The Journal of Genetic Psychology, 1949, **75**, 165-196.

CONSISTENCY AND VARIABILITY IN THE GROWTH OF INTELLIGENCE FROM BIRTH TO EIGHTEEN YEARS*

Institute of Child Welfare, University of California

NANCY BAYLEY

A. THE PROBLEM AND THE SUBJECTS

Various explanations have been offered for the changes which occur in the *IQ*'s of many children as they grow older. Among these explanations it has been suggested previously that irregularities may be due, at least in part, to innate differences in the tempos of children's maturational processes (4). However, the extent to which this hypothesis is true, if at all, is obscured by certain characteristics of the testing instruments on which we rely.

If we use several different tests of intelligence, the resulting variations in scores will be in part a function of the methods of standardization; including such things as the nature of the standardization sample, and the method by which the scores are obtained. They will also be in part a function of the kinds of intellectual abilities tested. That is, some scales test primarily verbal abilities; others weigh more heavily mathematical, or spatial functions, and so on. Another variable factor is the relative freedom of the test items from cultural and educational influences (11). There is also, of course, the further difficulty of determining the various effects of environment in stimulating or retarding intellectual development.

It is not proposed here to deal with the environmental aspects of the problem, but rather to examine some of the trends of intellectual development as found in some currently used tests of intelligence when applied to a small but constant sample, from birth through 18 years of age.

Ideally, for purposes of measuring the rates of intellectual growth in individual children, we should be able to measure the same children from birth to maturity on a single test which is applicable over the entire age range. Such a test, furthermore, should be calibrated in absolute units, so that velocities of growth in individuals and over different segments of the span may be compared directly. However, in spite of repeated efforts to produce them there are no existing intelligence tests which meet either of these

*Accepted for publication by Harold E. Jones of the Editorial Board, and received in the Editorial Office on December 29, 1948.

criteria. It now seems unlikely, from the very nature of the growth of intellectual abilities, that such a test can ever be devised. The mental behaviors which are developing during the first year of life are very different from those developing in the three-year-old who has learned to talk fluently, and these in turn are very different from the complex mental functions of later ages. From an examination of the nature of the intellectual functions available for testing, the growth of intelligence would appear to be the maturing of a succession of partially overlapping functions which become increasingly complex as they approach adulthood (4, 5).

We cannot, then, expect to have a single test of intelligence which is applicable at all ages. Such a test, for example, as the Stanford-Binet, which extends from two years to adult levels, though called one test, is made up of a series of *different* items which change in nature as they become more difficult. The extent to which these items and similar items in other tests are measuring the same things can be judged more adequately after large numbers of normal representative children have been tested and retested at successive ages, and their test scores compared.

We are beginning to accumulate such series of tests on the same children. Most of the groups of children on whom longitudinal test data are available are not average samples but tend to be superior. Nevertheless, much valuable information about the nature of intellectual growth has come and will continue to come from such studies because they are concerned with the growth of individuals through time. We may hope eventually to fill in the gaps with growth records from more average and below-average population samplings, as well as from more adequate tests.

The Berkeley Growth Study children, as reported previously (9), come, for the most part, from socio-economically superior homes. What is more, their intelligence scores tend to be well above the average. There were originally 61 infants enrolled: 40 of them have continued in the study through most or all of their 18 years. The principal contribution which the Berkeley Growth Study records can make to our knowledge about the nature of mental growth is in the length of the age span for which test scores are available. Although the number of children observed is not large, these same children have been tested repeatedly, at regular intervals throughout their lives. The further facts that the children were tested at most ages by the same examiner,[1] and that all had a similar program of testing

[1]All tests were given by the author, with a few exceptions. Occasional infant tests were given by Dr. L. V. Wolff, the pediatrician who participated in the program of infants' tests and measurements; most of the two-year tests were given by Dr. Marjorie Pyles Honzik; and the eleven-year tests were given by Dr. Mary Shirley.

experience given under the same general situational conditions, contribute to the comparability of the test scores. These conditions make it possible to study both the growth trends of individual children and the relations of age and test to scores for a constant sample.

The schedule of the study includes mental tests at most or all of 38 ages for the 40 children. The tests considered in this paper, with the ages at which they were administered, are as follows: The California First-Year Mental Scale (7), given at one-month intervals through 15 months; the California Preschool Scale (23), given at three-month intervals through three years, and at six-month intervals through five years; the Stanford-Binet, 1916 Revision, at six and seven years (35); the 1937 Revision (37), Form *L* at 8, 9, 11, and 14 years, Form *M* at 10, 12, and 17 years; the Terman-McNemar Group Test (36), Form *C* at 13 years, and Form *D* at 15 years; and the Wechsler-Bellevue (39), Form *I*, at 16 and 18 years. The scoring procedures for these various tests are different, and they are standardized on samples which were selected by different criteria, with resultant norms which are not equivalent in difficulty. Comparisons on this sample are made in respect both to the standard norms, and to methods adopted for the study of intra-group relationships.

Several aspects of these children's mental-test scores have been reported in previous studies, for the earlier ages up to and including nine years (4, 5, 6, 8). As shown in these studies, there was little or no relation between their mental test scores before two years of age and their scores at later ages. Similar results from other studies have convinced most investigators that existing tests of infant intelligence are inadequate for predicting children's later intelligence. Two alternative explanations of this inconsistency in early test scores have been suggested: (*a*) It may be that although we have not yet found the right tests, further search will reveal some infant behaviors which are characteristic of underlying intellectual functions, whose nature is such that they can be used for purposes of predicting the quality of intelligence at later ages. Or (*b*) early intellectual growth may be variable (either inherently so, or through environmental influences), making it impossible to predict later intelligence from any aspects of early infant behavior.[2]

B. The Selection of More Predictive Test Items

In the search for items of infant and preschool child behavior which may prove of predictive value, L. D. Anderson (3), Bradway (10), and Maurer (28) have made studies in which the scores made at a later age were used

[2]Except in cases of extreme retardation.

as criteria for selecting items or groups of items from tests given the same children at younger ages. Anderson compared 5-year *IQ*'s with test scores earned between three and 18 months. Bradway retested 10 years later children from the two- to five-year standardization sample of the 1937 Stanford-Binet. Maurer retested at 15 years children who had been given the Minnesota Preschool Scale at 18 to 54 months. The results of these studies are interesting but have not so far given us any adequately predictive batteries of tests. Both Anderson and Bradway found language or verbal items to be in general most predictive. Maurer found that the most predictive items required attention and adaptation, but that language entered in only after it had acquired the status of a well-developed tool. All three authors selected items of the type which they felt should be assembled for tests which might prove more useful than current tests in predicting intellectual growth.

As yet no complete item-by-item analysis has been made on the Berkeley Growth Study children. But various aspects of intelligent behavior, such as vocabulary and form-board performance, were compared over a period of years, as well as several different combinations of mental-test items (5). Recently a preliminary analysis of items has been made by comparing the six brightest with the six dullest 16-to 17-year-olds. A selection was made of those items in the First-Year Scale which were passed (on the average) at least two months younger by the bright group than by the dull group. Thirty-one items met this criterion. Cumulative point scores composed of these 31 items still did not reliably differentiate the bright from the dull ones during the first year. For the 12 ages (months 3-14) at which scores were computed, only six of the 12 children made scores which were consistently in the same general direction (i.e., above or below the average for the 12 cases) as their 17-year scores. It seems unlikely that correlation coefficients for the entire group would be significantly above zero.

In all of the comparisons so far made on the Berkeley Growth Study children, little consistency in relative scores could be found during the first two to four years. After this age, however, intellectual progress became fairly stable.

C. The Means of Mental Age and *IQ* Scores from One Month Through 18 Years

The data for the first three years have heretofore been reported in the form of point scores and sigma scores. For purposes of comparison with other data, mental ages have been computed for the First-Year Mental

Scale. To do this the mean cumulative point score at each age tested was called the mental age for the corresponding chronological age. Then *MA*'s (in months and tenths of a month) were interpolated and assigned to each point score. *IQ*'s were computed by the usual *MA/CA* ratio. *IQ*'s were computed for the California Preschool Scale and subsequent tests according to the published directions for each scale.

