

Vol. XXIV
No. 1

PSYCHOLOGICAL REVIEW PUBLICATIONS

Whole No. 102
1917

THE Psychological Monographs

EDITED BY

JAMES ROWLAND ANGELL, UNIVERSITY OF CHICAGO

HOWARD C. WARREN, PRINCETON UNIVERSITY (*Review*)

JOHN B. WATSON, JOHNS HOPKINS UNIVERSITY (*J. of Exp. Psychol.*)

SHEPHERD I. FRANZ, GOVT. HOSP. FOR INSANE (*Bulletin*) and

MADISON BENTLEY, UNIVERSITY OF ILLINOIS (*Index*)

Two Studies in Mental Tests

I. Variable Factors in the Binet Tests

II. The Diagnostic Value of Some Mental Tests

BY

CARL C. BRIGHAM, PH.D.

Instructor in Psychology
Princeton University

PRINCETON CONTRIBUTIONS TO PSYCHOLOGY

PSYCHOLOGICAL REVIEW COMPANY

PRINCETON, N. J.

AND LANCASTER, PA.

AGENTS: G. E. STECHERT & CO., LONDON (2 Star Yard, Carey St., W. C.)
LEIPZIG (Hospital St., 10): PARIS (76 rue de Rennes)



ACKNOWLEDGMENTS

The writer is indebted to Prof. H. C. McComas, Mr. Prentice Reeves and Mr. Norbert J. Melville for their cooperation in examining the Princeton children. The four experimenters are indebted to Miss Mabel T. Vanderbilt, supervising principal of the Princeton schools, not merely for allowing the use of the subjects, but for her careful attention to the administrative details of the experiment which facilitated the testing and made the work a source of pleasure.

The writer is indebted to Superintendent E. Mackey for the privilege of working in the Trenton schools, and wishes to express his deep gratitude to Dr. J. M. McCallie, supervisor of special classes and principal of the Franklin school, who willingly sacrificed much of his time in making all the arrangements for the conduct of the Trenton experiment. Without the cooperation of Dr. McCallie, the principals of the Carroll Robbins, Columbus, Cooper, Girard, U. S. Grant, and Jefferson schools, and the teachers of the special classes, the Trenton experiment would have been impossible.

Both the Princeton and Trenton investigations were carried out under the direction of Prof. H. C. Warren and Prof. H. C. McComas who at all times gave their expert guidance in devising and conducting the experiments, in suggesting methods of treating the data, and in correcting and revising the manuscript. The writer is deeply indebted to them for their assistance.

VARIABLE FACTORS IN THE BINET TESTS

TABLE OF CONTENTS

I.	Introduction	I
II.	Subjects and Methods	8
III.	The Personal Equation	18
VI.	Grade Correlations	37
V.	Sex Differences	65
VI.	Summary	91

I. INTRODUCTION

During the past decade, the Binet-Simon measuring scale for intelligence has received considerable attention, and a large amount of literature has appeared on the subject. No attempt has been made in the following pages to review all the literature on this scale or other systems of intelligence testing. Kite (38) gives an excellent account of the history and nature of the scale. Kohs (41) has assembled a very complete bibliography on the subject up to June 1914. Schmitt (57) gives an historical account of the development of the various attempts to correlate psychological findings with general intelligence, particularly in this country and England. Bobertag (10) and Schmitt both give detailed descriptions and analyses of the individual tests. Stern (62) has devoted a monograph to the collection, exposition and critical analysis of the large amount of data bearing on the problem of intelligence testing, and in another work (61) has assembled the literature of cognate fields. The literature bearing on the Binet scale up to 1912 is largely descriptive of the scale itself, the standard methods of procedure, etc. The more recent literature has been critical and reveals a tendency at the present time for investigators to depart from the methods of the extensive application of the scale as a whole to the more intensive study of the individual tests.

All systems of intelligence tests may be classified as qualitative or quantitative. The qualitative system consists of an aggregation of tests designed to detect the capacities or incapacities of the subject in order to afford the experimenter an opportunity to make a diagnosis concerning the subject's mentality. This method throws the responsibility for the final diagnosis on the experimenter. The system of tests proposed by Healy and Fernald (34) are of this type. Quantitative systems of tests necessitate a final score of some sort, whether that score be in the form of a mental age, a mental quotient, a certain number of points,

a coefficient of intellectual ability, a percentile rank or what not. The essential characteristics of the quantitative systems are the interpretation of the total scores in terms of the age of the subject, and the placing of the responsibility for the final diagnosis on the tests rather than the experimenter.

Binet and Simon's 1905 scale (5 and 6) was of the qualitative type. A series of 30 tests of approximately increasing difficulty was published with directions for their application. The authors reported in a general way that from their experience in examining a few selected normal children of different ages, and other subnormal children in the schools and at the Salpêtrière, approximate levels of performance could be found characteristic of the development of normal children of 3, 7, 9 and 11 years chronologically, the performance of idiots, imbeciles and morons corresponding roughly with that of normal children of 3, 7 and 9. Although the reference to chronological ages introduced the quantitative element, at no place were the authors insistent on this point, merely stating that they had found the series of tests exceedingly valuable in diagnosing and classifying defectives, and in their opinion others would also find it valuable.

The 1908 scale (7) was quantitative in character owing to the introduction of the concept of "mental age". It included a list of 56 tests grouped according to ages from 3 to 13, each group containing from four to eight tests. Most of the tests of the 1905 series were included, the additions including in a large measure tests of a scholastic nature. The authors gave directions for applying the series and for computing the resultant "mental age". A child testing three years below his chronological age was to be considered defective.

Although the scheme of the 1908 series was entirely quantitative, the authors did not discard the qualitative idea, and they cautioned against the application of the scale in the manner of a measure of height or weight. The border line between the idiot and the imbecile was fixed by the ability to use and comprehend spoken language. The imbecile was differentiated from the moron by the use of written language, illiteracy being dis-

ferentiated from imbecility by certain tests. The authors stated that the moron could be defined only in terms of the environment in which he lived, and they considered six tests important in differentiating the moron from the normal individual of the Paris population. Any system of tests which throws more weight on some tests than on others in making a differential diagnosis is fundamentally qualitative in kind, for the responsibility is placed not on the score but on the judgment of the experimenter. The idea of a quantitative measuring scale of intelligence however met with instant favor. The interest that actuated the psychologists of the "early nineties" to correlate the measurements of reaction time, motor ability, sensory discrimination, etc. with intelligence was revived. The scale was translated into several languages and applied to individuals of many classes and types.

In 1911, the authors published a revised scale (8) in which many of the tests of scholastic ability were discarded, and the remaining tests shifted about so that there were five tests for every year except one from III to X with similar groups for "twelve year", "fifteen year" and "adult" mentality. In the same year, Binet published an article (4), his last word on the subject, in which he discussed many of the criticisms which the scale had received, and again sounded the note of warning against the mechanical interpretation of results. However, as one traces Binet's thought on the subject through his writings, he may see the idea of a qualitative system of tests gradually dropping into the background, and more and more weight placed on the "scientific" (quantitative) measure of intelligence.

That Binet did not depart entirely from the qualitative standpoint is shown by his discussion of the test of comprehending difficult questions. "Sometimes after an examination one hesitates on a diagnosis. The child has failed in one or two tests, but this does not seem to be convincing. Failure to give the day and date and the months of the year are excusable errors, which may be caused by distraction or by lack of education. But the questions for comprehension dissipate all doubts. We recall

several instances when teachers brought us children, desiring to know whether or not they were abnormal; occasionally, in this way they set a trap for us, but we did not object, it was fair play. Our questions for comprehension decided us every time. We remember one child who was very slow in answering as though dull, his face was expressionless and unprepossessing; he knew neither the day nor the date, nor what day comes after Sunday, and he was $10\frac{1}{2}$ years old; his reading was syllabic. But when we asked question 5: Why do we judge a person by his acts rather than by his words? he gave the following answer: Because words are not very sure and acts are more sure. This was enough—our opinion was formed, that child was not so bad as he seemed." (Town's (72) translation, page 48.)

The popular interest that was manifest before the advent of the 1911 scale was tremendously reinforced in this country by Goddard's (30) publication of the results of the application of the scale to "two thousand" non-selected school children in Vineland, N. J. Popular interest increased rapidly, and the scale continued to have wider and wider application in the hands of less and less experienced investigators. The concept of "mental age" was exceedingly easy of comprehension, no apparatus was needed, and the scale has now become the common property of all. This development or overdevelopment has taken place in spite of the warnings of the authors themselves and the psychological fraternity in general. The very fact of overdevelopment however is striking evidence that persons interested in the social sciences need a quantitative scale for measuring intelligence.

The question whether the Binet scale is an accurate measure of intelligence can be decided only by the study of the individual tests and the factors underlying them. A study of this sort will show the errors that underlie the total score or "mental age", and at the same time will show the direction in which the correction of the scale should take place. The proper understanding of the individual tests involves the theory on which the measuring scale was constructed.

The method which Binet and Simon used in constructing their

measuring scale of intelligence was entirely empirical. A large number of tests were given to children of a certain social status. Certain tests could be shown to be correlated with age, and in the authors' opinion were correlated with intelligence. The fact that at a certain age a test could be passed by a certain proportion of the subjects was taken to mean that the test in question was characteristic of that age. Tests that were characteristic of the same age level were then combined into one age group. In this way a scale was built up with a number of tests for each age group. By a certain arbitrary system of scoring the reactions of a subject to all or part of the scale of tests, the "mental age" of the subject was obtained. The comparison of the "mental age" with the chronological age of the subject would show him to be advanced, at age or retarded, and the amount of acceleration or retardation would afford a quantitative index of his intelligence.

A person could construct a scale on the same basis and arrive at an age score using entirely different tests. A scale could be constructed containing tests of height, weight, vital capacity, strength of grip, circumference of the head, etc. and the results interpreted in terms of age. In this case however the age obtained would be more physical than mental. A scale of tests could also be constructed which involved the subject's knowledge of geography, spelling, history, grammar, etc. but in this case the resulting age would be determined very largely by the amount of training the subject had received.

The assumptions that a child at a certain age should weigh 25 pounds, at another age 50 pounds, etc., that a child can repeat 3 digits at one age, 5 digits at another and 7 digits at another, and that a certain percentage of children at one age can enumerate the months, and a higher percentage at another age, differ only in the possible determiners to which the growth may be referred. In the first case the growth is referred to certain physiological processes which are supposedly independent of intelligence and training. Binet believed that the principal determiner of growth in the last two cases was intelligence, but the possibility

remains that they might be more or less independent of intelligence, and more or less dependent on training and other variable factors.

The principle on which the scale was constructed involves three assumptions, (1) that the individual tests are correlated with age, (2) that the individual tests are correlated with intelligence, and (3) that intelligence is correlated with age—three distinct assumptions any one of which does not necessarily involve the others. The purpose of this investigation is to study the correlation of the individual tests with age, to determine the variable factors that might operate on the tests to produce an apparent correlation with age that was not a real correlation, or that might alter the real correlation in some way.

There is a possibility that an error might occur in the statistical treatment of the results, so that figures which would apparently indicate a correlation with age of a certain degree might actually represent a correlation of another degree. Another variable factor is the personal equation of the experimenter, who might alter the procedure in giving a certain test so that the correlation of that test with age might be different from the correlation obtained by another experimenter. If the subjects of various ages had received different school training, this difference might introduce another factor which would vary independently of the age of the subjects. If the tests used depended on any inherited or acquired differences between the sexes, then the correlation of the tests with age might be different for the two sexes. If any or all of the variable factors mentioned prove to be present in the correlation of the tests with age, then certain allowances will have to be made for these factors in making a diagnosis of the subject's intellectual ability on the basis of his total score or "mental age", and the scale becomes qualitative rather than quantitative.

At the Fourth International Conference for School Hygiene held in Buffalo in the summer of 1913, several persons of unquestioned authority in the field of mental tests held an informal

conference on the Binet-Simon scale, reporting the results in 1914 in the form of recommendations and suggestions (15). The question, "How much is the outcome of the testing influenced by the personal equation, both of the examiner and examinee?" was answered, "Undoubtedly there is some influence and it may be a serious source of error." Another question, "How much do previous environment and school training effect the outcome of the tests?" was left unanswered by the opinion, "The experimental evidence thus far available is conflicting. Further investigation is needed." The question, "Should the scale be divided, in the upper years at least, to furnish separate standards or separate tests for the two sexes?" was answered, "We do not know, and recommend this a subject for investigation." The following study is in part an attempt to answer these questions.

The method used in this study is that of studying the individual tests, disregarding entirely the total score or "mental age". There are at present so many revisions and editions of the Binet scale, that the term "mental age" has no meaning outside of the particular scale in question. The tests that are used in the various standardizations are however approximately the same, so that conclusions concerning the factors underlying the individual tests have a wider significance than those drawn from the "mental ages". Furthermore variable factors in the individual tests may balance each other in the total score so that their influence might be obscured.

The subjects and methods will be described first, and in connection with the methods of treating the results a statistical error will be pointed out. The problems of the personal equation, grade correlations and sex differences will then be taken up in detail.

II. SUBJECTS AND METHODS

SUBJECTS

The data which are here analysed to determine the influence of the personal equation, of grade training and of sex differences, are derived from all the boys and girls below the seventh grade in the Princeton, N. J., Model School. This group includes 422 subjects of the following age distribution,—

CHRONOLOGICAL AGES.

4	5	6	7	8	9	10	11	12	13	14	15	16
4	17	62	52	56	42	53	49	36	32	11	6	2

Each of the first six school grades was divided into a plus and minus grade, the latter division being under a different teacher, and containing those who were either backward, or, on account of illness, change of school, or for reasons not necessarily related to their mental development, were not sufficiently advanced to perform the work of their grade. The school also contained a special class for defective and exceptionally backward children. The subjects were distributed in the school grades as follows,—

SCHOOL GRADES.

Spec. Kind.	I—	I+	II—	II+	III—	III+	IV—	IV+	V—	V+	VI—	VI+	
18	32	38	51	12	40	12	45	15	35	15	49	11	49

39 or 9.2% of the subjects were children of non-English speaking parents, this group including 6.6% of the children in the Kindergarten and first six regular grades, and 15.7% of those in the special class and minus grades.

The selection of subjects is only fairly typical of the general run, for Princeton has no manufactories. The children examined came, for the most part, from the homes of laborers, domestics, artisans, farmers, tradesmen, clergymen and college professors. The selection is atypical in that none of the children came from homes of the manufacturing class, while an unusually large pro-

portion came from the homes of those engaged in domestic, personal, and professional service.

TESTS

The scale used was Goddard's (28) 1911 revision of the Binet-Simon scale. The methods used in giving the tests were, as far as possible, the same as those outlined by Goddard in the original revision, incorporating the rules and suggestions for standardized scoring published by that writer (29) in 1913. The methods used will not be discussed in detail, for the data are not used in obtaining age norms and standards for children generally. For the analysis of the data in terms of grade and sex it is not necessary that the procedure should be absolutely standardized, but that the experimenters who gave the tests should have used the same procedure. Differences in the technique of the experimenters will be discussed in the chapter on the personal equation.

One variation from the usual procedure was adopted. In no case did the experimenter know the chronological age of the child being tested. The influence of any prejudice or bias on the part of the experimenter is therefore eliminated from the problem of the correlation of the tests with age. The three experimenters who gathered the material in the spring of 1913 examined the sixth grade first and the remaining grades in decreasing order. During the school year 1913-1914, the fourth experimenter examined all children at that time in the kindergarten and first grades, and others who were not examined in the spring of 1913.

The tests in the "three year", "four year", "five year", "fifteen year" and "adult" groups were given so infrequently that the data from them are not treated. The tests used are as follows. The figure at the right shows the total number of times each test was given.

AGE VI

1. Distinguishing between morning and afternoon.....	108
2. Defining in terms of use.....	333
3. Executing three commissions.....	100

4. Showing right hand and left ear.....	107
5. Choosing the prettier of given faces.....	117

AGE VII

1. Counting 13 pennies.....	217
2. Describing pictures.....	219
3. Indicating omissions in pictures.....	217
4. Copying the diamond (in pencil).....	225
5. Naming four colors.....	218

AGE VIII

1. Comparing remembered objects (butterfly and fly).....	271
2. Counting backwards from 20 to 0.....	251
3. Enumerating the days of the week.....	277
4. Counting stamps.....	258
5. Repeating 5 digits.....	413

AGE IX

1. Making change.....	271
2. Defining in terms superior to use.....	333
3. Giving the day and date.....	307
4. Enumerating the months.....	284
5. Arranging five weights.....	334

AGE X

1. Recognizing pieces of money.....	282
2. Copying designs from memory.....	252
3. Repeating 6 digits.....	413
4. Comprehending easy and difficult questions.....	250
5. Using three words in sentence (two ideas).....	279

AGE XI

1. Detecting absurdities in statements.....	226
2. Using three words in sentence (one idea).....	279
3. Giving 60 words in three minutes.....	233
4. Giving rhymes with day, mill and spring.....	213
5. Reconstructing dissected sentences.....	190

AGE XII

1. Repeating 7 digits.....	413
2. Defining abstract terms.....	144
3. Repeating a sentence of 28 syllables.....	169
4. Resisting suggestion (length of lines).....	203
5. Solving problems from various facts.....	123

The tests in the "six year" group, with the exception of defining in terms of use, and the tests in the "twelve year" group, with the exception of repeating 7 digits, were given so infrequently or so irregularly that the data from them could not be treated. The apparatus used in the test of arranging five weights was not constant throughout the experiment, the standard cubes

and weighted pill boxes being used at different times by different experimenters. On this account, the data from this test are not included in the subsequent discussion.

METHODS OF TREATING RESULTS

The chronological age of each subject was taken as that at the last birthday, one tenth of a year being allowed for each 36 days beyond the birthday. The subject that was 10 years and 35 days would be rated 10.0 years, while ten years and 36 days would be 10.1 years. A subject one day short of 11 would be rated 10.9 etc. The teachers of each grade submitted the dates of birth of all pupils after the grade had been tested. These data were later checked up from the entrance cards. Since the purpose of this study is to analyze the factors involved in the individual tests, no "mental ages" or total scores were figured. The classifications of the subjects are all made independently of the tests.

Two measures of central tendency will be used in the subsequent discussion, the average and the median. The measure of variability from the average, that will be used, is the mean variation (or average deviation), the average of the differences, regardless of signs, between the separate measures in the series and the average of the whole series. The measure of variability from the median that will be used is the semi-interquartile range (Q), or half the difference between the measure with three times as many measures above as below it and the measure with one third as many measures above as below it, i. e. half the difference between the 25 percentile, and the 75 percentile. Any coefficients of correlation used will be stated in terms of the formula applied. The reader is referred to Thorndike (70) for the discussion and explanation of the statistical measures used.

The measures of ability in most of the tests are in the "all or none" form—the tests are either passed or failed. The only measure that can be obtained from data of this sort is the percentage that an ability is present in a defined group. This method of treating the results has as many "pit-falls" as the tests themselves. Before undertaking the analysis of the Prince-

ton data to determine the effect of the personal equation of the experimenter, and the age, grade, and sex of the subject upon the results of the individual tests, it is necessary to consider an error which underlies incomplete data, or those data derived from experimenting in which every test is not given to every subject.

No uniform instructions were given to the experimenters concerning the order in which the tests should be given, nor the number of tests that should be tried. The experimenters attempted to determine the mental age of the child according to the scale. In doing this they would start with some test which they considered would be interesting to the child, and, at the same time, well within his reach. The tests given first were usually those of describing pictures and arranging five weights. The experimenter would then gradually explore the subject's range of ability, varying the order of the tests so as to maintain the subject's interest, and to ward off fatigue. In this way the experimenter would eventually establish the basal age of the subject (that age in which he passed all five of the tests), and by the end of the examination would have tried all the tests above the basal age which, in his judgment, there was any possibility of the subject's passing. This method of experimenting will be called incomplete. The other method of experimenting, in which a certain number of tests are adopted and all of the tests are tried on each subject, will be called complete. Each experimenter in the Princeton investigation averaged 19 or 20 tests to a subject. In the Trenton investigation all the tests were given to all the subjects.

The incomplete method is more desirable from the standpoint of the subject who is not unnecessarily fatigued, and from the standpoint of the experimenter, as well, who saves in the expenditure of time and energy. However, the data derived from the incomplete method are subject to an error, which, unless it is properly considered, will completely vitiate the results.

When the experimenter does not try a test above the basal age because he believes that the subject will not pass it, he implies that the subject will fail it. This amounts to a failure,

for the subject receives no credit. However, a failure of this sort, due to the experimenter's assumption, is not the same as an actual failure in which the test is tried, for there is always the possibility that the assumption was unjustified. In like manner when the experimenter does not try tests below the basal age, he actually gives credit for passing the test without the actual trial.

In some cases the assumption on the part of the experimenter is quite justified. Obviously if a subject can make change, he can count up to thirteen; if he can repeat seven digits, he can repeat five and six digits; if he knows the names of the months, he will know the days of the week; and, conversely, if he cannot repeat the days of the week, he cannot repeat the months. Other assumptions are less justifiable. Since very intelligent persons, lacking in particular sorts of abilities, might fail in tests such as drawing the design from memory or arranging five weights, there is no reason for supposing that a subject making basal "eleven" or "twelve" will pass these tests. At the same time there is no reason for assuming that a subject failing to establish basal "seven" for instance, will fail to pass a test such as the line suggestion test in "twelve". The assumptions of the experimenters, then, are more or less justifiable and it is impossible to estimate the amount of the justification, since this is dependent on the nature of the individual tests.

The manner in which this error works out in the statistical treatment of the results may be shown by examining any test which has been tried through a number of chronological ages. Table 1 shows the results of the 60 word test obtained from subjects 7 to 13 years of age.

TABLE NO. 1

Analysis of the Results from the Test in Naming 60 Words in 3 Minutes.

Chronological ages	7	8	9	10	11	12	13
No. of times given.....	11	18	25	42	44	31	28
No. of time passed	4	10	10	24	34	19	21
Actual percentage passed.....	36%	56%	40%	57%	77%	61%	75%
Total number of subjects.....	60	52	42	54	48	36	28
Percentage of subjects to whom							
test was given.....	18%	35%	60%	78%	92%	86%	100%
Theoretical percentage passed....	7%	19%	24%	44%	71%	53%	75%

An example will make the above table clear. The 60 word test was given to 11 subjects, age seven, 4 of whom passed. In all there were 60 subjects at this age, so that the 11 subjects to whom the test was given constitute but 18% (and probably the brightest 18%) of this whole number. The percentage passed would have been 7% had the test been given to all 60 subjects, and had all the subjects failed who the experimenters assumed would fail if they gave the test. The true per cent. which represents the ability of non-selected seven year boys and girls in passing the 60 word test therefore lies somewhere between 7% and 36%, probably nearer 7%. An accurate estimate of the real per cent. which will represent this ability is, however, impossible. In like manner, the ability of the 8 year subjects is represented by a percentage somewhere between 19% and 56%.

As the proportion between the number of subjects in the group and the number actually tested increases, the disparity between the actual and theoretical percentage passed becomes less, or, in other words, the results which express the ability of a group become more reliable as the number of individuals actually tested as a sample of this group becomes larger. The higher the percentage given, the more reliable the percentage passed, when the reliability is measured by the difference between the actual percentage passed and the theoretical percentage passed.

The source of error mentioned causes great difficulty in comparing the results of different investigators. For example, it is desired to compare the results of Terman and Childs (66) and Dougherty (23) with those of this investigation on the 60 word test. Table 2, derived from their published results, shows the percentage that the test was given of the number of times it was possible to be given, (%G), the actual percentage passed, (A%P), and the theoretical percentage passed, (T%P), or that percentage passed that would have resulted had all of the subjects failed, who it is necessary to suppose would have failed, had the test been given all the possible number of times.

TABLE NO. 2

Analysis of the Results of Three Investigators on the 60 Word Test.

Age	This investigation			Terman and Childs			Dougherty		
	%G	A%P	T%P	%G	A%P	T%P	%G	A%P	T%P
7	18	36	7	14	50	7			
8	35	56	19	47	35	16	15	0	0
9	60	40	24	86	57	49	35	60	21
10	78	57	44	100	67		78	53	41
11	92	77	71	98	83	82	89	79	70
12	86	61	53	97	82	80	91	95	87
13	100	75		100	94		94	88	83

It is very difficult, if not impossible, to make a comparison of these results shown in Table 2 for the years 7, 8 and 9. The ability of Terman's 7 year group is represented by a figure somewhere between 7% and 50%, while that of the 8 year group falls somewhere between 16% and 35%. Dougherty's 9 year group falls between 21% and 60%. In the older years where the results have greater reliability, it is probable that the discrepancies between the investigators could be accounted for on the basis of the inferiority of the selection of the older subjects in this investigation, the other investigations including children from the seventh and eighth grades.

In order to make a comparison between investigators, it is necessary to express the results in terms of a percentage or a proportion. The expression of the ability of a group by a percentage or a proportion is inaccurate if the data are incomplete, and in order to judge the accuracy of the data, it is necessary to know the degree of completeness. Unfortunately, the results of most of the investigations on the individual tests are not published in a form that enables one to estimate the accuracy of the data. The writers who have published their data in a form that will admit of this treatment, have not treated the sexes separately. On this account, the writer will not attempt a systematic comparison of the results of this investigation with those of other experimenters.

Before analysing the Princeton data the following problem should be answered:—What proportion of a given group must actually be tested for an ability in order that the results may be

considered as typical of the ability of the whole group? The proper proportion to select as typical of any one group depends upon the characteristics of the group itself. If the members of a group are similar, a smaller proportion would stand for the ability of the group than would be necessary for a group composed of unlike individuals. A smaller number of individuals would be necessary to stand for the ability of all the 12 year boys in the sixth grade, for example, than for all the 12 year boys coming from a great many grades. This proposition operates directly counter to actual practice, for the members of a group of similar individuals will be given similar tests, while unlike individuals will receive different tests, inasmuch as the experimenter adapts his procedure to the need of the individual being examined. The proposition actually means, then, that selected results from incomplete testing are more reliable than non-selected results, if each group has the same range of testing. The proportion of a group that must be tested to stand for the whole group will also vary from test to test. In some tests of particular abilities, no proportion will accurately stand for the whole group—the entire group must be tested. In other tests that are easy for the group, the results of a very small proportion would not be altered by examining the remainder of the group.

The problem of deciding what proportion of a given group must actually be tested for an ability in order that their results may be considered as typical of the ability of the whole group has, therefore, no answer in the work. The writer will decide arbitrarily what the proportion will be. The actual magnitude of the proportion between the number actually tested and the number in the whole group (the percentage given) will always be published as an index of the reliability of the percentage that the group passes the test in question.

It is not possible to obtain reliable results showing the growth of an ability with age, if the data on which the results are based are of the incomplete sort. A test for any age will be given to a superior selection of subjects below that age, and an inferior selection of subjects above that age, so that the growth curve

will appear flatter than it actually is. For this reason, the Princeton data may not be used for the purpose of standardizing age norms.

Binet (4) recognized the fallacy of calculating proportions from the actual number of times a test was given and passed when the test had not been given all the possible number of times. In calculating the proportions from Levistre and Morle's data, Binet used what the present writer would call the "theoretical proportion passed".

It has been shown that the reliability of the theoretical percentage passed rests on the accuracy of the experimenters' assumptions, and that according to the nature of the tests and the character of the groups to which they are given these assumptions vary from complete certainty to absolute uncertainty. Inasmuch as these assumptions are not equally certain, the conclusions drawn from them are not equally certain, and the logic of scientific method demands that an investigator establish the degree of certainty of his conclusions. In this case the measure of the degree of certainty is the magnitude of the percentage given.

The use of the theoretical percentage passed without reference to the percentage given ignores the dictum that an investigator establish the degree of certainty of his conclusions, and sets up all conclusions as equally valid, a procedure which in actual practice results in making all conclusions equally invalid when the fact of degrees of certainty is admitted. The investigator who draws conclusions from incomplete data should always state the percentage given and the actual percentage passed. This much at least is experiment. The only legitimate use of the theoretical percentage passed is when it is compared with the actual percentage passed as a probable limiting value. The theoretical percentage passed alone has no claim to reliability.

III. THE PERSONAL EQUATION

Before attempting to correlate the individual tests with age, grade and sex, it is necessary to demonstrate the presence or absence of the effect of the personal equation. By the term "personal equation" is meant the complex of variable factors which are independent of the mental make-up of the subject and the environmental conditions at the time of the examination. The term includes such widely different factors as the experimenter's ability to obtain the cooperation of the subject, his procedure in giving the tests, his criteria in deciding whether a subject's response should pass or fail, and the tests used, insofar as the selection of tests and the construction of the apparatus were occasionally left to his discretion, apart from the uniform procedure.

The only method of detecting the influence of the personal equation in most of the tests is that in which the responses of similar groups of subjects to different experimenters are compared. On account of the wide variations in the character of the subjects examined, it is not possible to compare similar groups. On some tests, however, it is possible to determine the effect of the personal equation independently of the method of group comparison. The results of the tests that may be studied independently will be discussed at some length, in order to demonstrate the fact that certain tests are susceptible to this influence.

The examinations of the Princeton subjects were made by four experimenters, called for convenience A, B, C and D. None of the experimenters was highly trained in giving the tests, although they had all been trained in the methods of psychological experimentation, one experimenter being an assistant professor of psychology, and the other three graduate students of psychology of at least one year's standing. B, C and D performed their experiments at the same time, in the spring of 1913, while A experimented one year later. B, C and D studied

the scale together so that it was possible to secure a correspondence in method. At the close of practically every day's testing, B, C and D would confer on the questions brought out by the day's work, and as far as possible would adopt uniform methods of procedure and scoring. A was subsequently trained in these same methods.

In spite of the attempt to adopt uniform methods, there were a few tests which always caused difficulty, and concerning which the experimenters could reach no definite agreement. One of the tests that caused the greatest difficulty was that of defining in terms of use and in terms superior to use. The hierarchy of responses to this test could be fairly arranged as follows. To the question "What is a chair?" the following typical responses would be obtained,—1, "A chair is a chair." 2, "This is a chair." 3, "A chair is to sit on." 4, "A chair is what you sit on." 5, "A chair is a thing you sit on." 6, "A chair is a piece of furniture you sit on." 7, "A chair has four legs, a back, etc." 8, "A chair is a piece of furniture with four legs, a back, etc." Any of the objects for which a definition is asked (fork, table, chair, horse, mother) may be defined by repetition, by demonstration, by indicating the use to which it is put, by showing the class to which it belongs, by describing its parts, or by the combination of any or all of these methods.

The only problem is to decide, arbitrarily, how definitely the class must be indicated (i. e. by "what", "a thing" or "a piece of furniture") in order to have the definition considered as one of classification. The rule adopted in this study was to consider "thing" as indicating the class. Nos. 1 and 2, definitions by repetition and demonstration, received no credit in "six years". Nos. 3 and 4 were given credit in "six years" as definitions by use, and nos. 5, 6, 7, and 8 were given credit in "nine years" as definitions in terms superior to use.

In studying the ranks given to the responses of the subjects in this test, it was found that the experimenters did not record the responses all of the time. A gave the test 94 times, and recorded the responses 66% of this number. B gave the test 98 times, recording the subject's answer 67% of the time. C

gave the test 65 times, and recorded the answer in 95% of the cases, while D gave the test 76 times and recorded the response only once.

By ranking the recorded responses of A, B and C according to the rules shown above, it is possible to obtain an estimate of the relative severity of their criteria in marking these responses plus or minus. 19% of A's definitions were corrected, the correction in all cases being from minus to plus. 11% of B's definitions were corrected, all of the corrections being from plus to minus. 17% of C's definitions were corrected, three fourths of them being changed from plus to minus, and one fourth from minus to plus. C's standards changed during the course of the experiment, so that at first, with older subjects, he was too lenient, while later, with younger subjects he was slightly too severe. The tendencies of A and B remain constant throughout the experiment, A marking too severely and B slightly too leniently. The differences between the experimenters hold constant for both sexes. The experimenters agreed on all definitions by use, the cases of disagreement coming on the definitions superior to use.

One test in which variations between the experimenters might be expected is that of copying the diamond. In this test, although the apparatus and procedure were the same, the experimenters had very little to guide them in forming their judgments of passed and failed. The instructions given ("The result is considered satisfactory if it would be recognized as intended for a diamond shaped figure"), and the examples published furnish very vague criteria.

In order to determine the effect of the personal equation of the experimenters in giving credit on this test, all of the reproductions of the diamond obtained in the Princeton and Trenton experimenting, (311 in number), were first transcribed and then ranked. On the sheet containing the copy only the subject's number was placed, so that the person ranking the reproductions was in ignorance of the experimenter by whom it was obtained, the mark that the experimenter had given it, and the age, grade, sex, etc. of the subject. The 311 diamonds were then classified

into six groups by one observer. The classification, at best, was vague and indefinite, but it represented the unbiased judgment of a single person. Inasmuch as the reproductions were classified and re-classified a great many times, small errors in the classification would be counterbalanced.

The first group contained fairly accurate reproductions of the original, diamonds of approximately the same size as the copy, having the sides and opposite angles nearly equal, and with a proper proportion between length and width. The second group contained figures inferior to those of the first group in size or symmetry, but representing a fairly high grade of ability. The reproductions that were less symmetrical than those of the second group were classified in the third and fourth groups. Figures showing some inequality between length and width were classified in the third group, while those of approximately unit proportion, square shaped figures, were classified in the fourth group. The reproductions placed in the fifth group were figures less symmetrical than those of the fourth group, and figures which had curved sides and rounded corners. The sixth group contained all figures which it would have been difficult to have recognized as intended for a diamond, figures having three, five or more sides, circles, ellipses, unfinished lines and eccentric figures.

The above classification did not offer an opportunity for a sharp grading between one group and another, but in general, the reproductions placed in the various groups from the first to the sixth, represented a decrease in the ability to copy the diamond. The justification of the method was not in the accuracy of the classification, but in the fact that the material was all classified by one observer (B), in such a way that he was in ignorance of the original rank that had been given the reproduction, of the experimenter who graded it, and of the character of the subject.

16% of the reproductions were classified in the first group, 21% in the second group, 20% in the third group, 17% in the fourth group, 9% in the fifth group, and 17% in the sixth group. (The irregularity of the distribution is due to the presence of the diamonds drawn by the Trenton subnormal group.)

After classifying all of the reproductions the ranks given to them by the different experimenters were then compared with the group in which they were classified. That the sliding scale classification used represented real differences between the reproductions is shown by the relative certainty of the experimenters' judgments. None of the reproductions classified in the first and second groups were ranked as failed by the four experimenters, while only one reproduction in the third group was ranked minus. 18% of the fourth group, 45% of the fifth group and 77% of the sixth group were ranked as failures. All of the sixth group diamonds that were ranked plus (23%), were so ranked by one experimenter, A.

To obtain a general estimate of the relative severity of the experimenters' criteria in making their judgments of passed or failed, the diamonds obtained by each experimenter from boys and girls were classified according to rank, plus or minus, and according to their group in the classification. From this it was possible to obtain an estimate of the passing mark of each experimenter. For example, the boys of experimenter B passed the test 72% of the time according to his ranking. Had B given credit for the first five groups and failed only the reproductions in the sixth, i. e., had his passing mark been the fifth group, 88% would have passed. Had his passing mark been the fourth group, 81% would have passed. If it had been the third group, 72% would have passed, while only 56% would have passed had it been the second group. Since 72% of B's subjects actually passed the test, his passing mark was the third group—in the long run, he would pass all diamonds in the first three groups and fail all in the last three. The differences between the experimenters on this basis are quite marked. The passing mark for C and D was the fourth group, while A's passing mark was the fifth group. B was the most severe, and A was the most lenient, with C and D between the two. The results were the same for both sexes.

Another test in which the influence of the personal equation might be looked for is that of copying designs from memory. The experimenter must here use his own judgment in marking

the designs passed or failed. Very little guidance is given by Binet's rule, which reads, "The test is considered passed when one of the designs is reproduced exactly, and half of the other is correctly drawn", or by the interpretation of this "half right" as applying "when two component parts are transposed or one component part omitted".

In order to test the experimenters' judgments in ranking this test, a scoring system was devised, which may be explained by reference to Figure 1, which gives the original copy and various duplicated portions. In scoring the reproductions of the pyramid section, 5 points were given when the reproduction of the asymmetry of the figures was nearly exact, as in no. 1, 4 points for a less perfect reproduction as in no. 2, and 3 points for a reproduction in which the rectangle fell in the center of the figure, as in no. 3. 1 point was deducted from this score for each failure to connect the corners of the rectangles as in no. 4 (which is modified from no. 3 and would therefore receive only 1 point), and no credit was allowed for "boxes" (no. 5), and other eccentric figures.

In scoring the more complicated design, 4 points were allowed for each of the "posts", ABCDE and JKLMN, or no. 6. 2 points were deducted for turning them in the wrong direction as in no. 7, (which is "post" ABCDE turned in the wrong direction), 2 points for failure to make the line AB penetrate DE as in no. 8, so that a combination of these errors, as in no. 9, would receive no credit, along with other eccentric reproductions as in nos. 10 A, B, C, D, E and F. 1 point was given for each of the lines EF and IJ, and 5 points for the "hump", FGHI. A continuous line from E to J as in no. 11 would therefore receive no credit, while a division of the lines, without the portion FGHI, as in nos. 12 or 13, would receive 2 points. An accurate reproduction of the portion EFGHIJ, as in no. 14, would receive full credit for all parts, 7 points, no credit being allowed for eccentric reproductions of the "hump" as in nos. 15 A, B, C and D.

The maximum credit for the test is 20 points, divided between the two figures on the proportion of 5 to 15, a fair proportion

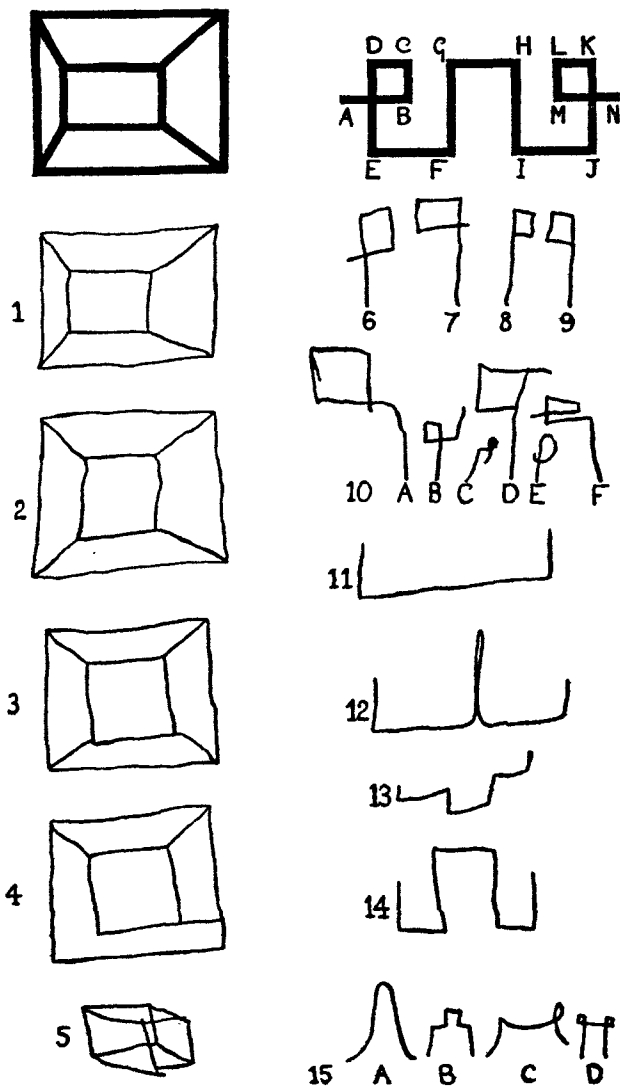


FIG. 1. *Method of Scoring Test of Copying Designs from Memory*

(in the writer's opinion) according to the relative difficulty of the parts. A design with "one component part omitted" would be scored 13 points according to this system, and one with "two

component parts transposed", 16 points, provided that the reproductions of the pyramid section were perfect in each case.

All the reproductions of the designs obtained from the Princeton and Trenton experimenting were then scored according to this system. The score of each subject of each experimenter in the Princeton series was then compared with the experimenter's ranking, which was recorded on the same sheet, and which was not seen at the time the designs were graded by the point system. From the number of times the test was given, and the number of times it was marked passed by the experimenter, the percentage passed was obtained for each experimenter for both sexes. The scores from all the designs from 0 to 20 were then classified according to the judgment passed or failed as given by each experimenter on subjects of both sexes. It was found that there were certain ranges where the experimenters' judgments coincided accurately, i. e. in the very low scores and in the very high scores. A certain range existed, approximately from 10 to 15 points, in which the same results would sometimes be ranked as passed and failed by the same experimenter at different times.

It was possible, however, to obtain a general estimate of the experimenters' criteria by a method similar to that used in the study of the diamond test. For example, B gave the test to boys 48 times, passing 40% of them. Had his passing mark been 18 (i. e. had he passed all subjects whose designs scored 18 points or better), 21% would have passed. Had his passing mark been 15 points, 35% would have passed. Had it been 13, 42% would have passed etc. B's passing mark would therefore fall between 13 and 15 points. In this way, by calculating the percentage passed at each score for each experimenter for both sexes, it was possible to obtain the passing mark of each group. The passing marks coincided very closely except in one case. With one exception the passing marks were around 12, 13, 14 or 15 points, for the boys and girls of all experimenters, i. e. the experimenters would, in the long run, rank all below this level minus and all above this level plus. The degree of correspondence was quite remarkable considering the fact that the

experimenters had very little on which to base their judgments.

The one exception is both striking and suggestive. C's passing mark for boys was 15 points, for girls 8 points. In order to receive a plus from C, boys would have to draw a much more accurate design than girls, or, in other words, a very faulty reproduction drawn by a girl would receive credit, while the same reproduction if drawn by a boy would invariably be failed. This deviation rests on a small number of cases. A gave the test to 24 boys and 21 girls, B to 48 boys and 33 girls, C to 28 boys and 22 girls, and D to 36 boys and 31 girls. A's results, although resting on a number of cases as small as C's, show no such deviation as those of the latter. On account of the small number of cases, this finding cannot be considered definite. It does, however, suggest the possibility of a difference in the experimenters' reaction to the sexes. An experimenter may show greater leniency to one sex than to the other, so that a supposed sex difference may be the results of an experimenter's reaction to the sex, rather than the sex's reaction to a test.

The test of using three words in a sentence ("Philadelphia, money and river") was given 279 times, and the sentences given by the subjects were recorded over half the time. Experimenter A gave the test 53 times, recording the result 36% of the time. B gave the test 95 times, recording the answer in 92% of the cases. C gave the test 56 times, recording the answer 23% of that number, and D gave the test 75 times, recording the response in 43% of the cases.

To obtain a check on the accuracy of the experimenters' scoring of this test, all of the recorded sentences were transcribed so that they could be studied and ranked without reference to the subject or the experimenter. The 162 recorded sentences were then marked plus or minus by one observer (B). This ranking was checked several times and then compared with the original ranking.

There was no disagreement between the judgments of the four experimenters and the one impartial observer in marking responses for the "ten year" credit. In marking for the "eleven year" credit, there were 8 disagreements out of the 162 judg-

ments, the 8 variations being evenly distributed among the experimenters. It may be concluded, then, that the influence of the personal equation is absent in this test, although there is ample opportunity for variation.

The detailed study of the foregoing tests has shown that the personal equation of the experimenters has a marked effect on the results of some of the tests. In the subsequent correlation of the tests with grade and sex the corrected score of these tests will be used. Only those definitions will be used which were recorded by the experimenters, and the ranking of the one observer will be followed. All reproductions of the diamond in the fifth and sixth group will be scored as failed, the others as passed. A reproduction of the designs scoring 15 or more points will be ranked as passed. The corrected results of the sentence test will be used.

To show that the effect of the personal equation of the experimenter is present or absent in the tests on which there is no actual record of the subject's response, is a more difficult problem. The most reliable method of showing the influence of this factor is that in which the reactions of similar groups of subjects, examined by different experimenters, are studied. The greater the similarity of the groups the more reliable the results. If two experimenters each examined 50 boys of 12 years of age from the sixth grade, their results should compare closely, and any difference could immediately be referred to a difference in the personal equation. However, if one examined boys from this grade and the other girls, the variations might be explained on the basis of sex differences. In the same way the results may vary with the age of the subject, and with his grade and nationality.

It is not possible in this study to obtain groups of a sufficient degree of similarity, in spite of the small number of children of non-English speaking parents, and the fact that the sexes may be treated separately. The subjects vary in age from 4 to 16, and in grade from the kindergarten to the sixth grade. A examined a very much younger run of subjects than B, C and D. The data of the four experimenters were treated by three meth-

ods, by comparing the per cent. that all boys and girls of each experimenter passed each test, by comparing the per cent. that selected subjects of each experimenter passed each test, and by comparing the per cent. that all subjects from 5 to 9 and from 10 to 13 passed each test. The sexes were separately treated in each method. None of the methods proved satisfactory, and it was found to be impossible to obtain an accurate quantitative estimate of the effect of the personal equation on each test. In certain of the tests, however, there were known differences of procedure which might have influenced the results, while the variations in the results of certain other tests were so striking that definite conclusions could be drawn.

One possible source of variation was the use of alternative questions in several of the tests. When an entire school system is examined, and the children learn that they will all be tested, the possibility is always present that they will inform each other of the nature of the tests, and the answers to some of the questions. The alternative questions were used to counteract the influence of this factor.

