JOURNAL OF EDUCATIONAL MEASUREMENT
VOLUME 21, NO. 3, FALL 1984, pp. 283–290

IQ GAINS AND THE BINET DECREMENTS

JAMES R. FLYNN
University of Otago

Thorndike compiled Stanford-Binet data which made it appear that children aged 6 and under have made greater IQ gains than older children and that this pattern dominated the whole period from 1932 to 1971–72. Therefore, he sought causal factors likely to affect preschoolers more than others, for example, TV in general and educational TV in particular. A wide array of data show that the atypical gains of young children are either an artifact of sampling error or totally antedate 1947, ruling out TV as an age-specific factor. This data also suggest that Americans have gained about 12 IQ points from 1932 to 1972 with verbal gains being a point lower and performance gains a point higher.

The 1972 Stanford-Binet is essentially the same in content and administration as the 1960 version, the Stanford-Binet LM, and the latter is a composite selected from the old Stanford-Binet L and the alternate form M, the tests that appeared in the 1930s. Therefore, all items on the current test date back to its origins, and it was possible to score the current standardization sample, tested in 1971–72, on the old norms set by the 1932 standardization sample. When Thorndike (1973, p. 359) did this, he found that the IQ gains American children enjoyed between 1932 and 1971–72 were age-specific: Children aged 2 to 6 had made large gains, those aged 7 to 16 modest gains, those aged 17 to 18 gains somewhere in between. For example, the gains of preschoolers were at least 8 IQ points higher than those of 11-year-olds.

Thorndike (1977, p. 197) used the pages of this journal to propose a number of sensible hypotheses about the causation of these age-related differences, including the hypothesis that they were an artifact of deficiencies in the 1971–72 standardization sample: Perhaps the preschoolers were an unrepresentative elite. The one hypothesis not proposed was that the differences were an artifact of deficiencies in the 1932 standardization sample, perhaps because there seemed to be no way of testing its truth.

With this hypothesis in mind, I decided to divide the 40 years spanned by Thorndike's study into two shorter periods to see whether an age-specific pattern emerged throughout. The first step was to calculate IQ gains as measured by the differential performance of the SB standardization sample of 1932 and the WISC standardization sample of 1947–48. The method was to utilize all studies in which the same subjects had taken both the Stanford-Binet, either the LM or the L or the M, and the WISC; the only studies rejected were those with a clear methodological flaw, such as those with too long an interval between testings, or those dealing with atypical subjects, primarily the mentally retarded. If subjects consistently scored higher on the SB than the WISC, the obvious explanation would be that the earlier test had norms easier to meet than the later test, thanks to the inferior performance of its standardization sample. For example, if subjects averaged 105.85 on the SB and only 100 on the WISC, this would evidence an IQ gain of 5.85 points between 1932 and 1947–48.

To render scores from various tests comparable, we need a uniform scoring convention and therefore, I have translated all results into deviation IQs with white mean and *SD* set at 100 and 15, respectively. The matter is straightforward enough although there are a few preliminary steps. Thanks to some alterations made in 1960, SB-L (or M) norms and SB-LM norms differ, the latter being about 2 IQ points tougher. I have chosen to translate SB-LM scores into SB-L norms (although converted into deviation IQs) rather than the reverse: first, because far more of our data comes from the SB-L; and second, because the SB-L norms probably give a truer picture of the performance of a representative sample of American whites circa 1932. However, it is essential to underline this: Whichever scale we use makes no significant difference to the age-specific pattern of IQ gains which caught Thorndike's eye and posed our problem of causal explanation. I have also translated results based on all races' norms into whites' only norms rather than the reverse. This too makes no difference, and actually, it is the only translation the data permits. Thorndike (1975, p. 6) concurs that this kind of translation is needed to secure comparable scores.