The relative status of the Berkeley group may be seen from the curve of their mean mental ages in Table 1 and in Figure 1*a*. These children constituted the standardization sample for the First-Year Scale,[3] and composed a part of the sample for the Preschool Scale: therefore the mean mental

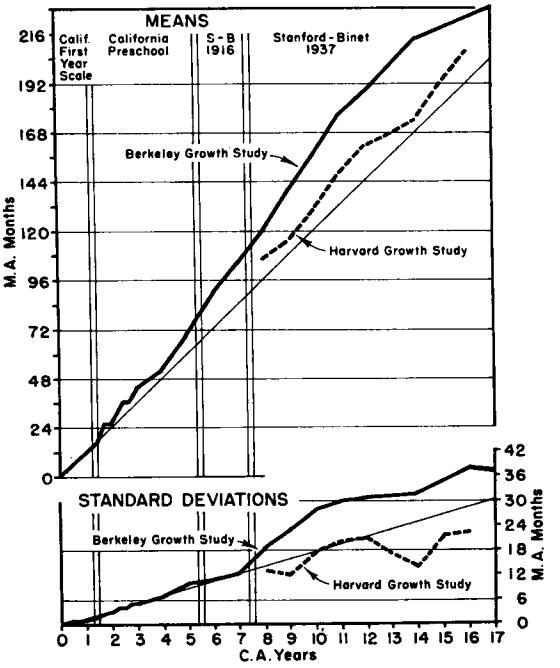


FIGURE 1

CURVES OF THE MEANS AND STANDARD DEVIATIONS OF MENTAL AGES FOR THE BERKELEY GROWTH STUDY CHILDREN FROM ONE MONTH THROUGH 17 YEARS, WITH COMPARABLE DATA FROM THE HARVARD GROWTH STUDY FOR YEARS EIGHT THROUGH 16

[3]No adjustment in these early mental ages was attempted. In view of the lack of correlation between earlier and later scores, we would not expect these children to show superior mental scores during the first year. The only other published data for the California First-Year Scale, those of Dubnoff, show the Russian infants she tested to be superior to our norms during the first nine months (13).

TABLE 1
MEANS AND *SD*'S OF MENTAL AGE AND *IQ*, BY AGE AND TEST
(Berkeley Growth Study)

Age		Test	*N*	Mental age in months* Mean	*SD*	*IQ* Mean	*SD*
Mo.	1	Cal. First-Year	52	1.04	.195	103.8	19.5
	2	Cal. First-Year	58	1.998	.34	101.8	16.9
	3	Cal. First-Year	61	2.92	.41	97.5	13.6
	4	Cal. First-Year	58	4.01	.51	101.0	12.9
	5	Cal. First-Year	58	5.00	.60	100.3	12.3
	6	Cal. First-Year	57	5.96	.79	99.1	13.2
	7	Cal. First-Year	52	7.03	.705	100.7	10.2
	8	Cal. First-Year	53	8.08	.77	100.9	9.7
	9	Cal. First-Year	56	9.01	.77	100.1	8.5
	10	Cal. First-Year	56	10.13	.75	101.3	7.6
	11	Cal. First-Year	52	11.03	.78	100.9	7.5
	12	Cal. First-Year	53	12.06	.82	100.7	6.7
	13	Cal. First-Year	53	13.04	1.07	100.3	8.4
	14	Cal. First-Year	46	14.08	1.12	100.7	8.1
	15	Cal. First-Year	52	15.00	1.38	100.0	9.3
	18	Cal. Preschool I	49	18.38	2.20	102.4	12.0
	21	Cal. Preschool I	52	22.59	2.47	107.6	11.7
	24	Cal. Preschool I	47	26.29	3.09	109.5	13.3
	27	Cal. Preschool I	48	30.48	3.69	112.6	13.6
	30	Cal. Preschool I	46	33.96	4.11	113.1	13.6
	33	Cal. Preschool II	44	37.04	4.87	111.6	15.0
	36	Cal. Preschool I	47	42.83	5.20	118.8	14.4
	42	Cal. Preschool I	39	49.39	5.50	117.6	13.2
	48	Cal. Preschool I	44	52.28	6.64	109.4	14.1
	54	Cal. Preschool I	43	62.28	8.03	115.0	15.2
	60	Cal. Preschool I	46	70.60	9.90	117.8	16.9
Yr.	6	Stanford-Binet '16	48	88.71	11.01	123.4	15.6
	7	Stanford-Binet '16	46	103.65	12.64	123.0	15.1
	8	Stanford-Binet L	47	120.00	18.91	122.6	20.1
	9	Stanford-Binet L	45	139.40	23.56	129.0	22.2
	10	Stanford-Binet M	47	157.96	28.75	131.9	23.6
	11	Stanford-Binet L	45	174.51	30.22	132.5	22.1
	12	Stanford-Binet M	43	186.93	31.71	130.3	22.1
	13	Terman-McNemar C	36	———	——	115.6	21.4
	14	Stanford-Binet L	37	213.08	31.85	129.9	19.2
	15	Terman-McNemar D	37	———	——	121.7	19.1
	16	Wechsler-Bellevue	39	———	——	117.4	16.2
	17	Stanford-Binet M	40	231.55	36.08	129.1	19.9
	18	Wechsler-Bellevue	37	———	——	122.1	16.1

*Data ungrouped

ages and *IQ*'s for the first five years cannot be used for estimating the representativeness of the sample. For school ages, we see that the group is superior to the Harvard Growth Study cases as reported by Dearborn and Rothney (12), and included in Figure 1 for comparison. It is far superior to the test norms, as represented by the straight diagonal line. Some of this superiority we may attribute to practice effect and test sophistication.

The means of the *IQ*'s are presented in Table 1 and Figure 2. It is

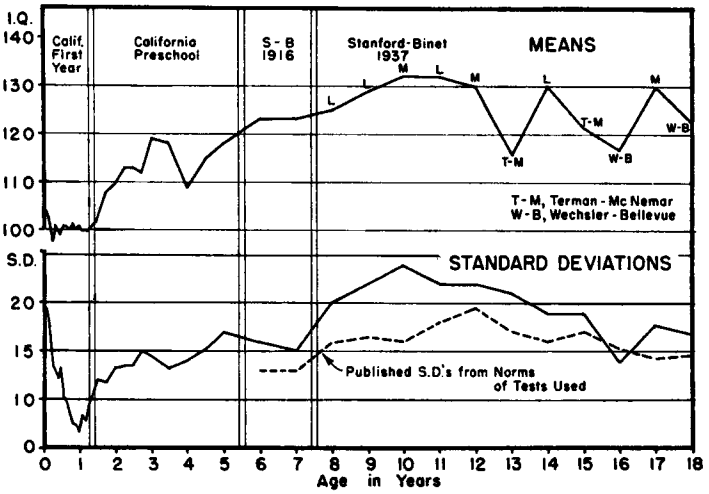


FIGURE 2
CURVES OF THE MEANS AND STANDARD DEVIATIONS OF *IQ*'S FOR THE BERKELEY GROWTH STUDY CHILDREN FROM ONE MONTH THROUGH 18 YEARS

obvious from their shifts, which range between 116 and 132 on the standard tests given after five years, that the norms used are not of equivalent difficulty at all ages. Stanford-Binet *IQ*'s average considerably higher than either the Terman-McNemar or the Wechsler.

Similar results are reported by other investigators. Sartain (33) for example, found that for 50 college freshmen, "*IQ*'s on the New Revised Stanford-Binet were significantly higher than those on the Bellevue Scale or the Otis Self-Administering Test of Mental Ability." He reported a Stanford-Binet *L* mean *IQ* of 129.48, *SD* 10.92, and a Wechsler-Bellevue Full Scale *IQ* mean of 117.48, *SD* 10.47.

The 1937 Revision yields higher scores for the Berkeley Growth Study than the 1916 Stanford-Binet. Ebert (14) has compared the 1916 and 1937 Stanford-Binets on a similarly selected superior group, and found con-

sistently higher means on the 1937 Revision. But Ebert also found a consistent tendency for the means of this last revision to increase with age from six to 10 years, as our means do from eight to 10. Therefore a part of the change in our means from the 1916 to the 1937 revision would seem to be a function of the *ages* at which the tests were given. Another factor which is probably operating here is the general superiority in intelligence of this group. The distribution of scores in this sample might very well be different for the two tests (1916 and 1937 Stanford-Binet). Although McNemar (29) found symmetrical distributions of *IQ*'s for the standardization sample, others (e.g., 32) have found that *IQ*'s above 100 on the 1937 Stanford-Binet are more variable than those below 100. If this is true it might account for both the higher means and the larger *SD*'s found for this test, as compared with the other tests, both for these children and for other above-average samples. (Our *SD*'s for the Terman-McNemar and the Wechsler-Bellevue are more nearly like those of the published norms.)

Scores on the second administration of both the Terman-McNemar and the Wechsler-Bellevue are higher than the first scores for each of these tests, even though the interval between the two administrations of a given test is two years. This might be due to specific practice effects.[4] Or it may indicate inadequate allowance in the standardization for intellectual growth during these late adolescent years. The *IQ*'s for both the Terman-McNemar and the Wechsler-Bellevue are not *MA/CA* ratios, but statistical equivalents, based on the means and *SD*'s of their standardization groups. When cross-sectional samples are used for standardization it is often difficult to secure groups of comparable abilities for successive years, especially at these ages when many children are dropping out of school. Although most test norms are based on the assumption that adult intelligence is reached by 16 or 17 years, a number of studies (18, 24, 25) indicate that intellectual growth continues, on the average through 18 years, and even at least for some persons, to around 21 or 22 years.

