CHAPTER XXIV

RETEST CHANGES IN THE IQ IN CERTAIN SUPERIOR SCHOOLS

ROBERT L. THORNDIKE
Assistant Professor of Educational Psychology

Teachers College, Columbia University
with the coöperation[1] of
CECILE WHITE FLEMMING
Horace Mann School
GERTRUDE HILDRETH
Lincoln School
MARGARET STANGER
Ethical Culture School
New York, New York

I. RETESTING IN THE UNIVERSITY OF IOWA SCHOOLS

Comparing two groups of children who had attended the University of Iowa preschool, Wellman[2] found that those who transferred to other schools showed no further gain in IQ, though maintaining the gains registered during preschool attendance, while those who continued to attend the University school showed an additional average gain in IQ of about 7 points. Elsewhere Wellman[3] has reported a correlation of .40 between performance on an intelligence test at college entrance and length of attendance at the University schools. The inference follows that something in the superior environment of the University School at Iowa produced a more rapid growth in intelligence, as tested by the Binet, than occurred in other school environments. One is led

[1] The coöperating authors supplied test results from the records of their respective schools and information about the conditions of testing. The analysis of the data and preparation of the report were carried out by Doctor Thorndike.

[2] B. L. Wellman. "Growth in intelligence under differing school environments." *Jour. Exper. Educ.*, 3: 1934-1935, 59-83.

[3] B. L. Wellman. "Mental growth from preschool to college." *Jour. Exper. Educ.*, 6: 1937-1938, 127-138.

to inquire whether this gain is unique to the University of Iowa school, or whether similar phenomena will be observed in other schools that have the advantages of superior children, superior facilities, and presumably superior teachers and curricula. The significance of the Iowa findings will be increased many fold if they are confirmed in data gathered from a completely different source. On the other hand, of course, even if we do not find such gains in the other schools here studied or in any other school in the country, we still cannot *prove* that the results obtained at Iowa are not genuine. However, negative evidence from other sources will tend to throw the burden of proof upon the Iowa experimenters.

II. General Conditions of Retesting in the Three Schools at New York

In order to provide further data for understanding the effect of schooling upon the IQ, we analyzed the Binet retest records that have been accumulated in the files of three of the best-known private schools in and around New York City—Ethical Culture, Horace Mann, and Lincoln. Some of the earlier retest material from one of the schools has already been reported by Hildreth[1] and by Rugg and Colloton.[2] These records have been accumulated over the past twenty-odd years. They represent retest data on a total of about 3,000 children. Over 1,100 of these retests had been given after an interval of at least 2½ years, and these records will be the ones on which most of our analysis will be based.

All the available records were used, with the following exceptions:

1. No test was considered when the child was over 14 years at the time of testing; this exception was made in order to avoid any possible question as to what chronological age to use in computing IQ's for older children.

2. All later tests were eliminated in which the M.A. predicted from the IQ on the first test would be over 18 years; this exception was made in an effort to eliminate records that might show a spurious decline in IQ because the child had 'gone through the roof' of the test.

The testing was incomplete, in that these 3,000 children represented only a fraction of the total school populations of these schools during

[1] G. Hildreth. "Stanford-Binet retests of 441 school children." *Ped. Sem. and Jour. Genet. Psychol.*, 33: 1926, 365-386.

[2] H. O. Rugg and C. Colloton. "Constancy of the Stanford-Binet IQ as shown by retests." *Jour. Educ. Psychol.*, 12: 1921, 315-322.

the period covered in the data. A great many children never received more than a single Binet.

We must recognize a possibility that the individuals who received more than one Binet and who make up the population we are dealing with were in some crucial way not representative of the school population as a whole. In an effort to check on this question, statements were requested from the psychologists in charge of testing at each of the three schools. Pertinent excerpts from these statements are these:

> *School A.* Each year at the beginning of the year my office makes up a list of the children who have not had tests for 3 years or more. Then . . . we systematically go through the list, taking children first for whom the time interval since the last test has been longest. However, we are not always able to stick to this schedule . . . Occasionally calls come from teachers or . . . principals to have certain children retested . . . I would say . . . that the total effect in selection, so far as IQ range is concerned, would be random.
>
> *School B.* During the last 8 years . . . selective factors have to a very slight extent determined which children should have retests. Our regular practice is to retest all pupils in the elementary school at an interval of two or three years . . . This practice of periodic retesting takes care of most of the children who . . . might otherwise have been selected for retests because of some special factor or condition . . . I cannot be sure to what extent special factors did operate in this school in earlier years. I am inclined to believe that the usual reasons for which teachers asked for retests were disparities between the child's work in school and his general aptitude as indicated by the test score on file; or a change in certain other aspects of the child's development or experience which made it desirable to check any consequent change in his mental function that might be revealed by the repeated Stanford-Binet examination. Such changes would include, obviously, improvement in physical condition, recovery from serious illness which might have affected the earlier test results, or any marked change in emotional control or habits of attention and work. The disparity between the evidence of ability as shown by school work and that indicated by the MA and IQ might point to a mental ability either greater or less than that recorded by the questioned IQ.
>
> *School C.* There are a few whose retests are requested by teachers because of doubt as to the validity of the old ones, or because of