In the test of detecting absurdities in statements, ten or eleven statements were used, the experimenter choosing the five that he would give the subject. The statements varied greatly in difficulty and the experimenters did not use the same selection throughout the experiment. This test was given by B to 26 girls whose average age was 10.6 years, while D gave the test to 25 girls whose average age was 10.9 years. 65% of the girls examined by B passed the test, while only 36% of D's group passed. The variation between the experimenters might be due to the selection of absurdities of unequal difficulty, or to different criteria in grading the responses. The sources of variation are too large to admit of obtaining any reliable results from this test in correlating it with grade and sex.

75% of the girls to whom B gave the test of reconstructing dissected sentences passed, while only 28% of C's girls passed. The average age of the 26 girls to whom B gave the test was 10.8 years, and the average age of C's subjects 10.5 years. Part of the difference between these two experimenters is due to the

fact that more of B's subjects came from the fifth and sixth school grades. Some variation might have been due to different apparatus, B using cards with the sentences printed on two lines, while C had the sentences typewritten on one line. The sentences used by B were more legible, and, being broken into two lines, it was easier to grasp the individual parts as discrete units. Each experimenter used six sentences of varying difficulty so that some variation might be expected from the selection of the three sentences for the test. Whatever the cause of the discrepancy between the results of the two experimenters, it is obviously impossible to obtain any reliable conclusions concerning the correlation of this ability with age, grade or sex, on account of the presence of so many variable factors.

Three problems were used in the test of making change, 20c — 4c, 25c — 6c and 25c — 9c, the process of subtraction involved in each being of unequal difficulty. Certain variations occurred in the tests of comparing remembered objects and comprehending easy and difficult problem questions. Alternative questions were used in both of these tests, and variations might occur due to the relative severity of the experimenters' judgments in marking the responses passed or failed. None of the tests in which alternative questions were used will be treated in the subsequent discussion of the results.

At the close of the experiment, it appeared that a difference of procedure had existed between A and B in the test of indicating omissions in pictures. A and B both showed the three faces first, and the figure with the arms missing last, according to the standard procedure, but A, if his subjects failed to detect the parts omitted from the faces, would give them another trial after they had detected the missing arms. A gave this test to 51 boys and 33 girls, B to 30 boys and 30 girls, his subjects averaging about a year and a half above those of A. The test was passed by 76% of A's boys and 97% of A's girls, but by only 60% of B's boys and 63% of B's girls, showing that the difference of procedure had a most striking effect on the results. It is interesting to note what the effect of a difference of this magnitude would mean if the material from this test were used

as a basis of assigning it to the proper "age group" in the scale. If a test is to be considered normal for a given age if it is passed by 75% of the non-selected school children of that age, the test of indicating omissions in pictures would be a "six year" test for A, and an "eight year" test for B. The data from this test will not be treated in the subsequent discussion.

In the analysis of the results of the definitions test, it was found that certain differences existed between A, B and C in scoring the responses of the subjects as superior to use. No estimates could be made concerning D, for he did not record the actual responses. B, C and D gave this test to approximately the same range of subjects, averaging about 9 years. The corrected results of B and C show, in all, 28% of their subjects giving definitions superior to use, while 65% of D's subjects pass this test. Obviously D was very much more lenient than B and C.

The influence of the personal equation may or may not be present in the remaining tests. In the opinion of the writer it is not present to any marked degree. The data of the four experimenters were treated in several ways, and in none of these was it possible to demonstrate this influence. The writer's opinion, however, is more or less certain according to the test. The tests of repeating digits might show a slight difference between C's results and those of the other experimenters, a difference which could be explained by reference to the rate at which the digits were spoken. The results of experimenter D are slightly lower than those of the other experimenters in the tests of naming 60 words in three minutes and naming rhymes. Whether these differences are real or not, the writer does not know. The data from these tests are included in the subsequent study.

In the subsequent treatment of the results in terms of grade and sex, the material from the following tests will be treated.

VI-2 and IX-2, Defining in terms of use and in terms superior to use.

VII-1, Counting 13 pennies.

VII-2, Describing pictures.

VII-4, Copying diamond.

VII-5, Naming four colors.

- VIII-2, Counting backward from 20 to 0.
- VIII-3, Enumerating the days of the week.
- VIII-4, Counting stamps (three singles and two doubles).
- VIII-5, X-3 and XII-1, Repeating 5, 6 and 7 digits.
- IX-3, Naming the day and date.
- IX-4, Enumerating the months.
- X-1, Naming the pieces of money.
- X-2, Drawing designs from memory.
- X-5 and XI-2, Constructing a sentence, containing one or two ideas from three given words.
- XI-3, Giving 60 words in three minutes.
- XI-4, Giving rhymes with "day", "mill" and "spring".

The treatment of the results of the definitions test will be confined to the recorded and corrected definitions of A, B and C. The results from the diamond test are based on the scoring system outlined, the passing mark being the fourth group unless otherwise indicated. The arbitrary point system of scoring the design test is used in the subsequent calculations, the passing mark, unless otherwise noted, being 15 points. The corrected scoring of the sentence tests will be used.

The foregoing study of the effect of the personal equation shows conclusively that in certain tests this influence is present to a very marked degree. The errors involved may be traced to three sources, to the apparatus used, to the technique of the experimenters in giving the tests, and to the experimenter's observation in marking the test passed or failed.

The error due to apparatus may result from a variation in the material itself, or from the calibration of different sorts of material as equal in difficulty, e. g.—alternative questions. The variation in the material used by B and C in the test of reconstructing dissected sentences illustrates the error due to defect in the material. The writer has seen apparatus for the line suggestion test in use, in which the last three pairs of lines were actually unequal, the difference between the pairs being above the threshold of discrimination. The subject with good discrimination will invariably fail this test when this faulty apparatus is used.

The error due to the use of alternative questions is more

common and therefore has more practical significance than defects in the material itself. There is a strong temptation for an experimenter, who believes a certain question to be unfair, to substitute another which seems to him to be of the same difficulty. In the study of the Trenton results, which will follow, it will be shown that the different questions included under the same test are not of the same difficulty. The question, "What would you do if you were delayed in going to school?" was passed by practically none of the normal children of 12, 13 and 14. If this question is changed to Goddard's (28) interpretation, "What ought one to do if he is afraid he'll be late for school?", the test is easily within reach of the 12 year children. The difficulty in the first test is caused by the word "delayed". Changing the structure of the test changes its nature completely. In this connection it is to be regretted that Town (72) in the appendix of her translation of Binet's 1911 scale, has changed the wording of some of the tests from that in the actual body of the translation. For example, the question "What would you do before taking part in an important affair?" (page 47) is changed to "Before taking part in something very important, what would you do?" (page 78), and "Why is a bad action done when one is angry, more excusable than the same action when one is not angry?" (page 47), becomes "Why do we more easily pardon a bad act done in anger than a bad act done without anger?", (page 79). The meaning is the same but the wording different, and in many cases success or failure in a test depends on the interpretation of a single word. If an experimenter using Town's translation were allowed to select his questions from the actual translation or the appendix indiscriminately, variations would, in all probability, result. The general proposition that there is no such thing as an alternative question, i. e. a question involving the same mental processes and having the same difficulty as another, could very easily be maintained. To avoid this error experimenters should adhere strictly to one wording and should never be allowed to substitute one question for another.

An example of the influence due to differences of the tech-

nique of the experimenters in giving the tests is afforded by the test of detecting omissions in pictures. This test is a "six year" test for A and an "eight year" test for B. Differences in procedure make it very difficult if not impossible to compare the results of one investigator with those of another. To eliminate this error, very careful and minute instructions should be published for the giving of each test. No edition of the Binet-Simon scale is entirely satisfactory in this particular.

Examples of errors due to the observation of the experimenters are afforded by the tests of copying a diamond and defining in terms superior to use. Errors due to observation may be avoided or minimized by increasing the number of grades of response with which the particular response in question may be compared. This principle is followed by Yerkes (82) in the arrangement of the Point Scale. In the diamond test, for example, Yerkes allows three grades of response while Binet allows but two—plus or minus. The accuracy of any measure increases with the number of gradations on the measuring scale, and the significance of the error of observation is diminished by decreasing the chances of wide displacement. In the tests in which a definite question is put to the subject, uniformity of scoring may be obtained by an accurate and painstaking cataloguing, and a subsequent classification and weighting of all the responses of a large number of subjects to each question. If the responses to a free association test may be classified into a relatively small number of groups, then the responses to a restricted association test could be classified into a much smaller number of groups. A sufficiently large number of responses will include practically all possible responses. In this way the chances of the error due to observation are diminished, while the adoption of a point system of scoring will minimize the effect of any errors that might be made.

The differences between the experimenters in this study are large enough to demonstrate the influence of the personal equation. Scientific procedure demands that the investigator who studies the results of the individual tests for the purpose of analysing the factors involved or for obtaining age norms should

demonstrate that the effect of the personal equation is not present in the results treated. The burden of proof should be on the person who maintains that the influence is not present. Negative results concerning the influence of the personal equation that are based on the method of comparing the total scores or "mental ages" of different experimenters should not be taken as conclusive, inasmuch as the experimenters may deviate in one direction in one test, and in the opposite direction in another, so that in a total score these deviations might equalize. In a study of this sort made on the basis of "mental ages," which has previously been reported, the writer (14) found no deviations between B. C and D, while deviations between these three experimenters do appear in the more detailed study of the individual tests. Studies of the individual tests can have no claim to reliability unless the personal equation has been eliminated.

The importance of the personal equation as a source of error in making diagnoses on the basis of the "mental age" of the subject is universally recognized by psychologists and almost universally ignored by medical men, field workers, school teachers and others who have had no experience in making mental measurements. Among psychologists there are two opinions concerning the solution of the difficulty arising from this source, the first, that of making certain allowances for the inexpert examiners or establishing limits within which their opinions are valid, the second, that of removing the scale from their hands entirely.

Doll (22) in discussing criticisms of the Binet scale on the ground that diagnoses of normality and feeble-mindedness are made by inexpert examiners urges "that those who are capable of doing good Binet testing of the mechanical sort without being clinical psychologists should report the findings of their examinations of children or groups in tables of related chronological and mental ages and not in terms of normality or abnormality. In their reports they can say with a high degree of certainty that those children who show an intellectual retardation of more than 3 years are feeble-minded, but they should not say that those who test less than 3 years retarded are backward or normal. In

the lesser degrees of retardation only the expert is capable of evaluating the details of a Binet test with any finality as to either diagnosis or prognosis." (page 607).

Doll also points out that Binet examiners who have worked in institutions give very reliable diagnosis, for they intuitively sense distinctions which inexpert laymen do not see. When the responsibility for the diagnosis is placed on the examiner in this way, the scale is treated as a qualitative instrument. This standpoint is quite different from that in which certain allowances are made for all inexpert examiners and the quantitative character of the scale preserved. Goddard (31) in a study of the personal equation based on re-testings of normal and feeble-minded individuals fixes the quantitative limits somewhat higher. "In all cases where a child tests four or more years behind his age, there is little danger of error in considering him feeble-minded, even though the test was made by a person who was not highly expert, provided such a person is able to use the test with reasonable intelligence. With the borderline cases, those who are two or three years backward, the best expert should be employed in the testing." (pages 76-77).

As early as 1910, before the scale had received very extensive application, Huey (35) took the stand that inexpert examiners should not use the scale. In discussing this point he said, "I would urge that these Binet tests must be used with judgment and trained intelligence, or they will certainly bring themselves and their authors into undeserved disrepute.—Results can be considered valid only when the tests are made by an experienced psychologist who has familiarized himself with Binet's directions, or by other competent persons who apply the tests under the direction and supervision of such a psychologist." (page 444).

Three years later, in referring to the reports that the medical inspectors in Pittsburgh were to take over the Binet testing in the schools, Whipple (78) says, "And we can only express our hopes that these reports are unfounded, or that at least those in authority may be led to understand that for a person, whoever he may be, without extensive psychological training to attempt to diagnose the precise mental status of a school child is about as

absurd as for a mere psychologist to attempt to diagnose incipient tuberculosis or any other obscure pathological condition." (page 302). The same position is taken by Whipple (77) in another editorial. "We have no quarrel with the use of the scale in the public school: properly used, it is of direct and practical value; but improperly used, it will become a farce which can but bring discredit upon psychology and retard the movement for its application to educational practise." (Page 119).

In defense of this position, Whipple calls attention to an error inherent in the procedure of all inexpert examiners. "There is nothing about the conduct of the Binet-Simon tests that is intrinsically difficult, yet there is a source of error inherent in the use of any psychological procedure, which, as experience shows, is surmountable only by drill in psychological experimentation. I refer to the difficulty of following directions. No one who has drilled students in the laboratory has failed to be struck with the impossibility of laying down fool-proof directions for the conduct by an amateur of a psychological test." (Page 119).

Kuhlmann (43) agrees with Whipple in this position. "The untrained examiner meets difficulties because he lacks the following: (a) Familiarity with the directions for giving the tests. (b) Familiarity with the rules for interpreting the responses of the children. (c) Ability to adapt the procedure in testing in special instances for which directions can not be given. (d) Ability to interpret responses in special instances for which rules can not be given. (e) Ability to adapt himself in attitude to the mental levels of children of different ages so as to obtain the best efforts from the child in each case. (f) General appreciation of the absolute necessity of adhering strictly to all rules of testing, and of careful, painstaking work. These deficiencies are of quite different degrees of importance. The last is, on the whole, the most serious and most frequent, and can be remedied only by extended laboratory training." (Pp. 255 and 256). In regard to the quantitative allowance that must be made for inexpert examiners, Kuhlmann's article affords the following, "The amount of error made by an examiner because of his lack of training seldom equals two years in the mental age; in the majority of cases it is less than one year." (Page 256).

IV. GRADE CORRELATIONS

The correlation between intelligence, as measured by the Binet scale, and school performance, as measured by age and grade standing, has been worked out by various investigators. In all cases intelligence was measured by the "mental age" or total score of the Binet tests, and pedagogical age by assuming that all children begin school at a certain age and should therefore be in certain grades at certain ages. Stern (62) has reviewed the work of Goddard (30), Binet (4), and Bobertag (10) in this field, with the general conclusion that the correlation is only moderately high. The number of children showing mental advance is in excess of those showing pedagogical advance, but very rarely do children showing pedagogical retardation show mental advance. The correlation is one-sided in that "inference from school performance to mental ability is safer than from mental ability to school performance." (Page 61). Stern accounts for the discrepancies on the ground that "performance in the school depends not only upon intelligence, but also upon certain other and quite different factors." (Page 63). These factors are strength of memory which plays a large part in school performance but correlates only to a moderate degree with intelligence, and other factors that have nothing to do with intellect but belong largely in the domain of the will—"the degree and duration of attention, industry and conscientiousness, sense of duty and capacity to fit into the social group." (Page 63). Stern concludes that "the lack of agreement between tests of intelligence and school performance is really calculated to increase our confidence in the psychological test-methods," (Page 64) that absolute correlation is not to be desired since that would mean that the tests were testing school performance only, and that the measure of intellectual ability was the school performance itself, the tests being superfluous.

More recently, Schmitt (57) has reviewed the work of God-

dard, Terman and Childs (66) and Dougherty (23) in correlating intelligence, as measured by the Binet scale, and school performance, and reaches conclusions quite opposite to those of Stern. The following quotations from Schmitt's monograph explain her view point. "Further doubt is cast upon the accuracy of the tests by the fact that judgments arrived at through their application do not coincide with that of the school concerning the same subjects." (Page 57). Concerning this lack of correlation Schmitt writes "The Binet tests, therefore, while professing to test native ability are concerned very little with the education which all normal children have the native ability to acquire, and which is of much importance in civilized life." (Page 60). To the investigations cited Schmitt has added one of her own, in which the lack of correspondence between the Binet "mental age" and school grade is shown.

The writer is of the opinion that the method of correlating school performance with "mental age" fails to demonstrate either the adequacy of the Binet tests according to Stern, or the complete inadequacy of the tests according to Schmitt. For the demonstration of this point Schmitt's investigation may be discussed, inasmuch as it shows the most striking deviations between the measures of the two performances. Schmitt applied Binet's 1911 scale (Town's translations with modifications) to 150 children of superior social status. The following quotations indicate the status of the subjects. "The children who served as subjects for the tests comprised the Kindergarten and first six grades of a private school in Chicago." "They were the children of the professional class mainly. A few were children of successful business men who sought the best obtainable type of education for their children." (Page 2). The tests were applied at the close of an examination with the Healy-Fernald tests under rather unfavorable conditions as indicated by the following quotations,—"In the conduct of the two sets of tests the Binet-Simon tests were reserved for the last. By the time they were reached the child had been doing tests for an hour or more. In some cases there was too much restlessness and fatigue to

carry the child as far as the majority of his comrades in his grade were able to go and the tests were then discontinued." (Page 68 and 69).

The tests in the various age groups given to each grade were as follows,—Kindergarten, tests for V, VI, VII, VIII and IX years; Grade I, tests for V, VI, VII, VIII, IX, X and XII years; Grade II, tests for VI, VII, VIII, IX, X, and XII years. Grades III and IV, tests for VIII, IX, X and XII years; Grade V, tests for IX, X, XII and XV years; Grade VI, tests for XII and XV years. The "Adult" tests were also given to Grade VI as a class-room test.

Schmitt compared three measures, chronological age, school grade age and "mental age". The "mental age", in case a subject passed all tests in one group and failed one or more in a lower group, could be reckoned from two basal ages, these alternative rating being included by Schmitt. The summary of the results is as follows,— Comparing the Binet age to the chronological age, 14 (or 20)% are retarded, 26 (or 24)% are normal, and 58 (or 54)% are advanced. Comparing the school grade to the chronological age, (using 5 to 6.5 years as the normal age for the Kindergarten, 6.5 to 7.5 for Grade I etc.) 38% are retarded, 56% are normal and 4% are advanced. Comparing the Binet age to the school grade age, 2 (or 4)% are retarded, 25 (or 35)% are normal and 72 (or 60)% are advanced. The essential discrepancies are indicated by Schmitt by the following,— "Where the school grading shows 4% advanced over the normal for the chronological age, the Binet grading shows 58% over the chronological age and 72% over the age normal to the school grade." (Page 80.) The discrepancies thus indicated, although much larger than those of other investigators, agree with the general trend of results in that more children are shown to be advanced according to the Binet mental age than according to the school grade age. The results disagree with those of other investigators in finding a higher per cent. advanced by Binet age compared to chronological age.

The inadequacy of the methods employed in the investigations of Schmitt and others is seen when the measures are separately

studied. The use of the normal grade age as a measure of scholastic ability is false inasmuch as it rests on the assumption that all children enter school at a certain age, which is not the case. The measure of scholastic ability is the measure of the child's reaction to the subject matter of the grades, and that measure may be expressed only in the fact of promotion, non-promotion or (very rarely) double promotion, in other words, it may be expressed only in the relation of grade to the length of time in school. Furthermore, the two measures of scholastic ability, the age in grade method, and the grade progress method, are measures of an historically past performance not of present possibilities, and the true measure of an ability must indicate potential ability.

As measures of scholastic ability in terms of actual reaction, these measures present a distribution of general ability that is skewed toward the lower end, or in the direction of no ability. If a child enters school late, he presents a picture of retardation according to the age and grade method, while through any number of causes independent of intellectual ability, a child may present a retardation of at least a year according to either method. The possibilities for advancement are not as great, however, for advancement means forcing a child through a mass of subject matter, a process which the school is generally unwilling to undertake and the parent is generally unwilling to sanction. The school therefore presents a picture of ability in which promotion is normal, and non-promotion far more frequent than advance. If general ability is to be considered as distributed over any sort of a frequency surface, that surface will not take the form presented by the school measure in which the modal ability is almost completely the upper limit.

The measure of "mental age" has been shown to be one which varies from one chronological age to another in the form of its distribution. Normal children of 6 or 7 test over age, while those of 11 and 12 test under age. This abnormal distribution is due to two facts. In the first place, the tests in the younger years are too easy and those in the higher years are too difficult.

In the second place, the younger children have a wider range of tests beyond their average ability, so that exceptional subjects may display exceptional ability in a manner that is impossible if ability is measured by school progress, while older children have only a few tests within their range, the picture of advancement being excluded as in the measure of school ability. If the mental ages of a run of subjects of different chronological ages are combined, the frequency surface is normal, the error of the extremities balancing.

The investigators who have compared "mental age" with grade age, have compared two distributions, one of which is markedly skewed, the other normal, but false. The resulting finding of mental advance in excess of pedagogical advance has significance only insofar as it shows that a measure of general ability that will admit of exceptionally high performance is a better measure than one that precludes the possibility of such performance. The only significant finding is that pupils who show marked retardation in school rarely if ever show mental advance.

Applying the foregoing discussion to Schmitt's results in particular, all that has been said concerning the inadequacy of the age in grade method applies to her results. The age for entering school being 5, none of the subjects in the Kindergarten could be advanced, while those who entered late would be retarded. It is difficult to see how these young children would be able to make up their work in such a way as to show advance during the first two or three school years. The normal age for the sixth grade is from 11.5 to 12.5 years. Inasmuch as no grades were tested above VI, none of the 37 subjects from 11.5 to 14.5 could show an advance, and all of the 19 subjects from 12.5 to 14.5 would necessarily show retardation. Schmitt's results differ from those of other investigators in finding more subjects advanced according to Binet age in relation to chronological age. This deviation is probably due to the fact that she examined a superior selection of subjects, and to the fact that the XV year and "Adult" tests were used, so that the older subjects, who in general fall below their chronological age, had an opportunity to

better their scores. The discrepancy shown by Schmitt between school standing and the Binet tests does not demonstrate the inadequacy of the tests.

The final demonstration of a correlation between the Binet scale and school grade, rests not in comparing the total score or "mental age" with school grade, for that is susceptible to the errors of over-estimation and under-estimation according to varying chronological age, but in comparing the results of subjects in each grade on the individual tests. The tests may vary in their correlation with grade. Inasmuch as there is a general growth in age with grade, and a corresponding growth of intelligence with age, a test, in order to be an adequate test of intelligence, must show a correlation with grade. If the correlation is too high, however, the value of the individual test is in question for it would then be testing, not intelligence, but grade training. This criterion was actually used, though not stated, by Binet in his discussion of the results of Decroly and Degand (19), and in his revision of the 1908 scale, in which many of the tests that he considered to relate to school training were eliminated.

Studies of the individual tests in the light of school grade are not available. Decroly and Degand published in 1910 the results of an investigation on 45 children in a Brussels school, similar in character to that studied by Schmitt in Chicago. Binet discussed these results and those of other minor investigations in the Paris schools in considering the effect of environment on the results of the tests. Although he referred to school training as a factor, and classified the tests in which Decroly and Degand's subjects were superior, he gave no quantitative demonstration of the effect of this factor. The results of Decroly and Degand are based on too few subjects to admit of quantitative treatment. Chotzen (18) studied the tests by comparing the performance of feeble-minded individuals of the same mental age but of different chronological age. Although this method shows the effect of environment and maturity on feeble-minded individuals, it does not bear directly on the factor of school training. The foregoing

investigations will be discussed in this chapter only in their relation to the results of the particular tests. Schmitt, in her monograph, published tables showing the reaction of each subject in each grade to each test, the tables being discussed in the text. Although it was not Schmitt's purpose to determine the correlations between the various tests and grade, her data are available for a study of this sort, and the writer has taken the liberty of figuring them in this light, indicating at the same time Schmitt's interpretation of the grade factor, contained in the accompanying text. These data will be compared with the results of the Princeton investigation.

422 subjects of this investigation were distributed in the kindergarten, first six regular grades, minus grades and the special class of the Princeton Model School. 301 of the subjects (161 boys and 140 girls) were in the kindergarten and first six regular grades. The data obtained from the examination of these 301 subjects were classified according to the grade in which the subjects were found, and the percentage that the subjects of each grade passed each test was calculated.

Only those tests were studied which showed themselves to be free from the influence of the personal equation of the four experimenters. The elimination of the unrecorded results of the definitions test left a number of cases too small to be studied. To avoid the influence of the error due to incomplete data, the writer has calculated the percentage from only those tests that were given from 75% to 100% of the possible number of times. The data from the tests of repeating 5, 6 and 7 digits have been combined into one weighted measure. The procedure of the experimenters in giving these tests was to start within the subject's range and continue till he failed. If 5 digits were successfully repeated, 6 were given, and if these were passed, 7 were given. The results have been combined into one measure for the sake of simplicity, 1 point being allowed for the successful repetition of 5 digits, 2 points for 6 digits and 3 points for 7 digits, the weighting being roughly in accordance with the weighting in Goddard's scale, the tests being in the age groups

VIII, X and XII respectively. The measure of the ability of a group to repeat digits is the per cent. that the number of points scored is of the number of points possible (i.e. 6 times the number of subjects in the group).

The number of subjects in each grade (boys and girls shown separately) the average age of the subjects in each grade, together with the mean variation from the average are shown in Table 3.

TABLE 3

Number of Boys and Girls in Each Grade, and the Average Age of All Subjects in Each Grade.

Grade	Number of Boys	Number of Girls	Total No. of Subjects	Average Age	Mean Variation
Kindergarten....	20	12	32	5.64 years	0.46 years
Grade I.....	27	24	51	7.05 "	0.50 "
Grade II.....	16	24	40	8.16 "	0.65 "
Grade III.....	21	24	45	9.31 "	0.75 "
Grade IV.....	20	15	35	10.46 "	0.91 "
Grade V.....	24	25	49	11.71 "	0.99 "
Grade VI.....	33	16	49	12.81 "	1.06 "

The above table shows an increase of a year or more (actually from 1.10 years to 1.41 years) in the average age of the subjects in each grade. From this it is reasonable to expect that there is a general growth in intelligence correlating with this increase in age, or, in other words, to expect a correlation between the results of the individual tests and the grade in which the performance occurred. If the correlation is too high, it will indicate a dependence of that particular test on the subject matter of the grade. In Table 4 are shown the percentages that the subjects in each grade passed each test. The notes referred to in the margin contain the proportions passed for all other subjects for whom the percentages are not given, the percentages being given only for those groups to whom the tests were given from 75% to 100% of the possible number of times.

A study of Table 4 shows that the tests in general correlate with grade. The combined score of the test of repeating digits, for example, shows a growth from 6% to 78%, more rapid in the first three grades than in the last four. The tests vary in

TABLE 4

Percentage that Subjects in Each Grade Passed Each Test. 301 Subjects.
Grades

Test	K	I	II	III	IV	V	VI	
VII-1, 13 pennies	72	96	100	.				Note 1
VII-2, Pictures	69	96	94					Note 2
VII-4, Diamond	46	75	88					Note 3
VII-5, Colors	72	90	97					Note 1
VIII-2, 20 to 0		9	53	80				Note 4
VIII-4, Stamps		13	50	78				Note 5
All digits, (combined)....	6	21	42	51	55	78	75	
VIII-3, Days of week....	16	45	90	100				Note 6
IX-3, Date		5	35	96	100			Note 7
IX-4, Months			28	84	90			Note 8
X-I, Money			20	36	57	82		Note 9
X-2, Designs				21	37	42	66	Note 10
X-5, Sentence (2 ideas)...				67	89	88	98	Note 11
XI-2, Sentence (1 idea)...				22	46	51	74	Note 12
XI-3, 60 words					63	63	87	Note 13
XI-4, Rhymes					67	63	76	Note 14

Note 1. Counting 13 pennies and naming colors given 20 times above II. Not failed.

Note 2. Describing pictures given 21 times above II. Not failed.

Note 3. Copying diamond given 25 times above II. Not failed.

Note 4. Counting from 20 to 0 given 18 times in K. Not passed. Given 31 times above III. Failed once.

Note 5. Counting stamps given 15 times in K. Not passed. Given 35 times above III. Failed 3 times.

Note 6. Naming days of week. Given 32 times above III. Not failed.

Note 7. Giving day and date given 5 times in K. Not passed. Given 56 times above IV. Not failed.

Note 8. Naming months. Given 26 times below II. Passed twice. Given 44 times above IV. Failed twice.

Note 9. Naming money. Given 26 times below II. Passed 3 times. Given 28 times in VI. Failed twice.

Note 10. Copying designs given 33 times below III. Passed 5 times.

Note 11. Sentence (2 ideas) given 32 times below III. Passed 12 times.

Note 12. Sentence (1 idea) given 32 times below III. Passed 4 times.

Note 13. Giving 60 words given 53 times below IV. Passed 19 times.

Note 14. Giving rhymes given 42 times below IV. Passed 26 times.

the number of grades taken to reach their maximum. The test of naming the day and date, for example, is failed by all subjects in the kindergarten, 95% of Grade I and 65% of Grade II, while only 4% of the subjects in Grade III and none of those in the

higher grades fail it. A sudden increase occurs between Grades II and III showing possibly the influence of grade training. The tests vary considerably in the degree of their correlation. An easily obtained measure of the degree of correlation is that of comparing the magnitude of the increases from grade to grade. For example, there is an increase of 61% (96%—35%) from Grade II to Grade III in the ability to pass the test of giving the day and date, and an increase of 16% (36—20%) between the same grades in the test of naming the pieces of money. The former test correlates higher with the influence of grade in this particular case than the latter.

In this manner the percentage difference between the performance of the subjects in each grade and that of the subjects in the preceding grade was obtained. All the increases or decreases in ability from one grade to another were thus obtained, these values serving as measures of the amount of correlation between the tests and the grades. 42 differences between the performance of the subjects in any grade and those of the next succeeding grade were thus obtained. In 4 cases there were actual decreases of 1, 2, 3 and 4% which were not significant. The difference ranged from -4% to +61%, the median being +19.5% ($Q=16.25\%$). Some of the differences between the grades might be due to the chance superiority of a particular grade. To overcome this chance variation, and to furnish another index of the growth of the various abilities, the differences were calculated by steps of two grades, i.e., subtracting the performance of the kindergarten from the second grade, the first from the third, etc. In this way, 26 differences were obtained varying from +9% to +91%, the median being +29% ($Q=18\%$).

Some of the differences noted are undoubtedly high enough to warrant the assumption of the effect of grade training on the tests. Just what tests show this effect is probably a matter of opinion. Allowance must be made for the growth of an ability independent of training. 25% of the highest increases from one grade to another were selected as being worthy of special consideration at least. A larger increase must be allowed be-

tween two grades. Those differences were considered worthy of special consideration that exceeded twice the value of the median of the one-grade differences or 39%. This manner of selecting the largest differences is quite arbitrary, but is justified by the outcome, for the tests that show the most significant increases according to this method show those increases in more than one step, so that the evidence is concentrated against a very few tests. In this way the significant values outweigh the less significant values and fair allowance is made for growth from one grade to another.

The following list includes the tests showing the greatest increases by one-grade and two-grade steps, together with the magnitude of the increases and the grades between which they occur.

One-grade steps. 25% of largest increases.	Two-grade steps. Increases greater than 39%.
+61% Date, II to III	+91% Date, I to III
+56% Months, II to III	+74% Days, K to II
+45% Days, I to II	+71% 20 to 0, I to III
+44% 20 to 0, I to II	+65% Stamps, I to III
+37% Stamps, I to II	+65% Date, II to IV
+30% Date, I to II	+62% Months, II to IV
	+55% Days, I to III
+29% Diamond, K to I	
+29% Days, K to I	+46% Money, III to IV
+28% Stamps, II to III	+42% Diamond, K to II
+27% 20 to 0, II to III	
+27% Pictures, K to I	

The above lists of increases are confined to but 8 tests. In all, there were 16 tests studied. According to the method of selecting the significant increases, 20 such values actually appeared. In this manner the evidence combines against a very few tests. Some tests appear in both lists and more than once in the same list. The most striking growth with grade is shown in the tests of giving the day and date, naming the months, nam-

ing the days of the week, counting from 20 to 0 and counting stamps. The tests of copying the diamond, describing pictures and naming money may or may not show this influence. The evidence is strongest in the case of the diamond test since that appears in both lists.

The foregoing method of selecting those tests which correlate with grade to such an extent as to indicate the influence of grade training is not conclusive, owing to the fact that there is also an increase in age from grade to grade. If a test showed a very rapid growth with age, and those ages fell for the most part in certain grades, then those grades would show an increase which might be wrongly assumed to be due to training. The tests of counting from 20 to 0 is a case in point. Yerkes (82) in Table 32, page 125, gives the percentage values for each test in the Point Scale, for English speaking boys and girls of each age. The test, of the twenty one tests included, that shows the most marked increase with age is that of counting backward, the values being as follows,— age 4=0%; age 5=3.5%; age 6=23.7%; age 7=45.7%; age 8=72.2%; age 9=96%; the values for ages above 9 being 97% or higher.

The age in grade distribution of the 301 subjects in this investigation is given in Table 5.

TABLE 5
Distribution of Subjects in Each Grade according to Chronological Age.
Grades

Age	K	I	II	III	IV	V	VI	Total
4	4							4
5	17							17
6	11	28	2					41
7		18	17	2	1			38
8		4	15	18	1			38
9			5	13	11			29
10		1		10	14	18	1	44
11			1	2	3	16	16	38
12					5	8	11	24
13						4	12	16
14						2	5	7
15						1	3	4
16							1	1
Total	32	51	40	45	35	49	49	301

The rapid growth of the ability in counting from 20 to 0, according to the method of comparing the subjects in each grade, was from 9% in Grade I to 80% in Grade III. From Table 5 it may be seen that practically all, (89%), of the chronological ages in Grades I, II and III were distributed in the ages 6, 7, 8 and 9, a chronological range coinciding with that in which Yerkes' results show the ability to develop. The growth of this ability might be due then either to age or to grade. For this reason, to arrive at any final conclusion, it is necessary to compare the subjects of the same age but in different grades. The treatment of the Princeton results according to this method follows, but the analysis of the data in this manner can have no great reliability owing to the small number of subjects in each group. The number of subjects in each group, (boys and girls shown separately), the average age and mean variation from this average are shown in Table 6.

TABLE 6

Number of Boys and Girls of Similar Ages in Different Grades, and the Average Age of the Subjects of Similar Ages in Each Grade.

Grade	Age	Number of Boys	Number of Girls	Total no. of Subjects	Average Age	Mean Variation
Kindergarten.	5	11	6	17	5.48	0.20
Kindergarten.	6	8	3	11	6.26	0.21
Grade I	6	14	14	28	6.59	0.17
Grade I	7	9	9	18	7.36	0.22
Grade II	7	7	10	17	7.56	0.24
Grade II	8	6	9	15	8.39	0.24
Grade III ...	8	8	10	18	8.60	0.22
Grade III ...	9	5	8	13	9.43	0.16
Grade IV	9	5	6	11	9.65	0.13
Grade IV	10	10	4	14	10.39	0.30
Grade V	10	7	11	18	10.54	0.25
Grade V	11	10	6	16	11.54	0.22
Grade VI ...	11	10	6	16	11.53	0.26
Grade VI	12	6	5	11	12.52	0.14

All chronological ages were computed in tenths of a year, so that a variation in age from 0.1 yr. to 0.9 yr. is possible within

TABLE 7

Actual percentage that each test was passed by the subjects of each age in each grade. 223 subjects.

AGE	Age 5	Age 6	Age 7	Age 8	Age 9	Age 10	Age 11	Age 12
GRADE	K	K	I	II	III	IV	V	VI
NO. OF SUBJECTS	17	11	28	18	17	15	18	13
Counting 13 pennies	71	91	96	94	100	100	Note 1	
Describing pictures	65	82	96	100	100	86	Note 1	
Copying diamond	50*	45	64	89	93	92	Note 2	
Naming colors	71	82	89	100	100	100	Note 1	
Counting from 20 to 0		16	0	63	43	83*	Note 3	
Counting stamps		13	12	29	60	85*	Note 4	
Repeating digits, (all)	6	6	21	19	42	46	47	51
Naming days of week	0	40	50	94	80	100*	Note 5	65
Giving the day and date			29	29	20	88	100	100
Naming months			25	25	27	88	85	100
Naming pieces of money				31	35	31	55	64
Copying designs						31	31	27
3 words in sentence. 2 ideas			Note 10			89	62	91
3 words in sentence. 1 idea			Note 11			33	8	45
60 words in 3 minutes				Note 12				44
Giving rhymes				Note 13				92
								56
								69
								87
								100

*The test of copying the diamond was given 59% of the possible number of times in K-5, counting from 20 to 0 and giving days of week, 66% in III-8, and counting stamps 72% in III-8. All other percentages are based on tests given from 75 to 100% of the possible number of times.

- Note 1. Tests of counting 13 pennies, describing pictures and naming colors each given 12 times above II-8. No failures.
- Note 2. Copying diamond given 15 times above II-8. No failures.
- Note 3. Counting from 20 to 0 given 16 times below I-6. Not passed. Given 31 times above III-8. Failed 4 times.
- Note 4. Counting stamps given 14 times below I-6. Not passed. Given 32 times above III-8. Failed 4 times.
- Note 5. Giving days of week given 32 times above III-8. No failures.
- Note 6. Giving date given 39 times below II-7. Passed twice. Given 36 times above IV-10. No failures.
- Note 7. Naming months given 24 times below II-7. Passed twice. Given 37 times above IV-9. Failed 4 times.
- Note 8. Naming pieces of money given 35 times below II-8. Passed 4 times. Given 14 times above V-11. Failed twice.
- Note 9. Copying designs given 26 times below III-8. Passed 5 times. Given 15 times above V-11. Failed 6 times.
- Note 10. Three words in sentence, 2 ideas, given 24 times below III-8. Passed 9 times.
- Note 11. Sentence, 1 idea, given same as 2. Passed 3 times.
- Note 12. 60 words in 3 minutes given 41 times below IV-9. Passed 10 times.
- Note 13. Giving rhymes given 37 times below IV-10. Passed 25 times.

each age group. That the subjects of the "same" age but in different grades are not exactly the same is shown in Table 6. The subjects of each age in the higher grades average from 0.01 yr. to 0.33 yr. different, with an average superiority of 0.19 yr. This difference, however, is about one fourth that between the subjects of different ages in the same grades, and may be called the same for practical purposes. For convenience, the groups will be referred to as K-5, II-7 etc., the first member referring to the grade, the second to the age. K-5 would mean the group of 5 year children in the kindergarten, II-7, the 7 year subjects in Grade II, etc. The actual per cent, that the subjects in each group passed each test was calculated and is shown in Table 7. Unless otherwise noted, the percentages are based on tests given 75% to 100% of the possible number of times.

Some of the groups from which results were obtained are too small to have great reliability, but the method is at least suggestive. The results of 14 groups are given. It is possible then to compare the results of subjects of 6 ages, (6, 7, 8, 9, 10 and 11), that are in different grades, and also to compare sub-

jects in all seven grades that are of different ages, and in this way to determine whether the dominating factor in the growth of any ability is that of grade or age. The reliability of the method rests only on its connection with that of the first method employed.

In answer to the question of whether the growth of ability in the test of counting from 20 to 0 is due to age or grade, a question which was unanswered by the first method, we may turn to the results shown in Table 7 in which the subjects of each age in each grade are shown. The test of counting from 20 to 0 was not passed by any of the 5 and 6 year subjects in the kindergarten. Comparing first the subjects of different ages in the same grade, the 7 year subjects in Grade I are 16% lower than the 6 year subjects in that grade, and the 8 year subjects in Grade II are 20% lower than the 7 year subjects in the same grade, the older subjects making a lower record in each case. Comparing the performance of the subjects of the same age but in different grades, the 7 year subjects in Grade II are 63% ahead of the subjects of the same age in Grade I, while the 8 year subjects are 40%¹ ahead of the subjects of the same age in Grade II. Allowing for the retrogression of the older subjects in each group, i.e. assuming that they should have done equally as well as the younger subjects in the same grade, the groups in Grades II and III are still 47% and 20% ahead of the subjects in the grades lower. The growth of ability in this test would therefore appear to be due to grade training.

A rapid growth of ability in the test of counting stamps occurred between Grades I and III ($37\% \text{ I-II} + 28\% \text{ II-III} = 65\% \text{ I-III}$), according to the first method, so that the same question arises as in the test of counting from 20 to 0. The test was not passed below group I-6. No growth with age is shown between

¹ This test was given to but 66% of the subjects in III-8, the experimenters assuming that the other 34% would pass. The score given, 85%, therefore represents the ability of the lowest selection of III-8 subjects, or the most conservative estimate of the ability of the whole group. The same applies to the other tests in III-8 given 66% and 72% of the time. In this way the hypothesis that the tests are not influenced by grade training is given the benefit of the doubt.

I-6 and I-7, but a growth of 31% appears between II-7 and II-8. A growth with grade of 17% is shown from I-7 to II-7 and of 25% from II-8 to III-8. This test shows therefore the operation of the two factors of age and grade training.

The improvement in ability in the tests of counting 13 pennies, describing pictures and naming colors, that was indicated between the kindergarten and Grade I by the first method, would refer to age rather than grade, for a greater increase in each test is indicated between K-5 and K-6 than between K-6 and I-6. Above I-6 these abilities are completely developed. It could be maintained that these tests are so completely within the ability of the groups that the effect of training would not be indicated. The test that is best adapted to show the influence of any factor on a group is one that is well within the ability of the group—the influence of the factor will be obscured if the measure is either too easy or too difficult. The test of copying the diamond is a case in point and one well worth study, for it has been attributed to the effect of training by various authors. All the reproductions of the diamond had been scored according to the arbitrary system outlined in the previous discussion of the personal equation. A control on the factor of difficulty was obtained by raising or lowering the passing mark in this test. The percentage passed was calculated for each group for each of the 5 possible passing marks. The relations indicated in Table 7, where the passing mark is Group IV, were not changed by this process of raising or lowering the passing mark. In all cases the influence of age was shown between groups I-6 and I-7, and the influence of grade shown between groups K-6 and I-6. The test was given to but 59% of the K-5 group, the experimenter assuming that the other 41% would fail, so that the percentages calculated represent the performance of the best selection of K-5 subjects, or, in other words, the benefit of the doubt is given to the hypothesis that the test is influenced by grade training. If the other members of K-5 had failed according to the experimenter's assumption, (and this assumption was quite justified for some had failed to draw the square), 29% of the group would have passed instead of 50%.

The influence of age indicated in this test is as great if not greater than that due to training.

The test of repeating digits, scored by the weighting system previously described, exhibits a slow but uniform progress throughout, the older subjects in each group making records that are about the same or slightly lower than those of the younger subjects in the same grade, an increase showing fairly regularly from grade to grade. The most marked increase in this ability appears between K-6 and I-6, and between I-7 and II-7, possibly indicating that the lack of familiarity with the use of digits in the lowest grades interferes with this test as a measure of auditory memory.

The test of naming the days of the week shows the most marked improvement with age (40%) from K-5 to K-6, practically no improvement (10%), from K-6 to I-6, no improvement from I-6 to I-7, a very marked increase with grade from I-7 to II-7, a drop from II-7 to II-8, group III-8 marking the complete development of the ability. The test would appear to be due to the combined effect of age and grade. The tests of giving the day and date and naming the months are passed only twice in the kindergarten and first grade, by about a quarter of the subjects in II-7 and II-8 without age increase, while the subjects in III-8 shows a most marked increase due to grade. Above III-8 these tests are seldom failed. The test of naming the pieces of money shows a slow growth from 8 to 11, the largest increases appearing from III-9 to IV-9 and from IV-10 to V-10, improvement with grade in each case. Copying the designs from memory shows a growth of 26% from 8 to 11, the development occurring in two age steps, from IV-9 to IV-10 and from V-10 to V-11.

The growth with age cannot be determined in the tests of constructing sentences from three given words, because they were given to too few cases below the third grade. The results do not show whether III-8 is exceptionally high or III-9 exceptionally low. Both tests show decreases in ability from III-8 to III-9 and from V-10 to V-11. The ability in the easier test is well within the range of the third and higher grades, showing, therefore, no

improvement. The improvement in the second test develops from 33% to 80% in three steps, correlating with Grades IV, V and VI in each case. The most vital question, that of determining whether or not the language training in the third grade helps to make the construction of a sentence possible, cannot be determined owing to the lack of material in the second grade. The experimenters' assumptions in not trying the test would indicate this fact, but this is not experiment. The same lack of material makes conclusions in regard to the rhyming test impossible. The performance of IV-10 is exceeded only by VI-11. The test of naming 60 words in three minutes shows two decided increases with age and one decided drop with grade.

The foregoing analysis is based on a number of subjects in each group too small to have any great significance. The general fact of the correlation of the tests with grade remains, and conclusions concerning what tests correlate too highly with training can be answered only by considering both methods of study, and by considering only the largest deviations. The two most striking instances are found in the tests of naming the months and giving the date. These tests undoubtedly relate almost entirely to training. Less striking but equally definite is the relation of the test of counting from 20 to 0 to training. The tests of naming the days of the week and counting stamps show the influence of age to an extent almost as marked as that of grade, so that while the development in these tests is rapid, the grade factor probably exerts only part of the influence. Conclusions concerning the other tests are largely a matter of opinion, and the opinion of the writer has been indicated in the detailed discussion.

A study of the tests in relation to grade by the first method employed may be made from Schmitt's results. The author gives, in Table I, II, III, IV, V, VI and VII on pages 70, 71, 73, 74, 75, 76 and 77 of her monograph, the results of each subject in each grade on each test. From these tables the present writer has calculated the percentage passed in each test. A study of this sort rests for its reliability on the accuracy of the published tables, and the facts indicated by the tables do not always coincide

with Schmitt's discussion.² The writer has followed the tables rather than the discussion in calculating the results. In the VIII-2 test where an alternative rank is given for counting from 10 to 0 instead of 20 to 0, the writer has considered success in counting from 10 to 0 as a failure in counting from 20 to 0. In the line suggestion test Schmitt recognizes two types of failure, the typical failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. The second type of response Schmitt marks as passed, using a special symbol. The writer has calculated these percentages separately, entering the first or Binet type of response under "Line suggestion A" in the table, and the second type under "B." The V year and Adult tests were omitted. All of the other tests were included that had been given over 70% of the possible number of times. Unless otherwise noted, each test was given 100% of the possible number of times. Table 8 shows the per cent. that Schmitt's subjects in each grade passed each test in Binet's 1911 scale (Town's translation with modifications). The table is given with the reservation that the tables from which the percentages were calculated might contain misprints, and that the writer's interpretation of the tables might be at fault.