D. Variability of Scores

1. *Mental Ages*

More significant than the means, it seems to me, is the trend of the standard deviations of mental ages from birth through 17 years (Table 1 and Figure 1*b*). It is plain that the *SD*'s do not increase at the constant rate

[4]All of these children are so accustomed to taking tests that we can attribute very little effect, at these ages, to any general learning experience in test-taking.

which is necessary if *IQ*'s are to remain constant during growth. The *SD*'s are too small during most of the first year and too large after seven years, and especially at 9, 10, and 11 years. These variations cannot be attributed to inequalities in the sampling of cases, as they are based on essentially the same cases throughout. But the Berkeley children are not alone in showing these age trends in variability. Although the Harvard Growth Study *SD*'s are smaller for the same ages (see Figure 1*b*), they agree in indicating greater variability in scores from 9 to 11 years, in a sample which is also primarily "longitudinal" (12, p. 170).

2. *IQ's*

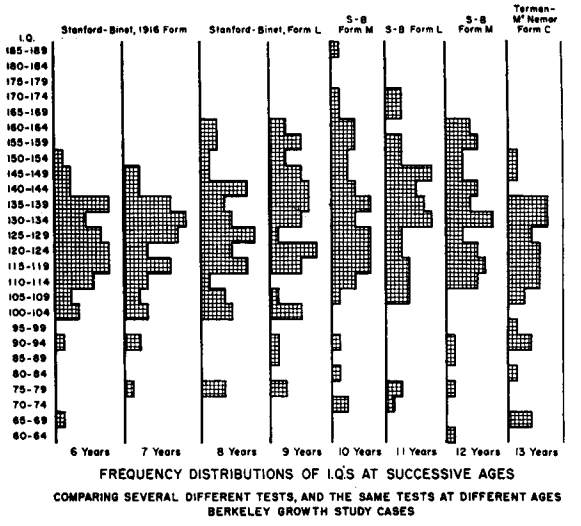
The *SD*'s of the *IQ*'s are given in Table 1 and shown graphically in Figure 2*b*. These standard deviations show strikingly why the *IQ* is a poor instrument to use in predicting later intelligence. When *IQ*'s are used these children's scores are most variable at one month (when the *SD* is 20) and around 9 to 11 years (when it goes as high as 24); and least variable around one year (when it drops below seven *IQ* points). The variability tends to diminish again as maturity is approached.

The distributions of *IQ*'s from six to 18 years are shown in Figure 3. Although statistical tests indicate that these distributions are within the limits of normal for samples of this size,[5] it is apparent that the high *IQ*'s are limited at the later ages. The usual interpretation of such a curtailment of high scores is that the tests used do not have enough "top" for the brighter children. Another possible explanation is offered later in this paper.

E. Variability of Scores in a Strictly Constant Case Sample

Although the data presented thus far are on the same children for the most part, a glance at the *N*'s in Table 1 shows that all 61 children were present at only one test age (three months). There is, thus, some fluctuation from age to age in the composition of the sample. It has been possible to select 21 ages, fairly well distributed over the 18-year span, at which the same 27 children were tested. The data on *IQ*'s for this sub-sample, for all of whom there are scores at all 21 ages, are given in Table 2 and Figure 4. We have here sacrificed cases and testing ages to gain constancy of sample. The same age trends in means and *SD*'s are found. This rules out the possibility that variations may be due to inconstant sampling of cases.

[5] Beta coefficients (30).



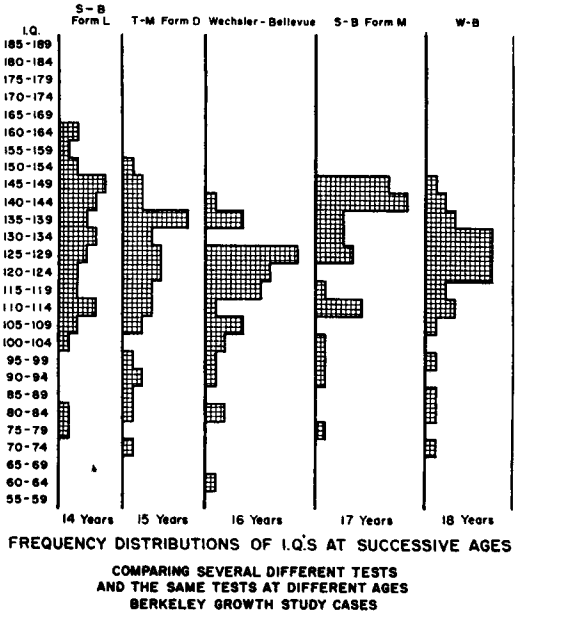


FIGURE 3

FREQUENCY DISTRIBUTIONS OF *IQ*'S AT SUCCESSIVE AGES (*a*) YEARS SIX THROUGH 13, (*b*) YEARS 14 THROUGH 18

TABLE 2
MEANS AND *SD*'s* OF MENTAL AGES AND *IQ*'s OF 27 SELECTED CASES

CA	Mental age in months		*IQ*	
	Mean	*SD*	Mean	*SD*
Mo. 3	2.97	.46	99.07	15.35
4	4.03	.45	100.59	11.02
5	5.07	.62	101.78	12.15
6	6.10	.74	101.56	12.28
8	8.28	.67	103.48	8.40
13	13.11	1.04	100.93	7.99
15	15.08	1.32	100.56	8.64
18	18.54	2.30	102.74	12.73
21	22.56	2.05	107.19	9.73
24	26.13	2.27	108.48	9.10
27	29.59	3.15	109.48	11.88
30	34.35	3.34	114.41	10.92
36	41.84	4.52	116.19	12.68
42	48.39	5.18	115.04	12.16
48	51.07	5.35	106.74	11.28
Yr. 7	105.26	10.08	124.96	11.85
9	143.63	21.09	132.81	19.52
11	180.96	27.52	137.15	20.75
14	217.33	28.92	132.59	17.70
15			122.70	18.37
16			120.52	12.12
17	236.67	25.94	131.52	14.43
18			124.44	12.28

*Data ungrouped.

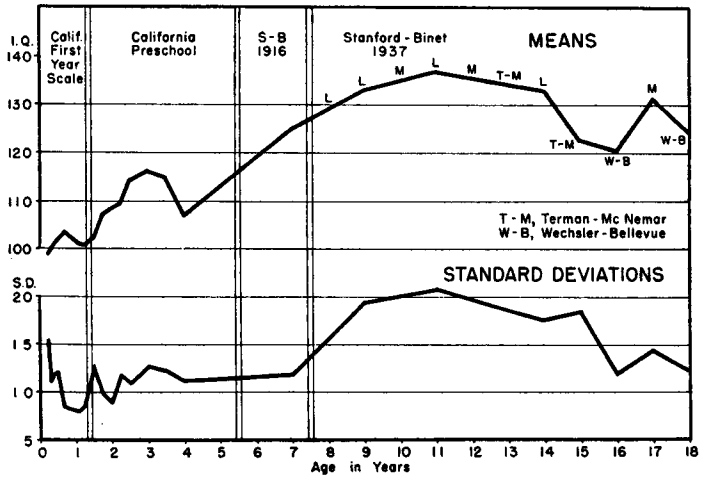


FIGURE 4
CURVES OF THE MEANS AND STANDARD DEVIATIONS FOR A STRICTLY CONSTANT SAMPLE OF 27 BERKELEY GROWTH STUDY CASES AT 21 TEST AGES

F. Age Changes in Variability for Different Tests

1. *Infant Tests*

The question arises whether changes in variability are due to the particular tests used. In Figures 5 and 6 some data are assembled on *SD*'s which have been published on tests of infants. The curves in 5*a* are *SD*'s of point scores for two groups of infants—the Berkeley cases (4) and Russian babies tested by Dubnoff (13)—who were given the California First-Year Scale. In 5*b* are *SD*'s of point scores reported by Fillmore for her Iowa Infant Scale (17) and by Nelson and Richards for the Gesell Schedules given to children in the Fels Foundation growth study (31). In Figure 6 are *SD*'s of *IQ*'s, for the Berkeley Growth Study, and *PE*'s of Kuhlman-Binet *IQ*'s as reported by Kuhlman[6] (26). For all tests and samples, and for different methods of scoring, there is decreased variability in scores at or

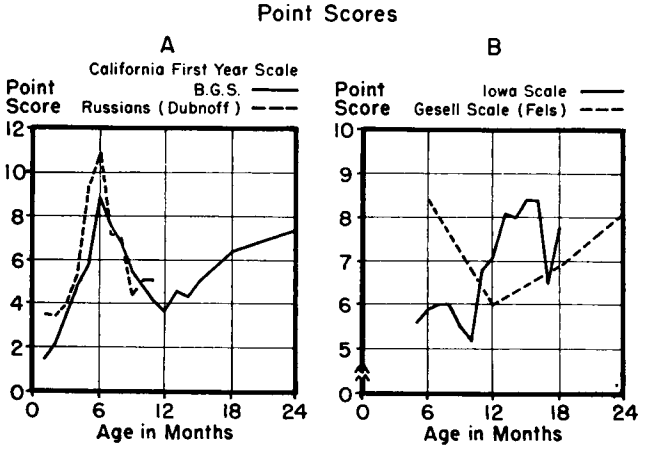


FIGURE 5

Age Curves of the Standard Deviations of Point Scores Reported for Several Different Infant Tests

near one year of age, with the *SD*'s increasing as we go either up or down the age scale from there. The consistency of these trends suggests that children are less variable in their behavior-maturity patterns at one year than earlier or later. An additional piece of evidence which may support such an hypothesis is given by L. D. Anderson (3). In his validation of infant test items by correlation with five-year *IQ* he found only five items (from a total of 97) at the one-year level which were "predictive." There were, by contrast, 16 items at six months and 18 items at 18 months.

