,266 subjects, and their results are presented in the second band of Table 1. The weighted averages come close to confirming the manual's values, except for a smaller gap between Verbal gains and Performance gains. It will be noted that these 26 studies had a disproportionate number of subjects with IQs below 100. For example, for Full Scale IQ, the WISC–R and WISC–III means are 92.94 and 87.75 respectively. Zimmerman and Woo-Sam list their studies from highest WISC–R Full Scale IQ to lowest. To approximate a normal distribution, I selected the top 13 studies (11 not counterbalanced, 2 counterbalanced) which collectively contain 1,033 subjects. The third band of Table 1 shows that these give WISC–R and WISC–III means of 103.28 and 98.16. Note that the 13 studies give differences close to those of the original 26, confirming that IQ gains are pretty uniform over the whole curve.

Table 1 ends by adjusting the WISC–III scores from Zimmerman and Woo-Sam's top 13 studies. The rationale is as follows. The 13 studies gave a difference (between WISC–R and WISC–III scores) for Full Scale IQ that was 0.18 points less than the manual's counterbalanced study. However, almost all of these studies were not counterbalanced but rather had subjects who took the WISC–III after the WISC–R: this would inflate the WISC–III scores because of practice effects. The intervals varied from 7 months to about 3 years, so the practice effects would be small. But they would be easily enough to account for 0.18 points. Therefore, it seemed sensible to deduct 0.18 points from the 13-study WISC–III scores, not just the Full Scale scores but the Verbal and Performance Scale scores as well. The resulting differences, of course, confirm the manual's estimate of Full Scale IQ gains at 5.30 points. But they alter the manual's estimates for Verbal and Performance gains, which shows how necessary the larger array of data was as a corrective. Performance IQ gains are now seen to be only slightly over two points (6.09 minus 3.71) higher than Verbal gains, while the manual's one study put them at fully five points (7.40 minus 2.40) higher. In sum, if the WISC–R and the WISC–III standardization samples were reliable, the following could be taken as firm estimates of IQ gains between 1972 and 1989: 5.30 points for Full Scale IQ, 3.71 points for Verbal, and 6.09 points for Performance. Over 17 years, these dictate rates of gain of 0.312, 0.218, and 0.358 IQ points per year, respectively.

Calculating The WAIS–III Estimate

The WAIS–III manual contains a single study of subjects who took

both the WAIS–R and the WAIS–III (Wechsler, 1997, Table 4.1). I have put the norming of the WAIS–III at 1995 because most of the members of the standardization sample had been selected and tested by the end of that year. I have put the WAIS–R at 1978, despite the fact that selection and testing was not completed until mid-1980. The testing was done over 4 years, so 1978 would have been about the midpoint (Wechsler, 1981, p. 18). Table 2 is based solely on the 192 subjects whose scores are given in the WAIS–III manual. The first section subtracts their mean IQ on the WAIS–III from their mean IQ on the WAIS–R, which gives the following estimated gains: 2.90 points for Full Scale IQ, 1.20 points for Verbal Scale, 4.80 points for Performance Scale. The period is again 17 years, dictating rates of gain of 0.171, 0.071, and 0.282 IQ points per year.

TABLE 2
U.S. IQ GAINS 1978 TO 1995 MEASURED BY WAIS–III VS WAIS–R

		Full Scale	Verbal	Performance
Differences before adjustment for extra low-IQ subjects				
WAIS–III manual ($N = 192$)	WAIS–R	105.80	103.40	108.30
	WAIS–III	102.90	102.20	103.50
	Difference	2.90	1.20	4.80
Rate of Gain 1978 to 1995: (IQ points per year)		0.171	0.071	0.282
Estimating the effects of the extra low-IQ subjects				
Data: WAIS–III Sample ($N = 2450$)		100.00	100.00	100.00
Low-IQ subjects ($N = 29$)		53.79	57.86	57.45
Example of calculations–Full Scale IQ				
$2421x = 2450\ (100.00) - 29\ (53.79)$				
$2421x = 243{,}440.09 \quad x = 100.554$				
Estimate: WAIS–III scores raised by 0.554 points (Full Scale), 0.505 points (Verbal), and 0.510 points (Performance)				
Gains after adjustment for extra low-IQ subjects				
Difference WAIS–R/WAIS–III		2.900	1.200	4.800
Plus effect low-IQ subjects		0.554	0.505	0.510
Difference Adjusted		3.454	1.705	5.310
Rate of Gain 1978 to 1995: (IQ points per year)		0.203	0.100	0.312