Inasmuch as there are many differences in procedure in giving the tests, and in the character of the schools tested, the results of the two investigations are not comparable in respect to the percentage passed in one grade in one study with those in the same grade in the other study. The method used in determining the

² In the discussion (page 69) Schmitt gives 15 subjects in the kindergarten failing test VII-4. Table I shows 13. On the same page she gives 24 subjects failing VIII-4. Table I shows 22 failing. In discussing the results of Grade I (page 72) Schmitt states that there is "more than 50% of failure with the discrimination of weight", while Table II shows 35% failure. Again, the tests referred to specific school instruction by Schmitt are VII-4, VIII-4, and IX 1, 2, 3 and 4. On page 72, in discussing the results of Grade I, she says "the tests below ten years which depend upon specific instruction are usually not passed except the VII-4 test. The percentages passed are as follows: VII-4 = 85%; VIII-4 = 45%; IX-1 = 35%; IX-2 = 75%; IX-3=90%; IX-4=30%. "Usually not passed" includes, therefore, tests passed 75% and 90% of the time.

TABLE 8

Per cent. that Schmitt's Subjects of Each Grade Passed Each Test. 150 Subjects.

	Grades						
	K	I	II	III	IV	V	VI
Number of subjects	25	20	17	21	22	22	23
VI-1, Distinguishing morning, afternoon	96	100*					
2, Defining in terms of use	92	94*					
3, Copying diamond	76	94*					
4, Counting 13 pennies	92	100*					
5, Choosing prettier of faces	92	100*					
VII-1, Showing right hand	92	80	100				
2, Describing pictures	72	65	81				
3, Executing 3 commissions	92	95	100				
4, Counting stamps	48	85	100				
5, Naming colors	96	100	100				
VIII-1, Comparing remembered objects	92	100	100	100	100		
2, Counting backwards from 20 to 0	40	85	94	95	100		
3, Indicating omissions in pictures	100	95	94	100	100		
4, Giving day and date	12	45	94	100	100		
5, Repeating 5 digits	64	85	94	100	100		
IX-1, Making change	6*	35	71	95	86	100	
2, Defining in terms superior to use	39*	75	65	100	95	100	
3, Naming pieces of money	28*	90	94	100	100	100	
4, Naming the months	6*	30	71	95	95	95	
5, Comprehending easy questions	61*	100	100	95	100	100	
X-1, Arranging 5 weights		65	41	57	50	64	
2, Copying designs		10	35	57	45	32	
3, Detecting absurdities		60	88	100	100	100	
4, Comprehending difficult questions		85	100	100	100	100	
5, Constructing sentence. Two ideas		65	76	100	100	100	
XII-1, Resisting suggestion, A. (Binet scoring)		64*	76	52	41	14*	100
B. Judgment error counted plus				100	86	100*	
2, Constructing sentence. One idea		57*	71	95	95	100*	100
3, Giving 60 words in three minutes		43*	82	62	100	95*	96
4, Defining abstract terms		7*	29	52	73	95*	100
5, Reconstructing dissected sentences		0*	6	10	23	81*	78
XV-1, Repeating 7 digits						62*	78
2, Rhyming words with "obey"						86*	70
3, Repeating a sentence of 26 syllables						10*	17
4, Interpreting pictures						14*	70
5, Solving problems from various facts						62*	70

Note.—All tests except those marked (*) were given all the possible number of times. The VI year tests were given 90% of the time in Grade I, the IX year tests 72% of the time in the kindergarten, the XII year tests 70% of the time in Grade I, and the XII and XV year tests 95% of the time in Grade V.

correlation of the tests with grade is the same as that used in the first method of treating the Princeton data, that of comparing the differences between grades by one-grade and two-grade steps, of selecting an arbitrary standard for detecting exceptional growth, and of comparing the resulting lists. The differences between the performance of each grade and the next succeeding grade were calculated. These differences, 100 in number, ranged from -24% to $+62\%$, the median being $+5\%$ ($Q=10.75\%$). The run of differences differs from that found in the Princeton study in two respects, in having a lower median and variability, and in containing more minus deviations. The lower median and variability is due to the fact that the tests were given over a wider range, the Princeton tests being given only on the "up slope" of the growth curve, or not being given when the tests were any distance above or below the probable range of ability of the group. The Princeton results showed only 4 minus deviations of 4, 3, 2, and 1% respectively, while Schmitt's results show 15 such deviations, 6 of them being 10% or over. These deviations are probably due to the smaller number of subjects, and if due to chance, should be counteracted by the precautionary measure of combining the indices of correlation into two-grade steps. 71 two-grade differences were obtained ranging from -25% to $+82\%$, the median being $+10\%$ ($Q=16.5\%$). 4 measures were still in the minus direction, one of these, -25% (Design III to V) is probably significant, the other values of -6% , -5% and -4% having no significance. Inasmuch as the variability of the series is lower, those differences were considered to be worthy of special study that had the value of $2Q+M$, or were in excess of the interquartile range plus the median. The lists of tests that appear as showing marked growth with grade according to the two methods are as follows:

One grade differences higher than 2Q+M	Two grade differences higher than 2Q+M
+62%, IX-3, Money, K to I	+82%, VIII-4, Date, K to II
+58%, XII-5, Dissected, IV to V	+71%, XII-5, Dissected, III to V
+49%, VIII-4, Date, I to II	+66%, IX-3, Money, K to II
+45%, VIII-2, 20 to 0, K to I	+65%, IX-4, Months, K to II
+41%, IX-4, Months, I to II	+65%, IX-4, Months, I to III
+39%, IX-5, Comprehension, K to I	+65%, IX-1, Change, K to II
+39%, XII-3, 60 words, I to II	+60%, IX-1, Change, I to III
+38%, XII-3, 60 words, III to IV	+55%, XII-5, Dissected, IV to VI
+37%, VII-4, Stamps, K to I	+55%, VIII-4, Date, I to III
+36%, IX-2, Definitions, K to I	+54%, VIII-2, 20 to 0, K to II
+36%, IX-1, Change, I to II	+52%, VII-4, Stamps, K to II
+35%, IX-2, Definitions, II to III	+47%, X-2, Design, I to III
+33%, VIII-4, Date, K to I	+45%, XII-4, Abstract Def., I to III
+29%, IX-1, Change, K-I	+44%, XII-4, Abstract Def., II to IV
+28%, X-3, Absurdities, I to II	+43%, XII-4, Abstract Def., III to V

A study of the above lists shows, as in the similar study of the Princeton data, that although the method of selecting the exceptional tests is an arbitrary one, the method is justified in practice, for only a few tests (13) appear in the lists as significant. In all, there were 34 tests³ studied, and 30 differences were considered large enough to be significant. These 30 differences were confined to 13 tests. The tests of naming 60 words and defining in terms of use drop out of the first list owing to the elimination of the errors of negative correlation. The design test is both positive and negative, the ability increasing from Grades I to III and decreasing after III. The test of defining abstract terms appears according to the second method because the ability increases with grade from 7% in I to 95% in V by

³ No differences were calculated from the line suggestion test owing to the possibility of misinterpreting the symbols. Schmitt notes the difference in the character of the responses from the suggestion error to the judgment error in passing from Grade II to III. The scoring of the suggestion error in the tables shows an inverse correlation with Grades II, III, IV and V, and a sudden change again from 14% in Grade V to 100% in Grade VI, so that there is probably a mistake. The scoring of the responses to this test according to the strict Binet ruling would make the "mental ages" lower, for many cases would then have basal X.

increases of approximately 25% in each grade. No conclusions may be drawn concerning the easy comprehension test and the absurdities test. The 20 remaining differences are confined to 7 tests, those of naming the day and date, naming the months, counting from 20 to 0, counting stamps, naming money, reconstructing dissected sentences, and making change. The first four were included in the five found to show the most marked influence of grade in the Princeton study. The test of naming the pieces of money did not show a marked relation to grade in the latter study, but this difference might be one of school curriculum. The test of naming the days of the week is not included in Binet's 1911 scale.

In the Princeton study alternatives were used in the making change question so that no data from this test were included in the quantitative study. These data show the ability in this test developing in the second and third grades, the test being passed only twice in the kindergarten and first grades, and generally passed above the third. The data in the test of reconstructing dissected sentences show very few passing the test below grade V with approximately three fourths passing in V and VI. In so far as the Trenton experimenting was applied to a few subjects in the regular grades below the seventh, this test was rarely passed in the third and fourth grade, passed about 5% in V, and almost universally passed in VI, VII and VIII. The number of subjects in each grade is small in the Trenton experiment, but each test was separately scored, i.e. each part of the dissected sentence test, each part of the absurdity test etc. Each of the three parts of the dissected sentence test showed the same growth between the same grades, and this growth was more marked than that in any other test. The evidence concerning these two tests, therefore, supports the evidence from Schmitt's results.

The quantitative analysis of the Princeton data and Schmitt's data would indicate that the tests of counting stamps, counting from 20 to 0, naming the days of the week, giving the day and date, naming the months, naming the pieces of money, making change and reconstructing dissected sentences were influenced to a considerable extent by grade training. The performance in

certain of these tests (days, date and months) may be the result of specific school training in the tests themselves, while others (perhaps the tests of counting stamps, counting from 20 to 0, and reconstructing dissected sentences) may involve a transfer effect in the application of the content of the grade in a new way. The fact that the tests correlate very highly with grade training does not show that the tests are worthless, but it does show that they should, perhaps, be placed in another scale, or should at least be placed on a different footing than those that test capacity irrespective of attainments.

One of the best tests⁴ of intelligence is the determination of what an individual can do with the training he has received, but tests of this sort rest on the assumption that the individual's opportunities have been determined. The importance of tests of information in cases of alienation presenting a picture of deterioration is recognized. The important change to be made is not the elimination of such tests from intelligence scales, but their standardization on a different basis. The diagnostic value of such tests rests not in the mechanical memorizing of a time series such as that of the months, but in the ability to apply such a series. In pointing out this fact Katzenellenbogen (37) suggests that the months test be given in some such manner as "If somebody asks you in November to return three months later, what month would it be?" Decroly and Degand also suggest that the mechanical tests of counting and naming the days of the week and months be modified in some such manner.

⁴ The writer recalls two cases in which the failure in tests which involved the application of training was very significant. The first was that of a woman of about 30, a parole patient in a hospital for the insane, who had never shown any marked symptoms other than a history of intellectual inferiority. This patient passed practically all of the Binet tests in the IX, X and XII year groups, but failed completely in the test of making change. This observation was later checked up. Another case of a woman of 22, in the same hospital, presented a border-line psychoneurotic picture perhaps, but no marked symptoms other than a history of intellectual inferiority. She passed in a great many of the difficult tests in the upper years but had great difficulty in telling time. Both cases had lived under very good home conditions and had mingled with people of ability. A great many tests of capacity were given, but the most illuminating evidence of their mental status came from the two tests mentioned.

Comparing the conclusions of this study with other investigations, the agreement is fairly close. Schmitt's results do not support her suggestion that the definitions test relates to specific school instruction. The other tests which she refers to this factor (stamps, date, 20 to 0, change, months and money) show the influence to a marked extent. Binet in classifying some of the tests referred the tests of copying a sentence, reading for memories, writing from dictation, copying a diamond, counting backwards and making change to scholastic training. The first three tests were not included in this investigation. The diamond test showed the influence of age to be as great if not greater than that of school training. The last two tests showed a marked influence of training. Binet referred the tests of counting 13 pennies, naming four colors, naming the days of the week and enumerating the months to home training. The last two showed a marked influence of school training. The results of the present investigation agree with those of Chotzen in finding no effect or very little effect of training in the tests of copying the diamond, repeating digits, describing pictures, counting 13 pennies, naming colors, comparing remembered objects, defining in terms of use and superior to use, and in finding marked influence of this factor in the test of naming the days of the week.

The methods used in analysing the results, especially the second method, reveal several suggestive relations between the tests and the school grades. There is a general correlation between the tests and the grades, a correlation that is very necessary to establish, for there is also a general correlation between intelligence and grade. In analysing the results of the individual tests by comparing the results of subjects of the same age in different grades, and of subjects of different ages in the same grade (Table 7), it was seen that, as a general rule, the growth in any particular ability occurred in passing from grade to grade, not in passing from age to age within one grade. In fact in only half of the cases in which the subjects of two ages in one grade may be compared do the older subjects make records that are higher than those of the younger ones, and only 10% of these gains are over 20%. If the groups were considered to be equal in all

cases in which their records were within 10% of each other, equality occurs in exactly 50% of the cases. Of the remainder, 20% of the groups were lower, while in only 30% of the cases are the older subjects actually higher than the younger subjects of the same grade. Some of the cases of retrogression could well be accidental, but they occur too frequently to be due entirely to chance.

Applying the same general method to the cases in which groups of the same age but in different grades were compared, 5% of the groups in a higher grade showed lower scores, the results correspond in 43% of the cases, while 52% showed definite improvement. This might indicate that there is a higher correlation between the tests and grade than between the tests and age. The fact that the comparison of children of different ages in the same grade showed the older children making lower records in 20% of the cases, equal records in 50% of the cases and higher records in only 30%, would confirm the general diagnostic value of the tests if Bonser's interpretation of this phenomena is correct. Bonser (12) applied various sorts of reasoning tests to children in the fourth, fifth and sixth school grades. In summarizing the results of the tests in the different grades, he says, "In the contrast with grade progress and progress with age, in the generally superior showing made by the younger groups of children of any grade when contrasted with the older pupils of the grade, and in the fairly substantial percentage of pupils from lower grades found in the highest quartile of ability for all, it is shown that native capacity is measured to a high degree by the tests."

In conclusion, the results shown in this chapter would indicate a correlation between the individual tests studied and the school grades, this correlation being high enough in some cases to show the actual effect of training. In answer to the general objection that since one demonstration of the accuracy of the tests rests on their correlation with school grades, the school grades are the real measure of intelligence and the mental tests superfluous, it is only necessary to point out that intelligence tests, besides affording the opportunity for accurate standardization,

also detect the subject's potential abilities independent of his past performance. The school measure indicates mental defect in cases of gross retardation, but it does not indicate exceptional ability.

Schmitt's contention that the school represents a standard environmental situation, and a measure of a subject's ability should include a measure of the adequacy of his reaction to this situation, is well founded. It is not, however, a criticism of the Binet scale, for the scale aims to test native capacity. At the Buffalo conference (15) on the Binet scale, the following question was raised,—“What is it, after all, that the scale aims to test?” The question was answered by “We believe that current misconceptions as to the aim of the scale should be removed. It is not intended to test the emotional or volitional nature, but primarily intelligence (judgment).” To this list might be added the assertion that the scale was not intended to test a child's reaction to the school situation, or to furnish an outline for taking a record of his life history.

Rogers and McIntyre (54) would also have mental tests include tests dependent on both school and home training. This general trend of present day discussion is a reversion to Binet's 1908 type of scale, a tendency to which Binet was in opposition. The probable solution rests in eliminating from the scale the tests involving training, and in constructing a standardized scale of another sort for the estimation of the individual's reaction to the school situation in terms of the length of time that he has met that situation. That such a scale is not a matter of speculation is shown by the number of scales now on the market for measuring handwriting, spelling, composition, arithmetical ability, etc. Tests of native capacity and tests dependent on school and environmental training cannot be standardized on the same basis, for they are essentially different measures. Measures of the first sort may perhaps be correlated with age, while measures of the other sort can be correlated only with opportunity.

V. SEX DIFFERENCES

The investigators who have studied the influence of sex differences on the Binet-Simon tests have used two methods, that of comparing the "mental ages" or total scores of subjects of each sex, and that of comparing the per cent. that the subjects of each sex pass each test. The first method throws no light on the individual tests, inasmuch as one sex may be superior in one test and inferior in another so that the total score will balance the influence of this factor. Inasmuch as the scale is founded on the principle that sex differences do not exist, it is important to study the individual tests, and to determine the accuracy of this assumption.

The Princeton data are available for a study of this sort. 352 subjects (187 boys and 165 girls) between the ages 6 and 12 were examined. The method of study adopted was that of comparing the results of non-selected boys and girls of each age, and, as a check on this method, of comparing the results of selected boys and girls of four ages.

Inasmuch as the subjects of each chronological age are distributed over a range of one year (the 6 year subjects for example being distributed from 6.0 to 6.9), the actual average age of the subjects of each age was computed to make sure that no differences might appear due to the chance selection of subjects at either extreme. These averages are shown in Table 9.

TABLE 9
Actual Average Chronological Age of Boys and Girls in Each Age Group.

	BOYS		GIRLS	
	Number of Subjects	Average Age (M. V.)	Number of Subjects	Average Age (M. V.)
Age 6	37	6.58 (0.20)	23	6.51 (0.20)
Age 7	29	7.50 (0.29)	31	7.39 (0.26)
Age 8	24	8.48 (0.29)	28	8.48 (0.22)
Age 9	20	9.46 (0.27)	22	9.54 (0.26)
Age 10	31	10.46 (0.25)	23	10.37 (0.30)
Age 11	28	11.59 (0.22)	20	11.52 (0.27)
Age 12	18	12.43 (0.30)	18	12.57 (0.24)

A perusal of this table shows that the subjects agree closely both in their average and in their variability. The 12 year boys are actually 0.14 yr. younger than the girls of the same age group. The 7 year boys are 0.11 yr. older than the 7 year girls. All other differences are less than 0.10 yr. The correspondence is close enough for all practical purposes, but these differences must be taken into consideration before drawing final conclusions.

The 352 non-selected subjects from 6 to 12 were distributed throughout the kindergarten, special class, and first six minus and plus grades as shown in Table 10.

TABLE 10
Age in Grade Distribution of 187 Boys and 165 Girls, 6 to 12 Years of Age.

Age	6		7		8		9		10		11		12		Totals
Sex	B	G	B	G	B	G	B	G	B	G	B	G	B	G	
Special Class	2		1		3		3				3		1		13
Kindergarten	8	3													11
Grade I-	13	4	8	9	1	1	1								37
Grade I	14	14	9	9	3	1			1						51
Grade II-			2	2	2	4		2							12
Grade II		2	7	10	6	9	2	3			1				40
Grade III-					3		4	2	1	1		1			12
Grade III			2		8	10	5	8	5	5	1	1			45
Grade IV-							1		5	1	2	2	3	1	15
Grade IV			1		1		5	6	10	4	1	2	3	2	35
Grade V-									1	1			3	3	9
Grade V									7	11	10	6	2	6	42
Grade VI-											1			1	2
Grade VI									1		10	6	6	5	28
Totals	37	23	29	31	24	28	20	22	31	23	28	20	18	18	352

It is generally conceded that a difference exists in the reactions of the sexes to the school curriculum, the girls in the long run making better progress in school work than the boys. A study of Table 10 shows that in general the girls have a slightly higher distribution than the boys, these relations being more clearly indicated in Table 11 in which the average grade of the subjects of each age and sex is shown. In computing the average grade, the kindergarten was counted 0; Grade I—, 0.5; Grade I+, 1.0; Grade II—, 1.5; etc. Each subject in the special class was assigned a grade 0.5 lower than the lowest subject of his age (0 being the smallest value given), on the theory that each subject in

the special class was less satisfactory than any of his comrades in the regular class. The fact that there were no girls in the special class would cause an unduly exaggerated difference between the average grades of the boys and girls. For this reason, the average grades of the boys, including and excluding the special class cases, were separately figured, these values being separately shown in Table 11 under Boys A (the average grade including the special class cases), and Boys B (the average grade excluding the classes). Had the special class subjects been in the regular grades, they would have lowered the average of each group, so that the two values may be taken only as limits, the values under "Boys A" being the lower limit, and those under "Boys B," the upper limit.

TABLE 11
Actual Average Grade of Boys and Girls in Each Age Group.

	BOYS A		BOYS B		GIRLS	
	No.	Average Age (M. V.)	No.	Average Age (M. V.)	No.	Average Age (M. V.)
Age 6	37	0.55 (0.34)	35	0.59 (0.33)	23	0.87 (0.35)
Age 7	29	1.24 (0.64)	28	1.29 (0.63)	31	1.31 (0.65)
Age 8	24	1.94 (0.91)	21	2.21 (0.65)	28	2.25 (0.59)
Age 9	20	2.48 (1.04)	17	2.91 (0.69)	22	2.98 (0.62)
Age 10	31	3.92 (0.71)	31	3.92 (0.71)	23	4.19 (0.80)
Age 11	28	4.66 (1.20)	25	5.04 (0.77)	20	4.83 (0.88)
Age 12	18	4.72 (0.91)	17	4.82 (0.88)	18	5.03 (0.59)

Table 11 shows that the scholastic ability of the girls as indicated by the average grade is uniformly higher than that indicated by the lower limit of the boys, and is below the upper limit of the boys in only one case (at 11 years). A slight sex difference in school work in favor of the girls may therefore be assumed at the outset. It is significant that the upper limit of the 11 year boys is higher than that of the 12 year boys, and that the lower limits show a difference of but 0.06. This would indicate a poor selection of 12 year boys, or a superior selection of 11 year boys. Both measures of the scholastic ability of the boys show a generally higher variability than that of the girls.

From Table 9 it may be seen that the growth in the actual average age of each sex is not uniform from year to year, the minimum increase for boys being 0.84 yr. (from 11 to 12), and

for girls 0.83 yr. (from 9 to 10), while the maximum increase for boys is 1.13 yr. (from 10 to 11), and for girls 1.15 yr. (from 10 to 11). A more marked lack of regularity in the growth of scholastic ability from year to year as measured by the average grade is shown in Table 11, no increase being shown by the 12 year boys over the 11 year boys, while the 10 year boys show an increase of 1.44 to 1.01 grades over the 9 year boys. In the same way the 10 year girls show an increase over the 9 year girls that is nearly three times that of the 7 year girls over the 6 year girls, while the increase of the 7 year girls over the 6 year girls is twice that of the 12 year girls over the 11 year girls. These relations indicate that the selection of subjects is not uniform at each age. The subjects of any one age may be either a superior or inferior selection of all children of that age, and there is no reason for supposing that this random sample of superior or inferior subjects of any age will correspond to a similar sampling of the subjects of the opposite sex of the same age.

The process of calculating the percentage that the boys and girls of each age pass each test is extremely simple, but the conclusion, that the differences found between the percentage passed by the sexes at each age may be attributed to sex differences, is not justified unless all the variable factors are known.

A previous chapter showed variations in the tests due to the influence of the personal equation of the experimenters. To avoid this variable influence, only those tests were studied that showed that they were free from the influence of this factor. Inasmuch as each experimenter examined approximately the same number of boys and girls of each age, any influence of this factor would be equalized, provided, of course, that there were no differences in the reaction of the experimenters to the two sexes. In the detailed study of the design test, it was found that experimenter C was more lenient in marking girls than boys. The possibility of a similar interpretation in a few other tests was suggested, but not demonstrated. In analysing the results for sex differences, however, the possibility of such an interpretation must be kept in mind.

Another possible source of error is that due to incomplete data.

The experimenters, in giving the tests, would give only those within the approximate range of the subject, so that each test would be given to a superior selection of children below the normal range of the test, and to an inferior selection of subjects above this range, a process tending to make the apparent growth of an ability less than the probable real growth. In comparing the results of the sexes, however, it is not necessary to have accurate results on the growth of an ability, but results which have the same determining factors. If the experimenters gave the test to approximately the same proportions of boys and girls at each age, a comparison of the percentage passed is legitimate, even if a small proportion of the whole group were actually tested, for the proportion would include the same selection of subjects. The number of boys and girls at each age, and the percentage that each test was given to these subjects are shown in Table 12. The test of counting 13 pennies, for example, was given 37 times to 6 year boys, or 100% of the possible number of times, while the test of counting from 20 to 0 was given 27 times to the same group, or 73% of the possible number of times. Column A shows the total number of times each test was given to all of the boys and girls. Column B gives the average age of all the boys and girls to whom each test was given. The average given in this case is not the actual average derived from the actual chronological age of each subject figured in tenths, but the weighted¹ average, the whole numbers 6, 7, 8, 9, 10, 11, and 12 being used.

Table 12 shows a very close correspondence between the percentage that each test was given to boys and girls of each age, so that the error due to incomplete data, though present, is present to the same extent in the results of both sexes, and may be disregarded. A fairly close correspondence in the average age of all the boys and girls to whom each test was given is also indicated in Table 12. In the test of counting stamps there is an

¹ For example, in the test of counting 13 pennies, the average age of the boys to whom the test was given is,—

$$\frac{(37 \times 6) + (28 \times 7) + (16 \times 8) + (8 \times 9) + (7 \times 10) + (3 \times 11) + (1 \times 12)}{100} = 7.33 \text{ years}$$

TABLE 12

Percentage that Each Test Was Given to Boys and Girls of Each Age, the Total Number of Times Each Test Was Given to Each Sex and the Average Age of All Subjects of Each Sex to Whom Each Test Was Given.

Chronological age		6	7	8	9	10	11	12	A Total number of times given	B Average age of subjects. (weighted)
Number of subjects	Boys	37	29	24	20	31	28	18		
Number of subjects	Girls	23	31	28	22	23	20	18		
Counting 13 pennies	Boys	100	97	67	40	23	11	6	100	7.33
	Girls	100	94	68	41	26	10	11	90	7.56
Describing pictures	Boys	100	90	67	45	26	11	6	100	7.38
	Girls	100	94	68	41	30	20	11	93	7.66
Copying diamond	Boys	100	93	63	60	32	14	17	108	7.30
	Girls	100	94	61	64	35	20	11	97	7.74
Naming colors	Boys	100	93	67	45	23	11	6	100	7.35
	Girls	100	94	68	41	30	15	11	92	7.62
Counting from 20 to 0	Boys	73	97	83	80	61	21	44	124	8.18
	Girls	74	71	79	77	52	35	28	102	8.25
Counting stamps	Boys	65	97	88	80	61	21	44	122	8.23
	Girls	83	87	79	82	52	35	39	112	8.23
Repeating all digits	Boys	95	100	100	100	100	100	100	185	8.75
	Girls	96	97	96	100	100	100	100	162	8.78
Naming days of week	Boys	92	100	83	80	61	21	44	132	8.04
	Girls	96	90	82	82	52	35	28	115	8.10
Giving day and date	Boys	43	76	88	95	84	64	89	138	8.10
	Girls	78	77	93	100	78	75	72	136	8.70
Naming the months	Boys	41	79	79	95	81	54	78	130	8.90
	Girls	39	65	93	100	70	65	61	117	8.84
Naming money	Boys	27	62	67	90	97	64	78	124	9.21
	Girls	43	39	86	95	100	80	67	118	9.11
Copying designs	Boys	16	31	67	85	94	79	78	113	9.56
	Girls	26	19	57	86	96	80	67	97	9.46
3 words in sentence	Boys	8	31	63	90	100	93	100	120	9.79
	Girls	26	26	68	86	100	95	94	111	9.36
60 words in 3 minutes	Boys	11	21	38	70	81	93	89	100	9.92
	Girls	30	10	32	50	74	90	78	79	9.75
Giving rhymes	Boys	8	21	25	50	74	89	94	90	10.08
	Girls	13	13	36	45	74	90	83	77	9.92
Defining "fork" etc.	Boys	38	62	50	55	48	25	17	80	8.35
	Girls	61	65	61	68	39	20	11	81	8.09

actual correspondence. The greatest difference is that of 0.6 yr. in the test of giving the date. The differences, on the whole, are small, but must be taken into consideration when comparing the percentages that all boys and girls pass each test.

Two methods are available for studying the influence of sex differences on the individual tests. The first is that of comparing the results of boys and girls of each age on each test. This method is affected by the chance selection of superior or inferior subjects, and the results can have no meaning unless the relations of the groups of each age of the same sex are understood. For example, the fact that the 12 year boys are 36% lower than the 12 year girls in the test of naming the months has no significance as an isolated finding, for its significance is modified by the additional fact that this group of 12 year boys is 10% lower than the 9 year boys, 12% lower than the 10 year boys, and 9% lower than the 11 year boys on the same test.

The other method is that of comparing the per cent. that all subjects of each sex pass each test. This method avoids the factor of variations in the results due to a chance superiority of one age group over the other of the opposite sex, but, at the same time, it tends to obscure the magnitude of the differences that might occur. The most reliable differential measure between two groups is one that is well within the range of ability of the groups. The difference will be obscured if the measure is too easy or too difficult. A comparison of the results of all subjects would, in this way, tend to minimize² the magnitude of the real difference between the groups. Furthermore, there is a possibility that one sex might acquire an ability first, but eventually be surpassed by the other. The per cent. that all subjects passed would show no deviation, because the two tendencies would balance.

² For example, if there were 20 subjects of each age and of each sex from 6 to 12, and a certain test were passed by 75% of the 6 year girls, and by all of the 7, 8, 9, 10, 11 and 12 year girls, by 50% of the 6 year boys, 75% of the 7 year boys and all of the remaining groups, the total percentage passed for all girls would be 96%, and for all boys, 89%. The differential character of the test is indicated by the value 7%, while its actual differential character, just within the range of ability of the groups, is 25%.

Neither method, then, is entirely satisfactory, the first because it would tend to exaggerate chance differences, the second because it would tend to obscure real differences. The method used in this study is that of comparing the results of non-selected and selected subjects of each age and sex, studying first the general growth of each ability from age to age within each sex, and using the per cent. that all subjects pass each test to determine the correlation between the results of non-selected and selected subjects.

Table 13 shows the percentage of proportion³ that the boys and girls of each age pass each test, the percentage that all boys and girls pass each test, the actual percentage that the boys are superior to (+) or inferior to (—) the girls of each age, the difference between the average age of all boys and girls to whom each test was given, and the difference between the percentage that all boys and girls pass each test.

The differences between the performance of the boys and girls at each age have no meaning unless the general growth of the abilities in each sex is first understood. Studying first the results of the 187 non-selected boys shown in the first seven columns of Table 13, it may be seen that the growth of ability in each test is rather irregular. The test of naming the months, for example, shows a slight decrease from 9 to 12. The differences between the percentage performances of the subjects of each age and those of the preceding age were calculated. The 12 year group, compared to the 11 year group, is +11% on the test of giving the date, —9% on the test of naming the months etc. 61 differences were thus obtained, varying in magnitude from —15% to +36%, the median being +8% ($Q=9.75\%$). 13 of the deviations (21%) were minus values. The largest negative deviations occurred in the tests of naming colors (—15%, 7 to 8), naming money (—15%, 11 to 12), and constructing a sentence containing two ideas (—13%, 8 to 9). The remaining 10 minus deviations were less than 10%.

³ The proportion given is the number of times a test was given over the number of times a test was passed. No percentages were calculated for tests given less than 12 times, and no percentages are given for the definitions tests on account of the small number of times they are given to all subjects.

An index of the growth from year to year was obtained by calculating the average percentage increase from one age group to another. For example, the 7 year boys were 26% higher than the 6 year boys in the test of naming colors, 5% higher in naming the date etc. The average of the 10 possible comparisons between 6 and 7 year boys shows that the latter averaged 16.1% higher than the former. The average increases in percentage passed from year to year are as follows,—6 to 7=16.1%; 7 to 8=13.5% 8 to 9=8.7%; 9 to 10=11.2%; 10 to 11=6.0%; and 11 to 12=0.2%. These figures show strikingly the irregularity of the growth from age to age. Comparing these average percentage increases in tests with the averages shown in Tables 9 and 11, there is no observable relation between this increase and the increase in average age from age to age, or the increase in average grade from age to age. The smallest increase in the tests (0.2%, 11 to 12) coincides with the smallest increase in average age from year to year (0.84 yr.), and the smallest increase in average grade from year to year. The other relations are varied.

The fact of the variability in the results of the non-selected boys stands out. The irregularity of the growth of the various abilities, and the fact that in 21% of the cases the boys of one age are actually lower than those of the previous age, point to the conclusion that certain allowances will have to be made for chance variations. It is not possible to account for the variations in growth by reference to the relative increase in average age or average grade from year to year.

The results of the 165 non-selected girls, shown in italics in the first seven columns of table 13, were studied in the same manner as the results of the boys. 60 differences between the percentage performance of the girls of each age and those of the preceding age were obtained. These differences ranged from -33% to +50%; the median being 7% ($Q=8\%$). 10 of the deviations (17%), were minus values. The largest deviations were shown in the tests of naming 60 words, (-33%, 11 to 12), counting stamps (-20%, 9 to 10), and drawing designs

(—14%, 8 to 9). The remaining 7 minus deviations were below 10%.

The average increases in the percentage passed from year to year are as follows,— 6 to 7=3.9%; 7 to 8=15%; 8 to 9=8.8%; 9 to 10=10.1%; 10 to 11=8.7%; 11 to 12=1.8%. Both boys and girls show the smallest average increase in the percentage passed in the step from 11 to 12, and the magnitudes of the increases agree fairly well except for the step from 6 to 7. The increase of the 7 year girls over the 6 year girls is 3.9%, the next to the smallest increase of one age group over any preceding group. The 7 year boys, however, show an average increase of 16.1%, over the 6 year boys, the largest increase of any group of boys over any preceding group. It will be difficult, then, to draw conclusions concerning sex differences from a comparison of the 6 year boys and girls, for the 6 year girls are either a superior selection or the 6 year boys are an inferior selection if the character of these groups be judged by the comparison with the 7 year subjects. The same comparison, on the other hand, might indicate that the 7 year girls were an inferior selection and the 7 year boys a superior selection from the general run. It is only possible to point out the irregularity, however, it is not possible to show the cause of the irregularity.

A comparison of the average increase in the percentage passed by girls from age to age with the increase in the average ages shown in Table 9 shows no demonstrable relation to exist. Comparing this growth in the ability on the tests with the growth in average grade, shown in Table 11, shows a very positive relation to exist between these factors. Where the increase in average grade is smallest (i.e. from 6 to 7 and from 11 to 12), the increase in the tests is smallest (3.9% and 1.8%), while the greatest increase in grade (from 9 to 10 and from 7 to 8) coincide with the greatest increase in the test abilities (10.1% and 15.0%). This relation was not indicated in the results of the boys. The explanation of this fact that a correlation between the increase in the tests with grade was found in the results of the girls but not of the boys is a matter of speculation. It has been shown that the boys have a higher variability in grade than

girls. This tendency of the boys to be distributed in a wider range of grades might nullify the grade correlation slightly, but probably not to any considerable extent. The fact that the causes of this variation are not determined serves to illustrate the dangers of comparing the results of two groups when the factors operating on the groups are not known.

The foregoing study of the growth of the various abilities from age to age in each sex, and the analysis of the causes influencing this growth, demonstrates the great variability of the results. This fact of variability must be considered before drawing conclusions concerning sex differences by the method of comparing the results of boys and girls of each age.

The percentage differences between the performance of non-selected boys and girls of each age are shown in Table 13. In actual magnitude, these differences vary from 0% to 36%, the median being 9% ($Q=5.5\%$). 75% of the differences are 17% or under, and only 16% are over 20%. In regard to sign, the differences vary from -36% to $+26\%$, the median being -3.5% ($Q=8.75\%$), showing a slight general superiority of the girls. If the number of possibilities of variation in comparing the results of small groups of non-selected subjects are taken into consideration, the presence of mental defectives, of subjects having language difficulties, of subjects in different grades influenced by different training, the possibility of a superior selection of subjects at one age group than at another, and the probability that similar chance samplings would not fall at the same age, the fact of correspondence indicated in Table 13 has more meaning than the fact of divergence.

The variability indicated in the study of the growth of abilities with age was so great that it makes interpretation of the results in terms of sex differences very difficult, and warranted conclusions impossible. It is legitimate to expect that the older subjects of either sex should make higher scores than the younger subjects of the same sex, but this was not found to be the universal rule. The boys' results showed minus deviations in 21% of the cases and the girls' results showed minus deviations in 17% of the cases. In one case the 12 year girls were 33% lower than

the 11 year girls. If this value (33%) be taken as the error due to chance variation, then only one value, that of —36%, (naming the months, age 12), may be taken as significant, and it has been seen that in this test the 12 year boys are 10% lower than the 9 year boys. The conclusion would follow, then, that there were no sex differences. This alternative, however, seems to place too much weight on one variation so that the truth probably lies in the assertion that the sex differences, that actually exist, are slight.

A study of the reactions of selected groups of boys and girls should throw light on the results from non-selected subjects, and make conclusions more certain. Subjects were selected by a process of elimination and selection. All of the subjects that were in the special class and minus grades were eliminated, along with all children of non-English speaking parents. From the following group of English speaking subjects in the regular grades all subjects were eliminated who had entered grade at an age very much above or below that of the general run of entrants.⁴ The remaining subjects ranged in age from 4.3 years to 14.4 years, but were found to group rather closely around certain ages. It was possible to find four groups of boys and girls of approximately the same chronological ages. The character of these subjects is indicated in Table 14.

The four groups of subjects, chronologically from 6.0 to 6.9, 7.6 to 8.9, 9.7 to 10.9 and 11.7 to 13.3 (which will be referred to as 6, 8, 10 and 12), were distributed in approximately the same grades, and had approximately the same average age and average grade. The results of these groups are shown in Table 15, which is arranged to show all the facts for selected subjects that were given for non-selected subjects in Tables 12 and 13. The first four columns show the percentage that each test was given to each group. The next four columns show the percentage or the proportion that the subjects in each group passed each

⁴ The ages on entering each grade of the subjects retained were as follows,—Kindergarten = 4, 5 and 6; Grade I = 5, 6 and 7; Grade II = 6, 7 and 8; Grade III = 8, 9 and 10; Grade IV = 9, 10 and 11; Grade V = 10, 11 and 12; Grade VI = 11, 12 and 13.

TABLE 14

Age in Grade Distribution, Average Grade and Average Age of 167 Selected Subjects. 86 Boys and 81 Girls.

		Age in Grade Distribution										Average Grade (M.V.)	Average Age (M.V.)
Age Group	Sex	K	I	II	III	IV	V	VI	TOTAL				
6.0 to 6.9	Boys	5	13						18	0.72 (0.40)	6.52 (0.22)		
	Girls	3	13	2					18	0.89 (0.39)	6.53 (0.22)		
7.6 to 8.9	Boys		7	13	3				23	1.83 (0.51)	8.09 (0.38)		
	Girls		2	13	5				20	2.15 (0.43)	8.32 (0.38)		
9.7 to 10.9	Boys				6	12	2		20	3.80 (0.48)	10.37 (0.36)		
	Girls				9	7	5		21	3.81 (0.69)	10.14 (0.32)		
11.7 to 13.3	Boys					2	8	15	25	5.52 (0.58)	12.35 (0.55)		
	Girls					3	8	11	22	5.36 (0.64)	12.41 (0.46)		

test. Column A shows the total number of times each test was given to all boys and girls, Column B, the weighted average age (the average ages given in Table 14 being used), and Column C the percentage that all subjects passed each test. The next four columns show the percentage that the boys are above (+) or below (—) the girls. Column D (derived from Column B), gives the difference between the average ages of all subjects to whom each test was given. Column E (derived from Column C), gives the differences between the percentages passed by all boys and girls on each test.

The growth of the various abilities with age in the selected groups of subjects is more uniform than that shown by the non-selected subjects. Only three cases appear in which the younger subjects make higher scores than those of older subjects, these exceptions occurring in the tests of describing pictures (—3%, girls 6 to 8), naming colors (—7%, girls 6 to 8), and naming months (—9%, boys, 10 to 12). In the comparison of the sexes 41 differences are obtained varying in magnitude from —28% to +26%, the median being 0% ($Q=9.5\%$). In actual magnitude the differences vary from 0 to 28, the median being 10% ($Q=4.75\%$), the median being 1% higher than that of non-selected data, and the variability 0.75% less. 75% of the differences were less than 14%.

TABLE 15
Results of 167 Selected Subjects. (86 Boys and 81 Girls).

	Percentage test was given.				Percentage or proportion seen 1931				Columns			Percentage that boys are higher or lower than girls				Columns	
	6	8	10	12	6	8	10	12	A	B	C	6	8	10	12	D	E
Counting 13 pennies.	100	87	15	0	94	95	3/3	1/1	41	757	95	-6	-5			-25	-5
Describing pictures.	100	70	24	5	100	100	5/5	1/1	38	782	100	+11	+14			-44	+10
Copying a diamond.	100	70	24	5	80	86	5/5	1/1	38	782	90						
Naming four colors.	100	83	40	0	56	100	8/8	1/1	45	787	82	-27	+8			-16	-8
Counting from 20 to 0.	100	83	15	0	83	92	9/9	1/1	40	803	90						
Counting stamps.	100	83	15	0	89	95	3/3	1/1	40	755	93	-11	+2			-27	-4
Repeating all digits.	100	94	75	12	100	93	5/5	1/1	38	782	97	-27	-28	+2		+10	-16
Naming days of week.	100	83	70	52	33	64	91	2/2	42	835	62						
Giving day and date.	100	78	100	75	14	48	93	3/3	55	850	55	+1	+8	+26		-08	+11
Naming the months.	100	89	75	23	13	40	67	5/5	48	858	44						
Naming pieces of money.	100	100	100	100	16	39	56	75	86	953	48	-11	0	-11	+4	+01	-4
Copying designs from memory.	100	100	100	100	27	39	67	71	80	952	52						
3 words in sent.	100	80	57	12	44	78	100	3/3	59	841	75	-23	+9	0		+14	-2
3 words in sent.	100	96	90	60	67	69	100	2/2	48	827	77	-12	-12	0	0	+03	-5
one idea.	100	91	85	52	10/1	29	94	85	66	937	60						
Naming 60 words in 3 minutes.	100	85	76	55	7/2	41	88	100	67	937	65	-12	-12	+6	-15	-20	-12
Giving rhymes with 3 words.	100	74	100	72	8	27	100	100	66	934	65						
Defining by use.	100	65	100	77	20	39	100	100	61	938	56	-7	-7	+8	-4	+03	+2
Defining superior to use.	100	61	100	68	10/1	31	52	82	60	985	50						
	100	61	100	68	2/0	14/3	30	53	53	1026	34						
	100	61	100	68	5/0	9/3	24	63	51	1018	35						
	100	61	100	68	1/1	13/6	75	88	59	1004	75						
	100	61	100	68	6/0	10/9	71	100	57	1024	77						
	100	61	100	68	1/1	13/2	50	68	59	1004	51						
	100	61	100	68	6/0	10/3	33	45	57	1024	33						
	100	61	100	68	2/0	7/3	69	68	47	1085	71						
	100	61	100	68	7/0	5/3	56	68	49	1033	53						
	100	61	100	68	1/0	5/3	69	71	43	1112	67						
	100	61	100	68	3/1	6/4	63	74	44	1063	66						
	100	61	100	68	15/15	11/11	4/4	4/4	38	887	100						
	100	61	100	68	8/8	11/10	9/9	3/3	38	887	94						
	100	61	100	68	11/10	11/10	9/9	3/3	38	887	94						
	100	61	100	68	8/1	15/9	11/4	4/4	38	887	47						
	100	61	100	68	11/2	11/2	9/5	3/0	34	858	26						

The change of the median of the series of differences from -3.5% (non-selected) to 0% (selected) shows that the elimination of over age and special grade pupils has helped the boys more than the girls, and has altered the general relations between the sexes. This fact is also indicated by the average difference in the percentages that all subjects pass each test, the average for non-selected subjects being -1.4% and for selected subjects $+1.6\%$. The non-selected boys from 6 to 12 were given, in all, 2436 tests, these tests being passed 60.8% of the time. The non-selected girls were given 2195 tests, passing 61.6%, the advantage being 0.8% in their favor. The selected boys were given 1125 tests, passing 64.3%, an advantage of 0.1% over the girls who passed 64.2% of 1034 tests. The foregoing changes indicate clearly that the selection of subjects has changed the general relations between the sexes, helping the boys more than the girls.

The relations between the results of selected and non-selected subjects may be studied by a comparison of the differences between the percentages passed by all subjects. If the differences between the scores of the boys and girls are due to but one factor, that of sex differences, then the correlation between the two methods of study should be very nearly absolute. The correlation (Pearson product-moments formula) between the differences in the percentage passed by all boys and girls according to the two methods is 0.726 ($p=0.075$). This correlation between the two methods is high, but it would probably be high inasmuch as the 167 selected subjects are included in the 352 non-selected subjects. The results of the two methods show certain large discrepancies. The changes of the greatest magnitude are those shown by the 60 words test ($+4\%$ by the first method to $+18\%$ by the second), the tests of defining in terms superior to use ($+7\%$ to $+21\%$), of naming the days of the week, (-16% to -2%), giving rhymes, (-10% to $+1\%$), naming colors, (-14% to -4%), copying the diamond, ($+1\%$ to -8%), and counting from 20 to 0 (-8% to -16%). The comparison of the median differences shows that the selected method tends to improve the results of the boys more

than those of the girls. All of the changes in the results of the two methods are not in favor of the boys, however, the total scores on the diamond and 20 to 0 tests showing changes in favor of the girls. If the cause of the variations shown by the first method is the presence of a few children of non-English speaking parents, to special class and minus grade children, then the elimination of this source of error should change the results in only one direction.