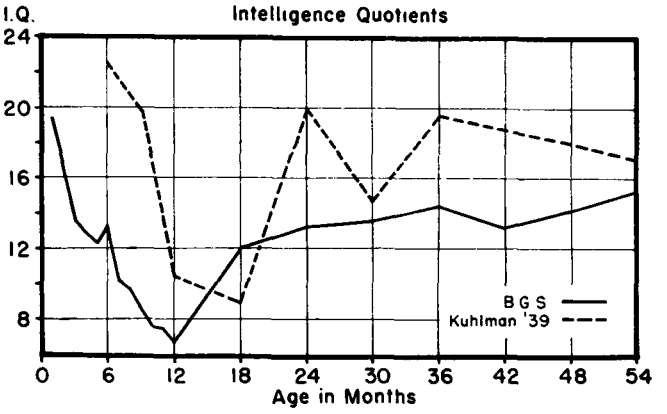


FIGURE 6

AGE CURVES OF THE STANDARD DEVIATIONS OF *IQ*'S: THE BERKELEY GROWTH STUDY COMPARED WITH THE KUHLMAN-BINET

2. *Tests From Two to 18 Years*

There is, furthermore, evidence from other studies indicating changes in variability at other ages. Goodenough (19) has called attention to the trends in the 1937 Stanford-Binet norms. The *SD*'s of *IQ*'s in the standardization sample, as reported by Terman and Merrill (37), show trends which Goodenough suggests are inherent in the tests, and not due to chance variations in sampling as Terman and Merrill had assumed. These *SD*'s tend to decrease from two and one-half years to six years, when they are smallest, then to increase to a high level from 11 to 15 years, after which they drop again. McNemar (29) agrees that the changes in variability are probably not due to chance, and has set up a table for correcting *IQ*'s at the ages where the *SD*'s are smallest and largest.

We have made one check on the relation of case sampling to variability in the 1937 Stanford-Binet, for a part of its range, by comparing the 34 Berkeley Growth Study children who took the test at all seven ages from eight through 17 years. Form *L* was given at four ages, and *M* at three ages. The means and *SD*'s are given in Table 3. Whether we regard these as the same test, or two different tests, the trends are evident. The age changes in variability do appear to be characteristic of the test.

This characteristic trend, however, is not confined to the 1937 Stanford-Binet tests. Such other published material as the Harvard Growth Study

[6]The Kuhlman-Binet *PE*'s, as he uses them, are interquartile ranges (26). They are from his Table 28 and Figure 1.

TABLE 3
MEANS AND *SD*'S STANFORD-BINET (1937 REVISION) MENTAL AGE AND *IQ* FOR 34 BERKELEY GROWTH STUDY CASES

CA	Mental age, months		*IQ*	
Years	Mean	*SD*	Mean	*SD*
8.0	119.76	18.67	124.33	19.75
9.0	139.32	22.63	129.18	21.83
10.0	159.18	27.71	132.12	23.47
11.0	174.03	30.27	131.97	22.49
12.0	189.47	29.79	131.53	20.88
14.0	209.62	32.54	127.85	19.61
17.0	230.65	32.41	128.00	18.23

(See Figure 1) which adapts scores from several tests (12, 34), the studies of Freeman, *et al.*, of mental growth in Chicago children (1, 18), Ebert and Simmons' report on the Brush Foundation children of Cleveland (15), and data reported by Goodenough on Minnesota children (19, 20), all give greater *SD*'s for mental test scores around 10 to 12 years of age than in the periods just before or after. The *PE*'s (and hence the *SD*'s) of Kuhlman-Binet *IQ*'s tend to drop from two to six years, and to rise after six but become large and erratic after 13 years (26).

These studies include a variety of testing instruments, and both cross-sectional and longitudinal samples. The trends in variability are, of course, to some extent peculiar to the particular tests used. But there is enough concomitance in these trends to merit an investigation of the possibility that the tests may be reflecting underlying growth processes.

G. VARIABILITY: THEORETICAL CONSIDERATIONS

Although the age changes in variability[7] may be artifacts of current methods of selection and standardization of test items, they may equally well describe tendencies which are inherent in intellectual growth. It seems quite probable that both of the clear-cut periods of restricted variability in the Berkeley Growth study intelligence scores—toward the end of infancy and of adolescence—are due to the approach to maturity of the particular processes being measured. The mental processes which are developing during the first year are largely sensory-motor in character (2, 4). And although they form the basis for further intellectual development, precocity or retardation in them is not necessarily related to rates of development in the more complex processes which we call intelligence in school-age children

[7]The coefficient of variation, as used by Ellis (16) and Henmon and Livingstone (21) for example, seems inapplicable here. *V* only seems to minimize or obscure changes in variability which are of practical significance.

and adults. By one year of age most of the slow developers have caught up with those who were precocious in these simple coördinations. The *SD*'s thus become restricted to individual differences in mature functions.[8] In the same way the approach to mature intellectual status after 11 or 12 years could reduce the variability of performance as the children whose mental growth is more accelerated reach their own "ceilings."

On this interpretation the ceiling is a function (at least in part) of the child's changing growth rate, rather than of a scarcity of difficult items at the upper levels of the test. This is shown very clearly in the study of Freeman and Flory (18, pp. 38-41), who were concerned over the reduced *SD*'s on their *VACO* tests after 15 years. They attempted to increase the variability of scores on the upper levels of the Analogies test by adding top in the form of more difficult items. However, they did not succeed in changing the trend. An analysis of their Opposites test likewise indicated that its reduced variability at later ages was not due to a lack of differentiating items at the upper end of the scale.

It thus seems likely that the test scores are reflecting actual changes in variability which are inherent in the processes of development of any given function. During growth of a structure or function variability increases, in part because of increasing individual differences in capacity, and in part because of individual differences in the speed with which the maturing process takes place. These two factors are known to be operative in physical growth, and it seems reasonable to expect that they may be characteristic of many growth processes. During the stage of development when both factors operate freely, the variability of measures or scores will become greater with the general increments in the structure or function concerned. But as an increasing number of individuals stop growing, and the means level off to a constant value, the individual differences which remain become restricted to those of the achieved mature state. On this hypothesis, we should assume that in the present series of tests of mental growth we have scores on at least two types of function which develop successively, resulting in alternating periods of increasing and decreasing variability. These large general trends may well obscure similar tendencies, which are occurring more or less simultaneously, in more specific functions which develop in various parts of the growth span. The *VACO* tests are examples of this, as is seen in the varying trends of means and *SD*'s of the four tests in the Freeman and Flory Study (18). Thurstone (39) in testing five- and six-year-olds, found that

[8]This point has been discussed in detail in my monograph on Mental Growth during the First Three Years (4).

certain factors seemed to mature much earlier than others. Another example is found in the study of Jones and Conrad (24) for the subtests of the Army Alpha between the ages of 10 and 60 years. This last study indicates wide variations in rates of decline of different intellectual functions, as well as in their rates of growth. It is reasonable to expect that similar differences will be found in any broad sampling of mental functions.

H. Consistency of Growth in Intelligence

1. *Method of Scoring*

We have thus far discussed three different conditions which militate against a child's maintaining a "constant *IQ*" throughout his growth. First, differences in standardization from one test to another, with differences in relative difficulty, cause spurious changes in the *IQ*'s. This is shown in the considerable differences in mean *IQ*'s of the Berkeley Growth Study children for the different tests used. Second, we have found age changes in variability of the tested mental functions, so that if relative intellectual status is expressed either by scaled point scores[9] or by the ratio *MA/CA*, the scores of exceptional children are necessarily brought closer to the average during periods when variability is reduced. Third, it would appear that different functions are being measured on different segments of the mental growth span.

To eliminate, as far as possible, changes in the scores for our sample which may be due to either of the first two factors, we have transposed all of their mental test scores into Sigma Scores computed from the means and *SD*'s of the points earned by this group of children at each age tested.[10] Using these Sigma Scores, or Standard Scores, we can determine both for the group as a whole, and for the individual child, the extent to which the children maintain constant positions in a total group which has had similar testing experience.

2. *Relation to Age and Test-Retest Interval*

We have computed several series of correlation coefficients between tests given at successive ages, to determine the extent to which predictions can be made for the children in the group, for different ages and for different intervals between tests. Samples of these *r*'s are shown graphically in

[9]Freeman and Flory (18).