However, the norming of the WAIS–III marked a departure for the Psychological Corporation—which D. Tulsky and J. J. Zhu (Project Directors) called to my attention. For the first time, they used marketing research firms (over 20) to select subjects for the standardization sample. And they also, for the first time since norming the WISC in 1947, added "level of performance" to the usual stratification variables. That is, they checked the subjects of low-IQ who had been selected against 107 cases classified on clinical grounds as mentally retarded. They decided that the sample as selected had too few subjects at low-IQ levels and therefore, included an additional 29

(Tulsky & Zhu, 1997). This is much to be commended, indeed, it may be that the WAIS–III is the first adult IQ test whose norms below 70 are reliable. But it creates a problem of comparability between the WAIS–III standardization sample and earlier Wechsler samples. A similar problem arose when 55 subjects with low IQs were added to the WISC sample in 1947-1948. When calculating gains between the WISC and the WISC–R, the recommended procedure was to adjust WISC scores (by deducting 0.86 points) to achieve comparability (Flynn, 1987, p. 117). The rationale was that, if one sample (the WISC) contained additional subjects at low-IQ levels, while the other sample (the WISC–R) did not, then the norms of the altered sample would be less demanding.

The present problem is complicated by the fact that no technical manual exists for the WAIS–R that would allow us to assess whether its sample needed, yet did not get, additional subjects at low-IQ levels. Therefore, Table 2 gives scholars a choice. The first section calculates IQ gains from the WAIS–R and WAIS–III samples with no adjustment. As noted above, this yields gains of 2.90 points for Full Scale IQ, 1.20 for Verbal IQ, and 4.80 for Performance IQ. The second section indicates how including additional subjects of low IQ affected test scores. As the "Example of calculations" shows, including 29 subjects with a mean Full Scale IQ of 53.79 would raise WAIS–III scores by 0.554 points. Since this would reduce the difference between WAIS–III and WAIS–R scores by that amount, the third section adds 0.554 points to that difference, thereby boosting the estimate of Full Scale IQ gains up to 3.454 points. Similar adjustments were made for Verbal and Performance IQ gains, putting them at 1.705 points and 5.310 points, respectively. The period between norming the two tests remains at 17 years, so the latter estimates give rates of gain of 0.203, 0.100, and 0.312 IQ points per year.

Analyzing The WISC–III And WAIS–III Estimates

The WAIS–III manual's estimates, based on one study with 192 subjects, may have to be modified some years from now, when we collate many studies representing 1,000 or more subjects. But for now, these estimates are all we have and clearly, they are lower than the WISC–III estimates. The discrepancies suggest four hypotheses.

The first hypothesis is that the WISC–III and WAIS–III estimates are both valid and that the latter simply means that IQ gains stopped in 1989. Arithmetic shows this is mathematically possible. Working with Full Scale IQ, assume the WISC–III rate of 0.312 from 1978 to 1989 and a zero rate from 1989 to 1995: the total IQ gain from 1978 to 1995 would be 3.432 points. This is too large to reconcile with our unadjusted estimate for the WAIS–III (2.90 points), but it does match the adjusted estimate (3.454

points). However, how likely is the assumption on which it rests, that U.S. IQ gains went at a robust rate from say 1932 to 1989 and then stopped dead in that year?

This brings us back to the WAIS–III manual because it includes another study of crucial significance: 184 subjects took both the WISC–III normed in 1989 and the WAIS–III normed in 1995 (Wechsler, 1997, Table 4.3). This study can be interpreted in two ways: either as a measure of IQ gains between 1989 and 1995 or as a vehicle for comparing the WISC–III and WAIS–III standardization samples to see if one or both suffered from sampling error. Our third and fourth hypotheses assign it the second role. Note what assigning it the first role would entail. As a measure of IQ gains between 1989 and 1995, it gives a nil result and as indicated above, it is simply not plausible that IQ gains stopped dead in 1989. Moreover, two samples only six years apart have no credibility as a measure of IQ gains. Anticipated gains over six years would be only about 1.5 points and sampling error can easily create a difference between two standardization samples of that magnitude. In other words, six years is just not enough time for gains to swamp sampling error and render it insignificant.

The second hypothesis is that the WISC–III and WAIS–III estimates are both valid and signal gains that are age-specific. Given that the periods the estimates cover largely overlap, this hypothesis has strong *prima facie* plausibility. During 1972 to 1995, there is no reason why the rate of gain should not have been 0.312 points per year for 6- to 16-yr.-olds, and drop to 0.171 or 0.203 points for 16- to 74-yr.-olds. However, the plausibility of this hypothesis must be assessed against that of the remaining hypotheses. An aside, positing age-specific gains does nothing to resurrect the WISC–III to WAIS–III study as a measure of IQ gains over time. This study consisted solely of 16-yr.-olds because that is the only age the two tests have in common (The Psychological Corporation, 1997). Therefore, treating it as a measure of IQ gains would imply that gains were nil or worse than nil at age 16 and significant at every other age from 6 to 74. That is hardly plausible.