Table 1 shows how things appear when we are limited to Thorndike's data, that is, data covering the whole period from 1932 to 1971–72. Taking age 11 as a low point of IQ gains, young children exhibit far greater gains, an excess of about 3 points at age 6 rising to an excess of over 8 points for age 2. Just to allay suspicion, I have included values (within brackets) from Thorndike's data as he presented it to show that my manipulations have not much affected the pattern. Given Thorndike's assumption, that the unusual gains of young children extended over the whole period up to 1972, it was quite reasonable to seek causal factors that might have an unusual impact on preschoolers, such as TV and particularly educational TV of the "Sesame Street" type (Thorndike, 1975, p. 6; 1977, p. 197). However, Table 1 also includes values for IQ gains based on a simulation of how the WISC standardization sample of 1947–48 would have scored against 1932 Stanford-Binet norms. This is based on 24 studies with 2,187 subjects, and those who wish to replicate the steps between these studies and the simulation should see both the References and the Appendix. Ideally, each study would give us a WISC and SB mean for each age, that is, average scores for 5-year-olds, averages for 6-year-olds, and so forth; in fact, we usually get an overall mean for a group of subjects whose ages range over a few years. But still, the ranges are narrow enough so that the studies collectively, when merged, should reveal an age-specific pattern if it exists.

The pattern that emerges from the WISC data is quite astonishing: Even though we cannot go below age 5, it is clear that the unusual gains of young children were fully intact prior to 1947. It will be recalled that Thorndike also saw a pattern of unusual gains for teenagers, and these are at least hinted at in the pre-1947 data, although Table 1 separates teenagers off by a gap to emphasize that by here we are losing reliability due to smaller samples. Table 1 leaves us with two possibilities: Either age-specific gains occurred with a vengeance before 1947, at a far greater rate than Thorndike suspected because they now achieve their total effect in 15 years rather than almost 40 years, and then stop abruptly after 1947 in favor of similar gains for all ages; or even the age-specific gains prior to 1947 are nonexistent and an artifact of sampling error, perhaps defects in the 1932 Stanford-Binet sample which is the oldest and most primitive of the standardization samples involved. The latter possibility seems the more probable, but even if we take the unusual gains of young children pre-1947 seriously, we can rule out TV which reached most Americans only after that date. TV may have the power to engender IQ gains, but there is no evidence that it affects one age group more than another.

Table 1

Comparison of IQ Gains during Two Periods

IQ Gains 1932 to 1971-72				IQ Gains 1932 to 1947-48			
Age	N	Gains[a]	Gains x age[c]	Age	N	Gains[b]	Gains x age[c]
2	95	15.00	8.25 (8.3)				
3	186	15.70	8.95 (8.5)				
4	235	14.11	7.36 (8.3)				
5	242	11.29	4.54 (6.8)	5	298	10.09	6.14
6	131	9.47	2.72 (4.9)	6	244	8.19	4.24
7	150	8.44	1.69 (2.8)	7	335	5.90	1.95
8	128	7.78	1.03 (1.1)	8	270	5.01	1.06
9	110	7.59	.84 (-.1)	9	308	4.72	.77
10	130	7.40	.65 (-.3)	10	281	4.15	.20
11	135	6.75	.00 (.0)	11	217	3.95	.00
12	137	7.03	.28 (.3)	12	117	4.34	.39
13	128	6.47	-.28 (.7)	13	73	5.57	1.62
14	129	6.84	.09 (1.1)	14	22	6.14	2.19
15	135	8.34	1.59 (1.7)	15	22	6.32	2.37
16	94	10.03	3.28 (2.5)				
17	100	11.90	5.15 (3.5)				
18	86	13.97	7.22 (4.7)				
	2351	9.89			2187	5.85	

[a] Gains based on how SB standardization sample 1972 scored against 1932 norms, using SB-L scale converted to deviation IQs and translated into uniform scoring convention with white mean and SD = 100 and 15.

[b] Gains based on a simulation of how WISC standardization sample 1947-48 would have scored against 1932 norms, using SB-L scale as above.

[c] The number of points by which each age exceeds the IQ gains of age 11; the values in brackets refer to Thorndike's original data before rescoring and translation.

Note. Sources for this table are in Flynn, 1984, Table 2, Numbers 1, 2, 3, and 7.