[10]Sigma Scores have for some purposes been transposed into their equivalent Standard Scores by multiplying by 10 and adding 50, thus eliminating all minus figures.

TABLE 4

CORRELATION COEFFICIENTS BETWEEN AGE-LEVEL STANDARD SCORES OF INTELLIGENCE*

Av.of months	4, 5, & 6	7, 8, & 9	10, 11, & 12	13, 14, & 15	18, 21, & 24	27, 30, & 36	42, 48, & 54	Years 5, 6, & 7	8, 9, & 10	11, 12, & 13	14, 15, & 16	17, 18
1, 2, & 3	.57	.42	.28	.10	—.04	—.09	—.21	—.13	—.03	.02	—.01	.05
4, 5, & 6		.72	.52	.50	.23	.10	—.16	—.07	—.06	—.08	—.04	—.01
7, 8, & 9			.81	.67	.39	.22	.02	.02	.07	.16	.006	.20
10, 11, & 12				.81	.60	.45	.27	.20	.19	.30	.23	.41
13, 14, & 15					.70	.54	.35	.30	.19	.19	.09	.23
18, 21, & 24						.80	.49	.50	.37	.43	.45	.55
27, 30, & 36							.72	.70	.58	.53	.46	.54
42, 48, & 54								.82	.71	.64	.70	.62
Years												
5, 6, & 7									.92	.85	.87	.86
8, 9, & 10										.94	.92	.89
11, 12, & 13											.96	.96
14, 15, & 16												.96

*These scores are the means of standard scores for three consecutive test-ages, e.g., months 1, 2, & 3; 4, 5, & 6, etc., and years 5, 6, & 7, etc. The last level is composed of only two test ages, 17 & 18 years. Each child's score is the average of all tests taken by him for the ages included in that level.

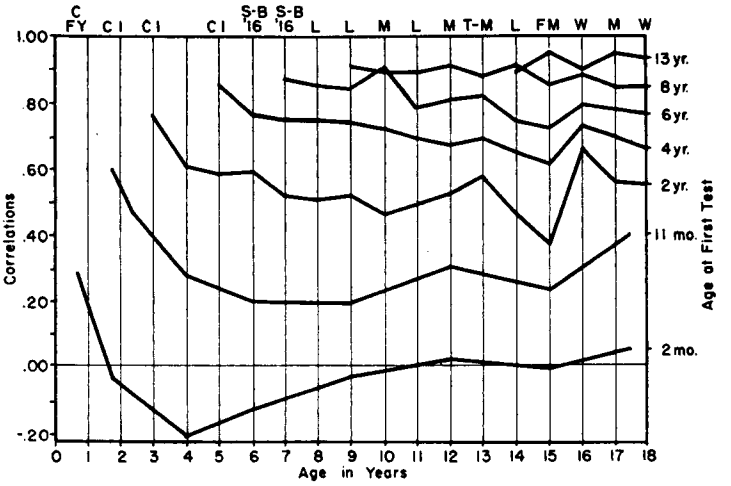


FIGURE 7

AGE CURVES OF CORRELATION COEFFICIENTS BETWEEN SCORES ON SELECTED INITIAL TESTS AND SUBSEQUENT TESTS GIVEN AT YEARLY INTERVALS

Figure 7. Table 4 gives the *r*'s for consistency of mental test scores for successive age levels in which each child's Sigma Scores for three successive test ages have been averaged. This particular set of *r*'s, by the use of averages for three tests, eliminates most of the chance variations which occur in single test scores. Furthermore, the use of Sigma Scores eliminates the age changes in variability which would tend to alter the magnitude of the *r*'s. For comparison, Table 5 gives the *r*'s between single test *IQ*'s for ages six through 18 years. Table 6 and Figure 8 give consistency correlations (single test point scores) for the 27 cases who make up a constant sample for a wide range of ages.

From these correlation coefficients we may see the extent to which the children's relative mental status remains constant. It has now become fairly well accepted that the size of a test-retest correlation for *young* children is a combined function of the age of the children and the length of the interval between tests.

The correlation coefficients, as we have pointed out in earlier publications (4, 5), indicate that these children's scores on the tests given before two years of age are quite unrelated to their test scores during school ages. They indicate, further, however, increasing stability of scores with increas-

TABLE 5

CORRELATIONS BETWEEN MENTAL TEST *IQ*'S FOR TESTS GIVEN AT DIFFERENT AGES (6 THROUGH 18 YEARS)

Age at test	1916	Stanford-Binet Form *L*	Stanford-Binet Form *L*	Stanford-Binet Form *M*	Stanford-Binet Form *L*	Stanford-Binet Form *M*	Terman-McNemar Form *C*	Stanford-Binet Form *L*	Terman-McNemar Form *D*	Wechsler-Bellevue	Stanford-Binet Form *M*	Wechsler-Bellevue
	7	8	9	10	11	12	13	14	15	16	17	18
6	.86	.85	.84	.90	.78	.81	.82	.74	.72	.79	.78	.77
7		.88	.83	.87	.82	.83	.88	.79	.75	.83	.83	.80
8			.91	.89	.89	.91	.88	.91	.85	.88	.84	.85
9				.88	.90	.82	.87	.86	.82	.87	.85	.87
10					.92	.90	.88	.92	.83	.88	.86	.86
11						.93	.91	.93	.89	.89	.92	.93
12							.87	.94	.85	.88	.90	.89
13								.89	.95	.90	.94	.93
14									.87	.92	.89	.89
15										.88	.89	.88
16											.89	.94
17												.90

TABLE 6
CONSISTENCY CORRELATIONS BETWEEN MENTAL TEST POINT SCORES AT INDICATED AGES FOR 27 SELECTED CASES

Age at test	Months						Years					
	6	13	18	24	36	48	7	9	11	15	17	18
Mo.												
3	.35	.02	—.05	—.13	.05	—.03	—.15	.08	.08	—.04	.12	—.03
6		.63	.35	.08	.13	.09	—.12	.04	—.07	—.26	—.04	—.24
13			.60	.47	.41	.23	.13	.13	.02	—.18	.002	—.14
18				.50	.54	.41	.33	.14	.11	—.02	.20	.03
24					.74	.47	.60	.43	.43	.27	.41	.39
36						.64	.53	.55	.48	.33	.56	.40
48							.71	.76	.69	.54	.71	.52
Yr.												
7								.79	.74	.71	.79	.68
9									.90	.77	.84	.80
11										.89	.92	.87
15											.88	.84
17												.79

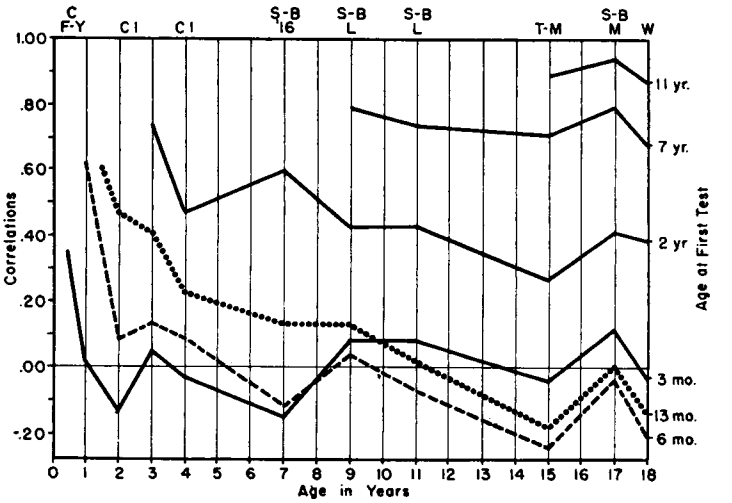


FIGURE 8
AGE CURVES OF CORRELATION COEFFICIENTS FOR A CONSTANT SAMPLE OF 27 CASES

ing age.[11] By two years the r's with tests at later ages hold up fairly well, rarely dropping below .50. The school-age correlations drop off only slightly as the interval between tests is increased for higher age levels. Studies on other children such as those of Honzik (22), Goodenough and Maurer (20), Ebert and Simmons (15), and Anderson (2) show very similar correlational trends.

3. *Correlations between Scores on Different Tests*

It has been suggested (2) that the consistency of the test scores will be affected by the use of different tests at different ages. In very few studies has the same test been given to the same children at all ages. One reason for this is that no test has been standardized for the entire age span. Furthermore, even if something had been named the same test, it would necessarily be comprised of very different items at the different age levels. Especially do the infant and preschool tests differ from the later ones. Perhaps the closest approach to this desirable condition of similar functions in a single testing instrument given to the same children over a wide age range, is to be found in the study of Freeman and Flory, in which the *VACO* tests were used from six through 18 years (a period of relative stability). This study shows individual variations in growth which are similar to our data, even though in the Berkeley Growth Study we do not have this constancy of testing instrument. Three forms of the Stanford-Binet, the Terman-McNemar Group test, and the Wechsler-Bellevue were given at various ages during this same age span.