The analysis of the results of selected subjects, therefore, does not lessen the difficulty of the interpretation of the results in the light of sex differences. The rate of growth of the various abilities with age is irregular. The analysis of the irregularities points to the fact that the boys or girls of any age may be a chance selection of superior or inferior subjects at that age. The method of comparing selected subjects would tend to eliminate the inferior selection of subjects, but would not eliminate the possibility of a superior selection.

The comparison of the results of the sexes shows differences at certain ages and on certain tests that are as high as 20%. The problem involved is that of deciding whether these large differences are due to chance or to differences in the reactions of the sexes. Certain tests show large deviations first in favor of one sex and then in favor of the other. If a difference of a percentage of any magnitude on any test is to be attributed to a sex difference, then the same line of reasoning will show that in certain tests the abilities change from one sex to the other. The analysis of the tests that show this crossing of ability should throw light on the other tests.

Three tests show substantial differences in favor of both sexes according to both methods. In the test of copying the diamond, the non-selected girls lead at the start, age 6, and the boys are ahead at 7, 8 and 9, the same relations being shown by selected subjects of 6 and 8. In the test of copying the designs from memory, the non-selected girls are 24% below the boys at age 9 and 21% above the boys at age 12, the same relations being shown by the selected subjects of 10 and 12. In the test of naming 60 words in three minutes, the non-selected girls are

19% above the boys at 9, and 19% below at 12. The selected boys of 10 and 12 are in advance of the girls in this test.

These three tests are crucial in the consideration of the problem of whether differences shown between the boys and girls are due to actual sex differences or due to accidental causes. Each of these tests may be studied by a method more accurate than that of comparing the percentage passed at each age. The reproductions of the diamond were arbitrarily sorted in six groups according to their merits by a method described in the discussion of the personal equation. The first group contained the best reproductions, the sixth, the poorest. The reproductions of the designs were graded from 0 to 20 by an arbitrary point system described under the discussion of the personal equation. A measure of the ability in the 60 word test is the actual number of words given in three minutes, a measure recorded by the experimenters in each case. Table 16 shows the average score made by the non-selected and selected boys and girls of each age in these three tests.

TABLE 16

Average Score (Mean Variation) of Subjects of Each Age on Three Tests.

	Copying the Diamond Average Group of the Reproductions.		Drawing the Designs Average number of points scored.		Naming 60 words Average number of words given in three minutes.	
	Boys	Girls	Boys	Girls	Boys	Girls
unselected subjects	6	4.27(1.28)	3.57(1.24)			
	7	2.85(1.04)	3.17(1.37)			
	8	2.20(1.15)	3.24(1.57)	8.06(6.19)	9.00(5.25)	
	9	2.33(0.89)	3.00(1.29)	10.29(5.30)	5.32(4.61)	52.93(11.20) 59.91(10.10)
	10			9.17(5.33)	9.18(6.73)	68.12(13.12) 61.76(11.25)
selected subjects	11			8.64(6.73)	10.94(7.06)	73.65(13.35) 71.28(14.25)
	12			8.64(6.02)	11.08(6.08)	68.75(12.28) 58.14(12.57)
	6	4.27(1.20)	3.33(1.26)			
selected subjects	8	2.32(1.00)	3.00(1.17)			
	10			9.55(5.60)	7.29(6.42)	67.31(12.74) 62.13(11.39)
	12			12.53(5.38)	13.56(5.55)	75.33(10.92) 66.84(13.87)

The relations indicated by the percentage passed are also indicated by the more reliable method of comparing the average scores. In the test of copying the diamond, the 6 year non-selected girls average 0.70 group better than the boys, while the

selected girls are 0.94 ahead. The comparison of the 7, 8 and 9 year subjects shows the boys ahead in all cases, the 8 year non-selected boys averaging over one group higher. The non-selected boys show an improvement of two groups from 6 to 9, while the girls show an improvement of only half a group. One sex shows a decided growth of ability, the other practically none. If the differences indicated are to be taken as real, it will be necessary to assume that the girls pick up the ability to draw a diamond easier than the boys, but that this ability once obtained remains constant—that the effect of maturity operates on one sex but not on another. The number of cases on which this assumption is based (174 subjects from 6 to 9) is so small, and the chances of variation in the selection of subjects of different intellectual status in each age group is so large, that the assumption is not substantiated.

The relations indicated in the test of copying the designs are more variable than those of the diamond test. The 9 year non-selected boys show an improvement over the 8 year boys, but from 9 to 12 there is a gradual decrease in the ability, so that the 11 and 12 year boys are only slightly ahead of the 8 year boys. The relations shown by the non-selected girls are exactly the reverse of those of the boys. The 9 year girls are very much lower than the 8 year girls, and a gradual increase appears from 9 to 12 instead of a decrease. The comparison of these opposite relations gives a maximum difference in favor of the boys at 9 and the girls at 12. If the relations indicated in this test are to be considered definite, the assumption is involved that the influence of increasing age on one sex is exactly opposite to that on the other sex, an assumption that is not substantiated in view of the small number of cases (183 subjects from 8 to 12) and the possibility of selecting subjects of chance superiority in the small groups at each age.

The relations indicated in the test of naming 60 words are more constant than those shown in the diamond or design tests. Both sexes show a growth of ability from 9 to 11 and a decrease from 11 to 12. The growth is irregular, however, the girls showing less growth from 9 to 10, and a greater drop from 11 to

12, so that a comparison of the sexes shows a deviation in favor of the girls at 9 and of the boys at 12. The assumption of any large sex differences in this test involves the assumption that 12 year girls have less ability in this test than 9 year girls, and that the influence of maturity operates differently on the two sexes, an assumption that is not substantiated in view of the many variable factors.

The conclusion that a definite crossing of ability between the sexes occurs in the tests of copying the diamond, copying designs and naming 60 words, is not substantiated. It is not justifiable to attribute a difference of 20% between the sexes to a real sex difference on one test and not on another. If the differences shown between the results of the sexes in the tests of constructing a sentence containing one idea, of naming the months, naming the days of the week, counting stamps and naming colors are to be attributed to sex differences, then the variations in ability shown in the diamond, design and 60 word test must be assumed to be definite. These assumptions were not found to be substantiated, however, so that it is not possible to draw any conclusions concerning sex differences from a study of the percentage that selected or unselected subjects of each age pass each test.

The variable influences due to the selection of subjects of different status at each age are eliminated or counterbalanced to some extent by combining the subjects of all ages. The differences between the percentages that all boys and girls pass each test are to some extent influenced by the ages of the subjects to whom each test was given. The correlation (Pearson product-moments formula) of the differences between the percentages that all non-selected boys and girls passed each test with the difference between the average ages of all the non-selected boys and girls to whom each test was given is 0.394 ($p=0.134$). The correlation between the same arrays from selected subjects (i.e. between Columns D and E of Table 15) is 0.388 ($p=0.135$). These correlations between the tests and age are high enough to indicate that the factor of age is present to some extent. The close correspondence in the correlations from the two methods

indicates that the age factor is present to the same extent in both methods. The tests vary in the degree with which they correlate with age, so that it is not possible to estimate the amount of the influence of this factor. Furthermore, it has been seen that the results from the two methods are not in strict accordance, that the elimination of inferior subjects caused changes in the results in both directions. For these reasons, it is not possible to draw any conclusions concerning sex differences from a comparison of the percentages passed by all subjects.

Certain negative conclusions are, however, possible. The number of subjects at each age in both methods is comparatively small. The chances of variations due to factors other than sex differences has been shown to be very large. The fact of correspondence between the results of the two sexes is therefore of more importance than the fact of divergence. 75% of the differences between the non-selected boys and girls are 17% or under, while the same proportion of the differences between selected boys and girls falls under 14%. If it is assumed that the subjects of any age should not test lower than those of any preceding age, and allowance is made for differences between the sexes that are exaggerated on account of the chance falling off of ability with older subjects, only 9% of the differences between the non-selected boys and girls are over 20% (derived from Table 13).

The evidence from the foregoing methods of study points to the conclusion that the sex differences, if present, are under 20% or 25% as a maximum, and that deviations of this magnitude are marked exceptions to the general run of differences. The conclusion that the differences that might possibly be attributed to the sex factor are slight, has no meaning unless the word "slight" is defined independently of the writer's personal opinion. The differences shown between the results of the sexes are smaller than those that were attributed to the factor of the personal equation in the study of the results of the four experimenters. It was concluded that certain tests were influenced by grade training. These tests showed from 40% to 60% improvement from one grade to another, so that the greatest influence that may be attributed to the sex factor is only approximately

one half that due to grade training. The following study of the diagnostic value of the tests will show that the deviations that might be attributed to the sex factor are insignificant when compared to the differences between the reactions of normal and retarded children to the individual tests.

Most of the investigators who have studied the factor of sex differences in the Binet tests, have studied them from the standpoint of the "mental ages" or total scores made by the subjects of both sexes. A few investigators have studied sex differences in the light of the individual tests. Descoeudres (20) reports the results of the application of the Binet tests to 24 subjects, one good and one poor pupil of each sex from each of six school grades, drawing conclusions from this investigation concerning the diagnostic value of the individual tests and the sex differences involved. Obviously the number of subjects is too small to allow any conclusions to be drawn. Chotzen (18) compared the percentage that all feeble-minded boys and girls passed each of 15 tests, finding differences varying in magnitude from 1% to 20%. The largest deviations were those of 20% in favor of the boys in the test of copying the diamond, 13% in favor of the girls in the test of executing three commissions, 12% in favor of the boys in naming the pieces of money, 11% in favor of the girls in the test of repeating a sentence of 16 syllables, and 10% in favor of the girls in detecting omissions in pictures. All other differences were less than 10%.

Bloch and Preiss (9) examined 155 normal Volksschule children (79 boys and 76 girls) varying in age from 7 to 13. Bober-tag's translation was used. These investigators found very striking differences in the reaction of the sexes to the individual tests, the differences running as high as 52%, most of them in favor of the boys. The differences between the performances of the boys and girls of each age were calculated, without reference to the many sources of variation. The factor of the personal equation is not treated, and this factor alone might cause these variations. If a more careful analysis of the results had been made, it is very probable that the conclusions would have been modified to some extent. The fact that the 11 year

boys are 37% higher than the 11 year girls on the test of criticising absurdities is most certainly modified by the fact that the 11 year subjects are 30% lower than the 10 year subjects in the test of repeating 7 digits. The small number of subjects (in five cases less than 10), would tend to emphasize chance variations. The fact that the number of subjects is too small to warrant definite conclusions is pointed out by the authors. Stern (62) in commenting upon these results, points out the significance of the fact that the inferiority of the girls extends to so many different kinds of tests. The results of Bloch and Preiss are in almost complete contradiction to the results of the present investigation. They find large differences, and find practically all of these differences in favor of the boys. This investigation shows a general run of differences very much smaller, and a slight general superiority of the non-selected girls. The mere fact of contradiction in the results of the two investigations would indicate that the differences were not produced by the common factor of sex. Rogers and McIntyre (54) give no figures, but report that they have studied their results in the light of sex differences, and have found no correlation between their results and those of Bloch and Preiss.

The results of the investigators who have compared the "mental ages" or total scores of children of different sexes are somewhat at variance. Goddard (30) reports that there are more backward boys than girls. Stern notes that Goddard's results do not bear out his statement, for the percentage of boys and girls testing two or more years retarded is the same (18.5%). The accuracy of Goddard's statement depends on the criterion⁵ used for measuring backwardness. Although Goddard's state-

⁵ If the criterion is four or more years retarded, there are more backward boys than girls (boys = 3.7%, girls = 3.1%). If the criterion is three or more years backward, there are more girls than boys (boys = 8%, girls = 9.1%). If the criterion is two or more years backward, the proportions are the same, as Stern notes. If the criterion is one year or more retarded, there are more backward boys than girls (boys = 41.4%, girls = 35.6%). There are more girls than boys testing at and above age according to Goddard's results. 34.7% of the boys and 36.6% of the girls test at age, while 23.8% of the boys and 27.7% of the girls test one year or more above age.

ment concerning the backwardness of the boys may be interpreted differently, his figures leave no doubt concerning the fact that there are more girls than boys at and above age, and therefore indicate a general superiority of the girls.

Bobertag (10) computed the average "mental age" of 90 boys and 90 girls regularly distributed from 7 to 12. The subjects were selected according to school grades, so that the average grade of each group differed by exactly one grade. His results show the boys ahead 0.06 yr. at 7, 0.14 yr. at 8 and 9, 0.20 yr. at 10, 0.19 yr. at 11 and 0.14 yr. at 12. These findings cannot be considered entirely out of harmony with those of Goddard, for, as this investigation shows, there may be a change in the relation of non-selected boys and girls and selected boys and girls.

Yerkes and his co-workers (82), scoring some of the Binet tests according to the point system, show that the girls of English speaking parents are superior to the boys of the same parentage between 5 and 7, that they fall below with minor variations till 11, where they again surpass the boys at 12 and 13, falling below at 14 and 15. The differences between the sexes are smaller and of less practical importance than the differences due to the language factor, but the authors suspect "that at certain ages serious injustice will be done to individuals by evaluating their scores in the light of norms which do not take account of sex differences." (page 73). In contradiction to these results are those of Terman and his co-workers (67), who, scoring the Stanford revision of the Binet scale according to "intelligence quotients," find differences of but 2% to 4% in these quotients in favor of the girls, and who conclude from the basis of their studies of sex differences that the conclusions of Yerkes are unjustified. These two investigations used tests different in character and differently weighted, so that the results would not necessarily have to correspond.

The one common feature of most of the researches on sex differences in the Binet-Simon tests is that the differences are small. Burt and Moore (17) summarize the work of various investigators in the general field of sex differences, and report an investigation of their own on 67 boys and 63 girls, 12½ to 13½

years of age. They discuss their results and those of the other authors in the order of the complexity of the mental processes involved. They find a high correlation between the size of the sex difference and the simplicity of the capacities compared—the higher the process, and the more complex the capacity, the smaller the sex difference.

The general trend of the investigations on sex differences indicates that no very large differences are to be expected in the application of intelligence tests, and that the differences to be expected will vary according to the nature of the tests. The results of this investigation are in agreement with the general trend of the investigations in showing only slight differences that might be attributed to the sex factor. The results do not show on what tests, if any, these differences occur. Conclusions concerning the amount of influence of this factor must be drawn from more exhaustive investigations on the individual tests. The research of Bateman (3), for instance, is conclusive in the test of naming colors. Bateman shows that there is a difference of 14% in favor of the girls in this test, showing furthermore that the factor of school training causes an improvement of but 18%. The results would indicate that the test should be placed in the fifth or sixth year, but the sex difference of 14% would probably not warrant the placing of the test in a different age group for boys and girls.

The investigations of Bolton (11) and Wooley (79) would show that small differences in favor of the girls are to be expected in the tests of repeating digits, and possibly in all memory tests. The investigations of Gilbert (27), Thompson (68), Burt and Moore, and Peterson and Doll (51) would indicate that a slight difference in favor of the boys should appear in the test of arranging five weights. Ruger's (55) finding of striking differences in favor of men in a series of puzzle tests, and Wooley and Fisher's finding of large differences in favor of the boys in the Healy puzzle-box test would show that rather large differences might appear in the general class of "puzzle" tests.

Even though the sex differences in intelligence tests may be shown to be small, scientific procedure should demand that the

investigator who standardizes any test or system of tests should treat his results in such a way as to demonstrate that the factor is present or not present. The burden of proof should still be on the person who maintains that sex differences are not involved. The knowledge of sex differences is especially important in diagnosing border-line cases of mental defect, where the diagnosis must often be made on the qualitatively different character of the responses to individual tests.

VI. SUMMARY.

One of the fundamental assumptions in the construction of the Binet-Simon scale is the correlation of the individual tests with age. The correlation of the tests with age is affected by the error due to incomplete data, by the influence of the personal equation of the experimenter, and by the training the subject has received in school.

The influence of the personal equation of the experimenter was found to be more marked in some tests than in others, the influence being most marked in the tests of copying the diamond, indicating omissions in pictures, defining in terms superior to use, drawing designs from memory, detecting absurdities in statements and reconstructing dissected sentences.

The variations between the experimenters could be traced to three sources,—

- 1) to the use of apparatus, variations in which were due to,
 - a) the construction of the test material, and
 - b) the use of alternative questions;
- 2) to the technique of the experimenters in giving the tests;
and
- 3) to observation errors made by the experimenters in marking a response passed or failed.

It is possible to eliminate all three sources of error.

The effect of school training was more marked on some tests than on others, the effect being most marked in the tests of counting stamps, counting backward from 20 to 0, enumerating the days of the week and the months, giving the day and the date, naming the pieces of money, making change, and reconstructing dissected sentences. Tests that involve school training should be standardized on a different basis than those relatively independent of this factor.

Although the comparison of "mental ages" and pedagogical ages gives no information concerning the general correlation be-

tween the Binet tests and the school grades, the study of the individual tests establishes the fact of a general correlation.

The correlation of the individual tests with grade is higher than the correlation of the tests with age, this fact being indirect evidence of the value of the tests as measures of intelligence.

Sex differences were found to be slight as compared with the influence due to the personal equation or grade training.

Since variations occur in the results due to the influence of the personal equation and grade training, certain allowances must be made for these factors in making diagnoses on the basis of the tests. The scale is therefore a qualitative rather than a quantitative instrument.

The investigator who wishes to use his results for standardizing age norms should use only those data based on the complete method of experimenting, and should treat his results in such a way as to demonstrate the presence or absence of the variable factors of the personal equation, grade training and sex differences.

THE DIAGNOSTIC VALUE OF SOME MENTAL TESTS

TABLE OF CONTENTS

I. Introduction	95
II. Character of Subjects	106
III. Tests and Procedure	116
IV. Variable Factors	126
V. Diagnostic Value of Binet Tests	140
VI. Diagnostic Value of Supplementary Tests	166
VII. Correlations of Abilities with Age	205
VIII. Results of Other Investigators	222
IX. Conclusions and Suggestions	232
Bibliography	249

I. INTRODUCTION

The Binet scale has been used to classify normal children, and as a means of studying mental differences due to race, sex and environment, but its most important function is the detection and classification of conditions of feeble-mindedness. The consensus of expert opinion would seem to show that the scale is unsatisfactory in this respect. Among the questions brought up at the Buffalo conference (15) was the following,—“Does the scale provide a reliable means of diagnosing feeble-mindedness?” The answer given was,—“It does not always furnish a sharp, nor a positive diagnosis of feeble-mindedness: in particular —*a*. A mental age of 10 or above is not necessarily indicative of feeble-mindedness, regardless of how old the examinee may be; and *b*. A young child may test almost at age and yet be feeble-minded as determined by other criteria.”

W. E. Fernald (26) discussing the question of detecting the higher grades of mental defect, writes “The Binet tests, in the hands of competent examiners, usually corroborate the results of clinical examination in the recognition of all degrees of mental defect in children under ten, and of pronounced defect in older persons. These tests are not so effective in detecting slight mental defect in world-wise adolescents and adults. In other words, the Binet tests corroborate where we do not need corroboration, and are not decisive where the differential diagnosis of the high grade defective from the normal is in question” (page (747)). And again, “The Binet test does not register as defective certain persons who present plain evidence of mental defect in their personal history, school history and performance, social and economic relations, etc., while on the other hand, certain individuals who fail to come up to the requirements of the Binet test do not present the usual personal social and economic reactions of mental defect” (page 748).

These opinions are corroborated by the lack of agreement be-

tween investigators who have made studies of groups of adolescents. In the studies of delinquency for example there is a wide disagreement between investigators. M. Otis (50) examined 172 girls, ages from 10 to 20, in a state home for delinquent girls to which commitments were made by the courts, and reports 25% "presumably normal," 30% morons, and 45% defective, i.e., 75% feeble-minded. Morrow and Bridgman (47) report the results of an examination of 60 girls of a similar run of ages in a similar institution in a neighboring state, finding 10% normal, 66% feeble-minded and 24% "doubtful." Bridgman (13) reporting the results of the examination of 118 girls at the same institution, the general run of admissions, gives 5% normal, 6% backward and 87% feeble-minded. Of the 104 girls committed as sexual delinquents, Bridgman finds 3% normal and 97% feeble-minded. These investigators used the same instrument, Goddard's 1911 scale. Healy (33) working in the same general field with a much larger group of delinquents and with more reliable methods, finds a much smaller percentage of feeble-mindedness. In fact, his group of feeble-minded is but 11.3% of the total number of cases (823), and the entire group of defective types (including cases of feeble-mindedness, poor native ability, mental subnormality, dullness from physical causes including epilepsy, and specialized defects including defects of self-control) is but 25.3% of the whole group of delinquents studied.

That the disagreement between the various investigators in the proportion of feeble-minded individuals reported among delinquents is due to the tests is shown by Kohs' (42) study of 335 cases at the Chicago House of Correction. Kohs used Goddard's 1911 revision of the Binet scale. He also used other criteria for deciding whether his subjects should be classified as normal or feeble-minded. He found that the feeble-minded individuals tested from 6 $\frac{1}{5}$ to 11 $\frac{2}{5}$, the normal individuals from 10 $\frac{4}{5}$ to 12 $\frac{2}{5}$. In other words the results of the Binet scale, instead of showing a positive differentiation between normal and feeble-minded, showed a marked over-lapping of per-

formance. That this error in the scale in failing to make a complete differentiation is a serious one is shown by the fact that 30% of Kohs' cases fell within the range of over-lapping (10 4/5 to 11 2/5).

Diagnoses of feeble-mindedness will probably never be made on the basis of mental tests alone, but the reliability of these diagnoses will most certainly be increased if the tests are improved. The lower grades of feeble-mindedness cause little difficulty in diagnosis. Idiots are self-diagnostic, and imbeciles rarely reach adolescence without detection. The moron group is the most difficult to diagnose. Inasmuch as this group is also the most difficult to treat socially, it would seem worth while to perfect the instruments of diagnosis.

Competent authorities maintain that the differences between the normal individual and the idiot are not differences of quality or species, but differences only of quantity or amount of intelligence. On this theory intelligence will be found in varying amounts from idiocy to genius. As intelligence is a means of adjusting the individual to his environment, individuals will be found varying in the degree of adjustment from the idiot who can not feed himself to the individual competent to control his environment in a number of ways. Some authorities, for instance Witmer (74) would hold that the diagnosis of feeble-mindedness is a social diagnosis. According to Witmer, the diagnosis is not made concerning the subject's mentality, but is concerned merely with the advisability of freedom or segregation. This view would seem to confuse the disease and its treatment, for the defective social adjustment of the feeble-minded is always referred to defective intelligence, just as the defective social adjustment of the blind person is referred to his lack of eye-sight. The most profitable method of increasing the accuracy of the diagnosis of feeble-mindedness would therefore seem to be that of increasing the reliability of tests of intelligence.

If the Binet scale fails to detect the higher grades of mental defect, it is legitimate to ask why it fails. Binet (7) suggested the answer in discussing the means of differentiating the moron

from the normal. He considered that six tests (arranging five weights, comprehending difficult questions, using three given words in a sentence, defining abstract terms, interpreting pictures, and giving rhymes) were important in distinguishing the moron from the normal individual of the Paris population. Binet therefore considered certain tests more diagnostic of intelligence than others. However, he offered no proof that these six tests were better than any others and his assertion can be no better than an expression of opinion, with the possibility that his opinion was wrong. In the actual construction of the scale, the tests were weighted equally, the individual receiving the same amount of credit for passing the test of naming the months as he would for passing the comprehension questions, the test which according to Binet dissipated all his doubts concerning a final diagnosis. (See page 3.) Each test counts for one fifth of a year, and it makes no difference in the quantitative score whether a child reaches a certain "mental age" by passing the most diagnostic or the least diagnostic tests. The final diagnoses that Binet made in his own cases must have been qualitative rather than quantitative for he threw more weight on some tests than on others in forming his opinion.

Following Binet's cue on the matter, it would seem that the method of increasing the accuracy of the diagnosis of the higher grades of mental defect would be that of determining what sorts of tests were the most highly diagnostic of this sort of defect. In the study of the Binet scale itself, the problem of determining what tests are diagnostic of intelligence immediately arises. Binet demonstrated that the tests were correlated with age, but he did not go beyond this point. He never demonstrated that the tests were correlated with intelligence. In his opinion of course, all the tests were diagnostic of intelligence or he would not have included them in his "Measuring Scale of Intelligence," yet it is plain from his writings that he considered some tests more valuable than others in this respect. It remains for other investigators to check up Binet's work, and to establish the accuracy of the individual tests that he included in his scale. Studies of the

diagnostic value of the Binet tests have been made by Descoeudres and Chotzen.

Descoeudres (20) published in 1911 the results of the application of the scale to one good and one poor subject of each sex from each of six school grades, using the material from these 24 subjects as a basis of studying the relative diagnostic value of the tests, and of studying the sex differences involved. Grouping the subjects according to the teachers' judgments of good or poor, the tests that showed the most marked differences between the two intellectual levels were those of arranging five weights, interpreting pictures, detecting absurdities, defining in terms superior to use, counting backwards from 20 to 0 and indicating the omissions in pictures. The group with inferior intellectual endowment made lower scores on all tests except those of understanding easy and difficult problem questions.

Descoeudres (21) also published in 1911 the results of an investigation on 14 backward and defective boys and girls from 6 to 14 years of age. 15 tests were used; 6 of them (describing pictures, defining, comparing remembered objects, naming as many words as possible in three minutes, comprehending questions, and recognizing coins) being taken from the Binet scale, the other tests being tests of motor dexterity, tactile ability, auditory and visual imagination, puzzle solving, cancelling a's, calculation, and auditory and visual memory. The obvious differences between the subjects made the independent (rank) estimation of their intelligence possible, so that the correlation of each of the tests with intelligence could be determined. In the list of 15 tests arranged according to the magnitude of their correlation with intelligence, comparing remembered objects showed the highest correlation, describing pictures and comprehending questions stood third and fourth, defining sixth, naming four pieces of money tenth, and naming words fifteenth. Analyzing the factors involved in each, Descoeudres concluded that the tests of reasoning show the highest correlation with intelligence, imagination next, while tests of memory, particularly auditory memory, show the lowest correlation. Although these

two investigations of Descoeudres are suggestive, the small number of subjects precludes the possibility of forming any definite conclusions concerning the diagnostic value of the tests.

In 1912, Chotzen (18) published the results of the application of the Binet scale to 280 backward and defective children. The subjects, 157 boys and 123 girls varying in age from 7 to 14, were either enrolled in the special classes for backward and defective children in Breslau, or were candidates for admission to these classes. The largest proportion of the subjects were distributed in the ages 8, 9 and 10. Bobertag's standardization of Binet's 1908 scale were used, so that Chotzen could compare his results from backward children with those of Bobertag from normal subjects. The tests used were largely those in the V, VI, VII, VIII, and IX year groups. The tests for "ten years" were given to some extent, but those in the higher groups were very rarely used.

Chotzen subjected the results of his investigation to a very careful analysis, part of which deals directly with the relative value of the tests in the diagnosis of feeble-mindedness. He used three methods of analysing the data,—(1) the comparison of his results from feeble-minded children with those of Bobertag from normal children; (2) the comparison of the results of feeble-minded children of the same mental ages but different chronological ages; and (3) the comparison of the results of groups of children of approximately the same ages but with different final medical diagnoses.

In studying the results according to the first method, Chotzen found that the feeble-minded children were in general from 2 to 4 years backward as compared to normal children, the deficiency being more marked in some tests than in others. The test of describing pictures, for example, placed by Bobertag in year VII, was passed by 73% of Chotzen's feeble-minded children of 8 years, while the test of counting backwards from 20 to 0 in Age VIII was passed by only 8% of the 8 year, 16% of the 9 year, 29% of the 10 year, 55% of the 11 year and 72% of the 12 year feeble-minded children. The first test would be a "seven"

or an "eight year" test for feeble-minded children, while the second would be a "twelve year" test.

Backwardness was more marked, then, in the latter test. On this basis, the tests in which backwardness was most marked were those of repeating a sentence of 16 syllables, making change, counting backwards from 20 to 0, defining in terms superior to use, comparing remembered objects, recalling a story read, naming the months, repeating five digits and arranging five weights. Backwardness was least apparent in distinguishing between morning and afternoon, defining in terms of use, describing pictures, counting 13 pennies, giving age, choosing the prettier of given faces and counting the fingers.

The second method gave Chotzen an opportunity to study the effect of school training and physical maturity on the tests. According to this method, the results of children of the same "mental age" but of different chronological ages were compared. Taking all subjects of the "mental age" of eight, those of the chronological ages 8 and 9 passed the tests of repeating five digits in 60% of the cases, while 50% of those of the chronological ages 10 and 11 passed the test. The same group of subjects of the "mental age" of eight showed much more improvement in the test of copying a sentence, this test being passed by 35% of those age 8, by 65% age 9, by 94% age 10, and by 100% ages 11 and 12. The effect of maturity was much more marked in the second test.

The tests that showed the greatest increase with age were those of copying a sentence, writing from dictation, recalling a story read, and enumerating the days of the week. A slight increase was shown in the tests of playing the game of solitaire, knowing age, executing three commissions, counting backwards from 20 to 0, and showing the right hand. The increase was still less in the tests of repeating a sentence of 16 syllables and copying the diamond, and practically negligible in naming 5 coins. No increase was found in the tests of describing pictures, detecting omissions in pictures, counting 13 pennies, counting the fingers, comprehending easy problem questions, repeating five

digits, counting 9 "pfennig" (3 doubles and 3 singles), naming colors, comparing remembered objects, and defining in terms superior to use. According to Chotzen, the tests that show the greatest increase with age relate almost entirely to school training, those that show a slight increase depend to some extent on school training and other environmental experience, while those that show no increase with age involve largely factors of judgment and memory.

In treating his results according to the third method, Chotzen classified his subjects according to the final medical diagnosis he had made, as "not feeble-minded," morons (*Debilien*), imbeciles and idiots. The last group was dropped out, being small in number, and each of the other groups was divided into two parts according to age, one group being composed of subjects aged 8 and 9, and the other of subjects from 10 to 13. Certain tests stood out as differentiating very sharply the lines between the groups. For example, the test of comparing remembered objects was passed in the 8 and 9 year groups by 62% of the "not feeble-minded," 28% of the morons, and 5% of the imbeciles, while in the older groups it was passed by 93% of the "not feeble-minded," 69% of the morons and 10% of the imbeciles. The test of showing the right hand was passed in the younger group by 76% of the "not feeble-minded," 80% of the morons and 60% of the imbeciles. The first test showed a difference of 57% between "not feeble-minded" and imbecile children of 8 and 9, while the second test showed a difference of but 16% between these groups. The first test would therefore appear to be more diagnostic of intelligence.

The tests that showed the greatest diagnostic value in differentiating the groups of the younger children were copying the diamond, naming five coins, comprehending easy problem questions, comparing remembered objects and repeating five digits. The last two tests also differentiated the members of the older groups together with the tests of reproducing an item of a newspaper, arranging five weights, making change, defining in terms superior to use and naming the coins. Certain tests then are more val-

uable than others in the diagnosis of feeble-mindedness itself, and in diagnosing the different degrees of defect.

The results of the three methods of studying the individual tests agreed very closely on some tests. The tests of repeating five digits, comparing remembered objects and defining in terms of use, for example, showed a marked difference between normal and feeble-minded subjects, revealed no growth with age in children of approximately the same mental level and proved to be highly diagnostic of the different degrees of mental defect. In some cases the agreement was not as close. The test of remembering a story read showed a most decided increase with maturity independent of intelligence, but at the same time proved to be highly diagnostic of the mental status of "not feeble-minded," morons and imbeciles, and was one of the tests in which backwardness of the feeble-minded was most marked. The test of describing pictures, on the other hand, showed little difference between the performance of feeble-minded and normal subjects, and no growth with maturity. It can not be maintained then that the tests which show the greatest increase with maturity are those that are least dependent on intelligence, or conversely that those that show the highest correlation with intelligence are least dependent on maturity.

The lack of correlation between the results of the different methods is probably due to the faults of the methods themselves. The comparison of the results of feeble-minded children with those of normal children was made without any adequate statistical justification as to what percentage of a given age group must pass a test in order to have that test considered to be well within the ability of the group, or how large a difference there must be between the performance of feeble-minded and normal subjects in order to have that difference indicate more or less backwardness on that particular test. The personal equation of the investigator must necessarily play a large part in forming these judgments for the figures are quite irregular. The results of the second method are not conclusive, for Chotzen's figures do not

prove that the tests that show the greatest increase with maturity are least dependent on intelligence. The method demonstrates the influence of maturity and school training, but gives no information concerning the diagnostic value of the tests. The method of comparing the results of the different groups according to the final medical diagnosis was not entirely satisfactory, for the results themselves played some part at least in this diagnosis. If certain tests were given more weight than others, consciously or unconsciously, in making this diagnosis, then those tests would of necessity stand out as those that had the highest diagnostic value in differentiating the groups. This method depends to some extent at least on the personal equation of the investigator.

In general, however, although there was not a high correlation between the results of the different methods of study, certain tests were definitely shown to be dependent on factors of training and others were shown to be independent of this factor. Certain tests also stood out as diagnostic of feeble-mindedness while others were not as valuable in this respect. The investigation, though most suggestive, is not conclusive, and needs checking up and elaborating. The tests in the "ten," "eleven" and "twelve year" groups were not given enough to furnish any available data. As this is the region in which the scale breaks down, i.e. in diagnosing the higher grades of mental defect, it is most important that the work be elaborated here, and the relative diagnostic value of these tests determined.

The consensus of expert opinion shows that the Binet scale is not a reliable instrument for diagnosing the higher grades of feeble-mindedness. The problem before present investigators is the correction of this defect in the scale. The solution of the problem was indicated by Binet who, although he made all tests quantitatively equal, considered some tests more valuable than others in making a diagnosis. Descoudres showed that the tests varied in the magnitude of their correlation with intelligence. Chotzen showed that all the tests were not of equal difficulty for feeble-minded children. The Binet scale has been shown to be

composed of some tests that are effective in diagnosing intelligence, and others that are ineffective. To make the scale more effective it is necessary to determine the relative diagnostic value of the individual tests.

During the process of analysing the Princeton data, the writer compared the results of 49 children who had been rated "dull" by the teachers and 46 children of the same age who had been rated "bright." The results were striking. Some tests were as easy for the dull children as for the bright children. Other tests were easy for the bright children but practically impossible for the dull children. The results of this study led the writer to undertake a much more extensive investigation in the Trenton, N. J. public schools, where on account of the large number of children in the schools it was possible to obtain two groups of known differences of intelligence. To these groups the Binet and other tests were given in order to determine what sorts of tests were most highly diagnostic of intelligence. The results of this investigation are contained in the following chapters.

II. CHARACTER OF SUBJECTS

Two methods are open to the investigator who wishes to determine the relative diagnostic value of mental tests. The first is that of giving the tests to a group of individuals who can be independently and accurately rated in their rank order of intelligence. From the standing of the individuals in the various tests, the correlation of the tests with intelligence may be obtained. The other method is that of comparing the results of two or more groups of known differences of intelligence. The differences in the performances of the groups on the tests will indicate the diagnostic value of the tests. The second method was followed by the writer as it was thereby possible to obtain a more objective indication of the intelligence of the children examined.

In speaking of intelligence, the writer means by the term exactly what Stern (62) meant: "general mental adaptability to new problems and conditions of life" (page 3). The objective indication of intelligence that was used was that of school standing. The public school presents a situation that must be met by all children. Some children meet the situation adequately and pass through the schedule of grades in the required time. Others are unable to adapt themselves to the changing conditions and drop farther and farther behind their fellows. A retardation in school of one or possibly two years may be due to a great many causes independent of the mental status of the individual, but a retardation of more than two years would probably indicate, in the absence of a more obvious explanation, an inferior intellectual endowment.

The experiments reported here were conducted on two groups of children, one group composed of boys in the regular grades of the Franklin school in Trenton, N. J., the other group of members of the special classes for backward and defective children or candidates for admission to these classes in the same city. The

law of New Jersey makes it mandatory that children showing three or more years pedagogical retardation be placed in special classes. The pedagogical retardation that is meant is that indicated by the age of the individual and the grade he should be in at that age. Inasmuch as the school grade age used in this way does not take account of late entrance or enforced absence, the criterion of retardation used in this investigation was that of grade progress or in other words the relation between the grade of the individual and the number of years he had been in school.

In order to find a group of retarded children the writer examined nearly all of the boys and girls in the Trenton schools who were either in special classes or who were candidates for admission to these classes. 229 boys were examined, 203 being in the classes while the remaining 26 were candidates for admission to the classes. All of these subjects were given the Binet tests, and those whose retardation showed no obvious explanation other than inferior intellectual endowment were given the complete series of tests.

Concerning each subject examined, the following information was obtained:

1. AGE. Obtained from school record and checked up by questioning subject. In cases of doubt, the families were consulted or the age was based on certificates of birth or baptism.

2. NUMBER OF YEARS IN SCHOOL. Obtained by the teacher and checked up from school records and by questioning subject. The actual number of years in school was used, exclusive of absence on account of long sicknesses, etc. Time spent in parochial schools where foreign languages were taught was not counted as time in school, so that the number of years in school actually means the number of years in English speaking schools.

3. GRADE. Estimated by teacher in case of children in the ungraded special classes. The teacher estimated the grade that the pupil could enter if he were to be transferred to the regular grades. The basis of the teacher's judgment was the work of the pupil in the reading books, spelling books, arithmetic books, etc., standard for the different grades. All of the teachers had taught in the regular grades, and all of the pupils at one time

or another had been in the regular grades. The teacher's estimate was checked up from previous estimates independently made, and from the monthly and annual reports of the pupil.

4. NATIONALITY, LANGUAGE USED AT HOME, PLACE OF BIRTH. Obtained from teacher and checked up by questioning subject.

5. PHYSICAL DEFECTS. Obtained from record of medical examination.

93 boys in the special classes were given the complete examination. The subjects ranged in age from 9 to 16 with the majority at the ages 12, 13 and 14. This latter group of 59 boys 12, 13 and 14 years of age is used for comparison with a group of 58 boys of the same age in the regular grades. The first group will be spoken of as the retarded group, the second as the normal group. 4 members of the retarded group were candidates for admission to the special classes, the remainder being regularly enrolled in these classes.

The character of the normal and retarded groups may be studied by comparison with the general run of all children in school. The members of the two groups had been in school 5, 6, 7, 8 or 9 years. The grade progress distribution of all children in the Trenton schools who had been in school 5, 6, 7, 8 or 9 years is shown in Table I which is derived from the "Grade and Progress" table on page 104 of the Report of the Trenton Board of Education for 1914.

TABLE I.

Grade Progress Distribution of 4323 Trenton Children in School 5, 6, 7, 8 and 9 years.

GRADE	Number of Years in School.					Totals
	5	6	7	8	9	
I	3	1				4
II	25	10	4			39
III	175	45	5	4		232
IV	460	181	61	6	1	709
V	586	376	130	28	9	1129
VI	157	479	272	62	4	974
VII	22	108	374	169	37	710
VIII	5	25	145	233	118	526
Totals	1433	1225	994	502	169	4323

By inspection of Table I it may be seen that the largest proportion of the 4323 children are "at grade" or in the grade corresponding to the number of years that they have been in school, i.e. the fifth grade for those who have been in school five years, the sixth grade for those in school six years, etc. The percentage distribution of the children "at grade" and at each year above and below grade is as follows:

+3	+2	+1	"at grade"	-1	-2	-3	-4	-5
0.1%	1.1%	9.5%	38.8%	32.2%	13.5%	3.8%	0.8%	0.2%

This distribution is shown graphically in the top portion of Fig. 1. By far the largest proportion (80.5%) are "at grade" or one year above or below grade, the range of the distribution of the 58 normal subjects. 18.3% are two or more years retarded, the range of the distribution of the 59 retarded subjects. Only 1.2%

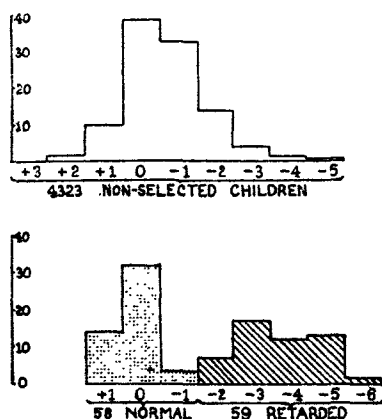


FIG. 1. Grade Progress Distribution of Normal and Retarded Subjects and of all Children in School 5, 6, 7, 8, or 9 years.

are two or more years ahead of their grade. This distribution can not be taken as typical of the entire school course, for there are no grades above the eighth included, so that there are no above grade groups possible for children 8 or 9 years in school, and no "at grade" group for those in school 9 years. The distribution has value only for comparison with the children actually tested. The grade progress distribution of the 117 subjects whose data are used in the subsequent report is shown in Table 2.

TABLE 2.

Grade Progress Distribution of 117 Subjects of This Investigation.

Grade	Years in School					Totals
	5	6	7	8	9	
I		1				1
II	1		3	1		5
III	1	3	4	5	1	14
IV		4	11	8	6	29
V			3	5	2	10
VI		2				2
VII		5	21	1		27
VIII			11	14	4	29
Totals	2	15	53	34	13	117

It may be seen by inspection that Table 2 is composed of two characteristically different groups, 59 boys who have been in school from 5 to 9 years and are in Grades I, II, III, IV, and V, and 58 boys who have been in school the same length of time but are in grades VI, VII and VIII. The percentage distribution of the children "at grade" and at each year above and below grade is as follows:

+1	"at grade"	-1	-2	-3	-4	-5	-6
13.7%	31.6%	4.3%	6.8%	17.1%	12.0%	12.8%	1.7%

This distribution is shown graphically in the lower part of Fig. 1, in which the dotted portion represents the normal group and the shaded portion the retarded group.

Comparing the distribution of the subjects tested with that of the general run of children, the range of the normal group, i.e. "at grade" or one year above or below grade, corresponds with 80.5% of the 4323 cases. The range of the retarded group corresponds with that of the lowest 18.3% of these 4323 cases. The two distributions are actually farther apart, for they represent the extreme samples of the children tending toward normal or accelerated grade progress and toward retardation. The largest proportion (86%) of the retarded group are 3 or more years retarded or within the range of the lowest 26% of all the 794 children two or more years retarded. The largest proportion (92%) of the normal group are either "at grade" or one year ahead of grade, a range falling within the highest 61% of the 3529 children one year retarded, "at grade" or accelerated.

The same facts concerning the extreme divergence of the two groups examined are brought out by a study of the ages in grades. From the "Age in Grade" table on page 113 of the Report of the Trenton Board of Education for 1914, the grade distribution of all 12, 13 and 14 year boys was obtained. From these data the average grade was computed for all 12, 13 and 14 year boys. The average grade of the normal and retarded groups was also computed at these ages. These averages are shown in Table 3.

TABLE 3.
Average Grade of 12, 13 and 14 year Boys.

Age	All Trenton Boys		Normal Group		Retarded Group	
	No.	Ave. Grade (MV)	No.	Ave. Grade (MV)	No.	Ave. Grade (MV)
12	656	5.21 (1.57)	18	6.94 (0.15)	18	3.22 (0.91)
13	644	5.66 (1.92)	20	7.50 (0.50)	21	3.90 (0.43)
14	354	6.85 (2.05)	20	7.90 (0.18)	20	3.95 (0.58)
all subjects			58	7.47 (0.53)	59	3.71 (0.72)

The average number of years that the retarded group have been in school is 7.42 yrs. (MV=0.83 yrs.), while the average number for the normal group is 7.28 yrs. (MV=0.61 yrs.). In spite of the fact that both groups have been in school the same length of time, the normal group averages over three and a half grades higher than the retarded group. The retarded group have gone just half as far as the normal group in the same length of time.

The age in grade distribution of the 117 subjects 12, 13 and 14 years of age is shown in Table 4.

TABLE 4.
Age in Grade Distribution of 117 Subjects.

GRADE	12	13	14	Total
I	1			1
II	4		1	5
III	5	5	4	14
IV	6	13	10	29
V	2	3	5	10
VI	2			2
VII	15	10	2	27
VIII	1	10	18	29
Total	36	41	40	117

Although the retarded subjects have been in school as long as the normal subjects, none of them are above the fifth grade. The largest proportion of the normal group (96.5%) are in the seventh and eighth grades.

Such then are the differences between the two groups. It will be seen subsequently that when tests are given to the two groups, some of the tests are equally easy or equally difficult for both groups, while others differentiate the groups sharply. In order to consider these tests as diagnostic of intelligence it is necessary to show that a real intelligence difference exists between the two groups. The writer's thesis is that the gross differences in school progress of the two groups indicate a real difference in intelligence. The retarded group although in school as long as the normal group have made only half the progress. The average retardation of three years and a half indicates an inferior intellectual endowment. To show this it is only necessary to prove that this retardation is not due to any other cause.

The most frequently mentioned cause of retardation other than intellectual defect is that of illness, which causes irregularity of attendance and consequent failure to keep up with the class-room work. As far as the writer was able to learn from the medical records, none of the retarded group had had any serious illness which kept him out of school any considerable length of time. Nor were there any uncorrected physical defects. The special classes contained children with physical defects such as very poor vision or deafness, but none of these cases was included in the retarded group. All of the special class children had annual eye examinations and the defects were corrected. As far as could be ascertained the members of the retarded group were physically normal.

Another frequently mentioned cause of retardation is an educational curriculum which is maladapted to the needs of the individual. Further than this it is claimed that the teaching is inferior and the lack of individual instruction a handicap. Faults in the educational regime would cause retardation in both groups of subjects, for both groups were exposed to the same regime. These causes of retardation would not explain why one group was retarded and the other not retarded.