As to why I link the possibility of sample bias to Stanford-Binet 1932 rather than the WISC sample, it is because the former seems to supply the necessary ingredient to produce our age-specific pattern. Take the SB 1932 with the WISC and you get the pattern; take the SB 1932 with the SB 1971–72 and you get the pattern. On the other hand, as we shall see, if you take the WISC with the WISC-R standardization sample, the pattern disappears. Thorndike (1977) gave us additional evidence with a follow-up study after 3 years of SB 1971–72 subjects aged 3 to 6, which showed that sample bias would be unlikely to account for the larger part of the advantage of younger over older children. The reader is asked to take it on faith, only for the moment, that Americans have been gaining about .300 IQ points per year, for this would reduce the latitude for sample bias further to only 1 or 2 points, and even that is not certain. In other words, when we take the SB 1932 and SB 1971–72 samples together, we have good reason to exonerate the latter for producing our age-specific pattern and indict SB 1932. Therefore, if SB 1932 and WISC samples together produce the very same age-specific pattern, it is more likely that it is SB 1932 which is at fault.

The above conclusion, however tentative, is subject to an objection very powerful at first glance. If the 1932 norms were at fault, the most parsimonious hypothesis is that the norms for ages 2 to 5 were substandard at least to some degree: perhaps not by a full 3 to 8 points, for after all the norms above age 5 could be a bit elite, but by a few points at any rate. This conflicts with the usual views of testing organizations that preschoolers, being less accessible for testing than older children, tend to be an elite sample rather than substandard. The flaw in this objection is the assumption that substandard *norms* for 1932 preschoolers signal the existence of a substandard *sample*. In fact, the sample for ages 2 to 5 was known to be elite, and the testers decided to compensate in setting their norms (McNemar, 1942, pp. 20 & 35–37; Terman, 1942, pp. 12–13). They put their norms for preschoolers 3 to 7 points below the performance of the standardization sample (Terman & Merrill, 1937, p. 35), and it is not unlikely that this was a few points too many.

Returning to Table 1, it implies that IQ gains since 1947–48 have been similar for all ages, and it seemed desirable to collect independent evidence on this point. Table 2 supplies that evidence and shows that IQ gains between 1947–48 and 1972 have been remarkably uniform for all children aged 5 to 17 and even for a group of adults aged 35 to 44. It is based on 22 studies with a total of 1,204 subjects all of whom took two Wechsler tests normed at different times (see both References and Appendix as a guide to replication). The method remains the same as already described: If subjects did better on the WISC (normed 1947–48) than the WISC-R (normed 1972), it was assumed that the earlier standardization sample was inferior to the later one; and if they did 8.20 points better, that was assumed to measure IQ gains over this 24½-year period. Unfortunately, these Wechsler studies each taken singly tend to have subjects covering a fairly wide age-range and this would work to blur age-specific gains when they are merged. But the data does divide nicely at certain points, and Table 2 uses gaps to separate those age-groups between which differential gains should appear if they exist, namely, ages 5–6, 7–10, 11–15, 16–17, and 35–44. We see that IQ gains do not vary substantially across these age-groups, indeed, the uniformity is so striking it cannot be explained by the imperfections of the data.

Table 2 supports Table 1 in its implication of uniform IQ gains since 1947. However, taken together, the two tables pose a problem serious enough to demand attention. First,

Table 2

IQ Gains 1947-48 to 1972

Age	N	Gains[a]	Test combinations[b]
5	54	7.08	WISC & WPPSI plus
6	60	7.74	WPPSI & WISC-R
7	55	8.05	WISC & WISC-R
8	114	8.26	
9	128	8.23	
10	132	8.36	
11	156	8.43	WISC & WISC-R
12	114	8.13	
13	112	8.27	
14	92	8.46	
15	75	8.71	
16-17	40	8.81	WAIS & WISC-R
35-44	72	8.04	WAIS & WAIS-R
	1204	8.20	

[a] Gains based on simulations of WISC-R (ages 5-17) and WAIS-R (ages 35-44) standardization samples scored against 1947-48 norms, WISC scale; also translated into uniform scoring convention with white mean and SD = 100 and 15.