The effect of changing tests on the Berkeley Growth Study group's relative status may be seen from Table 7. In this table the r's are grouped according to the tests involved. For 12 comparisons between repeats of the same test, the mean of the r's is .89.[12] For 26 comparisons between different forms of the Stanford-Binet the mean of the r's is .87. For 40 comparisons between unrelated tests the mean of the r's is also .87. The lowest r in this last group is .72 between the 1916 Stanford-Binet at six years and the Terman-McNemar at 15 years. It is likely that the age at first testing and

[11] Honzik's (22) findings that "the magnitude of a correlation between tests varies directly with the age ratio $\frac{CA \text{ at first test}}{CA \text{ at second test}}$ holds up fairly well to about five years. After this age, however, there is much greater constancy than the ratio would predict. See Figures 7 and 8.

[12] Computed by the formula $\frac{(N-3) \times Z's \text{ for } r's}{\Sigma(N-3)}$, see Lindquist (27, pp. 218-219).

TABLE 7

INTERCORRELATIONS BETWEEN TEST SCORES ACCORDING TO THE TESTS COMPARED (SIX THROUGH EIGHTEEN YEARS)

Intercorrelations When the Same Test is Repeated
(Mean of 12 *r*'s = .89)

S-B, 1916			*S-B*, *L* x *L*			*S-B*, *M* x *M*			*TMG*, *C* x *D*		
CA	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*
6x 7	44	.87	8x 9	45	.91	10x12	41	.90	13x15	33	.95
			8x11	44	.89	10x17	39	.86			
			8x14	36	.91	12x17	37	.90			
			9x11	43	.90				*W-B* x *W-B*		
			9x14	35	.86				*CA*	*N*	*r*
			11x14	37	.93				16x18	36	.94

Intercorrelations between Different Forms of the Stanford-Binet
(Mean of 26 *r*'s = .87)

1916 x *L*			1916 x *M*			*L* x *M*			*L* x *M* con't		
CA	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*
6x 8	45	.84	6x10	46	.90	8x10	45	.89	14x10	36	.92
6x 9	44	.84	6x12	42	.81	8x12	42	.91	14x12	35	.94
6x11	44	.78	6x17	39	.78	8x17	39	.84	14x17	36	.89
6x14	36	.74	7x10	45	.87	9x10	44	.88	11x10	43	.92
7x 8	44	.85	7x12	41	.83	9x12	40	.92	11x12	41	.93
7x 9	44	.81	7x17	40	.83	9x17	38	.85	11x17	39	.92
7x11	44	.82									
7x14	37	.79									

Intercorrelations between Unrelated Tests
(Mean of 40 *r*'s = .87)

S-B, L x *TMG*			*S-B*, L x *W-B*			*S-B*, *M* x *TMG*			*S-B*, *M* x *W-B*		
CA	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*
8x13	37	.88	8x16	40	.88	10x13	36	.88	10x16	40	.88
8x15	38	.85	9x16	39	.87	10x15	38	.83	12x16	36	.88
9x13	36	.87	11x16	40	.89	12x13	34	.87	17x16	40	.89
9x15	37	.82	14x16	36	.92	12x15	36	.85	10x18	36	.86
11x13	37	.91	8x18	36	.85	13x17	33	.94	12x18	34	.89
11x15	38	.89	9x18	34	.87	15x17	37	.89	17x18	36	.90
14x13	33	.89	11x18	35	.93						
14x15	36	.87	14x18	33	.89						

S-B 1916 x *TMG*			*TMG* x *W-B*			*S-B* 1916 x *W-B*		
CA	*N*	*r*	*CA*	*N*	*r*	*CA*	*N*	*r*
6x13	36	.82	13x16	32	.90	6x16	40	.79
7x13	36	.88	15x16	35	.88	7x16	41	.83
6x15	38	.72	13x18	31	.93	6x18	36	.77
7x15	39	.75	15x18	33	.88	7x18	36	.81

*See footnote 12.

the length of the interval between tests is at least as significant in causing this low *r* as the fact that they are two different tests. It would appear that for this group of children the consistency of their intellectual status relative to each other is very little influenced by the use of these different tests. This is true, even though the *IQ*'s as computed according to the several test norms, are often quite variable.

I. The Growth of Intelligence in Individuals

Individual age-curves of intelligence scores, as represented by Sigma Scores (or Stanford Scores) are very informative. In a previous study (5) the Sigma Score curves were presented for all 48 children who had completed the first nine years of the study. From inspection of the curves it was concluded that only a fifth of the group had maintained approximately the same relative status throughout the nine years. The others showed varying types of shifts in status, often consistent in their trends over long periods. While some grew more slowly and others more rapidly than the average, still others had successive periods of rapid and slow growth.

Examples of individual trends for the entire 18 years are shown in Figures 9 to 12, which present the mental scores of four different children in the study. For the purpose of comparing the *IQ* with the Sigma Score, which represents more accurately the interrelations of the children in this study, each child's scores are plotted in two ways. The broken line gives the *IQ*'s derived from the published norms for the tests used. (These charts are drawn to the scale of one *SD* to 15 *IQ* points, which approximates the average, for all the ages, of the *SD*'s of *IQ*'s in this group.) The solid line represents the Sigma Scores, which show the children's status in the Berkeley Growth Study group[13]

Inspection of the curves gives the impression of great instability of scores during the first year or two, regardless of the method of scoring. Usually the *IQ*'s are more variable, but sometimes, especially near one year of age, the Sigma Scores are more deviant. During the ages when the variability of the *IQ* is greatly restricted it is much more difficult to earn deviant *IQ*'s, even though relative to the group a child's score might be outstanding. Case 14 *F* (Figure 9) is an example: at 12 months she was the most precocious child in the study, earning a score three *SD*'s above the group mean (i.e., a Sigma Score of 3.00). Her *IQ*, however, was only 124, which would ordinarily be interpreted as about 1½ *SD* above average. When she was three years old, on the other hand, her Sigma Score had dropped to .80 while her *IQ* had risen to 132. Another case, 5 *M* (Figure 10), shows much greater variability in his Sigma Scores before five years, and in his *IQ*'s after this age. Although both of his curves indicate rapid growth and an upward trend in scores between 18 months and two years, the early retardation was much more marked in the Sigma Scores, and the later acceleration was by far greater in the *IQ*'s.

[13]The *IQ*'s are all higher than the Sigma Scores after the first few years. This is to be expected as the former are computed from the test norms, while the latter are computed from the means of *MA*'s or point scores for this superior group.

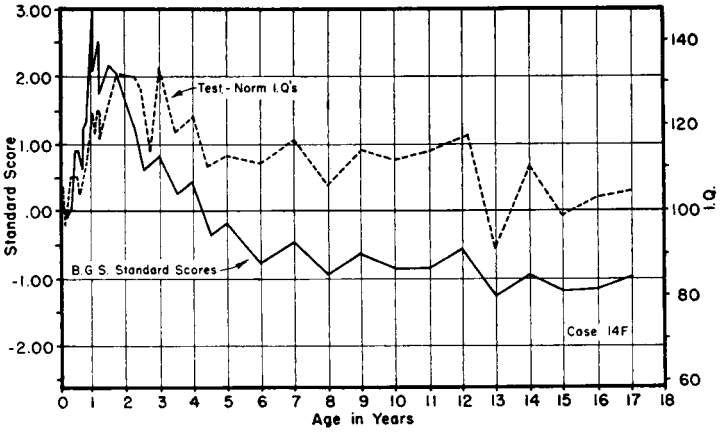


FIGURE 9

CURVES OF THE INTELLIGENCE SCORES OF CASE 14 *F*: THE SOLID LINE REPRESENTS HER RELATIVE POSITION (STANDARD SCORE) IN THE BERKELEY GROWTH STUDY; THE BROKEN LINE GIVES *IQ*'S COMPUTED ACCORDING TO THE DIRECTIONS FOR THE TESTS USED