Deficiency in the use of language is frequently given as the cause of the retardation of the children of non-English speaking parents. Since the retarded group contains 20 children of non-English speaking parents, it is possible that these children were retarded on account of language deficiency. The average number of years that these 20 children had been in school is 7.7 yrs. ($MV=0.73$ yrs.). The average grade of the group is 3.7 ($MV=0.72$). If this retardation of four years is to be attributed to language deficiency, it is legitimate to expect that the same factor would influence the normal group. The normal group contained 22 children of non-English speaking parents. The average number of years in school of this group is 7.36 yrs. ($MV=0.65$ yrs.), and the average grade 7.58 ($MV=0.52$). Instead of showing a retardation these boys show a slight advance. Although they have been in school the same length of time as the retarded group, they have progressed twice as far. It is not possible then to account for the differences between these two groups on the basis of language deficiency. Both groups had the same language handicap, and the difference in school progress must be referred to the inferior intellectual endowment of the members of the retarded group. The reactions of retarded and normal, English and non-English groups on the individual tests will be studied later to determine the influence of language training on the tests.

It is legitimate to conclude that the gross differences between the school progress of the retarded and normal groups indicate a difference in intellectual endowment. The members of the retarded group are retarded because they are less intelligent. The writer would not take an extreme position, and say that in all cases retardation indicates intellectual deficiency, or that presence in the proper grade indicates intellectual normality. He is willing to admit that a child may become seriously retarded by a particularly unfortunate combination of circumstances or by a lack of push or interest quite independent of his intellectual make-up. These cases are exceptional however. On the other hand, the writer would be willing to admit that subnormal children by a combination of fortunate circumstances may be pushed

ahead, and may find themselves in a grade altogether beyond their ability. The writer is certain that the retarded group contains some perfectly normal boys, and that the normal group contains some members that are intellectually inferior to some members of the retarded group. These cases are exceptional however, and are far from the general run. On the whole, the retardation of the backward group indicates inferior intellectual endowment. Tests which sharply differentiate the normal and retarded group are those which are diagnostic of intelligence.

This position may be extreme, and it is impossible to justify it on any other grounds than personal opinion. If the school measure were an adequate measure of intelligence there would of course be no need of mental tests. No one who has had experience in diagnosing mentality will ignore the school record entirely, nor will he stress it above all other measures. The distribution of ability given by the school measure is most certainly not the distribution of general ability, for the school gives practically no opportunity for intellectual precocity to show itself. The lack of confidence that experienced persons show in accepting the school record as an absolute indication of intellectual inferiority is also evidence that retardation does not always indicate intellectual backwardness. The writer believes however that on the whole there are certain large intellectual differences between the normal and retarded groups. It was not possible to account for the presence of any member of this group in the special classes for backward and defective children on any other grounds than mental defect, and the histories of the individuals were very carefully studied. With the same physical and environmental opportunities they have made half the progress of their brothers in the regular grades. The tests which differentiate the members of the retarded and normal groups are at least diagnostic of pedagogical retardation. On the basis of the extreme pedagogical retardation, the writer believes such tests to be diagnostic of intelligence.

It is not necessary in this study to classify the retarded group in terms of the number of normal individuals, morons and imbeciles. The giving of diagnoses would involve the presentation

of case histories that would require considerable space. The definition of the group in terms of school progress is sufficient. The retarded group by no means represents the lowest selection of the general population, for idiots do not enter school, and the low grade imbeciles are usually removed from school before they are 12, the lower age limit of the retarded group. A few cases were examined (not more than five) who presented the proper qualifications of age, parentage and scholastic standing to be included in the retarded group, but who were mentally so low that it was useless to put them through the series of advanced tests used. The selection of retarded subjects is therefore rather high. No member of the retarded group tested below the "mental age" of eight. The distribution of ability in the retarded group may be fairly said to range from that of the low grade moron to that of the normal individual. The group was composed mostly of border-line cases of feeble-mindedness. The lowest members of the group probably correspond to the highest grade of our present institutional cases. The group is therefore singularly well adapted for comparison with a normal group to determine what tests are most efficient in detecting the higher grades of mental defect.

III. TESTS AND PROCEDURE

The arrangement of tests used was that of Binet's 1911 scale (8). Town's (72) translation was used, but not followed strictly throughout. The translations of some of the tests were taken over from Goddard (28), as the freer translation seemed better adapted to the linguistic training of the subjects. The instructions given were partly those of Binet, partly those of Goddard, and partly those of the writer. Inasmuch as there were many departures from the standard form of procedure, the procedure used is given in detail below. The close student of standardized procedure will probably find several startling heresies in the account. In applying the tests, however, the writer was not seeking to find the "mental age" of the subject, or to obtain age norms for the tests. He was primarily interested in testing the tests by comparing the results of normal and retarded children. In order to discover the factors involved in each test, it was very necessary that the subject should understand the nature of the task he was to perform. For this reason a fore-exercise or practise test was given wherever possible. The important factors in the experiment were to be certain first, that the subject understood the nature of the task, and second, that the instructions were uniform for both groups of subjects.

The detailed account of the instructions illustrates the method of providing for the first factor. The retarded group was examined first so that the procedure is largely adapted to their needs. Although all of the Binet tests were used, only those are reported that were given to all the selected retarded and normal subjects. The list of Binet tests used, the procedure adopted, and the criteria for scoring each follow :

AGE III

No. 3. ENUMERATING OBJECTS IN PICTURES. Binet's original uncolored pictures were used. The question asked,—“Tell me what you see in this picture?” The question

was repeated for each picture. The picture of "A man and a boy" was shown first, of "A man and a lady," second, and "A prisoner," third. The responses were scored according to the instructions in the categories of enumeration, description, interpretation and emotion. No time limit.

No. 5. REPETITION OF SENTENCES. Instructions,—
"I am going to read you a sentence and I want to see how well you can remember it. Say it just as I say it. Don't leave out any words." The following list of sentences was used:

10—My name is William. Oh the naughty dog.

12—It rains in the garden. John has finished his task.

14—We are enjoying ourselves greatly. I have caught a mouse.

16—Let us go for a long walk. Give me the pretty, little bonnet.

18—Mary has just torn her new dress. I have given two cents to that beggar.

20—It is not necessary to hurt the birds. It is night, all the world rests in sleep.

22—We expect to have a great time at the seashore, digging in the white, beach sand all day long.

24—My little children you must work very hard for your living, you must go to school every day.

26—The other day I saw on the street a pretty, yellow dog. Little Maurice has stained his nice new apron.

28—Ernest is frequently punished for his bad conduct. I bought at the store a pretty doll for my little sister.

30—There was a severe storm last night with much lightning. My comrade caught cold, and he now has a high fever, and coughs a great deal.

32—The car is less expensive than the omnibus, it costs but two cents. It is strange to see women acting as coachmen in Paris.

All of the sentences were taken from Town's translation except No. 22, which was taken from Whipple's Manual (75) pg. 494. The word "Mary" was substituted for "Charlotte" in sentence No. 18, and in No. 24, the word "day" was used instead of

"morning," in order to make 24 syllables. The subject was scored according to the number of syllables in the longest sentence that he succeeded in repeating without error. After arriving at this "threshold," the subject was given the next two longer sentences, to make sure that he had reached his limit. Near the close of the experiment, this procedure was found to be entirely inadequate, so that the results from this test are unreliable. No time limit.

AGE V.

No. 3. See III, 5.

AGE VI.

No. 2. DEFINING IN TERMS OF USE. The words used were "fork," "table," "chair," "horse" and "mother," being given in the order named. The question asked was "What's a fork?" or "Tell me what a fork is." The responses were scored as definitions by use or superior to use according to the criteria given in the previous discussion of the personal equation (see page 19). No time limit.

AGE VII.

No. 2. DESCRIBING PICTURES. See III, 3.

AGE VIII.

No. 4. GIVING THE DAY AND DATE. An error of four days was allowed on either side of the day of the month. Scores were recorded for each part of the test, the day of the week, the month, the day of the month and the year. No time limit.

No. 5. REPEATING 5 DIGITS. "Say these figures just as I say them." 47395, 51742, 83964. One out of three scored. The experimenter's rate of giving the digits was somewhat faster than two per second, but was constant through habit.

AGE IX.

No. 1. MAKING CHANGE. "How much is six cents from a quarter?" The coins were not given. No time limit.

No. 2. DEFINING IN TERMS SUPERIOR TO USE. See VI, 2.

No. 3. NAMING PIECES OF MONEY. "Tell me all of the pieces of money you know." If the subject stopped at 50c, he was asked "What comes next?" In naming the bills, if he said "Three dollars, four dollars" etc., he was asked if he had ever seen a three dollar bill. None of the coins or bills were shown. No time limit.

No. 4. NAMING THE MONTHS. "Say the months for me." One error of omission or inversion was allowed. Time allowed, 15 seconds.

No. 5. COMPREHENDING EASY QUESTIONS. The following questions were asked verbatim:

- a. What would you do if you missed a train?
- b. What would you do if you had been struck by a playmate who didn't mean to do it?
- c. What would you do if you had broken something that didn't belong to you?

The order of presentation was *a, c, b*. No time limit.

AGE X.

No. 1. ARRANGING 5 WEIGHTS. Stoelting's standard cubes were used. Three trials were given and the number of successful attempts recorded. As a control on this test, a series of definitely supraliminal weights was introduced. The weights were made of metal salve boxes, and were weighted at 20, 30, 45, 70 and 100 gms, roughly in accordance with Weber's law, so that the subjective differences between the weights was approximately equal. One trial with these weights was given before the standard weights. If the subject failed, the failure was obviously due to the intellectual inability to comprehend a serial arrangement, or to make the logically necessary comparisons. If he passed, the test served as a "warming up" or practice test. Subsequent failure on the Binet weights would obviously be due to failure in sensory discrimination. The instructions were "Put these boxes in a row with the heaviest one first, and then the next to heaviest, and then the next, and then the next and then

the next." The experimenter pointed at a spot on the table for each box. The weights were not touched, and no suggestion was given for the subject to lift them. No time limit.

No. 2. COPYING DESIGNS FROM MEMORY. "I am going to show you this card with two drawings on it: I'm only going to show it to you for a little while, so I want you to study it hard and remember what it looks like." While these instructions were being given, the card was turned over as rapidly as possible, but in such a way that the subject caught a glimpse of it and understood the test better. The reproductions were scored at the time of the experiment, and then independently re-scored according to the arbitrary point system outlined in the previous discussion of the personal equation. The duration of the exposure was 10 seconds.

No. 3. DETECTING ABSURDITIES IN STATEMENTS. "I am going to give you a sentence that's got something foolish in it, and I want you to tell me what's foolish.—Now what's foolish in that?" Absurdity *d* was given first. If the subject failed to see the point, it was explained to him. After the first one, no explanations. The order of presentation was *d, c, a, b, e*, each test being separately scored. The following absurdities were given, the same wording being adhered to throughout. No time limit.

a. An unlucky bicycle rider fell off his bike and broke his neck. They took him to the hospital, and they don't think he'll get well.

b. I have three brothers, Henry, Robert and myself.

c. Yesterday, there was a railroad accident, but it wasn't a bad one, only 48 people were killed.

d. Yesterday, the police found a body of a girl cut up into 18 pieces. They think that she killed herself.

e. A man once said, "If I were going to kill myself, I wouldn't do it on Friday, for Friday is an unlucky day, and it might bring me bad luck."

No. 4. COMPREHENDING DIFFICULT QUESTIONS. The order of presentation used was *a, b, d, e, c*. The following questions were given, the same wording being adhered to throughout. No time limit.

- a. What would you do if you were delayed in going to school?
- b. What would you do before taking part in an important affair?
- c. Why is it easier to forgive a bad action done in anger than it is the same action done without anger?
- d. What would you do if you were asked your opinion about someone you didn't know very well?
- e. Why should you judge a person by his actions rather than by his words?

No. 5. USING THREE GIVEN WORDS IN A SENTENCE CONTAINING TWO IDEAS. A practice test was given as follows,—“I want you to make me up a sentence, a good sentence, with the words ‘boy, play and sled’ in it.” After the subject had given the sentence and had shown that he understood what was expected of him, the words “Trenton, money and river” were given. The subject gave the sentence orally. No time limit.

AGE XII.

No. 1. RESISTING SUGGESTION. (LENGTH OF LINES). The lines were accurately drawn by a draughtsman, on Bristol board 5 by 28 cm. The question asked for each of the first three cards was “Which is the longer line here?”, and for each of the last three, “And here?” The number of correct responses was noted.

No. 2. USING THREE GIVEN WORDS IN A SENTENCE CONTAINING ONE IDEA. See X, 5.

No. 3. GIVING 60 WORDS IN THREE MINUTES. “I want you to give me all the words you can think of in three minutes. Any old word will do, like ‘man-beard-boy-shirt-carriage,’ now go ahead.” The actual number of words was recorded.

No. 4. DEFINING ABSTRACT TERMS. “Charity, justice and kindness” were used. The question asked was “What’s charity?” or “Tell me what charity is?” A concrete illustration of an act was considered acceptable. No time limit.

No. 5. RECONSTRUCTING DISSECTED SENTENCES. “I am going to show you a card with a sentence on it that’s all mixed up. The words are in the wrong order, and I want you

to tell me what the sentence would be if it were in the right order." The following sentences were used, presented in the order *b*, *c*, *a*. One minute was allowed for each.

- a.* started-the-for-an-early-hour-we-country-at
- b.* a-defends-dog-good-his-master-bravely
- c.* asked-paper-the-to-I-teacher-correct-my

AGE XV.

No. 1. REPEATING 7 DIGITS. 2964375, 9285164, 1395847.

No. 2. GIVING THREE WORDS RHYMING WITH DEFENDER. As an introductory test, the subjects were asked to give rhymes with "day, mill and spring." An example was given, "man, ran, can, fan" and then the subject asked to give rhymes with "cat." If he failed, rhymes with "cat" were suggested until he either succeeded or failed utterly to comprehend. If he succeeded, he was then given the stimulus word in the order D, M, S, and allowed half a minute on each. At the end of this list, the stimulus word "defender" was given, and the subject allowed one minute.

No. 3. REPEATING A SENTENCE OF 26 SYLLABLES. See III, 5.

No. 4. INTERPRETING PICTURES. See III, 3.

No. 5. SOLVING PROBLEMS FROM VARIOUS FACTS. "I am going to tell you a story and I want you to tell me what's happened." The following wording was used in the problems:

a. A lady was walking through a park out in Chicago, when she suddenly stopped, very much frightened, turned around and ran back to the nearest police station, and there she told them that she had seen, hanging from the limb of a tree—What did she see?"

b. My neighbor has been having strange visitors lately. One after the other, a doctor, a lawyer and a priest (minister) have called at his house. What's happened there?"

The word "minister" was substituted for Protestant children. The only answer accepted to the first question was "a man hanged." to the second question, "someone's very sick," "just

died" or "is dead." The subject was cross-examined to make sure he had taken all the facts into consideration. If not, the test was failed. Question *a* was given first, *b* second. No time limit.

The method used by the experimenter in obtaining uniform procedure was that of practice. Before undertaking the Trenton experiment, the writer had examined over 150 children with Goddard's 1911 scale, that experience serving as an adequate practice series at least. The accuracy and uniformity of technique in giving the Binet tests rests entirely on the previous experience of the experimenter. The Binet tests used were, however, only a small portion of the complete examination given to the retarded and normal groups. The greatest difficulty was experienced in obtaining uniform procedure in the tests other than those in the Binet series. The method of obtaining uniformity of procedure in the supplementary tests is outlined in a subsequent section devoted to the discussion of these tests.

The complete series of tests was very long, and contained many tests such as that of repeating sentences that were not attractive to the subjects, so that great care had to be used by the experimenter in arranging the series so as to maintain the subject's interest. The complete series included besides the Binet tests described several special tests of memory, suggestibility, discrimination, reasoning ability and a long series of puzzle tests. The entire list of tests was given to only 153 of the 289 subjects examined.

The tests that were interesting depended largely on the subject. Any test that was obviously too difficult for a subject was uninteresting to him. The converse is usually true in practice also, that a test that is well within the subject's range of ability is interesting to him. The free play of thought on a novel situation is in itself pleasant if that situation offers a ready solution—the unpleasant tone arises with the attitude of doubt. For this reason the factor of cooperation in the mental examination of normal individuals is not as difficult as persons who have not attempted it would have us believe. The factor of cooperation in the examination of pure defective types is by no means an insurmountable difficulty, if the examiner has tact and discretion.

It is only in cases of mental alienation that the factor occasionally becomes an insurmountable difficulty, and none of the subjects in the finally selected groups presented any developed aberrational tendency.

When an individual is examined by the incomplete method of testing, the experimenter can keep the questions well within the subject's range of ability, and by a judicious selection of the questions can maintain his co-operation without difficulty. In this experiment, however, all the questions were given to every subject, and it was necessary in the case of most of the retarded group to hammer them through a long series of questions that were obviously impossible for them to answer, so that the experimenting was exceedingly difficult.

The duration of the complete examination was from an hour and a half to two hours depending on the individual. The experiment was always broken up into periods usually of half an hour each. No subject was ever kept over 45 minutes, and only that long in exceptional cases. The most difficult questions were given at the beginning of the period. If the subject showed signs of fatigue, he was immediately dismissed. Although the character and duration of the examination was such that the factor of cooperation would appear to be very large, the writer found little difficulty from this source. The special classes were composed of about 15 children. The plan of attack in beginning a new class was to have the teacher send the most popular boy in the room first, and the experimenter would do his best to show him a good time. The puzzle series and similar tests proved an unfailing source of pleasure. The first examination of the most popular boy contained the easiest and most interesting tests of the series, and the decoy invariably worked. In every class except one, the teacher would send the boys to the examiner as a reward for good conduct in the class room. The experimenter in turn saved the puzzle tests and easy tests as a reward for careful work in the more difficult tests, always saving some of the attractive tests for the close of the examination so that the subject would return to his room in good spirits. The examinations were conducted for the most part during the school hours, and it was found that the subjects were glad enough to avoid their lessons

for a more attractive occupation. The examination was never allowed to conflict with the gymnasium hours. When an examination was conducted outside of school hours, the subjects were paid at the rate of twenty cents an hour. If the subjects showed the slightest lack of interest or the slightest unwillingness to return for a second examination, the experiment was abandoned. In one class after three or four boys had been successfully examined, one boy returned with an unfavorable verdict, the rest of the pupils showed a slight unwillingness to come, and no further experimenting was done. On the whole, the greatest difficulty was experienced not in enticing the boys to come, but in arranging waiting lists and keeping the boys out of the room when they were not wanted. Very little difficulty was experienced from the factor of cooperation.

The conditions under which the examinations were made were not ideal, but were uniformly satisfactory in every case. A separate room was always provided. A table was arranged so that the light coming from behind the subject fell directly on it. In a few cases, the experimenter could not prevent the presence of a third person in the room. If this proved a distraction, the examination was discontinued, or if the person wished to observe the test, the examination was completed and the record discarded.

The factor of information was encountered, but the nature of the tests was such that information played little part. The examination was so varied, and so long, that the subject would have difficulty in remembering any but a few striking tests. The test of naming 60 words gave the most difficulty on this score, the subjects invariably telling the experimenter that they were "all loaded" for this test. The special classes were in various schools throughout the city, but the normal pupils were all examined in one school, so that the factor of information would be more in their favor. As a matter of experience, however, the experimenter had no difficulty in finding out whether or not the subject had any previous information. The record on any test was discarded if this factor were discovered. The results of the 60 word test are given, however, with this reservation. In any test in which the presence of this factor was suspected, the writer will report it in the discussion of that test.

IV. VARIABLE FACTORS

When the results of the normal and retarded groups are compared, it is seen that some tests are equally easy or equally difficult for both groups, while other tests differentiate the groups sharply. In order that the latter type of test may be considered diagnostic of intelligence, it is necessary to show that the difference in the reaction of the two groups may not be attributed to any other factor than intelligence. In the preceding study, several variable factors were found which might influence the outcome of the tests, and these factors will be discussed in conjunction with those of sociological conditions and language training. The variable factor of sex differences of course drops out, since only boys were included in this investigation.

THE ERROR DUE TO INCOMPLETE DATA.

The error due to incomplete data can have no effect on the results, for practically all the tests were given to all the subjects of both groups. From time to time a test would be dropped out because it was incorrectly given, or a test would be accidentally omitted, but on the whole the tests were given very completely. Taking the Binet tests, it was possible to give to the retarded group 3363 tests. Of this number, 3314 tests or 98.5% were given. The normal group were given 3280 tests or 99% of the 3306 tests possible to be given. Question *a* of the comprehension questions was given to the normal group only 74% of the possible number of times. The date test was given to this group 95% of the time, and all the other tests were given over 98% of the possible number of times. The retarded group were given the date test 86% of the time, parts *d* and *e* of the definitions test 92% and 88% of the time, and all other tests over 97% of the possible number of times. The testing may be called complete, then, and the results cannot be influenced by the error due to incomplete data.

THE INFLUENCE OF GRADE TRAINING.

The results of the tests may be influenced by grade training, but there is no way to determine the influence of this factor. The only difference between the training of the children in the special classes and in the regular grades, consists in the larger amount of time devoted to manual training in the special classes. Very nearly a fifth of the time is given to work of this nature. The rest of the time is devoted to regular school work, and the only difference between the special classes and the regular grades in this work is that the former receive individual help and instruction almost entirely. They have more actual training in the school subjects than the children in the regular grades but they do only half as well. The difference in school standing between the normal and retarded groups must be attributed to a difference in intellectual endowment, for both groups had had an equal amount of training along the same lines. Regardless of how they took the training, the retarded group had been exposed to it as long as the normal group.

THE INFLUENCE OF THE PERSONAL EQUATION

The differences between the two groups may be due to the influence of the experimenter's personal equation, but there is very little check on this factor due to the fact that there was only one experimenter, the writer. The test of copying the designs from memory was ranked plus or minus by the writer at the time of the examination, and later scored according to the arbitrary point system described in Chapter III of the first study (see pages 23 to 26). It is possible to compare the experimenter's judgments at the time of the examination with the later arbitrary scoring which was made without knowledge of the original rank given. According to the experimenter's first judgments, 69% of the normal group and 41% of the retarded group passed this test. According to the point system of scoring in which 15 points is used as the passing mark, 67% of the normal group and 31% of the retarded group pass the test. The score of the normal group is reduced 2% by correction, the score of the retarded

group, 10%. The passing mark for each group was calculated as in Chapter III of the first study, and was found to be 14 points for the normal group and 13 points for the retarded group. The experimenter marked as failed in the normal group one design scoring over 15 points, and marked as passed two designs scoring under 15 points. In the retarded group, on the other hand, the experimenter marked as failed two designs scoring over 15 points, and marked as passed nine designs scoring under 15 points. The experimenter was therefore more lenient with retarded than with normal children in the test of copying designs from memory.

It is not possible to obtain a quantitative estimate of the influence of the personal equation in any of the other tests. It is the opinion of the writer that if the personal equation influenced the results at all, the retarded group were favored. The experimenter examined 266 children in the special classes before examining the normal children. The work of the special class children was uniformly so low that comparatively poor answers to some questions would seem to be very good. The tests of detecting absurdities in statements and comprehending difficult problem questions were rarely answered in the special classes. An answer that would have been considered doubtful if given by a normal subject, would probably have been ranked plus if given by a member of the retarded group, owing to the contrast of his answer with those of the other retarded children. As far as possible, the experimenter used a uniform system of testing and scoring, but the criteria for judging some tests are at best indefinite, and are therefore susceptible to the recent experience of the experimenter. The writer is not conscious of favoring the retarded group, but this tendency might have been present. Considerable tact and patience was demanded in giving the difficult tests to subnormals, and the writer is certain that if the tests had been given to the special class children after the subjects in the regular grades had been examined, he would have been much more abrupt and less patient.

The fact that there was only one experimenter would tend to minimize the influence of the personal equation as it would make

impossible any large differences of procedure or technique. In spite of this fact however it is possible for this influence to be present. The Binet tests are scored by the all or none method, they are ranked either plus or minus, while the responses on many of the tests are not all or none responses. The 60 word test, for instance, is ranked plus if the subject gives 60 words or over, and minus if he gives under 60 words. The responses of the subjects of this investigation actually varied from 31 to 196 words. The degree of merit on this test may vary through 165 steps by steps of 1, while the expression of merit used is merely plus or minus.

The tests vary in the degree in which the responses may be accurately rated by the all or none scoring. Some tests, such as that of making change, have an all or none response—they are either right or wrong. Other tests may admit of a slight grading of response. The 5 weight test has four grades of response according as the subject arranges the weights correctly 0, 1, 2, or 3 times. At the other extreme are tests such as the design test that admit of at least 20 grades of response. Some tests would admit of considerable grading, but there is no available method for grading them. The responses to the test of using three given words in a sentence containing one idea, for instance, would vary from "Trenton's river costs money" and "Trenton has lots of money on the river" to "Trenton paid a large sum of money to have the Delaware river deepened" and "The people in Trenton that have money live along the river." It is conceivable that the responses to this test, and to other tests such as defining in terms superior to use, defining abstract terms, and some of the absurdity and comprehension questions could be arranged on a scale of merit from 0 to 10.

Given a scale of merit from 0 to 10, the experimenter must express his judgment by plus or minus. At some point on this scale there is bound to be a range of uncertainty—a range of play for the personal equation. In the preceding investigation it was found that Experimenter C changed his criteria in ranking the definitions test during the course of the experiment (see

page 20), so that it is possible for the personal equation to enter even with one experimenter. If the experimenter's judgment changed in examining normal and retarded subjects, the diagnostic value of the tests would be influenced. If the experimenter were too lenient with retarded and too severe with normal children, the diagnostic value would be less than it should be. The design test showed this influence, and the diagnostic value was lowered 8% by a variation in judgment of 1 point on a scale of merit of 20 points. If the opposite tendency were present—if the experimenter were too severe with retarded and too lenient with normal subjects, the diagnostic value of the test would be exaggerated. It is possible that this should happen, for a range of uncertainty is present, and the influence of the personal equation too subtle to be eliminated entirely.

It is possible then for a person to maintain that the differences in the reactions of the groups which are used to determine the diagnostic value of the tests are merely expressions of the personal equation of the experimenter. The safest position to take probably lies between the two extremes—to bear in mind that the diagnostic value may be thrown one way or another by the personal equation, but also to bear in mind that more of the responses are bound to fall in the range of certainty than in the range of uncertainty, and that therefore the personal equation can not vitiate the results entirely.

THE INFLUENCE OF SOCIOLOGICAL STATUS.

In order that the differences between the performance of the retarded and normal groups on the tests might not be referred to environmental conditions and home training, the writer was very careful to select subjects of the same sociological status. The sections of the city in which the subjects lived were very much the same, the congested districts around the large manufacturing centers. The writer has listed the occupations of the fathers of the boys just as they were given in the school records. No attempt has been made to classify the occupations. The reader may glance over the lists, and form his own opinion on the similarity or dissimilarity of the groups.

Seven of the fathers of the boys in the normal group were dead. The occupations of the fathers of the other boys were as follows,—4 potters, 3 wire-drawers, 3 laborers, 2 foremen over laborers, 2 bakers, 2 policemen, 2 machinists, 2 carpenters, 2 firemen, 2 tailors, 2 contractors, 2 merchants, and one each of the following, milkman, weight-master, decorator, janitor, superintendent of a pottery, manager of a chemical company, mason, hardware dealer, engraver, pattern-maker, shoe-maker, grocer, box-maker, laundry driver, printer, furnace tender, pipe fitter, railroad yard master, boilermaker, iron worker, brass worker, iron moulder and foundry worker.

Eight of the fathers of the boys in the retarded group were dead and two were invalids. The occupations of the fathers of the other boys were as follows, 7 laborers, 2 foremen over laborers, 8 potters, 2 peddlers, 2 contractors, 2 teamsters, 2 fish merchants, 2 blacksmiths, 2 painters, 2 masons, and one each of the following, tinsmith, watchman, fruit dealer, electrician, carpenter, plumber, carriage-maker, junk dealer, piano tuner, wire tinner, dairyman, meat packer, boilermaker, huckster, crockery dealer, railroader and machinist.

THE INFLUENCE OF THE LANGUAGE FACTOR.

The statement that the Binet tests depend on language training appears frequently in the literature of the subject. Investigators frequently refer the failure of their subjects on certain tests to 'deficiency in this sort of training or experience. Although investigators and critics of the scale frequently mention the language factor, no one has actually given a demonstration of the influence of this factor by comparing children of the same mental status but different linguistic training.

In this investigation it is necessary that this be done, for, in order to refer the differences found between the two groups to the relative diagnostic value of the tests, it must be shown that the differences are not due to the language training of the two groups. The subnormal group contains several children of non-English speaking parents. It is possible to compare this group with other subnormal children of the same age, who have been

in school the same length of time, and who are in the same grade, but who are children of English speaking parents. The two groups are objectively the same except that one group has had the advantage of the English language in the home, while the other has not. It is also possible to compare a group of normal boys of non-English speaking parents with normal boys of English speaking parents, the groups being similar in regard to age and school progress. If any differences occur between these groups, they may be referred to the influence of language training in the homes. The influence of the personal equation is absent, for the members of the groups were examined by one person at the same time. The influence of grade training is absent for the members of the groups compared have been in school the same number of years, and are in the same grades.

22 boys, aged 12, 13 and 14, of non-English speaking parents in the special classes were given the complete examination. They had been in school from 5 to 10 years, and were in grades II, III, IV, V and VI. To compare with this group, 22 boys of English speaking parents were selected who had been in school approximately the same length of time and were in the same grades. The average age of the 22 non-English retarded children was 13.26 yrs. ($MV=1.11$ yrs.) The average age of the 22 English retarded children was 13.10 yrs. ($MV=1.01$ yrs.) The average number of years in school of the non-English retarded children was 7.59 years. ($MV=1.05$ yrs.), of the English retarded group 7.09 yrs. ($MV=1.07$ yrs.) The average grade of both groups was the same, 3.68 ($MV=0.80$.) The average age of the non-English group is slightly higher, and they have been in school half a year longer on the average.

Twenty boys, aged 12, 13 and 14, of non-English speaking parents were examined in grades VI, VII and VIII. To compare with this group, 20 boys of the same age, grade and number of years in school, but of English speaking parents were selected. The average age of the non-English normal group was 13.58 yrs. ($MV=0.72$ yrs.). The average age of the English normal group was 13.53 yrs. ($MV=0.73$ yrs.). The average number of

years in school for both groups was the same, 7.30 yrs. ($MV=0.62$ yrs.). The average grade of both groups was the same, 7.45 ($MV=0.61$).

The results of the four groups of children were tabulated, and the percentage that each group passed each test calculated. The results are shown in Table 5. Column A gives the per cent. that each test was passed by the 40 children of English and non-English speaking parents combined. Column B gives the per cent. that the non-English group are above (+) or below (—) the English normal group. Column C gives the per cent. that each test was passed by the 44 retarded children of non-English and English speaking parents combined. Column D gives the per cent. that the non-English retarded group are above (+) or below (—) the English retarded group. Column E gives the difference between columns A and C or the per cent. that the 44 retarded pupils are above (+) or below (—) the 40 normal pupils.

TABLE 5.

Percentage Differences between Normal and Retarded Children of English and Non-English Speaking Parents.

	A	B	C	D	E
	% passed Normal.	Normal non-Eng. ± Eng.	% passed Retarded	Retarded non-Eng. ± Eng.	Retarded Normal.
Comparing remembered objects.....			97	— 6	
Counting backwards from 20 to 0.....			83	+22	
Indicating omissions in pictures.....			86	— 6	
Giving day and date.....	100	0	64	+16	—36
Enumerating the months.....	95	0	63	+ 8	—32
Naming the pieces of money.....	100	0	88	+13	—12
Making change	100	0	79	+22	—21
Arranging five weights.....	75	0	61	+ 5	—14
Copying designs from memory.....	65	—20	23	0	—42
Repeating five digits.....	100	0	95	0	— 5
Repeating seven digits.....	68	—25	27	— 9	—41
Using three words in a sentence. 2 ideas	100	0	59	+28	—41
Using three words in a sentence. 1 idea.	80	+20	43	+23	—37
Resisting suggestion	40	0	14	+27	—26
Naming 60 words in 3 minutes.....	90	+20	60	+33	—30
Giving 3 rhymes with "defender".....	0	0	0	0	0
Comprehending easy questions					
a. Train	100	0	98	— 5	— 2
b. Playmate	93	+ 5	61	— 6	—32
c. Broken	100	0	100	0	0
Any 2 out of 3.....	100	0	100	0	0

TABLE 5 (continued)

	A	B	C	D	E	
Comprehending difficult questions						
a. Delayed	3	- 7	16	+ 4	+13	
b. Important affair	88	+ 5	14	- 9	-74	
c. Forgive easier	48	-15	9	- 9	-39	
d. Asked opinion	98	+ 5	50	-10	-48	
e. Actions vs. words	98	+ 5	20	-23	-78	
Any 3 out of 5.....	88	+ 5	20	- 5	-68	
Detecting absurdities in statements						
a. Bicycle rider	83	+ 5	36	-18	-47	
b. 3 brothers	70	0	18	- 9	-52	
c. Railroad accident	95	0	48	-32	-47	
d. Suicide	93	- 5	64	0	-29	
e. Friday unlucky	78	-15	11	-23	-67	
Any 2 out of 3.....	93	- 5	30	-23	-63	
Defining terms superior to use						
a. Fork	87	- 6	30	+ 6	-57	
b. Table	80	-22	40	+16	-40	
c. Chair	74	-12	37	+ 2	-37	
d. Horse	85	- 1	26	+ 9	-59	
e. Mother	67	-17	23	-21	-44	
Any 3 out of 5.....	77	- 6	33	- 1	-44	
Defining abstract terms						
a. Charity	45	+10	16	+ 4	-29	
b. Justice	45	-20	5	0	-40	
c. Kindness	48	-25	23	- 9	-25	
Any 2 out of 3.....	48	-15	7	- 4	-41	
Reconstructing dissected sentences						
a. early-hour	83	+15	9	0	-74	
b. teacher-correct	95	0	39	+23	-56	
c. dog-master	93	-15	32	+ 9	-61	
Any 2 out of 3.....	95	-10	27	+18	-68	
Repeating sentences of						
18 syllables	58	+ 5	31	-24	-27	
20 syllables	15	-30	0	0	-15	
22 syllables	3	- 5				
24 syllables	61	-37				
26 syllables	0	0	0	0	0	
Solving problems from various facts						
a. Hanging from a limb.....	50	0	20	-23	-30	
b. Neighbor's visitors	73	- 5	11	- 5	-62	
Both correct	43	+ 5	7	-14	-36	
Responses to pictures						
No. 1 {	Description	70	-20	73	+18	+ 3
	Interpretation	33	-15	30	+ 5	- 3
	Emotion	3	- 5	0	0	- 3
No. 2 {	Description	70	-20	66	+14	- 4
	Interpretation	18	- 5	16	-14	- 2
	Emotion	10	0	9	- 9	- 1
No. 3 {	Description	70	-20	75	+22	+ 5
	Interpretation	33	-15	30	+13	- 3
	Emotion	3	+ 5	0	0	- 3
Summary of picture test						
Describing 2 out of 3 pictures						
(Age VII)	70	-20	75	+14	+ 5	
Interpreting 2 out of 3 pictures						
(Age XV)	23	-15	20	- 5	- 3	

The 62 differences between the normal English and the non-English groups, shown in column B, vary from -37% (i.e. 37% in favor of the English group) to $+20\%$ (i.e. 20% in favor of the non-English group). The median of the differences is 0% ($Q=7.5\%$). The average difference is -5.53% ($MV=9.33\%$). In the long run then the normal English group is about 5% above the normal non-English group.

The 63 differences between the retarded English and non-English groups, shown in column D, vary from -32% to $+33\%$. The median of the differences is 0% ($Q=11.5\%$). The average of the differences is $+0.83\%$ ($MV=11.22\%$). In the long run then the retarded non-English group are slightly better than the retarded English group.

From the figures given it is necessary to draw conclusions concerning what tests are influenced by language training, but this is very difficult on account of the lack of correspondence between the results of the two groups of English and non-English subjects. If one general cause, the language factor, were in operation in producing the divergencies in the results, then the results of the two groups should show a high correlation. The correlation (Spearman foot-rule method) of the differences between the normal English and non-English groups and the retarded English and the non-English groups (i.e. the correlation between columns B and D) is -0.06 ($pe=0.055$) or no correlation. This would indicate that the differences were due to chance rather than to the one general factor, language training.

Inasmuch as the number of subjects (20 or 22) in each group is small, there is a very strong possibility that the differences might be due to chance. A glance at columns B and D shows that in some cases the results agree, and in other cases they are exactly opposite. In the test of defining terms superior to use, for example, the non-English group are 17% below the English group in the case of the normal subjects and 21% in the case of retarded subjects in defining "Mother." In defining "Table," however, they are 22% below in the normal group but 16% above in the retarded group. The frequency of occurrence of

like and unlike signs in columns B and D is as follows: oo occurs 5 times; o+, 10 times; o—, 10 times; +—, 21 times; ++, 3 times and ——, 11 times. If the results were due to chance, the number of unlike signs should be twice that of like signs. The actual number of like signs (oo, ++, and ——) is 19, and the actual number of unlike signs (o—, o+ and +—) is 41, the proportion expected. Taking only the plus and minus signs, they are like in 14 cases and unlike in 21 cases. These figures would also indicate that the results were due to chance, but they are not conclusive, for the magnitudes of the differences are not taken into consideration.

The results in general would indicate that no one factor was in operation in producing the differences found, but the possibility remains that the language factor might influence some of the individual tests. If this factor were in operation on the individual tests, then the sum of the differences between the normal English and non-English groups and the retarded English and non-English groups would indicate what tests were influenced, for those differences with unlike signs would cancel out in combination, while the differences with like signs or a common tendency would be exaggerated. Combining the differences in this manner (i.e. taking the algebraic sum of the differences shown in columns B and D) the 60 sums of differences obtained vary in magnitude from —38% to +53%. The average of the sums of differences is —4.32% (MV = 13.29%). There is no method of deciding which of the tests are influenced by language training. The lists of tests are given below so that the reader may form his own opinion. The tests found in the highest 25% in favor of the non-English groups, and in the highest 25% in favor of the English groups are given.

THE 15 TESTS SHOWING THE LARGEST SUMS OF DIFFERENCES
IN FAVOR OF THE ENGLISH GROUPS.

- 38% Absurdity e. (Friday unlucky)
- 38% Definition e. (Mother)
- 34% Abstract definition c. (Kindness)
- 34% Repeating 7 digits.
- 32% Absurdity c. (Railroad accident)

- 30% Repeating sentence of 20 syllables.
- 28% Passing 3 out of 5 absurdities.
- 24% Comprehension c (forgive easier).
- 23% Problem a (Hanging from a limb).
- 20% Interpreting 2 out of 3 pictures.
- 20% Abstract definition b (Justice).
- 20% Copying designs from memory.
- 19% Interpreting picture 2.
- 19% Repeating a sentence of 18 syllables.
- 19% Passing 2 out of 3 abstract definitions.

THE 15 TESTS SHOWING THE LARGEST SUMS OF DIFFERENCES
IN FAVOR OF THE NON-ENGLISH GROUPS.

- +53% Giving 60 words.
- +43% 3 words in sentence (1 idea).
- +28% 3 words in a sentence (2 ideas).
- +27% Resisting suggestion.
- +23% Dissected sentence b (teacher-correct).
- +22% Making change.
- +16% Naming date.
- +15% Dissected sentence a (early-hour).
- +14% Abstract definition a (Charity).
- +13% Naming money.
- + 8% Definition d (horse).
- + 8% Passing 2 out of 3 dissected sentences.
- + 8% Enumerating the months.
- + 5% Emotional interpretation of picture 3.
- + 5% Arranging 5 weights.

The problem now arises of how large a difference may be taken to indicate the influence of the language factor. This problem has no answer outside of personal opinion. Taking the differences in favor of the English groups, the test of copying the designs from memory shows a difference of 20% in favor of these groups, but it is hard to see how this test can be influenced by language training. Again, the test of repeating 7 digits shows a difference of 34% in favor of the English groups. It is also hard to understand why this test should involve the language factor as both groups had on an average over seven years of experience in using digits. Taking the results that are in favor of the non-English speaking groups, the date test shows these children ahead 16%, the test of making change 22% and

the line suggestion test 27%. It is difficult to understand how any of these tests may be influenced by language training. Using the tests enumerated as a limit, it might be concluded that some of the absurdities tests and some of the definitions tests show the influence of language training in favoring children of English speaking parents. If this conclusion is drawn, however, it is also necessary to conclude that the tests of naming 60 words and constructing a sentence from three given words show a larger influence of language training in favoring the children of non-English speaking parents. This conclusion is certainly possible, for the training of this group of subjects in two languages may be a positive help. The reader may draw his own conclusions.

Although no definite conclusions may be drawn concerning the presence of the language factor, it is possible to estimate the importance of this factor in the present investigation by comparing it with another factor, that of the intellectual differences between the groups. Columns A and C show the per cent. that the 40 normal children and the 44 retarded children (English and non-English combined) pass the individual tests, and column E indicates the percentage difference between these groups. The 60 differences between the normal and retarded group vary from -78% (i. e. 78% in favor of the normals) to $+13\%$. The average difference is -30.48% ($MV=20.72\%$). The largest difference between the English and non-English groups is 37% or only 7% higher than the average difference between the normal and retarded groups. The average difference between the English and non-English normal groups is about 5% or one sixth of the average difference between normal and retarded children. It is therefore possible to conclude that the language factor has very little importance as compared to the intellectual differences between the groups.

In the present investigation, the conclusion that the language factor has very little importance as compared to the factor of the intellectual differences between the two groups does not mean that the language factor plays no part in the Binet tests, but that it may be disregarded in this study. It was not possible to demonstrate whether the differences found between the language groups were due to chance or training. The absence of

correlation between the results of the retarded and normal language groups would indicate that the differences were due to chance. Certain of the individual tests may show the influence of this factor, but this influence may be disregarded for it should be the same in both groups, the normal and retarded. Any effect of language training would be equal in both groups, for both groups contain approximately the same number of children of non-English speaking parents, both groups come from the same sort of homes, and both groups have had the same amount of linguistic training in the homes and in the schools. The differences may be related to the intellectual differences between the groups, and these differences may therefore be used as indices of the diagnostic value of the tests.

V. DIAGNOSTIC VALUE OF BINET TESTS

In the Introduction it was shown that the Binet scale is not a reliable instrument for diagnosing the higher grades of mental defect, for, as suggested by Binet (7) and demonstrated by Descoeudres (20) and Chotzen (18), it is composed of some tests that are effective and others that are ineffective in diagnosing intelligence. To determine what tests were most effective, the writer gave the tests to two groups of children who had had the same physical and environmental opportunities but who showed different school progress. It was not possible to account for the fact that one group had progressed only half as far in school as the other without assuming a difference in the general intellectual endowment of the groups. The differences in the reactions of the groups to the tests could not be due to differences in sociological status or school training, nor could the personal equation or the language factor influence the results to any extent. These differences may therefore be referred to the intellectual differences between the groups, and the magnitude of the differences used as indices of the diagnostic value or effectiveness of the tests.

The logic of this method of measuring the effectiveness of the individual tests is the same as that of the method which Binet (5) proposed for estimating the importance of stigmata in the diagnosis of subnormality. According to Binet's proposed method of calculation, if a certain stigma were always found among subnormals and never among normal individuals, this stigma would have the value of 100%. Another stigma found among all subnormals and 50% of normals would have a value of 50%. In this way the principle of calculation which Binet proposed would attach to each stigma its "coefficient d'importance," and the relative certainty of these diagnostic indices would be measured on a scale of 100. The present investigation simply reverses the process. If a certain test ability were present in all normals and not present in any subnormals, its diagnostic value would be

100%. If any ability occurs with the same frequency in normals and subnormals, the percentage performance of each group would be the same and its diagnostic value would be 0.

The per cent. that each test was passed by the 58 normal and 59 retarded subjects, and the differences between these percentages, or the diagnostic value of each test, are shown in table 6.

TABLE 6.
Diagnostic Value of Each Test.