> seeming 'slipping.' However, I should roughly estimate this class as comprising not more than 2 percent at the very most. I think fully 98 percent, in other words, are retested simply because it is our aim not to have to trust any one test.

These statements leave one with the impression that any selective effects were probably slight and not of great importance in determining the results. The question of selection will be considered further when the results have been presented and are being discussed.

The quality of the testing represented in these records is probably somewhat variable; part of it was done by advanced graduate students and part of it by experienced examiners. In School C, about nine-tenths of the tests were given by one or the other of two individuals who have served the school as psychologist during the last 20 years. In School A, over two-thirds of the tests were given by one or another of the three examiners. In School B, however, the testing was spread more widely among a number of examiners, and a goodly bit of the testing was done by advanced graduate students and internes. These students were carefully supervised, and their work checked, but of course they did not have the richness of experience of a school psychologist with years of testing experience. It is hard to see why the number of examiners should affect the amount of *constant* change in IQ, although we should expect the variation from test to retest to be somewhat greater with a greater number of examiners, especially if this included some less-experienced examiners.

III. Analysis of Data Secured at New York

First let us consider all the children from a given school who had a retest after an interval of 2½ years or more. In the case of any individual who had received more than two tests while in the school in question, the first and the last test were the ones considered. In every case, the differences with which we will work are obtained by subtracting the IQ on the first test from the IQ on the last test. The general results are shown in Table I.

Table I shows us that there has been a tendency to gain in IQ in each school studied. However, in Schools A and C the gain has been so small that it cannot be considered significant either statistically or practically. In School B, on the other hand, there has been a marked

gain, about comparable to that reported in the Iowa studies. The size of the gain is much too large to be attributed to chance, and the dif-

TABLE I.—DISTRIBUTION OF DIFFERENCES BETWEEN INITIAL AND RETEST IQ

Amount of Change (Difference)	*School A*	*School B*	*School C*	*Total*
48 to 52	..	2	..	2
43 to 47	..	2	1	3
38 to 42	2	4	1	7
33 to 37	1	9	1	11
28 to 32	3	7	5	15
23 to 27	3	22	20	45
18 to 22	14	31	15	60
13 to 17	26	46	33	105
8 to 12	42	56	56	154
3 to 7	48	62	69	179
− 2 to 2	49	55	72	176
− 7 to − 3	47	48	71	166
−12 to − 8	27	31	65	123
−17 to −13	20	18	42	80
−22 to −18	7	5	9	21
−27 to −23	3	2	3	8
−32 to −28	1	2	1	4
−37 to −33	1	2	3	6
−42 to −38	..	..	1	1
−47 to −43	..	..	1	1
Number	294	404	469	1,167
Mean difference	+ 1.40	+ 6.17	+ 0.65	+ 2.77
S. D.	11.65	13.75	12.36	12.89
S. D. of the mean	0.67	0.68	0.57	0.38

ference between this school and each of the others is also statistically significant. This marked discrepancy obviously calls for further analysis and study.

In order to improve our understanding of the gains in School B, we analyzed all the retest data from this school in terms of the interval between test and retest. If the gains in School B indicate a genuine improvement resulting from the school experience, we should expect them to be cumulative and to become progressively larger, the longer the interval between test and retest. The information on this point is given below in the accompanying tabulation.

Interval in Months	*Number*	*Mean Change*	*S.D.*	*S.D. Mean*
6 to 17	171	5.0	12.50	0.96
18 to 29	77	5.1	12.05	1.37
30 to 41	111	6.6	13.55	1.29
42 to 53	133	6.4	12.70	1.10
54 to 65	107	6.4	15.50	1.50
66 to 77	29	4.8	12.00	2.27
78 to 89	15	7.1	13.40	3.59
90 to 102	9	−1.4	12.30	4.35

From this tabulation we see that the amount of gain in School B is not significantly related to the length of time between tests. One year of exposure to the school environment results in as much gain upon retest as 6 years does, and there is no reliable difference between the gain at the end of one year and the gain after any longer period. Further evidence on this point comes from the study of 54 records in School B where there were 3 tests. The average gain from the first to the second test, with an average time lapse of about two years, was 7.5 points. The average gain from the second to the third test, with an average time lapse of about 3 years, was 0.3 points. Apparently, then, the gain in School B cannot be attributed to any cumulative effect of the school experience. If the school experience has produced the change, it must be in terms of some initial adjustment that has had its full effect within a year or so.