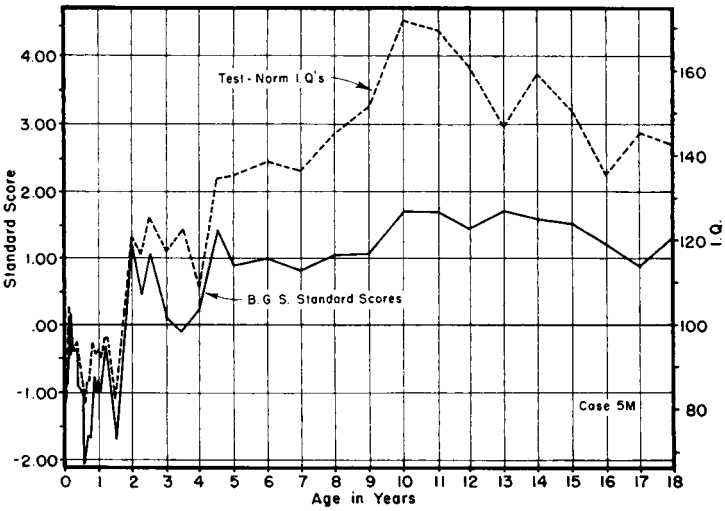


FIGURE 10

STANDARD SCORE AND *IQ* CURVES FOR CASE 5 *M*

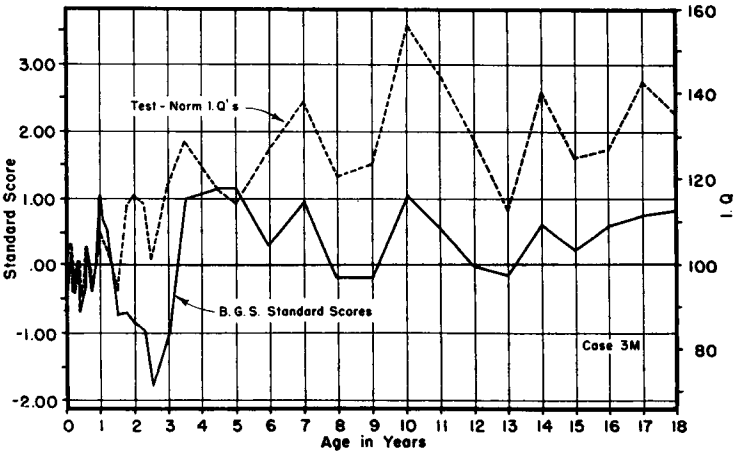


FIGURE 11
STANDARD SCORE AND *IQ* CURVES FOR CASE 3 *M*

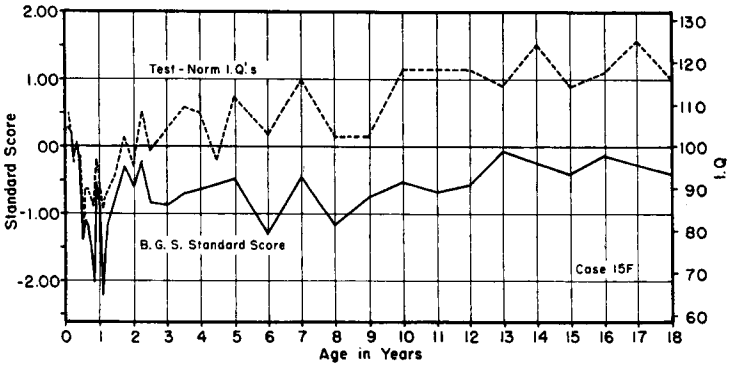


FIGURE 12
STANDARD SCORE AND *IQ* CURVES FOR CASE 15 *F*

Further examination and comparisons of the individual Sigma Scores reveals the normal variations in individual mental growth. We have quantified the individual differences in "constancy" by assigning "Intelligence Lability Scores" to all of the Berkeley Growth Study children. This was done by computing, for each child, the mean and *SD* of his Standard Scores earned over given age-intervals. A child's standard deviation from his own mean is his *Lability Score.* A high score, or large *SD,* signifies greater lability or variation from the child's own central tendency. Data on these scores are

TABLE 8
MEANS AND *SD*'S OF INTELLIGENCE TEST LABILITY SCORES FOR 40 CASES

	Infancy Months 1-21 I (17 test ages)	Preschool Years 2-5 II (8 ages)	School-age Years 6-18 III (13 ages)	Total Span 1 Mo. to 18 Yrs. IV (38 ages)
Means				
Boys	6.78	4.32	3.39	5.46
Girls	6.78	4.19	3.31	5.41
Total	6.78	4.25	3.35	5.44
SD's				
Boys	2.00	1.61	1.14	1.36
Girls	1.73	1.55	.99	1.05
Total	1.87	1.58	1.08	1.22

given in Table 8. For the 17-test period of one to 21 months, the Infant Lability Scores averaged 6.8, *SD* 1.9; for the eight-test Preschool period of two to five years, the mean Lability Score is 4.3, *SD* 1.6; and for the 13-test School-age period of six to 18 years, the mean is 3.4, *SD* 1.1. This is another way of showing that the children maintain their own relative status more closely as they grow older. Both the Lability Scores and the individual differences in Lability (*SD*'s) decrease with age. For the entire 18-year span (with a maximum of 38 tests per child) the mean Lability Score is 5.4, *SD* 1.2. Individual scores range from 12.23 for a boy in the Infancy period to 1.21 for another boy in the Preschool period.

Whether or not a Lability Score such as this will have value in describing characteristics of growth in children, or in differentiating children in any significant way, should be interesting to investigate. A few preliminary comparisons have been made. For example, we found no sex differences in Intelligence-Test Lability at any age-period, the largest critical ratio being 0.24 for the Preschool period.

Intelligence Test Lability has been correlated with level of intelligence at the several age-periods (see Table 9). The *r*'s are all practically zero,

TABLE 9
CORRELATIONS SHOWING THE RELATION OF INTELLIGENCE TEST LEVEL TO LABILITY SCORES FOR 40 CASES

Age at lability score	*r* with Intelligence Level at same age	*r* with mature Intelligence Level*
Months 1-21	—.02	—.005
Years 2-5	—.08	—.14
Years 6-18	.12	.18

*Mean Standard Score for years 16, 17, and 18.

the largest being that of mature intelligence with School-age Lability. This r of .18 is not significant but it is in line with McNemar's (29) finding of small significant relations between the magnitude of the *IQ* and test-retest differences. For the School-age period, the upper Quartile (10 cases) in intelligence has a mean Lability Score of 3.9, *SD*, .84; the middle 50 per cent is intermediate in Lability with a mean of 3.3, *SD*, 1.2; while the lowest intelligence quartile has a mean Lability of 3.0, *SD*, .85. The critical ratio between the means of the first and fourth quartile is 1.16. On McNemar's interpretation, this slight difference is inherent in the methods of test construction, and does not indicate that the brighter children are any less stable in their abilities over a period of time than those whose intelligence is mediocre or inferior. It does mean, however, that in interpreting the scores we should allow for some greater variability of *scores* at the higher levels of intelligence.

There appears to be little tendency for a given child to have a characteristic Lability pattern at all ages. Intercorrelations between the scores earned for the three age-periods are: Infancy with Preschool, .26; Infancy with School-age, .19; Preschool with School-age, —.29. As may be seen from the r's in Table 10, the score for the total 18-year span is determined almost

TABLE 10
CORRELATIONS SHOWING THE CONSISTENCY OF INTELLIGENCE TEST LABILITY SCORES FOR 40 CASES

Periods compared	r
Infancy with Preschool	.26
Infancy with School-Age	.19
Infancy with Total Span	.97
Preschool with School-Age	—.29
Preschool with Total Span	.38
School-Age with Total Span	.28

entirely by the Infancy scores, where the lability is so much greater than at the later ages.

Another approach is to select for study those children who are characteristically *labile* or *stable*. For this purpose we have called *labile* the 10 children (25 per cent) with the largest Lability Scores, and *stable* the 10 with the smallest scores, for any given period. Of the 40 children in the study there were four (two boys and two girls) who were *labile* for the total 18-year span and also for two of the three shorter periods. Similarly, there were two boys and two girls who were *stable* by the same criterion. Thirteen children (seven boys and six girls) were both *labile* and *stable*

at different periods. For example, a child would be very stable (in the lower Lability quartile) in his intelligence scores for several years, yet at another time he would become labile, with considerable change from test to test (i.e., in the upper Lability quartile). Only six (five of them girls) maintained moderate scores (i.e., in the middle 50 per cent) for all three periods as well as for the total 18-year span.

Whether the four children who can be characterized as generally labile are significantly different in any other respects from the four stable children, or whether these eight in turn are different from the six moderately labile, will have to await a more complete analysis of cases. The differences are not related to adult intelligence level: only one of the four labile children falls in the upper quartile of intelligence; the other three, as well as the four stable children are in the middle 50 per cent. It would appear, from inspection of the individual curves, that a high Lability Score is often the result of a consistent shift in relative mental status during the period covered by the score. Possibly a more fruitful measure of lability would be one which rules out consistent shifts in intelligence level by measuring the deviations of scores from a smoothed curve. As for the present method of measuring lability, it shows that not only are there wide individual differences among these children with respect to the lability of their intelligence test scores, but also that the degree of lability at one stage is no indicator of lability at another stage in the mental growth process.

The impression gained from inspection of the individual Sigma Score curves is corroborated by the Lability Scores. The relatively great lability of scores during the first two years is also evidenced in the correlation coefficients. However, even at the later ages, when the *r*'s between tests are high, some individuals are more steady than others in their mental progress. What is more, a child who had been labile may steady down to consistent intelligence test scores, while another child whose progress had been stable may speed up or slow down, thus increasing his Lability Score.

J. Summary

It has been the purpose of this report to present the growth trends in intelligence for a group of 40 children who had been tested at most or all of 38 testing ages from one month through 18 years of age. Attention has been focused primarily on age changes in variability of intelligence test scores and on individual consistency in relative scores.