The third hypothesis is that the WISC–III standardization sample is valid but the WAIS–III sample substandard. Table 3 explores this hypothesis and compares it with its mirror-image, the fourth hypothesis. The table has three sections.

The first section of Table 3 confirms our assertion that the WISC–III versus WAIS–III study at best gives a nil result for IQ gains. When no adjustment is made for the effect of extra subjects of low IQ in the WAIS–III standardization sample, it gives IQ losses of 0.400 to 0.700 points. When such an adjustment is made, the losses essentially disappear for a nil gain. I have used the adjusted estimates throughout the rest of the table, but this does not affect the analysis.

TABLE 3
WISC–III vs WAIS–III Sampling Error Estimated by Comparative Data

	Full Scale	Verbal	Performance
Comparative Data: from 184 subjects who took both tests			
Difference WISC–III/WAIS–III	–0.700	–0.500	–0.400
Plus effect low-IQ subjects	*0.554*	*0.505*	*0.510*
Actual (or adjusted) differences	–0.146	0.005	0.110
Third hypothesis: WISC–III sample valid; WAIS–III sample substandard			
Predicted Difference	1.872	1.308	2.148
Actual Difference	–0.146	0.005	0.110
WAIS–III Substandard	2.018	1.303	2.038
Example of calculations–Full Scale IQ			
Data: WISC–III rate of gain equals 0.312 points per year (Table 1)			
Period between WISC–III and WAIS–III equals 6 years			
Therefore, 0.312 × 6 = 1.872, as predicted difference			
WAIS–III Gain (Table 2)	3.454	1.705	5.310
Plus effect sample substandard	2.018	1.303	2.038
WAIS–III Gain Adjusted	5.472	3.008	7.348
WISC–III Gain. (Table 1)	5.300	3.710	6.090

Conclusion.—Third hypothesis reconciles Full Scale values—note that the WAIS–III Gain Adjusted and the WISC–III Gain are virtually the same. It eliminates 65% of the Verbal discrepancy—falls 0.702 short of explaining a 2.005 point discrepancy. It merely turns the Performance discrepancy from a negative one to a positive one—a short-fall of 0.780 points becomes an excess of 1.258 points. The discrepancies referred to are the differences between the WAIS–III Gain (labelled Table 2) and the WISC–III Gain (labelled Table 1) given immediately above.

Fourth hypothesis: WAIS–III sample valid; WISC–III sample elite.			
Predicted Difference	1.218	0.600	1.872
Actual Difference	–0.146	0.005	0.110
WISC–III Elite	1.364	0.595	1.762
Example of calculations–Full Scale IQ			
Data: WAIS–III rate of gain equals 0.203 points per year (Table 2)			
Period between WISC–III and WAIS–III equals 6 years			
Therefore, 0.203 × 6 = 1.218, as predicted difference			
WISC–III Gain (Table 1)	5.300	3.710	6.090
Minus effect sample elite	1.364	0.593	1.762
WISC–III Gain Adjusted	3.936	3.115	4.328
WAIS–III Gain (Table 2)	3.454	1.705	5.310

Conclusion.—Fourth hypothesis reconciles Full Scale values reasonably well by eliminating 74% of the discrepancy—falls 0.482 short of eliminating a 1.846 point discrepancy. It eliminates only 30% of the Verbal discrepancy—falls 1.410 short of eliminating a 2.005 point discrepancy. It merely turns a Performance discrepancy from a positive one to a negative one—an excess of 0.780 points becomes a short-fall of 0.982 points. The discrepancies to be explained, the differences between the WISC–III Gain (labelled Table 1) and the WAIS–III Gain (labelled Table 2) immediately above, are the same for both hypotheses. But the fourth hypothesis explains them less well than the third hypothesis.

The second section of Table 3 treats the WISC–III sample as valid and uses its rate of gain to predict how much lower WAIS–III scores should have been than WISC–III scores. As the "Example of calculations" shows,

given that its standardization sample was tested six years later and the rate of gain was 0.312 points per year, WAIS–III scores should have been 1.872 points lower for Full Scale IQ. And, when a small difference in the other direction is taken into account, the implication is that the WAIS–III standardization sample was substandard by 2.018 points, and therefore, that it inflated WAIS–III scores by that amount. The second section closes with a conclusion that assesses the hypothesis that the WISC–III sample was valid. That hypothesis does an excellent job of explaining the discrepancy between WISC–III and WAIS–III estimates for Full Scale IQ gains, a good job for Verbal IQ but does not explain the discrepancy for Performance IQ.