	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value
Giving day and date.....	100	71	-29
Enumerating the months.....	98	55	-43
Naming the pieces of money.....	100	91	-9
Making change	98	83	-15
Arranging five weights	74	56	-18
Copying designs from memory.....	67	31	-36
Repeating five digits	100	98	-2
Repeating seven digits	64	34	-30
Using three words in a sentence (2 ideas).....	100	71	-29
Using three words in a sentence (1 idea).....	79	56	-23
Resisting suggestion	33	21	-12
Naming 60 words in 3 minutes.....	93	66	-27
Giving three rhymes with defender.....	0	0	0
Repeating a sentence of 26 syllables.....	0	2	+2
Comprehending easy questions			
a. Train	100	98	-2
b. Playmate	95	72	-23
c. Broken	100	100	0
*Any 2 out of 3.....	100	100	0
Comprehending difficult questions			
a. Delayed	7	17	+10
b. Important affair	88	12	-76
c. Forgive easier	50	14	-36
d. Asked opinion	98	47	-51
e. Actions vs. words	91	17	-74
*Any 3 out of 5.....	86	15	-71
Detecting absurdities in statements			
a. Bicycle rider	90	48	-42
b. 3 brothers	71	25	-46
c. Railroad accident	95	54	-41
d. Suicide	91	68	-23
e. Friday unlucky	81	19	-62
*Any 3 out of 5.....	95	42	-53
Defining terms superior to use			
a. Fork	84	35	-49
b. Table	84	46	-38
c. Chair	81	39	-42
d. Horse	86	33	-53
e. Mother	72	29	-43
*Any 3 out of 5.....	84	33	-51

TABLE 6 (continued)

	Per cent. passed by normal	Per cent. passed by retarded	Diagnostic value
Defining abstract terms			
a. Charity	53	12	-41
b. Justice	47	5	-42
c. Kindness	48	24	-24
*Any 2 out of 3.....	59	8	-51
Reconstructing dissected sentences			
a. early-hour	84	14	-70
b. teacher-correct	93	39	-54
c. dog-master	97	32	-65
*Any 2 out of 3.....	100	29	-71
Solving problems from various facts			
a. Hanging from a limb.....	53	32	-21
b. Neighbor's visitors	74	23	-51
Both correct	45	10	-35
Responses to pictures			
No. 1 { Description	74	71	- 3
{ Interpretation	33	24	- 9
{ Emotion	2	2	0
No. 2 { Description	69	73	+ 4
{ Interpretation	19	12	- 7
{ Emotion	12	10	- 2
No. 3 { Description	74	78	+ 4
{ Interpretation	28	34	+ 6
{ Emotion	3	0	- 3
Summary of pictures test			
Describing 2 out of 3 pictures (Age VII)...	74	76	+ 2
Interpreting 2 out of 3 pictures (Age XV)...	21	15	- 6

A glance at table 6 shows the large variation in the diagnostic value of the tests. The tests of reconstructing dissected sentences, for example, show differences between the groups of 54%, 65%, 70% and 71%, while the test of resisting suggestion shows a difference of but 12%. The first test is passed by all the normal group, but by only 29% of the retarded group. The second test is passed by 33% of the normal and 21% of the retarded groups. Both of these tests are "twelve year" tests. The suggestion test, although quantitatively equal to the sentence test, is almost equally as difficult for both groups, while the sentence test is universally passed by one group and is difficult for the other. These figures would show then that the test of recon-

*The scores "Any 2 out of 3" or "Any 3 out of 5" refer to Binet's method of counting certain tests passed if the subject passes a certain number of the parts of the test.

structing dissected sentences is highly diagnostic of intelligence, while the line suggestion test has no value as an intelligence test.

The differences between the normal and retarded groups vary from -76% to $+10\%$, the median being -29% ($Q=23\%$). At first thought it might seem that the tests could be arranged immediately in the order of their diagnostic value on the basis of the figures given in table 6. More careful study shows that the method, although conclusive concerning certain tests, is inconclusive concerning others. The truth of this statement is very clearly shown by referring to the tests that stand out at the extremes of the list—the tests that show the highest and the lowest diagnostic values.

The following list contains all the tests that show a diagnostic value over 50% :

- 76 Comprehension b (Important affair).
- 74 Comprehension e (Actions vs. words).
- 71 Any 3 out of 5 comprehension questions.
- 71 Any 2 out of 3 dissected sentences.
- 70 Dissected sentence a (early-hour).
- 65 Dissected sentence c (dog-master).
- 62 Absurdity e (Friday unlucky).
- 54 Dissected sentence b (teacher-correct).
- 53 Any 3 out of 5 absurdities.
- 53 Definition d (Horse).
- 51 Any 3 out of 5 definitions superior to use.
- 51 Problem b (Neighbor's visitors).
- 51 Any 2 out of 3 definitions of abstract terms.
- 51 Comprehension d (Asked opinion).

The 14 tests in the above list are parts of six tests, comprehending difficult questions, reconstructing dissected sentences, detecting absurdities, defining in terms superior to use, defining abstract terms and solving problems. Turning to these tests in table 6, it is seen that their diagnostic value is in general very high. This would indicate then that these tests were the most effective ones in the scale for differentiating the groups in question.

Conclusions as definite can not be drawn concerning all the tests at the other extreme. The following list contains the tests that show a diagnostic value under 10% :

- + 10 Comprehension a (Delayed).
- + 6*Interpretation. Picture 3.
- + 4*Description. Picture 3.
- + 4*Description. Picture 2.
- + 2*Describing 2 out of 3 pictures (Age VII).
- + 2 Repeating a sentence of 26 syllables.
 - o Giving 3 rhymes with defender.
 - o Easy comprehension c (Broken).
 - o Any 2 out of 3 easy comprehension.
- o*Emotion. Picture 1.
- 2*Emotion. Picture 2.
- 2 Easy comprehension a (Train).
- 2 Repeating 5 digits.
- 3*Description. Picture 1.
- 3*Emotion. Picture 3.
- 6*Interpreting 2 out of 3 pictures.
- 7*Interpretation. Picture 2.
- 9*Interpretation. Picture 1.
- 9 Naming the pieces of money.

The above list of 19 tests contains 11 tests of one sort, so that there is strong evidence that this test, describing and interpreting pictures, has no value in diagnosing intelligence. It is not possible to draw conclusions concerning all the other tests in the list, because it is not possible to determine the relation between the difficulty of a test and its diagnostic value. If, for example, the members of the groups studied had been asked if they were little boys or little girls, 100% of both groups would have passed, and the diagnostic value of the test would have been zero, or, if they had been asked to translate a passage of Greek, none of them would have passed, and the diagnostic value would be zero again. The tests of naming the pieces of money and comprehending easy questions show no diagnostic value, but that does not prove that these tests would have no diagnostic value in differentiating groups with less ability than the ones examined. In the same way, the tests of repeating a long sentence and giving rhymes with "defender" are like the passage in Greek, and the fact that they show no diagnostic value for these groups does not prove that they would not be effective in differentiating groups of higher intelligence than the ones examined.

Between these extremes of tests that are entirely below or entirely above the range of ability of the groups examined the tests may be distributed in more or less uniform steps. Some tests may be just above the lowest ability of the groups, others just below the highest ability, and others nearer the median of these extremes. Concerning every test it is possible to raise the question of how the diagnostic value would be changed if the test had been more or less difficult, or how this value would be changed if the groups to whom it had been given had been more or less intelligent. It is not possible to alter the difficulty of the tests after they have been given, but in two cases it is possible to change the passing mark, and to calculate the per cent. that would have passed had the passing mark been more or less severe.

In the 60 word test, a subject is required to give at least 60 words in three minutes to pass the test. As a matter of fact, the subjects gave anywhere from 31 to 196 words in the required time. The retarded group gave from 31 to 148 words, the median being 69 ($Q=15.5$). The normal group gave from 43 to 196 words, the median being 83 ($Q=11.5$). If the passing mark had been 75 words instead of 60 words, 77% of the normals and 37% of the retarded would have passed, and the diagnostic value would have been 40% instead of 27% as shown in table 6. In this way it is possible to calculate the percentage passed and the diagnostic value for each passing mark. These values are shown for 12 passing marks in table 7. The test of copying designs from memory was scored according to the arbitrary point system described in chapter III of the preceding section. The scores of both groups varied from 0 to 20 points, the median of the retarded being 10 ($Q=6$), and of the normal 18 ($Q=4.5$). The percentage passed and the diagnostic value for 10 passing marks are shown in table 7.

The diagnostic value of the 60 word test rises from 10% (where the passing mark is 110 words and the test too hard) to 40% and falls again to 12% and eventually zero when the test is too easy. The diagnostic value of the design test rises from 23% when the test is too difficult to 36% and down to zero when the test is too easy. Every test then has a value which will be

TABLE 7.

Relation between the Difficulty of Two Tests and their Diagnostic Value.

Naming 60 words in 3 minutes				Copying designs from memory			
Passing mark.	Per cent. passed by	Per cent. passed by	Diagnostic value	Passing mark.	Per cent. passed by	Per cent. passed by	Diagnostic value
No. of words	normal group	retarded group		No. of points	normal group	retarded group	
110	14	4	—10	20	26	3	—23
100	19	7	—12	19	34	9	—25
95	31	14	—17	17	55	21	—34
90	38	16	—22	15	67	31	—36
85	47	19	—28	13	71	41	—30
80	64	30	—34	10	76	52	—24
75	77	37	—40	7	83	60	—23
70	84	47	—37	5	88	71	—17
65	88	53	—35	3	98	88	—10
60	93	66	—27	0	100	100	0
50	97	77	—20				
40	100	88	—12				

called the *Maximum Diagnostic Value*. This value is 40% for the 60 word test, and 36% for the designs test.

In the two tests discussed, the groups were constant and the difficulty of the tests varied. In cases where the difficulty of the test may be varied, the Maximum Diagnostic Value may be obtained, provided of course that the method of scoring the test is an accurate expression of the intellectual factors involved. In other tests it is not possible to alter the difficulty, and the Maximum Diagnostic Value can not be determined unless the test be given to groups of varying degrees of intelligence. In this case, the test is constant and the intellectual level of the groups must be varied in order to find the Maximum Diagnostic Value. In this experiment, the groups are constant, so that the method, although conclusive in regard to certain tests, is inconclusive in regard to others. It is conclusive in regard to tests that show a high diagnostic value, and those in which the scoring allows of the determination of the Maximum Diagnostic Value, but inconclusive concerning most of the tests that show no diagnostic value. The method admits of many positive, but few negative conclusions.

It is possible to use two measures of the diagnostic value of the tests, the absolute difference between the per cent. that the normal

and retarded groups pass each test, and the relative difference between these percentages. The dissected sentence test was passed by 100% of the normal group and 29% of the retarded group, the five weights test by 74% normal and 56% retarded, the test of defining "charity" by 53% normal and 12% retarded, and the test of giving an intellectual interpretation of picture 1 by 33% normal and 24% retarded, making the absolute differences between the groups 71%, 18%, 41% and 9% respectively. The difficulty of the tests for the normal group varied as shown by the percentages passed, 100%, 74%, 53% and 33%. Had all of the tests been equally within the range of the groups, it is possible that the diagnostic values would have been different. The relative measure would be the per cent. that the absolute difference was of the per cent. passed by normals, the values in the case of the four tests cited being 71%, 24%, 77% and 27%.

The use of the relative differences would imply that the diagnostic values would have changed if the intellectual level of the groups had been lower or higher, just as the diagnostic value varies if the difficulty of the test (the passing mark) is raised or lowered. This would undoubtedly have been the case, yet we are not warranted in making inferences from the performance of the two groups tested to the performance of any other groups. The data from the two groups give no information concerning the growth of abilities with age or with intelligence. The percentages merely indicate the actual performance of the groups tested. The absolute differences are used as measures of the diagnostic values of the tests in this study, and contain no implications concerning the nature of the performance of other groups.

The diagnostic value of the tests is not the only criterion that the figures in table 6 afford for judging the relative merits of the individual tests. The tests used were in the VII, VIII, IX, X, XII and XV year groups. Inasmuch as the normal subjects were 12, 13 and 14 years of age, it is legitimate to expect that all or practically all of these subjects should pass the VII, VIII, IX and X year tests. The VII year test of describing pictures is failed by 26% of the normal group so that something would ap-

pear to be wrong with this test. The VIII and IX year tests are almost universally passed with the exception of the definitions test which is failed by 16% of the normal group. In the X year group, the absurdity and sentence tests are almost universally passed, the comprehension test is failed by 14%, the 5 weights test by 26%, and the design test by 33%.

The normal group is composed of 18 subjects aged 12 and 20 subjects of 13 and 14. According to Binet's procedure in calibrating the tests for the different years, about 70 or 75% of the 12 year and all of the 13 and 14 year normal boys should pass the XII year tests, or approximately 90% of all the group should pass. This percentage is approximated by the sentence test (79%), the 60 word test (93%) and the dissected sentence test (100%). The definitions test would appear to be too difficult (59%) and the suggestion test entirely too difficult (33%).

If the theoretical curve of growth of the XV year tests could be guessed at, it would probably approximate 0% at 12, 25% at 13, 50% at 14, 75% at 15 and 100% at 16. Roughly then it would be expected that about 25% of the 12, 13 and 14 year normal children would pass these tests. None of these subjects passes the tests of repeating the long sentence or giving rhymes, 21% interpret pictures, 45% solve the problems and 64% repeat 7 digits. The results of the normal group should probably not be used to criticize the "fifteen year" tests, for these subjects were all under 15. The writer has available the results of 10 boys in the Princeton High School, three of whom were 15, one 16, four 17 and two 18 years of age. Three of these boys failed to repeat 7 digits and to interpret pictures, four failed to repeat the long sentence and only one gave three rhymes with "defender." The writer has given these tests to many normal adults but has not recorded them systematically. He has available however the results of seven graduate students all of whom had at least a bachelor's degree. All 7 repeated the 7 digits and the sentence of 26 syllables, two failed to give 3 rhymes, and one gave the "three year" response of enumerating the objects in the pictures, failing in VII and XV. These groups of subjects also throw interesting side lights on other tests. Of the 10 high school

students, 2 failed to copy the designs, 2 failed to arrange five weights, and 6 failed the line suggestion test. Of the 7 advanced university students, 1 failed the design test, 1 failed the weights test and 2 failed the line suggestion test. If these tests are for "ten" and "twelve" year mentality it is right to expect that all of these groups of subjects should pass them.

The figures given in table 6 also show whether the different sub-questions under the various tests are of the same difficulty. In general the results of the sub-questions are about the same. The most marked exceptions appear in the test of comprehending difficult questions, where question *a* is practically impossible, and questions *b*, *d* and *e* almost twice as easy as question *c*.

For the convenience of the reader, the tests shown in table 6 are arranged in table 8 in the order of their diagnostic value as shown by this method of study.

TABLE 8.

Per Cent that Normal Group Pass Each Test.
(Tests Arranged in the Order of Their Diagnostic Value.)

	Diagnostic value	Per cent. passed by normal
1. Comprehending difficult questions (3 out of 5).....	-71	86
2. Reconstructing dissected sentences (2 out of 3).....	-71	100
3. Detecting absurdities in statements (3 out of 5).....	-53	95
4. Defining in terms superior to use (3 out of 5).....	-51	84
5. Defining abstract terms (2 out of 3).....	-51	59
6. Enumerating the months.....	-43	98
7. Copying designs from memory.....	-36	67
8. Solving (both) problems from various facts.....	-35	45
9. Repeating 7 digits.....	-30	64
10. Giving the day and date.....	-29	100
11. Using 3 words in a sentence (2 ideas).....	-29	100
12. Naming 60 words in 3 minutes.....	-27	93
13. Using 3 words in a sentence (1 idea).....	-23	79
14. Arranging 5 weights.....	-18	74
15. Making change	-15	98
16. Resisting suggestion	-12	33
17. Naming the pieces of money.....	-9	100
18. Interpreting 2 out of 3 pictures (Age XV).....	-6	21
19. Repeating 5 digits.....	-2	100
20. Giving 3 rhymes with "defender".....	0	0
21. Comprehending easy questions.....	0	100
22. Repeating a sentence of 26 syllables.....	+ 2	0
23. Describing 2 out of 3 pictures (Age VII).....	+ 2	74

Describing and Interpreting Pictures. These tests appear near the bottom of the list. The small diagnostic value of these tests is surprising in view of the fact that Binet considered the test of interpreting pictures one of the most important in differentiating normals from morons. Binet considered this the most valuable test in the scale. "We place it above all others; and were we limited to one test we would without hesitation choose this one." (Town's (72) translation, page 13.) The results of this investigation would show that there is something radically wrong with this test. Very few of the subjects pass it (21% of the normal group) but this is to be expected if it is a "Fifteen year" test. Three of the high school students failed it, and one of the college adults gave the characteristic "three year" response. Furthermore the results show that the retarded children are just as likely to give an emotional or intellectual interpretation as the normal children. The same holds of the test of describing pictures which shows no difference between the groups. In fact 26% of the normal group fail to describe the pictures and give the enumeration response which is characteristic of the "three year" level. The pictures are not of the same difficulty (see table 6), it being easier to give an intellectual interpretation of pictures 1 and 3 than of picture 2, while the emotional interpretation of the latter appears more frequently.

The explanation of the fact that this test shows no diagnostic value probably lies in the instructions. According to Binet's procedure the child is given the picture and asked "What is this?" If he says "It is a picture," the question is put "Tell me what you see there." In this experiment, the instructions were "Tell me what you see in this picture." The use of the word "what" probably induces the response by enumeration. At any rate the instructions do not seem to produce the same "Aufgabe" in all subjects, and it is most important that all subjects should have the same "Aufgabe" on every test. Persons able to interpret the pictures do not interpret them because they think something else is expected of them. In most of the cases the subjects' responses are not real measures of their ability. The writer believes that the ability to give an intellectual or emotional interpretation of pictures is diagnostic of intelligence. If this factor is correlated with intelligence, it would seem reasonable to try to test for it. This could be done fairly accurately by using more than three pictures, and by framing the questions so as to demand the answer by interpretation, saying "What has happened here?" or "What is the matter with these people?" The test as it stands now is worthless.

Repeating a Sentence of 26 Syllables. This test shows no diagnostic value because it is too difficult. None of the normal subjects and only one retarded subject passed this test. Four of the high school students failed it, so that it would appear too difficult for a "fifteen year" test. In this test a graded series of 12 sentences (11 of them from Town and 1 from Whipple's (75) manual) were given, the sentences varying in length from 10 to 32 syllables. The procedure used was that of starting within the subject's range and continuing up the scale until he had failed two sentences in succession, the number of syllables in the last sentence being taken as the measure of his ability in the test. When working with normal subjects, this

procedure was found to be inadequate for some subjects would fail two or three sentences in succession and then pass the next.

Taking the results of the normal subjects, in one case a subject failed four sentences in succession and then passed the next. In 19% of the cases the subjects failed three sentences in succession and passed the next, and in 45% of the cases they failed two in succession and passed the next. The procedure was entirely wrong then in taking the subject's threshold as the point beyond which he failed two tests. The error made by the writer was that of considering the order of the increasing number of syllables to be the measure of the increasing difficulty of the test. This was not the case. The tests of repeating sentences of 18, 20, 22 and 24 syllables were given to the normal group 90% of the possible number of times. 51% passed 18 syllables, 18% passed 20 syllables, 3% passed 22 syllables and 61% passed 24 syllables. The order of difficulty is therefore 24, 18, 20, 22. Of the 61% of the subjects who passed the 24 syllable sentence, 6% failed to repeat 16 syllables, 41% failed to repeat 18 syllables, 79% failed 20 syllables and 94% failed 22 syllables. The explanation of this probably lies in the fact that the 24 syllable sentence was logically simpler than the others. The factor of logical memory can be separated from tests of auditory memory only by the use of nonsense syllables. The 24 syllable sentence was not the only one in error, however. One subject repeated 22 syllables and failed to repeat 16, 18 and 20 syllables, while 17% of those who repeated 20 syllables failed to repeat 18 syllables. The order of increasing number of syllables is therefore not the order of increasing difficulty.

According to the procedure, the test was discontinued as soon as two sentences in succession had been failed. On this account, all of the sentences were not given all of the possible number of times, and, as the error in procedure was not discovered till normal subjects were examined, a different range of sentences were given to the two groups, so that it is not possible to compare their results on many sentences. The 26 syllable sentence proved too difficult for both groups. The 24 syllable sentence was only given to one member of the retarded group so that no comparison is possible. The 22 syllable sentence was given to all the normal subjects and only passed once, while it was passed but once by the 17 retarded subjects to whom it was given, and is therefore too difficult to show any diagnostic value.

Some indication of the diagnostic value of the 16, 18 and 20 syllable sentences may be obtained. The 20 syllable sentence was given to 75% of the retarded and 98% of the normals. It was passed by 2% of the retarded and 18% of the normals, making the diagnostic value 16%. The 18 syllable sentence was given to 90% of the retarded and 91% of the normals. It was passed by 28% of the former and 51% of the latter, making the diagnostic value 23%. The 16 syllable test was given to 81% of the retarded and 47% of the normals, it being assumed that all the subjects to whom the test was not given would have passed if the test had been given. The test was actually passed by 67% of the retarded and 74% of the normals, but if all the subjects had passed whom it was assumed would pass, 72% of the retarded and 88% of the normals would have passed. The diagnostic value of this sentence is therefore between 7% and 16%. The 14 syllable sentence was

too easy to show any diagnostic value. The number of syllables for the 16, 18, 20 and 22 syllable sentences is the correct measure of their difficulty as shown by the per cent. that normals passed them (74%, 51%, 18% and 3%). The 14 and 16 syllable sentences are too far below the ability of the groups to show any diagnostic value, and the 20 and 22 syllable sentences too far above this ability. The diagnostic value shown by the 18 syllable sentence (23%) may therefore be taken as the Maximum Diagnostic Value for the test of repeating sentences.

Comprehending Easy Questions. Nothing can be said concerning the diagnostic value of this test, because all members of both groups passed two of the three questions. Question b (Playmate) is more difficult than the other two.

Giving Three Rhymes with "Defender." This test shows no diagnostic value, because none of the normal or retarded subjects passed the test. The writer's experience has been that this test is practically impossible for any but exceptionally gifted adults, and is not a fair test of "fifteen year" mentality. The error lies in considering the process of finding three tri-syllabic English words ending in "ender" equal in difficulty to the process of finding three French words ending in "ance" (the Binet test word being "obéissance"). It would be better to admit that the test can not be translated. The normal subjects gave in all 30 rhymes with "defender" and the retarded subjects 16. If the subjects who succeeded in giving one or more words are considered as passing the test, 7 of the retarded and 26 of the normals passed, making the diagnostic value 33%.

Before asking the subjects to give rhymes with "defender," a practice test was given in which the words "day," "mill" and "spring" were used, half a minute being allowed for each word. The total number of rhymes given by both groups for "day" was 439, for "mill" 457 and for "spring" 273. The latter word is therefore much more difficult. The difference in the number of rhymes given by the two groups for "day" was 73, for "mill" 111 and for "spring" 65. The word "mill" would seem to have the highest diagnostic value. Taking the total number of rhymes given for all three words as the measure of ability, the normal subjects varied from 0 to 23, the median being 12.5 ($Q = 3.75$). The retarded group varied from 0 to 21, the median being 9 ($Q = 5.25$). Calculating the percentages of each group that would have passed had the passing mark been fixed at any number of words from 0 to 23, and subtracting to determine the diagnostic value at each passing mark, the Maximum Diagnostic Value for this test was 32% at the passing mark of 5 or 6 words. The value is very close to the value found for giving one or more rhymes with "defender" (33%), so that these figures probably express the general value of rhyming tests in differentiating the intellectual differences between the groups. Binet included the rhyming test in the list of six tests that he considered valuable in differentiating morons from normals.

Repeating Five Digits. This test shows no diagnostic value, because it is too far within the ability of the groups.

Repeating Seven Digits. This test shows a diagnostic value of 30%.

Naming the Pieces of Money. No conclusions may be drawn concerning

the differential value of this test, as it was failed by only 5% of the retarded group.

Making Change. Although this test is slightly more difficult for the retarded group than naming money, it is still too far within the ability of the groups to show its true diagnostic value.

Resisting Suggestion. This test shows practically no diagnostic value (12%), and is passed by only 33% of the normal group. This test admits of a more accurate scoring, inasmuch as there are three lines on which judgments must be made. The normal group in all gave 58 correct judgments out of 174 possible judgments or 33% correct. The retarded group gave 45 correct judgments out of 174 or 26%, making the diagnostic value 7%. The small percentage passed by the normal group is surprising in view of the fact that this test is a "twelve year" test. 6 of the 10 high school students failed this test and 2 of the adults. It is certainly not a test for "twelve years" then, and the writer doubts if it is a test for intelligence, as in his experience persons of ability fail it just as readily as persons of no ability. It is seen to be equally difficult for retarded and normal subjects. Schmitt (57) notes two types of failure on this test, the typical type of failure according to Binet of accepting the suggestion of the first three lines, and the failure due to the fact that the subject actually judges the lines unequal after studying them. In the writer's opinion, this analysis is correct. It means that intelligent persons may fail the test by actually misjudging the length of the lines—cases in which the factor of suggestion is entirely absent. Even if suggestion does influence intelligent subjects in this test, it is not a symptom of defective intelligence to have suggestion warp one's judgment on sensory data. An experiment on suggestion by means of the size weight illusion conducted by Dresslar (24) actually showed that the brighter children were more suggestible than the duller children. Terman (65) has eliminated this test from the latest Stanford revision. Persons who have conducted laboratory experiments on the thresholds of sensation know how difficult it is to rule suggestion out even with highly trained subjects. In doubtful cases of discriminative judgments the subject is apt to take any clue: As a general rule the intelligent person is quite ready to discredit the evidence of his senses. Very rarely will he discredit the conclusion of a reasoning process, however, and the influence of suggestion is a symptom of defect only when it warps one's intellectual judgments.

Arranging Five Weights. This test shows a low diagnostic value (18%). A more accurate method of scoring (taking account of the actual number of successes and failures) shows the normal group arranging the weights correctly in 70% of their 174 trials, and the retarded group in 55% of their 176 trials, making the diagnostic value 15%. This result is surprising in view of the fact that Binet included this test in the list of six tests that he considered most valuable for differentiating morons from normal individuals. Binet styles it "An excellent test which presupposes no schooling or acquired knowledge, and expresses intelligence in its most natural form," and says that "this test is one of those which best detect intelligence without culture, as it is absolutely independent of all instruction." (Town's translation,

pages 41 and 42.) The results of this investigation would show that the test as it stands is also independent of intelligence.

The reason for this lack of correlation with intelligence lies in the indiscriminate mixing of the many factors involved in the test. In discussing this test, Binet points out the various types of response—"Many children do not understand the explanation and remain motionless; so much the worse for them. Others place the boxes in any order without lifting them; and from the little attention that they give them, it is easy to see that they make no comparison. Others understand that the heaviest must be placed first; and they distinguish between the weights of the others most accurately, but they are incapable of arranging the other boxes in the order of their decreasing weight; this idea of decreasing weight is unintelligible to them. They do not lack in sensibility to weight, but in ability to arrange. Others finally grasp the idea of decreasing order, and they come a little nearer to applying it; they arrange such series as 15, 12, 9, 3, 6, where a single box is misplaced; they can do better, they fail from lack of attention and care. This is not a grave error. Nevertheless, we exact two absolutely correct arrangements." (Town's translation, page 42.) Other writers are in general agreement with Binet's analysis of the factors involved. Yerkes (82) classifies the factors as "Kinaesthetic discrimination, ideation (notion of series), attention."

In this experiment a control test of five definitely supraliminal weights (20, 30, 45, 70 and 100 gms.) was used. This test was passed by all the normal and retarded subjects showing that the intellectual ability to comprehend a serial arrangement or to make the logically necessary comparisons was present. The failures were therefore failures in sensory discrimination, and were of the sort that Binet characterized as "not grave." The results show then that the test as a test of sensory discrimination has no diagnostic value. Another proof of the test's lack of worth is the fact that 26% of normal 12, 13 and 14 year boys fail to pass what is supposedly a "ten year" test. Schmitt reports that half of a college class of twenty students failed to arrange the five weights correctly, and believes that "the grasp of the idea of arranging them serially, and an intelligent attempt to do so, is the significant part of the test." (Page 39.) In the writer's experience, the control test of 20, 30, 45, 70 and 100 gram weights proved to be very useful and highly diagnostic of the deficiency in the intelligence of low grade cases. It would seem then that the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons were diagnostic of intelligence while the sensory discrimination was not.

The above conclusion differs little from those of Peterson and Doll (51) who find that the sensory capacity of defective children in muscle sense is not noticeably below normal, and that the slight differences found may be accounted for on an intellectual rather than a sensory basis. They affirm that the test of discriminating lifted weights which they used was not diagnostically valuable except in types of success in following instructions. Smith (59) reports a high correlation between pitch discrimination and general intelligence, but believes that this correlation is due to the intellectual factors in the test rather than to any physiological factors of sensory dis-

crimination. These two factors, the intellectual and the physiological (or as Smith calls them, the "elemental"), enter into almost all sensory tests, and the writer believes that most of the correlations reported between sensory discrimination and intelligence are due to the intellectual factors in the tests.¹ Burt (16) finds no general connection between the capacity to discriminate lifted weights and intelligence as estimated by the school masters. The results of Thorndike, Lay and Dean (71) on 37 normal school women and 25 high school boys show low correlations between accuracy in reproducing lengths and intelligence as estimated by the pupils (25), by the teachers (12) and by the school records ($-.01$); and between ability in weighing boxes to standards and intelligence as estimated by the pupils (23), by the teachers (08) and by the school records (21). From Thorndike's results, Simpson (58) estimates the probable correlation between general sensory discrimination and general intelligence to be about .23. In the light of the results of this investigation and those of other investigators it is safe to conclude that the test of arranging five weights has no diagnostic value and should be eliminated from the scale. If the test were changed so as to rule out the factor of sensory discrimination, and involve only the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons, the test would probably prove very valuable in differentiating the intelligence of younger children.

Constructing a Sentence. The test of constructing a sentence containing three given words with the resulting expression containing one or two ideas shows diagnostic values of 23% (1 idea) and 29% (2 ideas). In a preceding chapter (see page 129) it was seen that large differences existed in the character of the sentences constructed. The sentences "Trenton has lots of money on the river" and "Trenton paid a large sum of money to have the Delaware river deepened" represent widely different logical constructions, yet both would have to be scored plus under the "twelve year" test, for according to Binet the fact that a child constructs a single sentence containing the three words proves that he has a "mental age" of twelve, even if the sentence given be devoid of sense. An example given by Binet of such a sentence that would receive a "twelve year" credit is "Paris is a city of fortune by a stream." The writer believes that if the test were scored according to logical rather than grammatical merit, the diagnostic value would be much higher. Meumann (46) in an extensive study of the ability of children to construct sentences from two words, three words, and from pairs of words, places the greatest emphasis in the analysis of the results in their relation to intellectual endowment on the character of the responses.

Naming 60 Words in 3 Minutes. This test shows a Maximum Diagnostic Value of 40% when the passing mark is 75 rather than 60 words. In this test, 21% of the retarded group exceed the median of the normal group, 32% exceed the lowest 14 of the normal group and 84% exceed the lowest one.

¹ This applies to other tests more elementary than those of sense discrimination. Some of the correlations between intelligence and vital capacity are undoubtedly due to the fact that there is a trick in blowing up a spirometer, and that dull and defective children loose a lot of "wind." These correlations refer more to the quickness in sizing up the apparatus and catching on to the method than to the cubic contents of the mouth and lungs.

Giving the Day and Date. This test which in a preceding section was shown to depend on school training shows a diagnostic value of 29%.

Solving Problems from Various Facts. This test shows a diagnostic value of 35%, but this is not a true expression of the merit of the test for it is the resultant score of an effective and an ineffective test. Problem a (Hanging from a limb) shows a low diagnostic value (21%), while the other problem (Neighbor's visitors) shows a value of 51%. The second test is twice as effective as the first yet its merit is obscured by the scoring of the test according to the ruling that the subject must solve both problems in order to pass the test. The fact that one poor test may in this way lower the effectiveness of another test illustrates one of the advantages of the partial credit system adopted by Yerkes in the Point Scale. The explanation of the fact that the first problem shows a low diagnostic value probably lies in the fact that a large number of the normal group gave the answers "a bear," "a snake," etc., answers which to intelligent subjects seemed to be perfectly rational, but which had to be scored minus according to Binet's rule that "a man hanging" is the only acceptable answer.

Copying Designs from Memory. This test shows a Maximum Diagnostic Value of 36%. 9% of the retarded group exceed the median of the normal group, 52% exceed the lowest 14 and 88% exceed the lowest one. The fact that 33% of the normal 12, 13 and 14 year boys fail this test for "ten year" mentality would show that it is not a real "ten year" test. In all probability the visual memory involved is of a particular sort so that no group of individuals randomly selected would ever succeed in 100% of the cases.

Enumerating the Months. This test shows a diagnostic value of 43%. In a previous section it was shown that this test depended on school training. The examination of Chotzen's (18) results showed that they did not prove that the tests that showed the greatest increase with maturity were least dependent on intelligence. A test that depends on training may have a high diagnostic value, but the previous training of the subject must be known of course for the subject's failure to have significance.

Defining Abstract Terms. This test shows a diagnostic value of 51%. It is interesting to see that in this test the diagnostic value obtained by scoring a subject passed if he defines two or three terms correctly is higher than the diagnostic value of any of the three parts taken singly. The three words are equally difficult to define yet the word "kindness" has the smallest diagnostic value. That the test of defining abstract terms is too difficult is shown by the fact that only 59% of the normal subjects passed it, when the proportion should be approximately 90%. This test is one of the six that Binet considered as diagnostic of the mental differences between morons and normals.

Defining in Terms Superior to Use. This test shows a diagnostic value of 51%. All of the words are of the same difficulty except "mother" which is slightly more difficult due to the occasional embarrassment reaction that is encountered. The fact that 16% of the normal 12, 13 and 14 year subjects fail this test is probably due to the amused attitude that some older subjects assume when given this test. The consequence is short and careless answers. It is certainly true that all of the normal subjects were able

to give definitions in terms superior to use. Since this test shows such a high diagnostic value it would probably be well to change it in some way so that all subjects would have the same "Aufgabe," and so that the test would not depend on the random interpretations of "What is a ———?"

Detecting Absurdities in Statements. This test shows a diagnostic value of 53%. Absurdity *e* shows the highest diagnostic value (62%) and absurdity *d* the lowest (23%). Absurdities *a*, *c* and *d* are the easiest, absurdity *b* the most difficult, with absurdity *e* between.

Reconstructing Dissected Sentences. This test shows a diagnostic value of 71%. This test was found to depend on school training, the effect being noticeable between the fourth and fifth grades. This test was passed by 3 of the 10 retarded boys in the fifth grade, by 13 of the 29 in the fourth grade and by 1 boy in the third grade. None of the normal subjects in the sixth, seventh or eighth grades failed the test. Two interpretations of this test are of course possible. The first is that the high diagnostic value shown is due entirely to grade training. The second is that the test is entirely dependent on intelligence, and that children who have not sufficient intelligence to pass the test never reach the fifth grade. The truth probably lies in the view that the test depends on both factors. It can not be training entirely for the groups had been in school the same length of time, and a larger proportion of fourth than fifth grade subjects passed the test.

In designing the dissected sentence test, Binet sought to detect the same abilities that were involved in the Ebbinghaus mutilated prose tests. The results of Stenquist, Thorndike and Trabue (60) on a completion test show a very marked increase in the performance of children in the fifth and sixth grades over those in the third and fourth grades, the sudden increase in performance which indicates school training appearing between the fourth and fifth school grades, where the influence of this factor appeared in the dissected sentence test. The results of Fraser, reported by Whipple (76) show a higher correlation between performance on the completion test and scholastic status than between performance on this test and chronological age. Ebbinghaus (25) believed that the completion test involved factors most intimately connected with intelligence. Simpson (58) found that a completion test differentiated his two groups almost completely. Wyatt's (81) results show a high correlation between intelligence as estimated by the teachers and performance on the completion test (0.85, $p < 0.04$), a correlation higher than that obtained from any of the other 15 tests used. There seems to be good evidence then that both the Ebbinghaus completion test and its mutant offspring, the dissected sentence test, depend on school training and also correlate highly with intelligence.

Comprehending Difficult Questions. This test also shows a diagnostic value of 71%. Question *a* is practically impossible because the subjects did not understand the meaning of the word "delayed." The fact that the retarded children were 10% ahead of the normal children on this test may be due to chance or to the personal equation of the experimenter. Question *c* was passed by only 50% of the 12, 13 and 14 year normal children, showing that it is too difficult for "ten years." Questions *b*, *d* and *e* are of approximately the same difficulty, but questions *b* and *e* have a higher diagnostic value than

question *d* (76% and 74% to 51%). This test is included in the list of six tests that Binet considered to be valuable in differentiating the moron from the normal individual.

In the foregoing discussion of the 23 Binet tests used in this investigation, it was not possible to draw any conclusions concerning the tests of comprehending easy questions, repeating 5 digits, naming the pieces of money and making change, because they were too easy for both groups. It was found that the other tests varied in their diagnostic value from -71% to $+2\%$, or in other words it was shown that the scale contained some tests that were very effective and others that were quite ineffective. To summarize the results it is best to classify the tests according to the mental processes involved, in order to determine what sort of tests correlate best with intelligence.

Any classification of the tests according to the mental processes involved is of course inadequate, for these processes can not be determined except by experiment. The fact that a test is classified as involving a certain process does not prove that that process is involved. In fact two subjects may use quite different mental processes in solving the same test. The classification given in table 9 is offered with these qualifications. For the most part, the writer has adopted the analysis given by Yerkes (table 1, pages 7 and 8). The writer has added the factor of school training to the dissected sentence test, and of logical memory to the test of repeating sentences. The tests of solving problems, rhyming, naming the months and giving the date are not included in Yerkes' list, and were classified by the writer. The list of tests arranged according to their diagnostic value with the factors involved in each is shown in table 9.

In table 10 the tests have been re-classified according to the main factors involved in each, and these factors arranged in the order of their apparent worth in diagnosing intelligence. The diagnostic values of each test and of each part of each test are shown. All these values except those given for the tests of rhyming and repeating sentences are taken from table 6.

From table 10 it will be seen that most of the large differences between the normal and retarded groups appear under the head-

TABLE 9.

Factors Involved in the Various Tests.

Diagnostic Value	Tests
—71	Comprehending difficult questions. PRACTICAL JUDGMENT involving memory and imagination.
—71	Reconstructing dissected sentences. IDEATION involving analysis, imagination, command of language forms, school training.
—53	Detecting absurdities in statements. LOGICAL JUDGMENT based on imagination, analysis and reasoning.
—51	Defining terms superior to use. IDEATION (association and analysis).
—51	Defining abstract terms. IDEATION involving vocabulary.
—51	Problem b (Neighbor's visitors). PRACTICAL JUDGMENT, reasoning inductively from a concrete situation.
—43	Enumerating the months. SCHOOL TRAINING, memory.
—40	Naming 60 words in three minutes. ASSOCIATION (free), vocabulary, attention.
—36	Copying designs from memory. VISUAL MEMORY, perception, attention, motor coordination.
—33	Rhyming words with "defender" and with "day," "mill" and "spring." ASSOCIATION (controlled), vocabulary, attention.
—30	Repeating 7 digits. AUDITORY MEMORY for words (digits), attention.
—29	Giving the day and date. SCHOOL TRAINING, memory.
—29 and —23	Using 3 words in a sentence containing either one or two ideas. IMAGINATION and command of language forms.
—23	Repeating sentences (18 syllables). AUDITORY MEMORY for sentences, logical memory, attention.
—21	Problem a (Hanging from a limb). PRACTICAL JUDGMENT, reasoning inductively from a concrete situation.
—18	Arranging five weights. KINAESTHETIC DISCRIMINATION, ideation (notion of series), attention.
—12	Resisting suggestion. SUGGESTIBILITY, visual perception, comparison.
—6 and +2	Interpreting and describing pictures. PERCEPTION (visual—of things, relations, meanings), apperception, association, imagination.

TABLE 10.

Diagnostic Value of Various Tests and Parts of Tests Classified according to the Factors Involved.

	Diagnostic Value of of Parts					
	total	a(1)	b(2)	c(3)	d	e
IDEATION						
Reconstructing dissected sentences*.....	-71	-70	-54	-65		
Defining in terms superior to use.....	-51	-49	-38	-42	-53	-43
Defining abstract terms.....	-51	-41	-42	-24		
JUDGMENT (logical and practical)						
Comprehending difficult questions.....	-71	+10	-76	-36	-51	-74
Detecting absurdities in statements.....	-53	-42	-46	-41	-23	-62
Solving problems from various facts....	-51	-21	-35			
SCHOOL TRAINING						
Enumerating the months.....	-43					
Giving the day and date.....	-29					
ASSOCIATION (free and controlled)						
Naming 60 words in 3 minutes.....	-40					
Giving rhymes with "defender".....	-33					
MEMORY (auditory and visual)						
Copying designs from memory.....	-36					
Repeating 7 digits.....	-30					
Repeating sentence (18 syllables).....	-23					
IMAGINATION						
Using three words in a sentence (2 ideas)	-29					
Using three words in a sentence (1 idea)	-23					
KINAESTHETIC DISCRIMINATION						
Arranging five weights.....	-18					
SUGGESTIBILITY						
Resisting suggestion	-12					
PERCEPTION						
Describing pictures	+ 2	- 3	+ 4	+ 4		
Interpreting pictures	- 6					
Emotional interpretation		0	- 2	- 3		
Intellectual interpretation		- 9	- 7	+ 6		

*The test of reconstructing dissected sentences also involves school training.

ings "Ideation" and "Judgment." Five of the 29 values under these headings are 70% or over, 14 are over 50% and 22 are over 40%. Under the remaining headings only two tests show a diagnostic value of 40% or over.

Two interpretations of these results are possible: the first, that the factors involved in the first six tests are those that are most intimately associated with intelligence; the second, that these 6

tests all involve the use of language, and that they are really diagnostic of the amount of linguistic training rather than intelligence. The 6 tests that show the highest diagnostic value most certainly deal with verbal material. One interpretation must hold that the normals handle the material better because they are more intelligent, the other that they handle it better because they have had more linguistic training. The position that the differences are due to training is supported by the fact that the dissected sentence test depends on school training, and that the test of enumerating the months also shows a high diagnostic value (43%). Against this position it may be said that the retarded children had been in school as long as the normal children, and had come from the same sort of homes. When the results of children of different linguistic training are compared, the differences are slight compared to those found between children of different intelligence. The normal children of English speaking parents average 5% higher than those of non-English speaking parents on these six tests, and the same difference is found between the retarded groups of different linguistic training. The normal group however averages 47% higher than the retarded group on these six tests. It is therefore legitimate to conclude that these tests involve other factors beside language—factors that are intimately connected with intelligence.

The reader may draw his own conclusions concerning the nature of the mental processes involved in the tests. In the writer's opinion, the tests that show the highest correlation with intelligence are those that involve reasoning—that demand the application of the subject's knowledge in a new way. It is safe to assume that a group of impartial judges would classify the mental processes involved in the six tests that show the highest diagnostic value among those commonly called the "higher thought processes," and would place the other processes of memory, imagination, sensory discrimination and suggestibility somewhat below these on a scale of complexity.

It is not necessary, however, for the purposes of this discussion to classify the mental processes involved in the tests. It is only necessary to note that certain tests show a very high diag-

nostic value while others show practically none. The tests of arranging weights, resisting suggestion and describing and interpreting pictures should either be changed so as to bring out their diagnostic value or thrown out of the scale entirely. The rest of the tests should be weighted in the scoring system according to their relative diagnostic value.

It has been seen that the Binet scale is inadequate in diagnosing the higher grades of mental defect. The reason for this is now obvious. Certain tests are diagnostic of intelligence while others are not. The subject receives the same credit for passing the tests that are not diagnostic as he does for those that are highly diagnostic. The subject who arranges the weights correctly receives the same credit (one fifth of a year) as the subject who answers all five of the comprehension questions or all five of the absurdity questions. In the same way a subject gets the same credit for passing the suggestion test as he does for defining abstract terms or reconstructing dissected sentences. A subject may pass all the tests of low diagnostic value and fail all the diagnostic tests and yet have the same "mental age" as a subject who passes the diagnostic tests and fails the others. For example, one subject may pass all the tests in VIII, all of the tests in IX except the definitions test, the weights, design and sentence tests in X, and the suggestion, sentence and 60 word tests in XII. He then has a basal age of 8, and having passed 10 tests above that, his "mental age" is 10. Another subject passes all of VIII and IX, the absurdity, comprehension and sentence test in X, and the definitions and dissected sentence test in XII. He then has a basal age of 9, and having passed 5 tests above that, his "mental age" is 10. The two subjects would have the same "mental age" yet the second would be far more intelligent than the first, because he passed all of the diagnostic tests while the other passed none of them. In some cases then the value of the diagnostic tests may be obscured by those that are not diagnostic. Such cases would of course be exceptional for according to chance normal children are just as apt to pass the tests that are not diagnostic as feeble-minded children. The fact that such cases are possible however certainly justifies the opinion of the

Buffalo conference (15) that "A mental age of 10 or above is not necessarily indicative of feeble-mindedness, regardless of how old the examinee may be."

The elimination of some of the tests of low diagnostic value would make the scale easier to apply and more accurate. This may be shown by comparing the total scores on the whole series with those on a few tests. There are two ways of scoring subjects on the Binet tests, by taking the "mental ages," and by subtracting these ages from the chronological ages and finding the age differences. The computation of "mental ages" over 10 is made difficult on account of the missing groups of XI, XIII and XIV. A child making basal 10 and passing 4 tests in XII has a "mental age" of $10\frac{4}{5}$. If he passed the additional test in XII his "mental age" would be 12, so that one test counts for a year and a fifth. In the same way a child passing all tests in XII and 4 in XV has a "mental age" of $12\frac{4}{5}$, but if he passes the additional test in XV his "mental age" would be 15, the extra test in this case counting for two years and a fifth. This difficulty may be overcome by weighting the tests in XII and XV, counting all tests in XII as two fifths of a year, and all tests in XV as three fifths of a year. The first case cited would then have a "mental age" of $11\frac{3}{5}$ which would become 12 if the additional test were passed, and the second case would have a "mental age" of $14\frac{3}{5}$ which would become 15 if the additional test were passed. The writer has scored the subjects in both ways, according to the conventional method and according to the method of weighting the advanced tests. The comparison of the "mental ages" with the chronological ages will also yield two measures depending on whether the conventional or weighted "mental age" is used. In treating the total scores, the writer has computed them according to all four methods in order that the most favorable method may be used for the purposes of comparison.

The scores of the subjects on five tests were used to compare with the total scores of the whole series. The scores of the subjects were computed for all five parts of the tests of defining in terms superior to use, all five of the absurdity questions, the last four comprehension questions, all three abstract definitions and

all three dissected sentences. The score was taken therefore on 20 parts of five questions so that the ability of each subject may be expressed anywhere on a scale from 0 to 20 points.