In an effort to determine why the gains appeared in School B, but not in Schools A and C, the data from each school were fractionated in various ways. Since the testing had been done over a period of more

than 20 years, the results were analyzed in shorter time samples. These results are presented in tabular form.

Time of First Test	School A		School B		School C	
	Average	Number	Average	Number	Average	Number
Before 1920	4.0	2	4.3	59	3.7	100
1920 to 1922	4.1	86	7.6	98	—0.9	73
1923 to 1925	1.0	68	2.8	32	—0.7	81
1926 to 1928	—3.7	47	3.8	53	4.2	113
1929 to 1931	2.6	51	6.1	90	—2.5	76
1932 and later	0.9	43	8.6	42	—3.0	25

All those cases in which the first tests were given within a particular time interval have been grouped together. There we see rather large, irregular changes in the average shift from test to retest. Some of these changes are too large to be attributed readily to chance. However, the irregularities in the ups and downs in any one school suggest that they were caused by changes in the testing personnel or fluctuations in the standard of testing rather than by any progressive real change in the effectiveness of the educational experiences being offered. In no one of the schools is there a clear trend for gains in IQ to become either progressively larger or progressively smaller.

A second analysis of the data was made in terms of age at the time of first test. The records were broken up into three groups—cases where the first test was given (*a*) before the age of 5, (*b*) between five and eight, and (*c*) after the age of eight. The results of this analysis are also given in tabular form.

Years Old at First Test	School A		School B		School C	
	Average	Number	Average	Number	Average	Number
Under 5	2.0	78	4.4	86	5.8	6
5 to 7-11	1.0	172	6.5	259	0.5	359
8 and over	0.8	57	6.2	17	2.0	112

It seems clear from this tabulation that there is no particular relation between the age at which the child was first tested and the

amount that he gained before retest. In School A, where the gains are small in any case, there is a tendency for them to be smaller if the child is first tested after he is 5 years old. In School B, this tendency is just reversed. In School C, very few tests were given below the age of 5, but comparing the 5- to 8-year age range with cases where the first test was given after age 8, we find somewhat larger gains for the older children. The age at which the first test was given is somewhat different for the populations in the three schools. However, the data of the preceding tabulation make it clear that we cannot explain the relatively large gains in School B on this basis. The largest gains in that school were for the large 5-year to 8-year group, which was also the largest group in each of the other schools and the group that showed insignificant gains in these schools.

A third possible explanation of the results in School B is that the examiners who gave the final tests were somewhat more lenient than those who had given the initial tests. Cattell[1] has shown that examiners may differ by as much as 10 or 12 points in the average severity of their grading when pairs of tests on the same children are compared. The examining personnel in the final tests differed somewhat from that in the initial tests in this school.

As a check upon this possibility, a sample of initial-test IQ's was obtained for each examiner (where these could be found), and his average was compared with the average initial-test IQ given by all examiners. Thus, if Examiner A gave higher IQ's on his initial tests than the average of all examiners, a rough correction could be applied to his IQ's.

Using the corrections thus secured and weighting the correction for each examiner by the number of examinations that he gave, an average correction was determined for the initial tests on the one hand and the final tests on the other. This procedure yielded no evidence that the retest examiners were generally more lenient than the initial-test examiners. The correction to be applied to the average of the initial tests was found to be substantially identical with the correction to be applied to the average of the final tests. It is to be admitted that the evidence here is inconclusive because first, the sample of initial tests available for some testers was very small, and second, there is no guarantee that the sample tested by any tester is a random one. As

[1] P. Cattell. "Stanford-Binet IQ variations." *School and Soc.*, 45: 1937, 615-618.

far as it goes, however, the evidence suggests that the gains from initial to final test were due to some factor other than leniency of testing.

In order to see whether the differences between the three schools might be due to differences in the intelligence level of the populations, the means and standard deviations of the IQ's on the first tests have been computed for the individuals upon whom we had retests after 2½ years or more. The mean IQ for Schools A, B, and C are 118.6, 117.5, and 118.8, respectively. The standard deviations of these IQ's are 13.1, 14.2, and 11.9, respectively. It is clear that the three schools are very similar in the superior level of intellect upon which they draw. In this respect they are also comparable to the University of Iowa Demonstration School.