Some evidence has been found which indicates that the distributions of intelligence test scores do not exhibit consistent trends in variability during

growth. There appear to be periods in which the abilities of children are relatively homogeneous, and others in which there are much greater individual differences. These periods are found in the scores obtained from a number of different tests and investigations, and thus seem to be inherent in the processes of mental development.

It is postulated that greatest homogeneity in scores occurs for a function when it is just starting to develop; that scores are most dispersed when that function is still growing rapidly but when those who are growing most rapidly in the function are not yet mature; and that as the slower-growing individuals reach maturity in the function the differences again become somewhat restricted. Consequently, if the tests are adequate measures of the abilities under consideration, fluctuations in the standard deviations of scores would be caused by the successive (and at times partially concurrent) developing and maturing of different types of intellectual ability.

If these postulates are valid, it would seem well worth while to direct studies, not only toward isolating, but also toward discovering the onset and course of development, of the different functions, or "factors," of intelligence. Furthermore, the tools with which we measure general intelligence should be fashioned with these considerations in mind.

Statistically, in order to increase the constancy of relative mental test scores (and to compare abilities in the same children through periods of time), it is important to use scores which do not fluctuate with the *SD*'s. It is also necessary to rule out differences due to the use of different tests with unequal standardizations. These sources of irregularity have been controlled for the Berkeley Growth Study by computing Sigma Scores (and Standard Scores) from the means and *SD*'s of the point scores or mental ages earned by these children.

The consistency of the mental test Sigma Scores is then studied by means of test-retest correlations, of individual age-curves, and of Lability Scores. The latter measure the extent to which each child fluctuates from his own intelligence level, in tests taken during a given age-span.

By all three methods of comparing, it is seen that children's scores are very labile during infancy, and become gradually more stable. By school age the prediction of the general level of intelligence is fairly stable. However, there are considerable individual differences in lability at all ages. This is true for our Sigma Scores, but when the test-norm *IQ*'s are used there is much wider fluctuation, especially for those children with the more deviant scores. Such deviant *IQ*'s should, in practice, be interpreted with great caution. These data point to the desirability of using some form

of Standard Score (or *IQ*'s derived from Standard Scores) instead of the ratio *IQ*.

The high *r*'s between scores on the Stanford-Binet, Wechsler-Bellevue, and Terman-McNemar Group tests indicate that these three tests measure much more nearly the same abilities than would be expected from the children's differences in *IQ*'s. Equivalent scores for these tests, based on comparable case samples would be useful in practice.

Boys and girls were found to be equally labile in their test scores. Children with high levels of intelligence were not significantly more labile than those with less intelligence, in this group.

For the school-age period which is definitely more stable than for younger ages, the children's Lability Scores averaged about one-third of a standard deviation, or roughly five or six *IQ* points. This figure is very similar to those given for earlier studies which emphasized the "constancy of the *IQ*." It must be kept in mind, however, that our Lability Scores are *SD*'s based on 10 to 13 tests per child (for the school-age period), and do not represent the extremes, but the central tendencies for a number of tests. Although many children maintain fairly constant levels of intelligence after six years of age, in some there are wide shifts in mental level. These shifts may occur at any age, and over a wide range of intellectual ability.

References

1. Abernethy, E. M. Relationships between mental and physical growth. *Monog. Soc. Res. Child Devel.*, 1936, **1**, No. 7.
2. Anderson, J. E. The limitations of infant and preschool tests in the measurement of intelligence. *J. of Psychol.*, 1939, **8**, 351-379.
3. Anderson, L. D. The predictive value of infancy tests in relation to intelligence at five years. *Child Devel.*, 1939, **10**, 203-212.
4. Bayley, N. Mental growth during the first three years: A developmental study of sixty-one children by repeated tests. *Genet. Psychol. Monog.*, 1933, **14**, No. 1.
5. ———. Mental growth in young children. *Yearb. Nat. Soc. Stud. Educ.*, 1940, **39**, 11-47.
6. ———. Factors influencing the growth of intelligence in young children. *Yearb. Nat. Soc. Stud. Educ.*, 1940, **39**, 49-79.
7. ———. The California First-Year Mental Scale. Berkeley: Univ. California Press, 1933.
8. Bayley, N., & Jones, H. E. Environmental correlates of mental and motor development; a cumulative study from infancy to six years. *Child Devel.*, 1937, **8**, 329-341.
9. ———. The Berkeley growth study. *Child Devel.*, 1941, **12**, 167-173.
10. Bradway, K. P. Predictive value of Stanford-Binet preschool items. *J. Educ. Psychol.*, 1945, **36**, 1-16.

11. DAVIS, W. A., & HAVIGHURST, R. J. The measurement of mental systems (can intelligence be measured?). *Sci. Mo.*, 1948, **66**, 301-316.
12. DEARBORN, W. F., & ROTHNEY, J. W. M. Predicting the Child's Development. Cambridge: Sci-Art Publishers, 1941.
13. DUBNOFF, B. A comparative study of mental development in infancy. *J. Genet. Psychol.*, 1938, **53**, 67-73.
14. EBERT, E. H. A comparison of the original and revised Stanford-Binet scales. *J. of Psychol.*, 1941, **11**, 47-61.
15. EBERT, E., & SIMMONS, K. The Brush foundation study of child growth and development: I. Psychometric tests. *Monog. Soc. Res. Child Devel.*, 1943, **8**, No. 2.
16. ELLIS, R. S. The "laws" of relative variability of mental traits. *Psychol. Bull.*, 1947, **44**, 1-33.
17. FILLMORE, E. Iowa tests for young children. *Stud. Child Welf.*, 1936, **11**, No. 4.
18. FREEMAN, F. N., & FLORY, C. D. Growth in intellectual ability as measured by repeated tests. *Soc. Res. Child Devel.*, 1937, **2**, No. 2.
19. GOODENOUGH, F. L. Studies of the 1937 revision of the Stanford-Binet scale: I. Variability of the *IQ* at successive age-levels. *J. Educ. Psychol.*, 1942, **33**, 241-251.
20. GOODENOUGH, F. L., & MAURER, K. M. The Mental Growth of Children from Two to Fourteen Years. Minneapolis: Univ. Minnesota Press, 1942.
21. HENMON, V. A. C., & LIVINGSTONE, W. F. Comparative variability at different ages. *J. Educ. Psychol.*, 1922, **13**, 17-29
22. HONZIK, M. P. The constancy of mental test performance during the preschool period. *J. Genet. Psychol.*, 1938, **52**, 285-302.
23. JAFFA, A. S. The California Preschool Mental Scale: Form *A*. Berkeley: Univ. California Press, 1934.
24. JONES, H. E., & CONRAD, H. S. The growth and decline of intelligence: A study of a homogeneous group between the ages of ten and sixty. *Genet. Psychol. Monog.*, 1933, **13**, 223-294.
25. KNEZEVICH, S. The constancy of the *IQ* of the secondary school pupil. *J. Educ. Res.*, 1946, **39**, 506-516.
26. KUHLMANN, F. Tests of Mental Development. Minneapolis: Educational Test Bureau, 1939.
27. LINDQUIST, E. F. Statistical Analysis in Educational Research. Boston: Houghton Mifflin, 1940.
28. MAURER, K. M. Intellectual Status at Maturity as a Criterion for Selecting Items in Preschool Tests. Minneapolis: Univ. Minnesota Press, 1946.
29. MCNEMAR, Q. The Revision of the Stanford-Binet Scale: An Analysis of the Standardization Data. Boston: Houghton Mifflin, 1942.
30. MILLS, F. C. Statistical Methods Applied to Economics and Business. (Rev. Ed.) New York: Holt, 1948.
31. NELSON, V., & RICHARDS, T. W. Fels mental age values for Gesell schedules. *Child Devel.*, 1940, **11**, 153-157.
32. PARKYN, G. W. The clinical significance of *IQ*'s on the revised Stanford-Binet scale. *J. Educ. Psychol.*, 1945, **36**, 114-118.
33. SARTAIN, A. Q. A comparison of the new revised Stanford-Binet, the Bellevue scale, and certain group tests of intelligence. *J. Soc. Psychol.*, 1946, **23**, 237-239.
34. SHUTTLEWORTH, F. K. The physical and mental growth of girls and boys age six to nineteen in relation to age at maximum growth. *Monog. Soc. Res. Child Devel.*, 1939, **4**, No. 3.

35. TERMAN, L. M. The Measurement of Intelligence. Boston: Houghton Mifflin, 1916.
36. TERMAN, L. M., & MCNEMAR, Q. Terman-McNemar Test of Mental Ability. Yonkers-on-Hudson: World Book, 1941.
37. TERMAN, L. M., & MERRILL, M. A. Measuring Intelligence: A Guide to the Administration of the New Revised Stanford-Binet Tests of Intelligence. Boston: Houghton Mifflin, 1937.
38. THURSTONE, L. Theories of intelligence. *Sci. Mo.*, 1946, **62**, 101-112.
39. WECHSLER, D. The Measurement of Adult Intelligence. Baltimore: Williams & Wilkins, 1944.

Institute of Child Welfare
University of California
Berkeley 4, California