To compare the five methods of scoring (the "mental ages" and age differences according to the conventional and weighted methods of scoring, and the score on five tests), the medians of the normal and retarded subjects were found, the Maximum Diagnostic Value was computed, and the per cent. of retarded subjects exceeding the median of the normal subjects, the lowest 14 of the normal subjects, and the lowest normal subject was determined. These values are shown in table 11.

TABLE 11.
Comparison of Various Methods of Scoring the Binet Tests.

	Median of normal group (Q)	Median of retarded group (Q)	Per cent of retarded group exceeding the			Maximum Diagnostic Value
			Median of normal group	Lowest 14 of normal group	Lowest 1 of normal group	
Mental ages.....	10.8(0.2)	9.6(0.8)	2%	3%	41%	69%
Weighted mental ages..	12.0(0.5)	10.0(0.9)	3%	14%	41%	64%
Age differences.....	-2.6(0.75)	-4.0(0.7)	10%	22%	63%	52%
Weighted age difference	-1.3(0.9)	-3.6(0.8)	7%	22%	49%	54%
Score on 5 tests.....	16(2.75)	6(2.5)	3%	7%	19%	83%

From table 11 it may be seen that the effect of weighting the "mental ages" increases the quantitative differences between the medians of the groups, or extends the distribution of the measures. The difference between the conventional "mental ages" 1.4 yrs. (10.8 to 9.6) is increased to 2 yrs. (12.0 to 10.0), and the difference between the age differences is increased from 1.4 to 2.3. The variability of the measures is of course raised in each instance. The dispersion of the measures has however no effect on their effectiveness in differentiating the groups. The conventional method of scoring the "mental ages" shows the greatest differentiation between the groups and the greatest Maximum Diagnostic Value (69%). The score of the subjects on five tests is more effective in differentiating the groups however and

shows a higher diagnostic value (83%). In all, 23 Binet tests were used. These results show the paradoxical situation that the scale would have been more effective if 18 of these tests had not been given. This result is valuable only for a demonstration, and does not prove that only five tests should be used in testing intelligence, for the accuracy of any measure of general intelligence increases with the number of tests, provided that the tests are all effective. The result does show however that the method of increasing the accuracy of the measures of higher degrees of mental defect is that of increasing the number of diagnostic tests and eliminating the tests that are not diagnostic.

VI. DIAGNOSTIC VALUE OF SUPPLEMENTARY TESTS

The preceding discussion has been confined to the Binet tests entirely. In order to make the investigation more complete a number of other tests were given. The principle used in selecting the supplementary tests was that of diversity of character. Ten different sorts of tests were used, eight of them being apparently independent of language training. In cases where standard tests were used the procedure was adapted to the needs of the experiment. Deviations from standard procedure are to be excused on the grounds that age norms were not being sought, and to find the diagnostic value it was only necessary that the subjects understood the nature of the task they were to perform, and that the instructions were uniform for both groups. The detailed account of the method is given in the description of each test.

TEST I. PUZZLE TESTS.

A series of ten puzzles tests was used. The apparatus of one test was changed during the course of the experiment so the data are given on but nine. The two bicycle bell puzzles were given to all of the normal group and to 97% of the retarded group. All of the other puzzles were given all of the possible number of times. In a previous section, the value of these puzzle tests in obtaining the cooperation of the subjects was noted (see page 124). The discussion of their diagnostic value follows.

HEALY CONSTRUCTION PUZZLE A.

The material used for this test was the standard apparatus manufactured by C. H. Stoelting Co., and described under Test III on pages 14 and 15 of Healy and Fernald's monograph (34), and on pages 93 to 96 of Schmitt's monograph (57). The pieces of the puzzle were disposed irregularly on the table and the subject told to "Put this puzzle together." Healy's method of scoring is to record the number of moves, the number of

obvious impossibilities, the repetition of such obvious impossibilities, and the time, the subject being marked as failed if the time exceeded 10 minutes. In the present experiment, the number of moves and the time were taken, the subject being marked as failed if he made over 30 moves or did not succeed in solving the puzzle in a minute and a half.

81% of the normal subjects and 63% of the retarded subjects succeeded in solving the test under these requirements. The normal subjects solved the puzzle in from 5 to 27 moves, the median of all normal subjects (including those who failed) being 11.5 ($Q=7$). The median of the retarded group was 15 moves, the variability being higher than 10.5 moves. The Maximum Diagnostic Value for the test was 18% at the arbitrary passing limit selected (30 moves).

Schmitt classifies the responses to this test under three types—planned, trial and error, and chance. The subject is considered to have done the test by the planned method if he solves the test with less than 6 errors, by the trial and error method if he makes from 6 to 11 errors, and by the chance method if he makes more than 12 errors. The method of scoring used in this investigation may be compared to that used by Schmitt by adding the smallest possible number of moves (5) to the number of errors, and in this way considering the difference between the smallest possible number of moves and the actual number of moves the number of errors. A subject would be considered to have done the test by the planned method if he made less than 11 moves, by the trial and error method if he made from 11 to 16 moves, and by the chance method if he made 17 moves or over. 45% of the normal subjects and 36% of the retarded subjects did the test by the planned method, 19% of both groups by the trial and error method, and the remainder (36% of the normals and 46% retarded) by the chance method. This method of scoring therefore does not show a diagnostic value higher than 10%.

HEALY CONSTRUCTION PUZZLE B.

The material used for this test was the standard apparatus manufactured by C. H. Stoelting Co., and described under Test

IV on pages 16 and 17 of Healy's monograph, and on pages 97 to 100 of Schmitt's monograph. In this investigation, the number of moves and the time were recorded, the subject being marked as failed if he made 35 or more moves, or did not succeed in solving the puzzle in three minutes.

81% of the normal subjects and 51% of the retarded subjects succeeded in solving the test under these requirements. The normal subjects made from 11 to 34 moves, the median being 19 ($Q=9$). The retarded group solved the puzzle in from 11 to 34 moves, the median being 34 (the other half failing). The maximum diagnostic value was 35% if the passing mark were 30 moves.

Schmitt considers the subject as solving the test according to the planned method if he makes 8 errors or less, by the trial and error methods if he makes from 9 to 16 errors, and by the chance method if he makes 28 or more errors. The smallest possible number of moves is 11. Comparing the two methods of scoring, the subjects of this investigation who performed the test in 19 moves or less were classified under the planned method, from 20 to 27 moves under the trial and error method, and in 28 or more moves under the chance method. 50% of the normal and 24% of the retarded group do the test by the planned method, 16% of the normal and 15% of the retarded by the trial and error method, and the remainder (34% normal and 61% retarded) by the chance method. This method of scoring therefore shows a diagnostic value of 26%.

The comparison of the results of this investigation with those of Schmitt can not be made, for what Schmitt considers an "error" may not be what the writer considers a "move." The method of scoring is adopted merely to throw light on the results of this investigation. The criticisms of these two Healy tests which the writer makes are not made against the tests given according to Healy's procedure, but according to the writer's procedure under the conditions of this experiment. At a matter of fact Healy considers the quantitative scoring of his performance tests of little importance as compared to the qualitative scoring—to the experimenter's judgment of the manner in which

the subject approaches the problems. Schmitt has of course converted a quantitative scoring (number of errors) into a qualitative scoring (planned, trial and error, and chance methods), so that her results can not be used to criticize Healy's method.

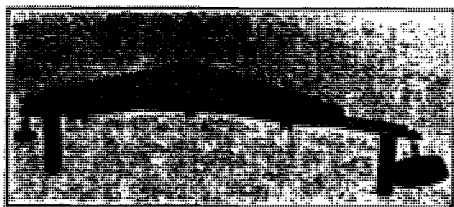
Schmitt's results however taken on their own merit show one rather striking feature. The results on these two tests, classified according to chronological ages in tables XVII and XX on pages 95 and 98, show a fairly steady gain in the excellence of performance from 5 to $11\frac{1}{2}$, and a decrease from $11\frac{1}{2}$ to $14\frac{1}{2}$. Taking the three age groups, $10\frac{1}{2}$ to $11\frac{1}{2}$, $11\frac{1}{2}$ to $12\frac{1}{2}$, and $12\frac{1}{2}$ to $14\frac{1}{2}$, 90% of the first group, 77% of the second and 60% of the third solve construction puzzle A according to the planned method, while 90% of the first group, 71% of the second and 65% of the third solve construction puzzle B by the planned method. The results show therefore an inverse correlation with age from $10\frac{1}{2}$ to $14\frac{1}{2}$. The meaning of this is not exactly clear. Taken in connection with the fairly low diagnostic value found for these two tests (18% and 35%) it would seem to indicate that these tests had little value in diagnosing the higher grades of mental defect.

Haines (32) classified 63 subjects age 12 to 18 into three groups according to their performance on the Binet scale and the Point Scale. 21 of the subjects were classified as high grade morons, 16 as showing doubtful mental defect, and 26 as showing no mental defect. He found the construction puzzle A to be valuable for distinguishing the not defective from the doubtful, and the high grade defective from the doubtful, but found that construction puzzle B showed no definite diagnostic value. These results are quite opposite to those of this investigation in which puzzle B shows a higher diagnostic value (35%) than puzzle A (18%).

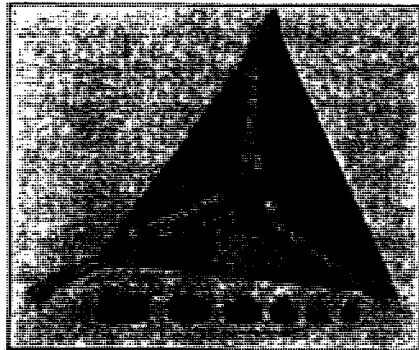
BICYCLE BELL PUZZLES A AND B.

Two mechanical puzzles were designed by the writer after the suggestion of Prof. H. C. McComas, bicycle bells being chosen to arouse the subject's interest. The various parts of the puzzles are shown in Fig. IA. Puzzle A consists of parts A, B, D, E and F. The position of parts D and E is shown inverted in B

to show the manner in which the cogs mesh. Part F slips over the center pivot of B, to which the cover A is then attached, the completed puzzle being shown under C. The minimum number of moves of puzzle A is four. Puzzle B differs from puzzle A in that part F of puzzle A is broken up into parts G, H, I, J, K and L of puzzle B. In puzzle B parts D and E fit the same as in puzzle A. Part H slides on the center pivot, G fits over H, K and L fit in the ends of G and are held in place by I and the



D



C



A



B

FIG. 1

- A. Bicycle Bell Puzzles.
 B. Balance Test.
 C. and D. Test of Lifting the Table Asymmetrically Balanced.

washer J. The minimum number of moves for puzzle B is eleven. The procedure in giving both puzzles was to place the parts in front of the child with the instructions "Put that bell together so it will ring." The subject was not allowed to see the bell as it was being taken apart. The time and the number of moves were recorded.

The subjects were considered as failing bell A if they took more than 20 moves or did not succeed in two minutes. 71% of the normal subjects and 47% of the retarded subjects passed the test. The median of the normal subjects was 15.5 moves, of the retarded subjects 27 moves. The Maximum Diagnostic Value was 24% when the passing mark was 20 moves.

Bell B was not given to the subjects who failed bell A, it being assumed that they would fail the more difficult one. Bell B proved entirely too difficult for both groups, being passed by but 22% of the normal subjects and 18% of the retarded subjects, the subjects being considered as failing if they took more than 30 moves or did not solve the puzzle in three minutes. The Maximum Diagnostic Value was 10% if the passing mark were fixed at 26 moves.

PUZZLES A, B, C, AC AND BC.

A series of puzzles was designed in order to test the subject's puzzle solving ability and his capacity to profit by experience. The puzzles are shown in Fig. 2 reduced one half. The puzzles were constructed of three-ply board painted white on one side and black on the other.

Puzzle A consisted of a triangular opening 9.8 cm. at the base, and 7.8 cm. on each side. Two right angle pieces $4.9 \times 6.05 \times 7.8$ cm. were provided to fill the opening. This puzzle is the same as the triangular portion of Healy's Introductory Picture Form Board described under Test I, pages 11 to 13 of his monograph.

Puzzle B consisted of the right angled piece fitting in the left side of puzzle A, and two other pieces, one $9.7 \times 6.05 \times 4.0$ cm., and the other $7.5 \times 4.9 \times 4.0$ cm. Puzzle B fitted in the same opening as puzzle A.

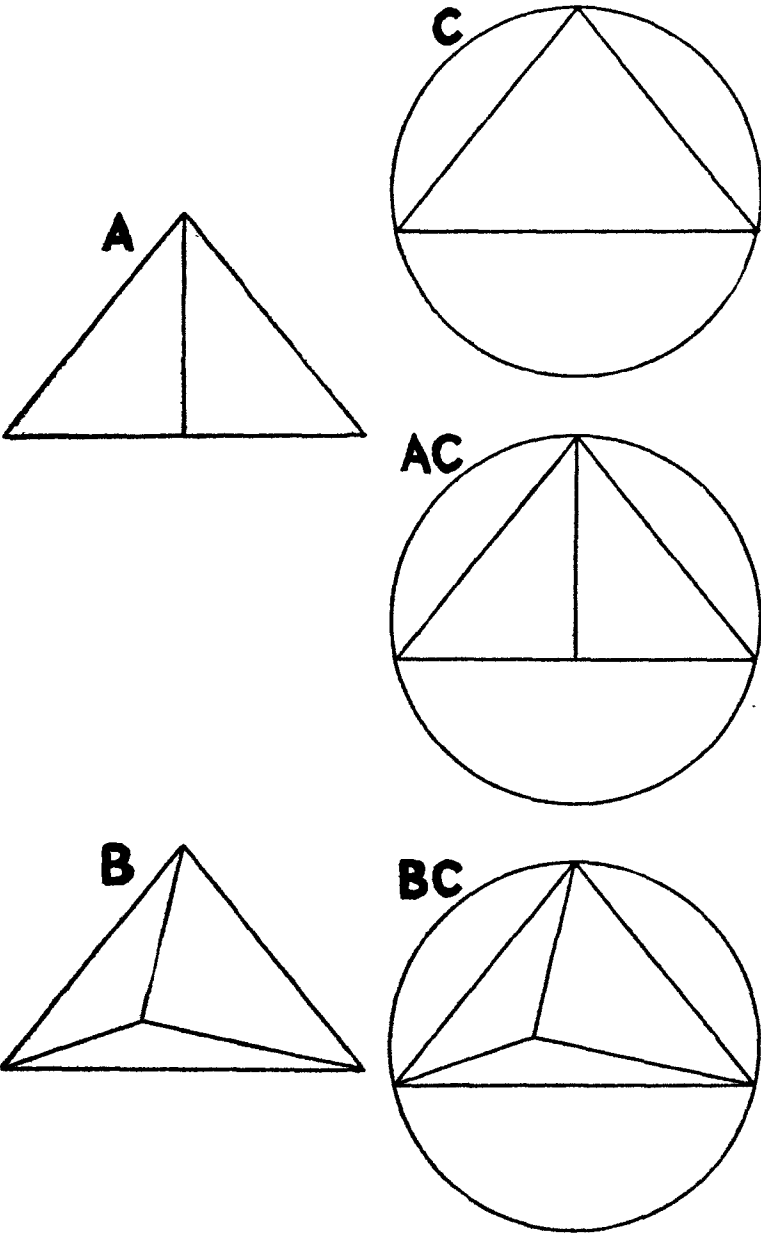


FIG. 2. Puzzle Series.

Puzzle C consisted of four pieces fitting into a circular opening 5 cm. in diameter. The triangular piece was the same size as the opening in puzzle A, 9.8 x 7.8 x 7.8 cm.

Puzzle AC consisted of the three curved portions of puzzle C and the two right angled pieces of puzzle A. In this test the subject has an opportunity to carry over the learning from puzzle A.

Puzzle BC consisted of the three curved portions of puzzle C and the three pieces of puzzle B, so that the learning from the latter puzzle might be carried over.

The instructions in every case were "Put those pieces in there. Keep the white side up." All five of the puzzles were given all the possible number of times. Puzzle A may be solved in two moves, puzzle B in three, puzzle C in four, puzzle AC in five and puzzle BC in six. Puzzles A, B, AC and BC were considered failed if the subject made 30 moves or over, or failed to accomplish the task in a minute and a half.

88% of the normal and 80% of the retarded passed puzzle A. The normal group varied from 2 to 28 moves, the median being 10 ($Q=7$). The retarded group varied from 3 to 26 moves, the median being 12 ($Q=7.5$). The Maximum Diagnostic Value was 14% when the passing mark was 5, 6, 7, or 8 moves. Schmitt classified a subject's response as trial and error if he tried each piece in more than two positions before finding the right one, and as planned if he tried one or both pieces in less than two positions before finding the right one. The subjects of this study are therefore classified as doing the test by the planned method if they make 6 moves or under, and by the trial and error method if they make more than 6 moves. 33% of the normal and 19% of the retarded used the planned method. The diagnostic value of this method of scoring is therefore 14%.

79% of the normals and 59% of the retarded passed puzzle B. The normal group made from 5 to 26 moves, the median being 15.5 ($Q=7$). The retarded group made from 4 to 27 moves, the median being 18 and the variability higher than 9 moves. The Maximum Diagnostic Value was 20% when the passing mark was placed at 25 moves.

Puzzle C was considered failed if the subject made over 15 moves. All of the normal group and 93% of the retarded group passed this test. The median number of moves of the normal group was 4 ($Q=1$), while that of the retarded group was 5 ($Q=2$). The Maximum Diagnostic Value was found to be 20% when the passing mark was 7 moves.

81% of the normal group and 61% of the retarded group passed puzzle AC. The normal group made from 5 to 28 moves, the median being 11 ($Q=6.5$). The retarded subjects made from 6 to 29 moves, the median being 19 and the variability higher than 9 moves. The Maximum Diagnostic Value was 27% when the passing mark was placed at 10 moves.

79% of the normal group and 59% of the retarded group passed puzzle BC. The normal group varied from 6 to 27 moves, the median being 13 ($Q=8.5$). The retarded group varied from 6 to 28 moves, the median being 22 moves and the variability higher than 9 moves. The maximum Diagnostic Value was found to be 26% when the passing mark was 15 moves.

To study the effect of learning from puzzles A to AC and from B to BC, the difference between the number of moves on the first test and the number on the second minus the three additional moves from the curved pieces was used. The coefficients of learning for all subjects were calculated according to the formulae,—

Coefficient of learning from A to AC $= A - (AC - 3)$

Coefficient of learning from B to BC $= B - (BC - 3)$

A positive value would therefore indicate that the subject had profitted by the experience of the first test, a negative value that he had not.

In the comparison of the scores from A to AC, the normal subjects varied from +25 to -21, the median being 0 ($Q=7.5$). The retarded subjects varied from +22 to -27, the median being 0 ($Q=9.5$). The Maximum Diagnostic Value was 11%.

In the comparison of the scores from B to BC, the normal subjects varied from +27 to -23, the median being +2.5 ($Q=4$). The retarded subjects varied from +27 to -26, the median being 0 ($Q=9$). The Maximum Diagnostic Value was 13%.

Although the series of puzzles was designed so that the practise on the first two puzzles might carry over to the last two, the medians indicate that very little if any of the learning was carried over. In only one case (normal group B to BC, +2.5 moves) is any learning shown. In the other cases the medians are zero indicating no learning. This is probably due to the fact that there is something in the way of a "trick" or a "catch" in these puzzles.

SUMMARY OF PUZZLE TESTS

The distribution of the number of moves made by normal (N) and retarded (R) subjects in solving each of the 9 puzzles is shown in table 12.

A study of table 12 shows that the two groups do not differ greatly in the character of their responses to the puzzles. The distribution of the number of moves is very much the same, and the writer can not find any marked differentiation between the groups in the character of their responses. The normal group is of course ahead of the retarded group in general. The median of the normal group is 3.5 moves lower than that of the retarded group on the Healy construction puzzle A, 15 moves on construction puzzle B, 11.5 moves on Bell A, 2 moves on puzzle A, 2.5 moves on puzzle B, 1 move on puzzle C, 8 moves on puzzle AC and 9 moves on puzzle BC. The highest diagnostic value found was 35% (Healy construction puzzle B).

The low diagnostic values found for the puzzle tests are surprising in view of the fact that the results from such tests are so frequently stressed as bearing considerable weight in mental examinations. Puzzle tests are used very largely at Ellis Island and at numerous clinics where mental examinations are made on individuals who have difficulty in the use of English. The reason that puzzle tests are stressed in such cases may be that no other tests can be used. Certainly the results of this investigation do not justify the confidence that is generally placed in such tests in the diagnosis of the higher grades of mental defect.

It may be argued that the element of chance operates in individual cases and might obscure a real correlation. If a subject

TABLE 12.

Distribution of Responses (Number of Moves) of Subjects in Solving Puzzle Tests.

M. D. V.	Healy A 18%	Healy B 35%	Bell A 24%	Bell B 10%	A 14%	B 20%	C 20%	AC 27%	BC 26%
Number of moves	N R	N R	N R	N R	N R	N R	N R	N R	N R
2					2				
3					5 3				
4			1 5		4 1	1	29 21		
5	4 4		2 1		4 3	1 2	11 13	4	
6	5 6		1 1		4 4	3 3	9 5	6 1	3 3
7	3 1		4		1 1	5 3	5 4	5 1	5 4
8	8 3		3		4 4	2 1	1 4	4 3	3 1
9	4 5		3 2		2 3	4	1 2	6 5	7
10	2 2		3 2		4 5	2 3	1 2	2 2	1 3
11	3	3 1	3 4	1	3 3	2 4	1 2	3 3	6 5
12	2 4	5 1	3 2	1	2 4	2 4	1	1 5	2 3
13	2	3 2	4 4	1	2 3	4 1		1 2	5 1
14	3 2	4 3		1	2	2 3	1	3 2	2
15	1 2	8 1	2		3	2 1		1	1
16	2 1	3 1	1 1	1 1	2 2	2 1		2	2 1
17	2 1	1 1	5 2		3	1 2		1	1 1
18	1	2 2	2 2	2 4	1	3 2		3 3	2
19	2	2		3	3 1	4 1		1 1	1
20		1 1	4 1	1	1	2		3	3
21	2	3		1	1			1	1
22	1 1	3 1				1 1		1	5
23	1	1		1	1 2	2 1		1	2
24		1 3			1	1			
25	1	1		1	1			1 1	1
26	2	1		1	2	1		1	4
27	1	2				1			1
28		1			1			1 1	1
29		2		3				1	
30		2							
31		1							
32		1 1							
33		1 2							
34		2 3							
Failed	11 22	11 29	17 30	45 47	7 12	12 24	0 4	11 23	12 24

has any such thing as a general puzzle solving ability, it would be unlikely that chance would operate against him in a series of nine puzzles, so that a combined score on all nine puzzles would guarantee the presence of any such general factor, while treating the combined score in the same manner as a single score would show

if this general factor were correlated with intelligence. The combined score of each subject was obtained by adding the number of moves that he made on each test. In case the subject failed the Healy construction puzzle A, bell B, puzzle A, B, AC or BC, his score was counted as 30 moves. Failure was counted 35 moves on construction puzzle B, 20 moves on bell A, and 15 moves on puzzle C. If a subject solved all the puzzles in the fewest possible number of moves his score would be 49, while if he failed all puzzles his score would be 250.

The total number of moves made by normal subjects varied from 93 to 220, the median being 138.5 ($Q=22.5$). The total number of moves made by retarded subjects varied from 96 to 245, the median being 171 ($Q=28.5$). The Maximum Diagnostic Value was found to be 34% when the passing mark was 150 moves. Combining the scores in order to give weight to any general puzzle solving ability therefore does not help matters any.

The fact that the puzzles used in this investigation did not show a high diagnostic value of course does not prove that all puzzle tests have the same low value. The results apply to this investigation only. In the writer's opinion however puzzle tests receive rather more weight in mental examinations than they deserve. The writer is frankly skeptical. The puzzles used in this series were all scored quantitatively, and it may be true that when they are scored qualitatively, when they are used merely to afford the experienced experimenter an opportunity to observe the subject's behavior on a concrete problem, they may have a high diagnostic value. Even then the experimenter should get his experience from examining a large number of normal individuals, for the writer has seen perfectly normal adults make the most impossible mistakes in solving puzzles, apparently because they had a definite attitude toward puzzles in general as being something to which reasoning does not apply, and in the solution of which the most rational method is that of trial and error. Of course intelligent persons do not place round blocks in square holes and leave them there, but this sort of ability is already satisfactorily tested by the form board.

The foregoing statements are simply an expression of the writer's opinion, and as such may be entirely false. The actual results only prove that the puzzle tests as given and as scored show a diagnostic value for the groups studied of but 34%. It is possible that if the shades of response scored were grosser the puzzles would be diagnostic of lower degrees of mental defect. The value obtained as an index of a test's merit by raising or lowering the difficulty of the test (the Maximum Diagnostic Value) is an accurate expression of the merit of a test only if the scoring is an adequate expression of the intellectual factors involved. As a matter of fact the puzzles used were well within the ability of the groups. The per cent. of subjects passing each test under the requirements is as follows,—

	Healy		Bell		Puzzles				
	A	B	A	B	A	B	C	AC	BC
Normal	81	81	71	22	88	79	100	81	79
Retarded	63	51	47	18	80	59	93	61	59

Healy says that most of his twelve year children solve construction puzzle A in from 12 seconds to 2 minutes, and he considers a subject as failing if he takes more than 10 minutes. The difference between the two minute response and the ten minute response cannot be expressed on a scale of 30 moves. It is quite possible that the puzzle tests might be diagnostic of the lower grades of defect if they were scored on a gross scale such as the number of minutes taken for the solution rather than the number of moves taken in a minute and a half. Under the conditions of the experiment it is not possible to determine the validity of this statement. It is only possible to show that the puzzle tests used have little value in diagnosing the higher degrees of mental defect.

TEST II. LIFTING THE TABLE ASYMMETRICALLY BALANCED.

A test was designed for testing practical judgment without the use of language in which the subject was required to lift with a hook a small triangular table asymmetrically balanced. A piece of board 1.4 cm. thick was cut in the form of an isosceles triangle 50.5 cm. on a side. 157 hooks numbered in rotation from left to right were attached to the upper surface of the board. On the bottom surface, legs 10 cm. long were attached to slides which could

be moved outward a distance of 12 cm., the distance being read on a millimeter scale. At the end of each slide a hook was fixed so that different sized weights could be suspended. The six weights used were ordinary 2 lb., 1 lb., 8 oz., 4 oz., 2 oz., and 1 oz. weights with hooks attached (the weights in their final form weighing 912 gms., 467 gms., 238 gms., 124 gms., 62 gms., and 33 gms. respectively). A button hook was furnished by which the subject was to lift the table. The apparatus is shown balanced in plate ID, the under surface being shown in plate IC.

By sliding the legs outward or inward and suspending different sized weights on the ends of the slides, the center of gravity of the apparatus could be thrown on any one of 43 hooks within a triangle whose outer surface ran about 13 cm. from the outer edge of the board. The table could be lifted by but one hook for each adjustment, one or two legs remaining on the floor for all of the remaining 156 hooks. 12 adjustments were used in this study, a scale for making them being attached to the under surface of the board. The table alone weighed about $1\frac{1}{2}$ kgms., and the heaviest possible combination of weights would make the whole apparatus slightly over 3 kgms.

The procedure in giving the test follows. The slides were adjusted and the weights attached while the subject was looking on. The experimenter then held the table a little above the floor and said, "You see this table is fixed with a heavy weight here and a little one here (etc.). I am going to give you a button hook, and you must pick up the table, so that it will balance, in just as few moves as you can. There is only one right hook to pick it up, and you must find that hook. If you get the wrong hook the table will tip up this way or that way (illustrating), if you get the right hook it will come up nice and level (illustrating). Now pick it up in just as few moves as you can." The table was placed on the floor, and the subject was not allowed to touch it with his hands. The instructions were given twice. The experimenter recorded the number of each hook in order as the subject tried them. No hint was given to the subject concerning the method but he was encouraged if he lost patience. The test was given to all the subjects, each subject having three trials.

The responses of the subjects varied from that of selecting the right hook immediately to that of selecting the hooks randomly and finally chancing upon the right hook. One subject made 97 attempts. Most of the subjects found the center of gravity in less than 10 moves. The distribution of the number of moves made by normal and retarded subjects in all three trials is as follows,—

No. of moves....	1	2	3	4	5	6	7	8	9	10	over 10
Normal	20	22	34	15	18	11	8	8	5	4	29
Retarded	20	18	20	17	13	12	9	5	10	5	47

The responses of the normal subjects varied from 1 to 42 moves, the median being 4 ($Q=2.5$), and those of the retarded subjects varied from 1 to 97 moves, the median being 5.5 ($Q=4$).

The factor of information was encountered and avoided by changing the adjustments. One 15 year subject made 23 attempts on the first trial and gave it up. On the second trial, after he had made 38 attempts the experimenter told him on which end of the board he could find the right hook, and he found it after 18 more attempts. On the third trial he found the center of gravity after 15 attempts. In spite of his poor performance he had enough intelligence to remember the number of the hooks by which he succeeded in lifting the table on his second and third trials, and he told the rest of the boys in his room. The experimenter used a different set of adjustments after this. The factor of information could enter by telling the subject to pick the table up near the end where the heaviest weight was suspended. There is no way of avoiding this factor.

The score of the subjects was not taken as the number of moves for the moves differed in their merit, 10 moves close to the right hook being a better response than 10 moves in various parts of the board. The distance in centimeters of each hook from each center of gravity was measured, and the score of the subject taken as the sum of the distances from each hook selected to the proper hook. For instance two subjects attempted to solve an adjustment which placed the center of gravity on hook no. 25. The first subject tried hooks 17, 33, 20, 15 and 25, the first four being 3 cm., 2 cm., 2 cm. and 5 cm. from no. 25, the solution. Another subject selected hooks 14, 17, 26, 29 and

25, these moves being 4 cm., 24 cm., 5 cm. and 3 cm. from no. 25. Both made five moves, but the total error of the first was 12 cm., and that of the second was 36 cm.

In the first trial the score of the normal subjects varied from 0 to 417 cm., the median being 31 cm. ($Q=35.5$ cm.). The retarded subjects varied from 0 to 1252 cm., to the median being 73 cm. ($Q=62$ cm.).

In the second trial the normal subjects varied from 0 to 371 cm., the median being 14 cm. ($Q=14.5$ cm.). The retarded subjects varied from 0 to 376 cm., the median being 15 cm. ($Q=23.5$ cm.).

In the third trial the normal subjects varied from 0 to 232 cm., the median being 7.5 cm. ($Q=9.5$ cm.). The retarded subjects varied from 0 to 1017 cm., the median being 17 cm. ($Q=21$ cm.).

The effect of practice is shown by the normal subjects in the reduction of the median from 31 to 14 to 7.5 cm. The retarded subjects improve from the first to the second trial (73 cm. to 15 cm.), but show no improvement from the second to the third trial (15 cm. to 17 cm.).

To study the diagnostic value the sum of the scores of each subject on each of the three trials was obtained. The sums of the scores of the normal subjects varied from 13 cm. to 798 cm., the median being 77 cm. ($Q=48$ cm.). The sums of the scores of the retarded subjects varied from 9 cm. to 2360 cm., the median being 121 cm. ($Q=83.5$ cm.). The Maximum Diagnostic Value was found to be 28% when the passing mark was 160 cm.

The test was given 131 times to 11 high school and college students. 14% of this group solved the problem in one move, 33% in two moves, 21% in three moves, 17% in four moves, 8% in five moves and 8% in more than five moves. The scores of these subjects varied from 0 to 94 cm., the median being 6 cm. ($Q=4.5$ cm.). The scores of the 12, 13 and 14 year normal subjects in their 174 trials, varied from 0 to 417 cm., the median being 15 cm. ($Q=16.5$ cm.). The maximum difference between the adult group and the normal group was found to be 35% when the passing mark was 15 cm.

The low diagnostic value found for this test (28%) indicates that it has little worth in differentiating the higher types of defect. It may be possible that the test would be serviceable in detecting grosser differences. 83% of the normal group and 73% of the retarded group solved the test in 10 moves or less so that the test was well within the ability of the groups. If a grosser means of scoring was used the test might detect lower degrees of defect.

TEST III. BALANCE TEST.

In order to test reasoning without language, a test was designed in which the subject was required to arrange weights subliminally different by means of a balance. The balance used was modified from C. H. Stoelting Co.'s Army Prescription Balance (catalogue no. 240). The balance was cut down and mounted on a new base as shown in plate IB. The weights used were five 200 gm. Stoelting Universal Laboratory Weights (catalogue no. 445), reduced as follows,—

no. 1	no. 2	no. 3	no. 4	no. 5
199.85 gms.	198.75 gms.	197.79 gms.	196.90 gms.	195.87 gms.

The first test given was to arrange three weights, the second five weights. In giving the test the experimenter would pick up two weights and say "These weights look alike but one is heavier than the other. I can't tell the difference by lifting them for they are too close together. If I want to find out which is the heavier I hang them on here this way (hanging the weights on the balance). Now which one is heavier?" After the subject had designated the heavier, he was given weights nos. 1, 2 and 3 with the instructions "I want you to put these weights in a row with the heaviest one first, then the middle one and then the lightest one, weighing them on the balance to find out which one is the heaviest, which one is middle and which one lightest." For the five weight test, the subject was given all five weights and told to "Put these in a row with the heaviest one first, and then the next, and then the next, etc."

The test as it stands is not worth standardizing, but it showed several suggestions that may be worth following out. The ob-

jectionable features to the test were due to the fact that the subjects were constantly on the alert for other criteria to judge the weights than the mere position on the balance. In the first place, they would try to sense the difference, and it was possible to discriminate between some pairs, although it was unlikely that they could be arranged in order. The experimenter would correct the subjects if they tried this method and tell them to use the balance. Occasionally the subjects would watch the pin and the scale, but as the weights were so unlike that one member of any pair always rested on the base of the balance, his method gave no help. If the subjects used this method they were told that it did not help. The third method was that of "watching the bounce" as the subjects called it. If weight no. 5 were on one side, and weight no. 1 were placed on the other, it would fall with greater force than weight no. 4 would in the same position. The subjects were also told not to use this method.

Inasmuch as the test was new the experimenter did not properly understand it, and gave the retarded group more credit than they deserved. The solution of the three weight problem depended somewhat on the chance selection of the pairs by the subject. Most of the subjects would leave the heavier weight of the first pair on the balance and compare the third with this. If the subject first compared nos. 2 and 3, and then compared 1 and 2, the problem was easy. If he first compared 1 and 3, and then compared 1 and 2, the problem was more difficult because another comparison (2 and 3) was necessary.

The results of this test are given with the reservation that they are not absolutely accurate. 90% of the normal group and 51% of the retarded group arranged the three weights correctly, making the diagnostic value 39%. This value should be higher for many of the retarded group were given credit for arranging the weights correctly even if the method were wrong, i.e., if they placed the weights in the right order without making all the necessary comparisons. 83% of the normal group and 34% of the retarded group arranged the five weights correctly, making the diagnostic value of this test 49%.

The writer has given the five weight test to many intelligent

adults, college professors, college graduates etc., and has found that the test is much easier for 12, 13 and 14 year normal boys than for normal adults. The child sees but one method of arranging the weights, that of elimination. He will make all the comparisons possible, always leaving either the heaviest or the lightest weight on the balance (usually the heaviest). When he has compared this one with all the others, he will sometimes try all of the other four again to make sure he has the heaviest one, and then place it to one side with the remark "There that one's heaviest" or "That one's king." He will then proceed through the remaining four in the same way, then through the three etc. His method is the longest one possible but it is absolutely certain. The intelligent adult not only tries to arrange the weights correctly, but he tries to do it in the fewest possible number of moves. He invents short cuts and tries to remember previous moves with the result that he frequently becomes lost in his own complications.

The three weight test is more valuable than the five weight test and more diagnostic if the conditions are controlled. The most important test is to give the subject nos. 1 and 3 for the first comparison, and nos. 1 and 2 for the second comparison, and see if he will make the arrangement without comparing 2 and 3. In the writer's experience intelligent children and adults will always refuse to make this unqualified generalization, while the unintelligent person is ever ready to accept it.

TEST IV. HEALY CROSS-LINE AND CODE TESTS.

The cross-line and code tests described under Tests IX, X and XI on pages 28 to 34 of Healy's monograph were given. The procedure was changed somewhat to meet the needs of the retarded group. The cross-line test A was used entirely as a practice test. A large "X" was drawn on the paper before the subject and the four numbers filled in. "You see I draw this cross and fill in the numbers. Now if I want to make a mark that stands for 1, I make it this way (drawing the symbol for 1), because the 1 here points that way. Now if I make a mark like this (drawing the symbol for 2), what would that stand for?"

In this way all four symbols were drawn and the subject practiced until he could name them all with the original diagram before him. The paper was then turned over and the subject required to name the proper digit for each of the marks as they were drawn before him in irregular order. If the subject named all four correctly, he was marked plus.

The cross-line test B was given with briefer instructions. The figure was drawn before the subject and digits filled in. The subject was asked to name the proper digit for the symbols of 7, 5 and 2. After he understood the arrangement he was told to study the figure carefully for the paper would be turned over and he would be asked to name all the figures. When the subject said he was ready, the paper was turned over and he was required to name all the symbols as they were drawn in irregular order. If he named any wrong, other symbols were given, and the wrong ones given later. If the subject named all nine digits correctly he was scored plus. The code test was drawn out for the subject in the same manner as the cross-line test and the system of dots explained. The subject was required to distinguish between the symbols for c and l, t and x. He was then told to study the figure, and when he said he was ready, the paper was turned and two or three symbols were given for each figure. The task was continued till the experimenter was certain that the subject could or could not perform the task. If the subject failed cross-line test B, he was marked failed on the code test without trying it. In this way the three tests were given to all the members of both groups.

98% of the normal subjects passed cross-line test A, 84% passed cross-line test B, and 57% passed the code test. 64% of the retarded subjects passed test A, 20% passed test B, and none passed the code test. The diagnostic value for the first test is therefore 34%, for test B 64%, and for the code test 57%.

Wyatt (81) found rather of a low correlation between the teachers' estimates of the intelligence of a group of 34 boys and girls and a cross-line test modelled somewhat after cross-line test B, the correlation (0.46, $p < 0.09$) standing twelfth in a list of fifteen tests. Wyatt's procedure probably accounts for the com-

paratively low correlation. After the cross-line test A had been drawn on the board of the class-room and explained, the second figure was exposed for 20 seconds and the subject required to fill in the symbols on a prepared blank. The short exposure probably made success in the test depend on visual memory rather than on the rational comprehension of the arrangement of the figure. This interpretation is supported by the fact that a test of memorizing letter squares correlated higher with the cross-line test than with any other.

Healy says of these tests "These three tests are especially noteworthy and valuable because their correct performance seems to demand mental powers which appear strongest in the normal adult mind and which are weakest in mentality of the child type" (Page 29). Concerning the cross-line test B he says "On account of the readily ascertained differences in performance between bright subjects and dull subjects, we have come to regard the test as extremely valuable" (Page 31). The results of this investigation certainly justify Healy's belief in the value of these tests. Only two of the Binet tests show a diagnostic value higher than 64%. Haines found the cross-line test B valuable in differentiating the not defective from the doubtful.

TEST V. MEMORY FOR COMMISSIONS.

Four tests of memory for commissions were used. The materials (a penny, key, knife, eraser, book, saucer, small card with 1c stamp, small card with 2c stamp, and two penny match boxes, one covered with blue and the other with red paper) were placed on the table in front of the subject. The following tests were used,—

Test 1. Give me the penny and then put the key in the saucer.

Test 2. Put the saucer on the book, then put the key in the saucer, and then give me the eraser.

Test 3. Put the eraser in the saucer, then put the penny on the book, then turn over the card with the red stamp on it, and then give me the blue box.

Test 4. Give me the red box, then put the eraser on the book with your left hand, then put the knife on the blue box, and then give me the card with the green stamp on it.

The commissions were read twice to the subject. During the reading the articles on the table were covered with a card board screen. Tests 1 and 2 were given to 95%, test 3 to 93% and test 4 to 75% of the retarded group. All four tests were given to all of the normal subjects. 100% of the normal subjects passed tests 1 and 2, 52% passed test 3, and 14% passed test 4. 98% of the retarded group passed test 1, 73% passed test 2, 22% test 3, and 2% test 4. The diagnostic value for the first test is therefor 2%, for the second 27%, for the third 30% and for the fourth 12%.

Inasmuch as the four tests are of the same sort they may be scored together on a scale of 10, each test being weighted according to its difficulty. 99% of all subjects passed the first test, 87% passed the second, 37% passed the third, and 9% the fourth. The tests were weighted on a scale of ten according to the formula,—

$$\frac{\text{reciprocal of per cent. passed} \quad x}{\text{sum of reciprocals}} = \frac{\quad}{10}$$

This formula gives the value of 0 for the first test, $\frac{1}{2}$ for the second test, 4 for the third test, and $5\frac{1}{2}$ for the fourth test. The scores of the normal and retarded subjects were distributed as follows,

No. of points....	10	6	$4\frac{1}{2}$	4	$\frac{1}{2}$	0
Normal	6	2	24	0	36	0
Retarded	0	1	10	2	30	13

Calculating the percentage passed by normal and retarded subjects at each passing mark from 0 to 10, the Maximum Diagnostic Value was found to be 35% when the passing mark was $4\frac{1}{2}$ points. This value is close to that found for the test of repeating digits (30%), and copying designs from memory (36%), but higher than that found for repeating sentences (23%). None of the memory tests shows a high diagnostic value.

Terman (65) reports "that after the age of 12 or 14 years memory for relatively meaningless material, like digits or non-sense syllables, improves but little; and that above this level it does not correlate very closely with intelligence (page 323).

Abelson (1) gave nine tests to 88 backward boys and 43 backward girls, averaging about 11 years in age, and found that a test of memory for commissions showed a higher correlation with "practical intelligence" than any of the other tests (.53 for girls and .65 for boys). However, he instructed the teachers to estimate the "practical intelligence" of the children by considering, in forming their opinion, which of the children they would soonest trust on an errand requiring the sharpest intellect. The correlation found possibly proves that a psychological test of ability to run errands correlates with the teacher's judgments of this ability, but proves nothing concerning the relation of this ability to intelligence.

TEST VI. DISTINGUISHING BETWEEN TERMS.

The subjects were asked to distinguish between three pairs of terms, "steam and smoke," "lie and mistake," and "laziness and idleness." The questions were asked in the form "What is the difference between—?" The answer required for the first pair, "steam and smoke," was that steam came from water while smoke did not. If the subject said that smoke was black and steam white, they were asked if they had never seen white smoke. The distinction between the second pair required was that a lie was intentional and a mistake accidental. It was not required that the words "accidental" and "intentional" be used, but that their meaning implied. For the third pair it was required that the subject imply that laziness was due to a subjective condition while idleness might be due to accidental circumstances.

98% of the normal group and 73% of the retarded group distinguished between "steam and smoke," making the diagnostic value 25%. 66% of the normal group and 27% of the retarded group distinguished between "a lie" and "a mistake," making the diagnostic value 39%. 45% of the normal subjects and 3% of the retarded subjects distinguished between "laziness and idleness," making the diagnostic value 42%. 85% of all subjects passed the first pair, 46% the second pair, and 24% the third pair. Weighting the tests roughly on a scale of 10 in the same manner as the commissions test was weighted, the first pair re-

ceives a value of 1 point, the second 4 points and the third 5 points. 1 retarded subject scored 10 points, 15 scored 5 points, 1 scored 4 points, 28 scored 1 point, and 14 scored nothing. 19 of the normal subjects scored 10 points, 7 scored 6 points, 18 scored 5 points, 1 scored 4 points, and 13 scored 1 point. Calculating the percentage passed at each passing mark, the Maximum Diagnostic Value was found to be 49% when the passing mark was 4 or 5 points.

TEST VII. SUBTRACTION TESTS.

In an article on the detection of higher grades of mental defect, W. E. Fernald (26) stated that feeble-minded individuals have difficulty in subtracting. A series of 10 subtraction tests was devised following the suggestion of Prof. H. C. McComas. On account of the time required by the test, only three of the tests were continued throughout the experiment. The three tests were as follows,—

Test 1. Subtract 3 from 16 till you get to 7, then subtract 2 till you get to 1.

Test 2. Subtract 4 from 30 till you get to 10, then subtract 3 till you get to 1.

Test 3. Subtract 3 from 43 till you get to 25, then subtract 4 till you get to 9, then subtract 2 till you get to 1.

A preliminary practice test was given in which the subject was asked to subtract such figures as 4 from 13, 4 from 31, 3 from 22, etc. If he could not do these the rest of the test was not given. The problems were repeated twice for the subjects.

98% of the normal group passed test 1, 90% passed test 2, and 60% passed test 3. 24% of the retarded passed test 1, 17% passed test 2, and 7% passed test 3. The diagnostic value for the first test is therefore 74%, for the second test 73%, and for the third test 53%. The retarded group had great difficulty in passing the first problem and even in doing any sort of work in subtraction. This ability would of course depend on school training, but if the amount of a child's training is known, the results show that this test would be very valuable in diagnosing intelligence. The retarded group on an average had been in

school over 7 years, yet only 24% of them passed the first test. The fact that the experimenter gave himself an opportunity of deciding on the basis of the preliminary problems whether or not the other problems should be tried, affords an opportunity for the influence of the personal equation. It is very possible that if the experimenter had exercised more patience in some cases more of the retarded group would have passed the easier tests, (nos. 1 and 2), and the diagnostic value been lowered. On this account the diagnostic values obtained (74% and 73%) are probably too high, but the fact remains that there is a very large difference between the normal and retarded groups in their performance on these tests. The writer has used simple subtraction problems in examining mature defectives and suspected cases of dementia, and has found the tests very helpful and suggestive. All individuals must do a certain amount of subtracting in their general daily experience in counting change etc. The general custom among store keepers of counting the change up from the amount of the purchase to the amount of the piece of money given is probably evidence that some of their customers have difficulty in subtracting.¹

TEST VIII. SUGGESTION BY PROGRESSIVE LINES.