An interesting side light is thrown upon the question of constancy of the IQ at different levels by comparing the variability in the initial tests with that in the final tests. The standard deviations of the distributions of final-test IQ's for the three schools were 14.2, 17.3, and 14.7, respectively. When these are compared with the standard deviations upon the initial tests, as reported in the last paragraph, an increase in variability is observed for each school. In other words, there is a tendency for high and low IQ's to draw apart, as has been indicated by Cattell.[1] Under these circumstances, it seems that some gain on retest should be expected in an above-average group. It may be that the small average gains found for Schools A and C are to be attributed to this effect.

IV. Discussion

We have studied the Binet retest results from three schools that, it seems safe to say, would be generally considered by educators to be superior schools. In no one of these did we find evidence of a cumulative increase in IQ related to the length of time spent in the school environment. In two of the schools (A and C) the gains were too small to be even statistically significant. In the third, appreciable gains were found, but they appeared as well at the end of one year as at the end of 5 or 6 years. Nothing in our data or our knowledge of the schools gives us any clear reason why the results in School B differed from those in the other schools. The difference between the schools cannot be explained in terms of differences in either the intelligence

[1] P. Cattell. "Do the Stanford-Binet IQ's of superior boys and girls tend to decrease or increase with age?" *Jour. Educ. Research,* 26: 1933, 668-673.

level of the populations or the age at which they were tested. Though there are ups and downs in each school, the difference between School B and the others tends to be maintained throughout a 20-year period.

We are at a loss to produce any convincing explanation of the difference. Several alternative hypotheses suggest themselves, however, and the merits of these will now be considered.

1. School B is a better school, and provides greater intellectual stimulation. We doubt whether this hypothetical explanation would seriously be upheld by one who knew the three schools. The resemblances between their clienteles, facilities, and approaches to education are much more marked than are any differences between them. But even more damaging to this explanation is the finding that the gains in School B are essentially as large at the end of one year as they are after any longer time lapse.

2. The testing at School B was more variable. If this contention were true, it would explain the greater variation from test to retest found in School B, but it is hard to see how it would explain a constant change running through the records. One might possibly argue that inexperienced testers tend to underestimate the intelligence of children who have not yet made an adjustment to the test or to the school environment. On the retest the child would be relatively immune to the inexperienced tester and consequently would show his abilities to better advantage. This explanation seems far-fetched to us, but may need to be considered if no better can be found.

3. In School B there was a marked general adjustment during the first year or few months, which also carried over to the Binet. There is no evidence to suggest why this should be the case in School B to any greater extent than in the other schools. A possibility, concerning which we have no satisfactory evidence, is that more children at School B were tested approximately at the time of admission, and thus before adjustment to the school environment had taken place. However, this requires the subsidiary assumption that children do generally test higher after a few months in school than at the time of admission —an assumption that seems open to question. What is more, the gains were also found for the older children at School B, who had become generally adjusted to the school environment.

4. In School B there was a selective factor determining which children should receive retests. If there was a tendency to give retests to

children whose school performance deviated from what would be expected from their first tests, and if these were primarily cases whose school work was better than expected, we would expect a gain on the second test. It is clear from our data that there is a large element of chance error in Binet IQ's for young children. Those who test at 100, for example, will cover a considerable range in 'true' ability. If there were a tendency for those who are overrated on the first test gradually to be eliminated from a rather exacting school environment, and a tendency for those who were underrated to be picked out sooner or later for retesting, an apparent gain upon retest would result, which might be independent of the interval between test and retest. We cannot demonstrate that any such influence did operate in School B and not in the other schools, but it may have been none the less.

If subtle and unidentifiable selective factors can operate to produce constant gains of the size and type that we found in School B. it seems at least possible that some analogous influence may have operated to produce the results reported by Wellman.[1] Such an explanation is rather unsatisfying, but does perhaps serve to unite otherwise discordant results.

V. Summary

1. A study was made of the Binet test-retest records for about 3,000 students in three well-known private schools in New York City. The analysis centered on the more than 1,100 cases for which the interval between test and retest was over 2½ years.

2. In two of the schools the average gain in IQ was negligible, while in the third it was appreciable, amounting to over 6 points.

3. The data suggested no satisfactory explanation for the difference reported above, and it remains something of a mystery to the authors, who have to content themselves at present with proffering a plausible, though unsupported, hypothesis to account for it.

[1] B. L. Wellman. "Mental growth from preschool to college." *Jour. Exper. Educ.*, 6: 1937-1938, 127-128.