[b] For ages 5-6, gains from WISC to WPPSI and gains from WPPSI to WISC-R were summed to estimate WISC to WISC-R gains. For ages 16-17, a gain was computed from WAIS to WISC-R to cover the 18-year period 1953-54 to 1972; this was increased by one-third to cover the 24 years from 1947-48 to 1972 under the assumption that gains over the longer period were constant year by year. For ages 35-44, the gain from WAIS to WAIS-R covers the 24 years from 1953-54 to 1978 and this was assumed to be identical with the gain over the 24 years from 1947-48 to 1972.

Note. Sources for this table are in Flynn, 1984, Table 2, Numbers 11, 13, 14, 15, and 17. Schwarting has been dropped because his results cover all ages and therefore are useless for our purpose.

Thorndike's data in Table 1 give an overall gain for 1932 to 1971–72 amounting to 9.89 points over a period of 39½ years. If we expand this on the assumption of equal gains year by year during the full 40 years from 1932 to 1972, we get a gain of 10.02 points. Now note that the SB to WISC data in Table 1 give an overall gain of 5.85 points from 1932 to 1947–48; and finally, that the Wechsler data in Table 2 give an overall gain of 8.20 points for 1947–48 to 1972. Add these two together and the result is a gain of 14.05 points covering the full 40 years from 1932 to 1972. In other words, there is a discrepancy: Thorndike's Stanford-Binet data suggest 10 points and my mainly Wechsler data suggest 14 points as the overall IQ gain for the same 40-year period.

We can narrow this discrepancy somewhat. The Stanford-Binet is primarily a verbal test, so it seems profitable to determine whether the Wechsler gains have been less for the verbal scale than for full scale IQ. One would expect no difference to emerge from this in the Table 1 Wechsler estimate, but I have analyzed the Table 2 Wechsler data, and it does show that verbal gains were 1.10 points less than full scale gains for 1947–48 to 1972. We have explained away about 1 point of our 4-point discrepancy leaving 3 to go. Some readers may be aware of the thesis that WISC to WISC-R data give inflated estimates of IQ gains due to differential practice effects and be tempted to posit an inflation of 1 or 2 points, all that could reasonably be argued, to take us further. However, I have elsewhere (Flynn, 1984, p. 32) presented considerable evidence against this thesis and therefore advise caution. I suspect that the remaining 3 points derive from sampling error: Perhaps the 1972 Stanford-Binet standardization sample was substandard by a point, the WISC-R sample elite by a point, and another point could come from the method of combining lots of little studies to simulate a comprehensive sample.

A final possibility: In 1960 the Stanford-Binet organization made some changes in the scoring and administrative practices of the 1930s and these changes were carried over to 1972 (Terman & Merrill, 1973, p. 25). If the 1971–72 standardization subjects were marked more strictly than their 1932 counterparts, this would depress their performance and do something to account for the lower estimate of IQ gains from Thorndike's data. What we really need are studies in which subjects actually take both the old Stanford-Binet L and the Stanford-Binet 1972 in counterbalanced order. Until then, the best we can do is estimate that Americans gained about 12 IQ points from 1932 to 1972, a rate of gain of .300 points per year, while acknowledging that the gain could well be 2 points higher or lower. Whatever estimate we eventually establish for full scale IQ, verbal gains were about a point lower and performance gains a point higher. But however tentative the above, our mass of data point in one direction without ambiguity: The evidence for unusual gains by young children is weak for the period prior to 1947 and nonexistent thereafter. Attempts to provide causal explanations for age-specific gains are premature.

Appendix

What follows will allow the replication of all necessary calculations, but those who have not read the background literature can easily go astray and correspondence with the author is invited.