The fourth hypothesis is that the WAIS–III sample is valid but the WISC–III sample elite. The third and last section of Table 3 explores this hypothesis. This hypothesis entails treating the WAIS–III rate of gain as valid and using it to predict how much lower WAIS–III scores should have been than WISC–III scores. The "Example of calculations" multiplies six years (the fact that the WAIS–III standardization sample was tested six years later of course remains the same) by the new and lower rate of gain (now 0.203 points per year); this shows that WAIS–III scores should have been 1.218 points lower for Full Scale IQ. Taking the small difference in the other direction into account, the implication is that the WISC–III standardization sample was elite by 1.364 points. The third section's conclusion assesses the hypothesis that the WAIS–III sample was valid. Compared to the other hypothesis (WISC–III sample valid), it explains the discrepancies between WISC–III and WAIS–III estimates for IQ gains less adequately. About 74% of the discrepancy for Full Scale IQ gains is explained (rather than all), 30% of the discrepancy for Verbal IQ is explained (rather than 65%), and the failure to explain the discrepancy for Performance IQ is much the same.

Fairness dictates adding that the Psychological Corporation is preparing an analysis inclusive of the study in which 184 subjects took both the WISC–III and WAIS–III, and that they believe they can defend both standardization samples against the suspicion of sampling error (Zhu & Tulsky, 1997). Their analysis will focus on differences between the two tests such as that there were penalties or bonuses for speed on the WISC–III but a lack of such on the WAIS–III; that the WISC–III discriminates better at low-IQ levels and the WAIS–III at high-IQ levels; and so forth. Their analysis cannot be assessed until complete, but two points appear relevant. First, the factors they name look like they would lower correlations between the two tests rather than produce mean score differences. Second, the sampling errors hypothesized are the sort to be expected from any stratified sample however carefully selected. If the methods of the Psychological Corporation make samples accurate within one or two points, their samples deserve the

respect they have been accorded: use as the standard vehicle for estimating IQ gains over time.

Compensating for Obsolete Norms

The crunch question is what are we to do between now and say the year 2008, when the Psychological Corporation may give us another relevant standardization sample? The hypotheses that posit sampling error cannot be dismissed. Unless one rejects both the WISC–III and the WAIS–III estimates of IQ gain, the scores of those who took both tests look like symptoms of sample differences. Unfortunately, the sampling error hypotheses, despite slightly favoring the validity of the WISC–III standardization sample, do not really tell us which sample and which estimated rate of IQ gain is correct. Either the WISC–III sample could be elite and a rate of 0.171 or 0.203 points per year be correct, or the WAIS–III sample could be substandard and a rate of 0.312 correct.

For the present, I recommended that we split the difference between the two rates. However unsatisfactory this seems, the only alternative would be to choose between them by the flip of a coin. This expedient gives either 0.241 points per year (using the unadjusted WAIS–III rate) or 0.258 points (using the adjusted rate). Both estimates round off to something like 0.25, a value easy to remember, for the years 1972 to 1995. Projected into the post-1995 period it should give a workable compensation for obsolescence. The estimate and projection apply, of course, only to U.S. gains on *these two* Wechsler IQ tests. Finally, the strong possibility of age-specific gains should be kept in mind for the day when WISC–IV and WAIS–IV standardization samples become available.

REFERENCES

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

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

The Psychological Corporation. (1997) *The WAIS–III/WMS–III technical manual.* San Antonio, TX: The Psychological Corp.

Tulsky, D., & Zhu, J. (1997) Lowering the floor of the WAIS–III. Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, Illinois.

Wechsler, D. (1981) *Wechsler Adult Intelligence Scale–Revised: manual.* New York: The Psychological Corp.

Wechsler, D. (1992) *Wechsler Intelligence Scale for Children–Third Edition: manual* (Australian Adaptation). San Antonio, TX: The Psychological Corp.

Wechsler, D. (1997) *Wechsler Adult Intelligence Scale–Third Edition: manual.* San Antonio, TX: The Psychological Corp.

Zhu, J., & Tulsky, D. (1997) Validity studies of the WAIS–III. Paper presented at the 105th Annual Convention of the American Psychological Association, Chicago, Illinois.

Zimmerman, I. L., & Woo-Sam, J. M. (1997) Review of the criterion-related validity of the WISC–III: the first five years. *Perceptual and Motor Skills*, 85, 531-546.

Accepted April 24, 1998.