First, the calculation of IQ gains from 1932 to 1947–48 as presented in Table 1. The method consists of using studies in which subjects took both the WISC and one of the relevant forms of the Stanford-Binet, either the SB-L or SB-M or SB-LM (1960). The WISC results from each study can stand unaltered in that they are already deviation IQs with white mean and *SD* = 100 and 15, respectively. The SB results can be translated into that convention by using the table Terman and Merrill (1973, pp. 339–341) have provided, but it must be used in this way: SB-L and SB-M scores, (Score − 100) *K* + 100; SB-LM scores, Score + (*Mean* − 100). The *K* and *Mean* used in these equations refer to the headings of columns in Terman and Merrill's table. The adjusted scores will all refer to a convention in which white mean circa 1932 was 100 and white *SD* was 16. As noted in the text, the resulting scale is not that of the SB-LM (1960) but rather that of the old SB-L, adjusted slightly to convert from "mental age" IQs into deviation IQs. Then we must convert from white *SD* = 16 to white *SD* = 15 but this is simple enough, for example, a score of 116 becomes 115.

Now that we have comparable SB and WISC scores, the next step is to divide the subjects of each study by age. Unless there was an indication to the contrary, this was done on the assumption that each age spanned by a group is represented by equal numbers, for example, a study of 32 subjects aged 5, 6, and 7 entails allocating 10.67 subjects to each of those ages. When the subjects of all 24 studies have been allocated, we calculate weighted mean scores for each age. For all subjects aged 5 who took both tests, the SB mean was 111.64, the WISC mean 100.44, a difference of 11.20 points. Thus, it is concluded that the WISC standardization sample of 1947–48 would have scored 11.20 points above the 1932 Binet norms, that is, SB-L norms adjusted as described so as to produce deviation IQs with white *SD* = 15. Finally, the results were smoothed with a 3-year moving average just as Thorndike's were, and this gave an IQ gain of 10.09, the value for 5-year-olds to be found in the relevant column of Table 1.

Second, the calculation of IQ gains from 1947–48 to 1972 as presented in Table 2. The method consists of using studies in which subjects took two Wechsler tests, normed about 25 years apart, although for ages 5 and 6, the WPPSI was used as a bridge between the WISC and WISC-R as noted in Table 2 itself. Despite the variety of tests involved, the calculations for all pairs of tests pose very much the same problems, and therefore, I will deal only with studies in which subjects took both the WISC and WISC-R. Once again, the WISC results from each study can stand unaltered in that they are already deviation IQs with white mean and *SD* = 100 and 15, respectively. The WISC-R results are another matter having been scored against the all races standardization sample of 1972, whose white members had a mean of 102.26 and an *SD* of 14 (Kaufman & Doppelt, 1976, p. 167); these values can be taken as roughly accurate for all Wechsler tests that have post-dated the WISC because all of them were normed in precisely the same way (Flynn, 1984, pp. 30–31). Therefore, to translate a WISC-R score of say 88.26 on the WISC scale, we must proceed as follows: 88.26 − 102.26 = −14.00 or 14 points below the white mean; −14.00 ÷ 14 = one white *SD* below the white mean. In other words, a WISC-R score of 88.26 translates into 85 on the WISC scale, a scale on which white mean and *SD* = 100 and 15, respectively.

The translating done, proceed as before and apportion the subjects of each study to the appropriate age, and when the subjects of all 16 studies have been so allocated, calculate weighted mean scores for each age. For all subjects aged 7 who took both tests, the WISC mean was 99.77, the WISC-R mean 90.83, a difference of 8.94 points. Thus, the WISC-R standardization sample of 1972 would have scored 8.94 points above the WISC norms of

1947–48. When smoothed with a 3-year moving average, this becomes 8.05, the value given for 7-year-olds in Table 2. Although the results for ages 5 to 15 were smoothed, the unsmoothed results were so uniform this had little effect.

REFERENCES

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

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

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

TERMAN, L. M. (1942). The revision procedures. In Q. McNemar (Ed.), *The revision of the Stanford-Binet Scale.* Boston: Houghton Mifflin.

TERMAN, L. M., & MERRILL, M. A. (1937). *Measuring intelligence.* London: Harrap.

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

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

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

THORNDIKE, R. L. (1977). Causation of Binet IQ decrements. *Journal of Educational Measurement, 14,* 197–202.

AUTHOR

JAMES R. FLYNN, Professor and Chairman, Department of Political Studies, University of Otago, Box 56, Dunedin, New Zealand. *Degrees:* BA, MA, PhD, University of Chicago. *Specializations:* Moral philosophy, psychology and race.