A test of suggestion by progressive lines, modelled somewhat after Test 42 of Whipple's Manual (76)² was used. The apparatus consisted of 13 cards 4 x 28 cm., each card having a line drawn 2 cm. from the left side equidistant from the top and bottom. The lengths of the lines on the first six cards were 10, 20, 30, 40, 50 and 60 mm. The lines drawn on the last seven cards were all 60 mm. in length. The subject was given a sheet

¹ In giving the subtraction tests, the writer has met many newsboys who can not do problems such as 4 from 22, 5 from 31, etc., but who can make change from some of the smaller pieces of money by means of finger counting systems, etc. The concrete problems are much easier.

² The test described by Whipple involves the use of a kymograph which is inconvenient. The test was shortened by using 13 lines instead of 20. The coefficient of suggestibility used in this investigation is more accurate than that given by Whipple because the measure of a subject's ability to reproduce lengths is obtained from six reproductions of lengths instead of one.

of 2 mm. cross section paper with heavy rulings at each centimeter. The instructions were "I am going to show you cards like this, and this (showing card no. 1, and one of the last seven), and I want you to draw lines on the paper just as long as the ones I show you." The subject was shown where to draw the lines on the paper, beginning at the left hand margin. The subject was allowed to study each line as long as he wanted to, and as soon as he started to draw a line and turned his attention to the paper, the card was dropped and the next one exposed.

The suggestion arising from the fact that the first six lines increase in length will usually cause the subject to increase the length of the last seven lines. The measurement of the amount of suggestion is made by comparing the increase in length of the last seven lines with the measurement of the subject's ability to reproduce the first six lines. The total actual length of the first six lines is 210 mm. The measurement of the subject's ability to reproduce lengths is the difference between 210 mm. and the sum of the lengths drawn. If no suggestion were present the length of the last seven lines would be to the length of the first six as 420 mm. (the actual length of the last seven) is to 210 mm., or the sum of the last seven should be twice the sum of the first six. The measurement of the amount of suggestion is therefore the difference between the sum of the last seven lines and twice the sum of the first six.

The lines drawn by the subjects were measured within 2 mm. and the length of each line recorded. The sum of the first six (SF) and last seven (SL) lines were then computed and the coefficients of accuracy and suggestion computed according to formulae explained,—

Accuracy in reproducing lengths= $SF - 210$

Coefficient of suggestion= $2SF - SL$

A coefficient of accuracy with a plus sign indicates that the subject over-estimated the lengths, a minus coefficient that he under-estimated the lengths. A minus coefficient of suggestion indicates the influence of suggestion while a positive coefficient indicates no influence or a negative influence.

In general the subjects under-estimated the length of the first six lines and accepted the suggestion of increasing lengths. One subject increased each line one centimeter so that the last line was 130 mm. in length, over twice the length of the line to be copied. The average length of each line was computed for normal and retarded subjects. These results are given in Table 13, and are shown graphically in Fig. 3.

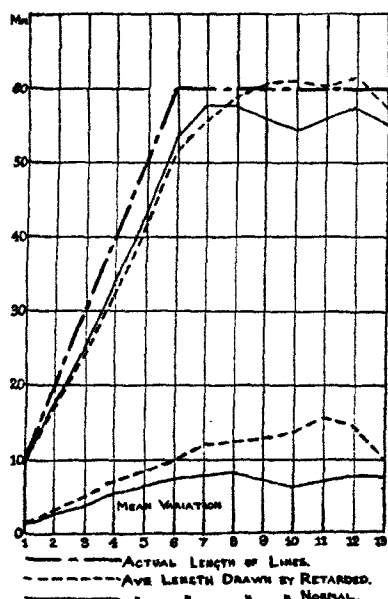


FIG. 3. Results of Normal and Retarded Subjects on Test of Suggestion by Progressive Lines

The normal subjects show the greatest influence of suggestion on the 7th and 8th lines, while the suggestion is still fresh. They then throw off the suggestion a little on the 9th line and more on the 10th but increase on the 11th and 12th, dropping back again on the 13th. The retarded subjects increase to the 9th and 10th, drop slightly on the 11th, increase on the 12th, and drop back on the 13th. The sum of the average lengths of the first six lines for normal subjects is 181.26 mm., and for retarded subjects 176.24 mm. The normal subjects underestimate

TABLE 13.

Suggestion by Progressive Lines.

Average Length and Mean Variation in Millimeters of Each Line Reproduced by Normal and Retarded Subjects.

Line no.	Actual length	Normal subjects	Retarded subjects
1	10	9.34 (1.12)	9.69 (1.29)
2	20	17.34 (2.59)	17.19 (3.10)
3	30	24.79 (3.79)	24.68 (4.88)
4	40	33.48 (5.38)	32.05 (6.85)
5	50	43.17 (6.19)	41.22 (8.34)
6	60	53.14 (7.31)	51.41 (9.83)
7	60	57.62 (7.69)	55.47 (12.00)
8	60	57.62 (7.95)	58.56 (12.20)
9	60	55.90 (7.21)	60.73 (12.56)
10	60	54.27 (6.31)	60.90 (13.52)
11	60	55.76 (6.93)	60.15 (15.44)
12	60	57.17 (7.66)	61.47 (14.78)
13	60	55.03 (7.61)	57.15 (9.95)

the length of the first six lines 28.74 mm., and the retarded subjects 33.76 mm. The sum of the average lengths of the last seven lines for normals is 393.37 mm., for retarded 414.43 mm. These totals compared to the ability to reproduce the first six lines show a coefficient of suggestion for normals of —30.84 mm., and for retarded of —61.95 mm. In general then retarded subjects underestimate the length of the first six lines and overestimate the length of the last seven lines more than normal subjects.

The coefficients of accuracy found for normal subjects varied from +28 mm. to —92 mm., the median being —30 mm. ($Q=23$ mm.). The coefficients of accuracy for retarded subjects varied from +54 mm. to —112 mm. the median being —24 mm. ($Q=29$ mm.). The Maximum Diagnostic Value was found to be 19% when the passing mark was taken at —50 mm.

The coefficients of suggestion found for normal subjects varied from +50 mm. to —210 mm., the median being —26 mm. ($Q=22$ mm.). The coefficients of suggestion found for retarded subjects varied from +86 mm. to —342 mm., the median being —56 mm. ($Q=45$ mm.). The Maximum Diagnostic Value was found to be 28% when the passing mark was —74 mm.

The diagnostic value found for the Binet line suggestion test was 12%, and this value taken in connection with the maximum diagnostic value shown by this test (28%) is fairly conclusive evidence that the correlation between intelligence and suggestibility is not very high, and that tests of suggestion have little value for diagnostic purposes.

TEST IX. ESTIMATION OF LENGTHS.

A test of estimating lengths was arranged after the suggestion of Prof. H. C. McComas for obtaining a simple intellectual judgment on sensory material. Four lines were drawn on each of ten cards 12 x 7½ cm. The top line was in all cases 100 mm. in length. The other lines were of lengths such that two of them if combined would be equal in length to the standard line. Ten cards were used, graded in difficulty so that the error of the wrong combinations varied from 25 mm. to 5 mm. The lines were drawn equidistant from the sides of the cards. Two of the cards (nos. 1 and 7) are shown in Fig. 4.

In giving the test the cards were presented in order. The instructions were "Here is a card with a long line at the top and three other lines here. Two of these lines if put together will make the top line. Which two lines would you put together to make the top line?" After the subject had judged the first card, whether he answered right or wrong, the lines were measured

TABLE 14.

Estimation of Lengths

Length of Lines Judged and Percentage of Correct Judgments.

Card no.	Length in mm. of lines no.			Per cent. passed by normal	Per cent. passed by retarded
	1	2	3		
1	75	50	25	88	71
2	75	25	5	98	86
3	90	70	30	76	64
4	70	30	10	100	81
5	65	50	35	60	54
6	80	65	35	57	47
7	60	40	30	93	75
8	70	60	40	41	29
9	55	50	45	28	29
10	60	55	45	21	10

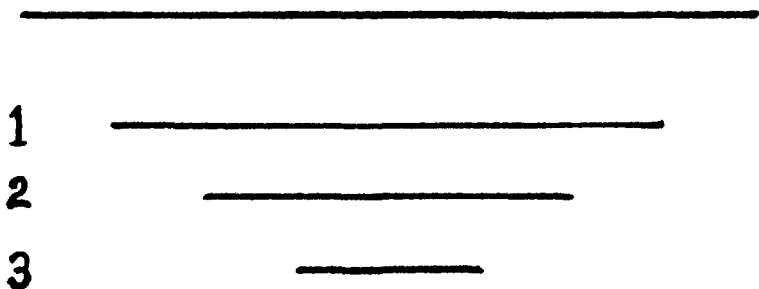
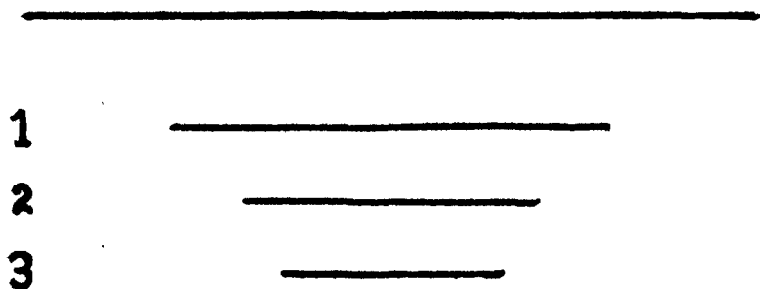
NO.1**NO.7**

FIG. 4. Estimation of Lengths Test.

off on a piece of paper, added together, and compared with the standard in order to illustrate the principle of the test. The subject was allowed all the time he desired to make the judgments. The measurements of the lines on each of the ten cards, and the per cent. of correct judgments on each card are shown in table 14.

From table 14 it may be seen that the difference between normal and retarded subjects is not more than 19% on any one set of lines. Scoring the test as a whole by counting the number of correct judgments in ten, the record of the normal and retarded subjects was as follows:

		no. of correct judgments									
		1	2	3	4	5	6	7	8	9	10
Normal	0	0	2	4	16	3	12	11	8	2
Retarded	1	0	3	12	14	16	6	6	1	0

Calculating the percentage that would have passed had the passing mark been fixed at any point, the Maximum Diagnostic Value was found to be 35% when the passing mark was 7 correct judgments in 10.

TEST X. CLASS-ROOM REASONING TEST.

A special test was designed to involve reasoning ability. The test included seven questions with two to four sub-questions under each. The questions varied greatly in difficulty so that the test could be given to a wide range of subjects. A special blank was prepared with questions printed on both sides of a paper 22 x 26 cm. Places were marked at the top of the test blank for the subject to fill in his name, school, grade, the date, and the date of birth. The directions printed at the top of the blank were as follows:

Write your name, school, grade, date of birth and date in the places marked at the top of the paper.

Answer each of the questions below in one or two words. Write your answer in the place marked "Answer" like this:

- a. When are days warmer, in February or June?

Answer—June.

- b. Is it because the sun's rays are slanting or straight at that time?

Answer—Straight.

All the questions may be answered in one or two words, such as "Yes," "No," "Right," "Left," "Not necessarily," "Wrong," etc.

Under each question a space was left after the word "Answer—." The questions asked were as follows:

- 1, a. When are days the longer, in January or in July?
b. Is it because the sun rises earlier or later?
- 2, a. If anything floats in water, is it lighter or heavier than water?
b. If a thing is heavier than water, will it float or sink?
c. If a thing floats in air, will it float in water?
d. If a thing sinks in air, will it sink in water, too?
- 3, a. Where is the sun in the morning, North, West, East or South?
b. If you face the South, will the West be on your right or your left?
c. If you face the North in the morning, which side will your shadow fall on, your right or your left?
- 4, If you put a stick in the water on a slant, it will look as if it were bent upwards.
a. If you were going to spear a fish from the side, would you aim above or below the fish?
b. Why? Is it because the fish seems to be above or below the place where he really is?
- 5, a. When are shadows longer, in summer or in winter?
b. Is it because the sun is farther North or farther South?
- 6, Supposing a clock starts running backwards at six o'clock.
a. In six hours, will it tell the right or the wrong time?
b. In twelve hours and a half, will it tell the right or the wrong time?
Supposing another clock starts running backward at half past six.
c. In six hours and a half, will it tell the right or the wrong time?
d. In twelve hours will it tell the right or the wrong time?
- 7, Water and cream both float on milk because they are lighter than milk. Water is heavier than cream.
a. Will cream float or sink in water?
b. If a thing floats on milk, will it float on water?
c. Supposing something sinks in cream, will it sink in milk?
d. Will a thing float on water if it floats on cream?

Questions 1a, 1b, 2a, and 5a were taken from questions 1, 4, 7 and 9 of set *a*, test IIIB of Bonser's (12) list of tests. The rest are new. The test was given individually and as a class room test. When given individually, the experimenter would give the subject a paper and read the instructions to him. The experimenter would then read each question and record the subject's answer, the subject in each case having a paper so that he might read and follow the experimenter. When given as a class room test, the teacher of the room would distribute the papers and tell the class to read the instructions and start to work. Any questions asked were referred to the instructions. As the children finished, they would bring their papers to the teacher who would glance over them to see that all the questions had been answered. If any questions were unanswered, the children were made to return to their seats and fill them in. No time limit was set in either the individual or group testing. The subjects were allowed all the time necessary. The test very rarely took over 20 minutes. The subjects were told that they might look at a clock or watch for question 6. The class-rooms had clocks on the walls, and the experimenter showed the individual examinees his watch.

The test was given individually to 56 members of the retarded group. It was given as a class-room test to all the children in the seventh and eighth grades of the Franklin School in Trenton, and to a class of college juniors and seniors in Princeton. The experiment at the Franklin School was conducted by the departmental teacher of geography who followed the instructions carefully. The experiment on college juniors and seniors was conducted by the professor in charge of the course with the assistance of the writer.

The age and grade distribution of the seventh and eighth grade girls is as follows:

	Age						Tot.
	11	12	13	14	15	16	
Grade VII	6	18	24	8	4	2	62
Grade VIII		2	15	19	15	1	52

The age in grade distribution of the seventh and eighth grade boys is as follows:

	Age					Tot.
	11	12	13	14	15	
Grade VII	2	22	24	21	5	74
Grade VIII		3	12	15	7	37

The average age (at last birthday) of the seventh grade girls is 12.87 yrs. ($MV=0.87$ yrs.), of the eighth grade girls 13.98 yrs. ($MV=0.67$ yrs.), of the seventh grade boys 13.07 yrs. ($MV=0.79$ yrs.), of the eighth grade boys 13.70 yrs. ($MV=0.74$ yrs.). The seventh and eighth grade boys included 53 members of the normal group.

The age distribution of the 41 college juniors and seniors was as follows:

Age	19	20	21	22	23	24	Tot.
No. of subjects..	1	11	12	14	2	1	41

The average age of the college students was 21.20 yrs. ($MV=0.86$ yrs.)

The results were calculated for each group on each question. In 18 of the 21 questions the right and the wrong answers were given in the questions. Questions 2d, 7b and 7c had only one answer possible, "Not necessarily" or some equivalent such as "Not always." "Sometimes" etc. Question 3a had four possible answers, so that 25% could answer correctly by chance. The remaining 17 questions could be passed by 50% by chance. The results were also calculated for combinations of parts of questions. One half of the subjects could pass either part of question 1, by chance, but only one fourth could answer both parts correctly. The scores were calculated for passing all parts of questions 1, 3, 4, 5, and 6, for parts a and b of question 2, for parts a, b, and c of questions 2 and for parts a and d of question 7. The chances are 1 out of 4 for passing 1ab, 2ab, 4ab, 5ab and 7ad, 1 out of 8 for passing 2abc, and 1 out of 16 for passing 3abc and 6abcd.

The results are shown in table 15 for each question and for the various combinations of questions as passed by the retarded and normal groups, by seventh and eighth grade boys and girls and by college adults.

If the personal equation figured in the results of this test it could only affect the retarded group as the test was given to all

TABLE 15.

Per Cent. that 322 Subjects Passed Reasoning Tests.

No. of subjects	Retarded group.	Normal group.	Grade VII boys.	Grade VIII boys.	Grade VII girls.	Grade VIII girls.	Adults
Test	56	53	74	37	62	52	41
1a	64	98	95	100	90	92	95
1b	71	89	88	89	85	85	88
1ab	46	89	85	89	82	83	88
2a	82	96	93	97	79	94	98
2b	96	98	95	100	84	92	98
2c	52	79	76	84	60	50	98
2d	7	17	15	16	11	29	93
2ab	80	96	92	97	74	92	98
2abc	46	75	70	81	47	46	95
3a	45	89	89	95	87	87	100
3b	48	72	70	81	50	63	98
3c	52	68	69	81	61	60	100
3abc	18	53	46	73	39	37	98
4a	41	55	66	62	52	50	83
4b	21	57	53	51	42	35	88
4ab	11	38	38	38	21	17	80
5a	16	57	53	51	40	56	78
5b	54	58	47	70	45	54	56
5ab	4	25	23	30	11	25	46
6a	27	72	74	81	68	75	95
6b	63	60	72	54	44	52	85
6c	54	49	45	46	37	44	78
6d	63	66	59	70	56	65	88
6abcd	5	30	27	27	6	19	66
7a	66	85	77	73	74	77	85
7b	0	8	4	5	6	8	76
7c	0	9	4	11	5	12	68
7d	66	91	76	97	63	75	76
7ad	48	75	59	73	45	65	66

other subjects in classes. If a question is too difficult for a group the percentage passed should at least approximate that expected by chance. According to chance 50% of all subjects should pass a question with two alternatives such as 5a. Table 15 shows that only 16% of the retarded group pass this test, so that it is fair to assume that this question was weighted against the subject in some way. Only 21% of the retarded subjects pass question 4b and only 27% of that group pass 6a, it being legitimate to expect 50% by chance. If there is anything in the form of the question, any popular misconception, or any other constant factor that would tend to make the wrong answer appear more frequently, the same question should show a deviation in the same

direction in the results of the other subjects. Minus deviations of 8% and 15% occur in the results of seventh and eighth grade girls on question 4b, and of 10% in question 5a. Minus deviations occur in the results of the groups to whom the test was given in class, but in no case are these deviations more than 15% below that expected by chance. It is right to assume then that questions 4b, 5a and 6a were influenced by the personal equation of the experimenter.

The test was given individually to the retarded group by the writer who is positive that there was no conscious intention to throw the results one way or another. However it seems legitimate to assume that there was something in the way the questions were read, a slight stress of the voice on the wrong alternative, a factor not consciously analyzable by the experimenter but one that was strong enough to throw the results in a definite direction in the long run. If this is not the case, and a minus deviation of 34% may still be attributed to chance, it is necessary to assume that a plus deviation of 34% may also be due to chance. Or again, if the experimenter unconsciously forced the results against the subjects in some questions, he might have favored the subject in others, and if 34% is the magnitude found in one direction, it is fair to expect an equal deviation in the opposite direction. The results of the retarded group are above 84% in only one question, 2b. It is therefore necessary to discard all the results of the retarded group on this test. No comparisons can be made with the normal group, and no indication obtained of the diagnostic value of these tests for the normal and retarded groups.

The comparison of seventh and eighth grade boys and girls shows slight sex differences in favor of the boys. Comparing the differences between the results of the sexes in both grades in all tests scored individually and in combination (58 comparisons), the median is 7.5% ($Q=6.5\%$) in favor of the boys. The differences are 20% or higher in 17% of the cases, and higher than 30% in 3% of the cases. The largest differences are in favor of the boys in question 2abc and 3abc.

The comparison of the results of seventh and eighth grade

boys and girls (range of ages from 11 to 16) with the results of college adults (range of ages 19 to 24) affords some indication of the value of these tests in differentiating adolescents from adults. Tests 7ad, 1ab, 2ab, 2abc and 3abc are too easy for adolescents to afford any diagnostic value. Test 5ab is apparently too difficult for adults and therefore worthless. Test 6abcd is rather hard for adults. The subnormal group might have guessed the answers to the clock questions, but the writer is certain on the basis of the individual examinations, that very few could figure the answers out. This ability seemed uniformly present in college adults, the mistakes being due to carelessness. Of the 164 questions in this test answered by college adults, 87% of them were answered correctly. Of the 900 questions in this test answered by the seventh and eighth grade boys and girls, 59% were answered correctly, only slightly better than the 50% to be expected by chance. If the question had been worded "What time would it be?" or the test arranged to bring out this factor, the diagnostic value of the test would probably have been demonstrated. Test 4ab seems to show considerable value in differentiating adolescents from adults. This test is passed by 28% of the seventh and eighth grade subjects, 25% being expected by chance, and is passed by 80% of the college adults.

The greatest differences between the adolescent and adult groups are shown in questions 2d, 7b and 7c, all of which involve the answer "not necessarily." The child almost always answers "yes" or "no" to these questions. It would seem from this that the child is willing to generalize too easily. The adult generally refuses to make an unqualified generalization. The child stumbles into it. This same factor appeared in the test of arranging three weights by means of a balance. Tests involving this factor would seem to be worth developing in the extension of measuring scales upwards.

It is interesting to note that Martin (44) finds this same type of response a valuable test for the upper years. Martin applied the Binet and DeSanctis tests to 212 normal children and 150 feeble-minded children. She noted the character of the responses of the subjects to questions 6a of the De Sanctis series

("Are large things heavier or lighter than small things?"), which was intended as a preliminary question to 6b ("How does it happen that sometimes small things are heavier than large things?"). It was found that children of the higher "mental ages" tended to qualify their statements by saying "that depends on the material," "large things are usually heavier," or by such words as "many," "sometimes," "often," etc.

Martin found that 24% of the 21 feeble-minded subjects of "mental age" 10 and 20% of the 10 normal subjects of the same "mental age" gave qualified answers. Only 2 of the 9 normal and feeble-minded subjects of "mental age" 11 failed to give qualified answers. The test was also given as a class-room test in several school grades. 27% of the children in the fourth grade, 68% of those in the fifth grade, and 81% of those in the sixth grade gave qualified answers. Martin concludes that "If note is made of the qualified answers, it would seem that the question is quite valuable in itself and might be used among the tests for the upper years." (page 102)

Inasmuch as questions 2d, 7b and 7c of the present investigation are obviously too difficult for seventh and eighth grade children, and Martin's data show that the qualifying response to question 6a of the De Sanctis series is well within the ability of the fifth and sixth grades, it would seem that the intellectual level is indicated by the refusal to make an unqualified generalization *from given material* rather than by anything in the nature of the process itself. The level is apparently indicated by the refusal to make a certain generalization, not by a change in the character of the reasoning process itself.

The form in which the whole reasoning test was given is unsatisfactory for individual testing. The method in which the right and the wrong answers are given is valuable for a class-room test for it saves time. The large number of subjects that may be obtained by the method shows whether an ability is present or absent. One can never tell however whether an individual child has not guessed the answer. There is no reason why the form of the questions should not be changed, and the tests used to bring out the factors found valuable in differentiating adolescents from adults.

SUMMARY OF SUPPLEMENTARY TESTS.

The list of supplementary tests used arranged in the order of their diagnostic value is shown in table 16.

TABLE 16.

Diagnostic Value of Supplementary Tests.

	% passed normal	% passed retarded	Diagnostic value
Subtraction Test no. 1*.....	98	24	74
“ “ no. 2*.....	90	17	73
“ “ no. 3*.....	60	7	53
Healy Cross-Line Test A.....	98	64	34
“ “ B.....	84	20	64
“ Code Test	57	0	57
Balance Test. 3 weights*.....	90	51	39
“ “ 5 weights*.....	83	34	49
Distinguishing between Steam and Smoke*.....	98	73	25
“ “ Lie and Mistake.....	66	27	39
“ “ Laziness and Idleness.....	45	3	42
Weighted score on all three pairs.....	(Max. diag. value)		49
Estimation of Lengths*.....	“	“	“ 35
Memory for Commissions. Test 1*.....	100	98	2
“ “ “ Test 2*.....	100	73	27
“ “ “ Test 3*.....	52	22	30
“ “ “ Test 3*.....	14	2	12
Weighted score on all four tests.....	(Max. diag. value)		35
Puzzle Tests.			
Healy Construction Puzzle A.....	“	“	“ 18
“ “ “ B.....	“	“	“ 35
Bicycle Bell Puzzle A*.....	“	“	“ 24
“ “ “ B*.....	“	“	“ 10
Puzzle A	“	“	“ 14
“ B*.....	“	“	“ 20
“ C*.....	“	“	“ 20
“ AC*.....	“	“	“ 27
“ BC*.....	“	“	“ 26
Learning from A to AC*.....	“	“	“ 11
“ “ B to BC*.....	“	“	“ 13
Pooled score on all nine puzzles.....	“	“	“ 34
Lifting the Table Asymmetrically Balanced*.....	“	“	“ 28
Suggestion by Progressive Lines			
Influence of Suggestion.....	“	“	“ 28
Accuracy in Reproducing Lengths.....	“	“	“ 19

Note.—Tests marked with an asterisk* are new.

VII. CORRELATION OF ABILITIES WITH AGE

The essential feature of all quantitative measuring scales of intelligence is that they relate the total score of the individual to his age. The measure of the individual's intelligence is the relation between his performance and the average performance of other children of his own age. The various Binet scales and revisions compute the total score of the individual in terms of his "mental age." The difference between the "mental age" and the chronological age is used as a quantitative measure of intelligence. A variant of this measure is that obtained by dividing the "mental age" by the chronological age, the resulting "mental quotient" or "intelligence quotient" serving as a quantitative measure—a method advocated by Stern (62) and used very largely by Terman (65). Still another quantitative measure of intelligence is that used by Yerkes and his co-workers (82), in which the total score of the individual in a group of tests is referred to the averages of the scores of groups of similar individuals of different ages. The score of the individual compared to that of other similar individuals gives the "mental age" which compared with the chronological age will give the "mental status" (age difference) or the "coefficient of intellectual ability" (mental quotient) as quantitative measures.

If intelligence is measured in terms of age, the correlation of the tests with age should throw light on the correlation of the tests with intelligence. When we say that tests are diagnostic of intelligence, do we mean that they are diagnostic of age? Are those tests that show the most rapid growth with age those that have the highest value in differentiating groups of different intelligence? From the results of this investigation and that of Chotzen it is seen that different tests have different diagnostic values. The comparison of these results with results showing the growth of the abilities with age should throw light on the problems mentioned.

Results showing the correlation of the test abilities with age

are influenced by the error due to incomplete data. As far as the writer knows, Yerkes' results are the only ones that are free from this influence, for under the conditions of the application of the point scale, all of the tests must be given to all of the subjects. The present writer has therefore incorporated an analysis of Yerkes' data bearing on the problem of the relation between the individual tests and age.

In tables 30 and 32, (pages 123 and 125) Yerkes gives the results of 468 English speaking boys and girls from 4 to 15. The run of the 449 subjects from 5 to 14 is as follows:

Chronological ages	5	6	7	8	9	10	11	12	13	14
No. of subjects	28	55	48	47	43	53	55	40	43	37

The present writer has combined these data into five groups, 5 and 6, 7 and 8, 9 and 10, 11 and 12, and 13 and 14. In table 17 the results of each group are given on each test, the results being expressed in the form of the per cent. that the average number of points scored by each group is of the total number of points possible.

TABLE 17.
Growth of Abilities with Ages.
Per Cent. of Points Scored by Children of Various Ages.

Test.	5&6	7&8	9&10	11&12	13&14
1. Repeating sentences	64	65	67	74	71
2. Describing pictures	53	63	72	79	84
3. Repeating digits	55	65	73	80	84
4. Comparing lines and weights.....	51	67	86	97	100
5. Copying diamond and square.....	28	51	76	82	93
6. Defining concrete terms.....	36	47	65	74	78
7. Aesthetic comparison	68	87	99	100	100
8. Indicating omissions in pictures.....	57	75	88	97	98
9. Naming words	14	26	52	72	83
10. Comparing remembered objects.....	19	48	80	92	95
11. Counting backwards	14	59	97	98	99
12. Comprehending difficult questions....	11	24	43	60	80
13. Using three words in a sentence.....	1	8	57	77	84
14. Arranging five weights.....	21	53	79	92	89
15. Detecting absurdities	1	12	33	50	69
16. Line suggestion	31	46	64	76	83
16a. Length of letters.....	0	9	30	49	64
17. Defining abstract terms.....	0	0	21	43	67
18. Analogies	2	9	25	38	52
19. Drawing designs from memory.....	5	13	32	54	63
20. Reconstructing dissected sentences...	0	1	29	45	68

An examination of table 17 shows that the tests differ greatly in the rate of growth with age. The growth of the ability to repeat sentences, for instance, is represented by 64%, 65%, 67%, 74% and 71%, while at the other extreme are tests such as the test of counting backwards in which the growth is 14%, 59%, 97%, 98% and 99%. Further study of table 17 shows a high degree of similarity in the results of some tests. Take for instance tests 10 and 14

10. Comparing remembered objects	19%	48%	80%	92%	95%
14. Arranging five weights.....	21%	53%	79%	92%	89%

or tests 6 and 16

6. Defining concrete terms.....	36%	47%	65%	74%	78%
16. Resisting suggestion	31%	46%	64%	76%	83%

or tests 1, 2 and 3.

The similarities in the growth of the various tests are more easily shown graphically. In Fig 5 the writer has drawn the graphs of the various tests, having classified them roughly according to the similarities shown. All of the percentages were taken from table 32 (page 125), except that for test 2 age 5 which is obviously a misprint (38.7 instead of 48.7). The wide variation in the growth of abilities with age is clearly shown in Fig. 5. Tests 10, 11 and 14 show a very rapid growth, which is in marked contrast with that shown by tests 1, 2 and 3. Tests 9 and 13 show very nearly as rapid a growth as tests 10, 11 and 14, but the growth occurs for the most part between 8 and 11 or somewhat later than the abilities in the first group which are almost completely developed at 9. Test 5 shows a slope somewhat more gradual than tests 9 and 13, but slightly sharper than tests 4, 7 and 8 which are extremely easy for younger subjects. Tests 12, 15, 16a, 17, 18, 19 and 20 show considerable similarity. Test 12 is the easiest of the group and test 18 the hardest. Test 16a shows considerable variation, and as this was an extra test, it is possible that it was not given all the possible number of times. The slope of the curves of these seven tests is similar to that of tests 4, 7 and 8, but the abilities develop very much later, being hardly better than 25% at 8, having the fastest growth between 11 and 12 and scarcely reaching 75% in any test. The slope

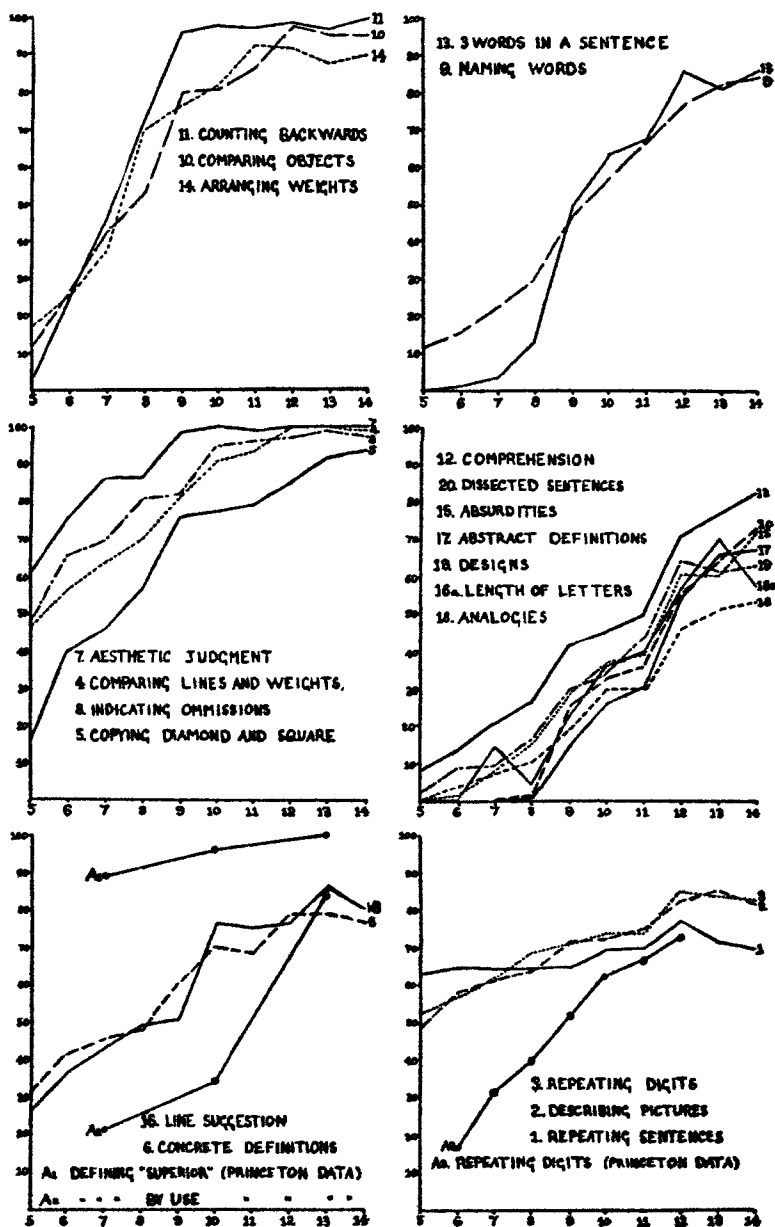


FIG. 5. Growth of Test Abilities with Age. Results of Yerkes from the Point Scale.

of the curve for test 18 is more gradual than that of the other six tests, and probably coincides more closely with that of tests 6 and 16 which in turn have a more rapid growth than tests 1, 2 and 3.

Test 1 consisted of three sentences, one of 5 words (part a), one of 10 words (part b), and one of 18 words (part c), two points being awarded for the correct repetition of each. The explanation of the fact that test 1 shows practically no growth is given by Yerkes. "According to the results of our analysis of data, this test is eminently unsatisfactory, because parts (a) and (b) are so easy that even the four- or five-year-old child has little difficulty with them, whereas part (c) is so very difficult that only a few of the children among the 750 examined obtained credit for it. Such being the case, it is obvious that the score for this test cannot increase either markedly or regularly with increasing age" (page 128). In revising the scale, Yerkes changed the third sentence to 20 words, and added another of 15 words, giving credit of 1 point for repeating 5 or 10 words, and 2 points for repeating 15 or 20 words. With this system the growth of the test with age should be more rapid.

Yerkes finds nothing wrong with test 3. "Test 3 (Memory span for digits) has proved eminently satisfactory, and we see no reason for making other change than in position" (page 129). Comparison of the curves for tests 1 and 3 in Fig. 5 shows that test 3 is hardly more satisfactory than test 1. Test 3 consists of five parts, repeating 3, 4, 5, 6 and 7 digits, one point being allowed for the successful repetition of each. The lack of growth is evidently due to the same cause as test 1. The smaller numbers of digits are too easy, and there is little opportunity to differentiate superior ability. In curve A₃ the writer has inserted a graph of the average of the percentages given for non-selected Princeton boys and girls shown in table 13 of the first study under the discussion of sex differences (page 73). This test was weighted on a scale of six points, three points being awarded for repeating 7 digits, two points for 6 digits and one point for 5 digits. The advantage of this method of weighting tests according to their relative difficulty is obvious from comparing

curves A₃ and 3 in plate IX. One method gives a differential measure of growth, the other does not.

It is fair to reason from a demonstrable error in tests 1 and 3 to a similar error in test 2. The three Binet pictures are shown in test 2, 1 point being allowed for enumeration, 2 points for description and 3 points for interpretation for each picture. The procedure in giving the test ("Look at this picture and tell me about it") was designed to avoid the response by enumeration suggested by the word "what," so that the diagnostic value of the test should in all probability be higher than that in the Trenton investigation. The lack of growth in the test is probably due to the fact that the gradations of response are not weighted according to their relative difficulty. Any child able to talk is able to enumerate, and this response hardly deserves credit. The present writer weighted the digits tests 1, 2 and 3 points because Goddard placed them in years VIII, X and XII. On the same line of reasoning, the three responses to the picture test would have the relative weight of 1, 3 and 7 rather than 1, 2 and 3, for they appear in III, VII and XV. This method is however entirely arbitrary, and it is a very simple matter to weight tests empirically according to their difficulty as shown in the discussion of the commissions test in the preceding chapter. If the three parts were weighted on a scale of 9 according to their relative difficulty as found in the Trenton investigation, the value for enumeration would be 0, for description 2 and for interpretation 7. With Yerkes' procedure in which enumeration is not suggested and description therefore easier, the relative weight of description and interpretation would be nearer 1 to 4 than 2 to 3.

The definitions test was given to 66 non-selected boys and girls age 6 to 7, and 50 non-selected boys and girls age 9 and 10 in Princeton, and to the normal and retarded 12, 13 and 14 year subjects in Trenton, the procedure and method of scoring being the same. The test of defining in terms of use was passed by 89% of the 6 and 7 year subjects, 96% of the 9 and 10 year subjects, and by all of the Trenton retarded and normal subjects. The test of defining in terms superior to use was

passed by 21% of the first group, 34% of the second group, and by 84% of the Trenton normal group, the performance of the retarded group (33%) being about the same as that of non-selected 9 and 10 year subjects (34%). Curves drawn for these three points are shown under A1 and A2 in Fig. 5. The test of defining in terms of use is probably as easy as test 7, while the test of defining in terms superior to use would seem to approach test 12 in difficulty. Curve 6 might be the resultant of two tendencies such as illustrated in curves A1 and A2. Yerkes gives 1 point for definitions by use and 2 points for definitions superior to use, 4 words being used. If the tests were weighed according to their difficulty, the proportion would be nearer 1 to 8 than 1 to 2. The effect of improper weighing is to obscure a real correlation with age.

The relations indicated by the growth of the other abilities are fairly definite. The test which shows the most rapid growth (test 11, counting backwards) was shown in the first section to depend upon school training. The curve of this test is not very different from the curve of tests 11 and 14 (comparing remembered objects and arranging five weights). It would seem that these tests were very valuable for indicating growth from 5 to 9, but as the abilities are practically developed at 9, they are useless as differential measures above this age. In the present investigation it was found that the five weight test was worthless for differentiating normal from retarded subjects, but there was reason to suppose that the test given to younger subjects would be diagnostic.

The test of constructing a sentence from three given words (test 13) shows a very rapid rise between 7 and 10 (60%) and a very much slower growth from 10 to 14 (23%). Yerkes gives credit of 2 points for a compound sentence and 4 points for a simple or complex sentence. His results indicate that very few children below 9 can construct a sentence at all. The part of the test that reveals the real growth is the mere construction of any sort of a sentence, the relative merits of the sentences constructed showing much less differentiation. This test needs checking up, for in the Princeton study there was

reason to suspect that the ability to construct any sort of a sentence possibly depended on the training of the third grade. The dependence of a test on school training may possibly be indicated by the rapidity of the growth of a test with age. Test 11 certainly shows a very rapid growth. Test 20 (dissected sentences) which depends partly on school training shows no growth till 9, a sharp rise at 9, but above this age agrees closely with other tests. This criterion of suddenness of growth is however very uncertain, as many other tests that have been shown to be independent of school training also show a very rapid growth. The proof of the presence of school training can only be shown by an analysis similar to that made in the first study.

It is probably not possible to reason from the rapidity of growth to the relative diagnostic value of the tests in a definite way. The curve of test 16, (the line suggestion test) which was shown to have a low diagnostic value is somewhat flatter than the curves of tests 12, 15, 20 etc., but this is an indefinite criterion. As a matter of fact the curve for test 16 shows two phases. There is a growth of 24% from 5 to 9, a growth of 26% from 9 to 10, and a maximum growth of 10% above 10. Yerkes gives credit of one point for each resistance, defining a "resistance" as saying "the same", or "equal", or for pointing to the left instead of to the right in the case of each of the last three pairs. In the present investigation this latter type of response was taken to indicate the influence of suggestion, a difference of procedure which probably accounts for the fact that the Trenton normal group resisted but 33% of the possible suggestions, while Yerkes' 12, 13 and 14 year subjects resisted 81%. Two characteristic responses to this test were pointed out (see page 56), the suggestion error and the discrimination error. It may be possible that subjects below 10 fall into the suggestion error, while those above 10 fall into the discrimination error. The character of the curve would indicate a change in the character of the response, or a change in the procedure in giving the test.

The comparison of the curve for test 18 (analogies) with that of test 12 (Comprehension) would show that the former was less useful than the latter in differentiating the intellectual growth from 5 to 14. Inferences from the slope of the curve to the diagnostic value are however uncertain. The design test was found to have a lower diagnostic value (36%) than the comprehension test (71%), dissected sentence test (71%), or the absurdities test (51%), yet the character of the curves for tests 12, 15, 19 and 20 are very much the same. The diagnostic value of the tests necessitates other evidence than that from the growth of the abilities, in the same manner as the demonstration of the presence of school training needs other evidence. However, it is perhaps possible to obtain corroboration of the diagnostic values found by comparing the performances of children of different ages.

It was noticed that the performance of the Trenton sub-normal group on the test of defining in terms superior to use was comparable to the performance of the 9 and 10 year non-selected Princeton subjects. Inasmuch as intellectual defect is usually regarded as a slowing up of mental development, the comparison of non-selected subjects age 9 and 10 with non-selected subjects age 13 and 14 might possibly throw some light on the relations found by comparing 12, 13 and 14 normal and retarded subjects. In the latter case the question was—what tests differentiate normal subjects of 12, 13 and 14 from retarded subjects of the same age? In this case the question is—what tests best differentiate children of 13 and 14 from children of 9 and 10? The comparison may readily be made by subtracting the third from the fifth column of table 17. The list of tests in the order of their value in differentiating 13 and 14 year subjects from 9 and 10 year subjects is as follows:

- 46% No. 17. Defining abstract terms.
- 39% No. 20. Reconstructing dissected sentences.
- 37% No. 12. Comprehending difficult questions.
- 36% No. 15. Detecting absurdities.
- 34% No. 16a. Length of letters.

- 31% No. 9. Naming words.
- 31% No. 19. Drawing designs from memory.
- 27% No. 13. Using three words in a sentence.
- 27% No. 18. Analogies.
- 19% No. 16. Line suggestion.
- 17% No. 5. Copying diamond and square.
- 15% No. 10. Comparing remembered objects.
- 14% No. 4. Comparing lines and weights.
- 13% No. 6. Defining concrete terms.
- 12% No. 2. Describing pictures.
- 11% No. 3. Repeating digits.
- 10% No. 14. Arranging five weights.
- 10% No. 8. Indicating omissions in pictures.
- 4% No. 1. Repeating sentences.
- 2% No. 11. Counting backwards.
- 1% No. 7. Aesthetic comparison.

For 13 of these tests (nos. 1, 2, 3, 6, 9, 12, 13, 14, 15, 16, 17, 19 and 20) values were obtained indicating their relative merit in diagnosing differences between normal and retarded subjects (shown in table 10.). The correlation (Pearson products-moments method) between the diagnostic value of these 13 tests found in this investigation and the value of these tests in differentiating 9 and 10 year subjects from 13 and 14 year subjects shown by Yerkes' results is 0.71 ($p = 0.09$). Four of the tests (nos. 1, 2, 3 and 6) show an error in the method of scoring. The correlation between Yerkes' results and the Trenton results for the other 9 tests is 0.81 ($p = 0.04$). The correlations are probably high enough to indicate that as a general rule the tests that most successfully differentiate normal subjects of 12, 13 and 14 from retarded subjects of the same age, also show the largest differences between the performances of 9 and 10 year subjects and 13 and 14 year subjects.

Concerning the results of the younger subjects, a similar question may be asked,—what tests most effectively differentiate 5 and 6 year subjects from 8 and 9 year subjects? The list of tests in the order of their value in making this differentiation is as follows:

- 83% No. 11. Counting backwards.
- 61% No. 10. Comparing remembered objects.

58%	No. 14.	Arranging five weights.
56%	No. 13.	Using three words in a sentence.
48%	No. 5.	Copying diamond and square.
38%	No. 9.	Naming words.
35%	No. 4.	Comparing lines and weights.
33%	No. 16.	Line suggestion.
32%	No. 15.	Detecting absurdities.
32%	No. 12.	Comprehending difficult questions.
31%	No. 7.	Aesthetic comparison.
31%	No. 8.	Indicating omissions in pictures.
30%	No. 16a.	Length of letters.
29%	No. 6.	Defining concrete terms.
29%	No. 20.	Reconstructing dissected sentences.
27%	No. 19.	Drawing designs from memory.
23%	No. 18.	Analogies.
21%	No. 17.	Defining abstract terms.
19%	No. 2.	Describing pictures.
18%	No. 3.	Repeating digits.
3%	No. 1.	Repeating sentences.

It is possible to compare these results on the easier tests with those of Chotzen. Chotzen (18) found that the backwardness of his feeble-minded children was most marked on nine tests, making change, naming months, recalling a story read, repeating sentences, repeating digits, defining in terms superior to use, counting backwards, comparing remembered objects and arranging five weights. The first three of these are not included in Yerkes' scale. There is an error in the scoring of the second three so that their value as a measure of growth is obscured. The last three show the highest value of any of Yerkes' 21 tests in differentiating 5 and 6 year subjects from 9 and 10 year subjects.

Chotzen also named nine tests as showing the greatest value in differentiating the groups of feeble-mindedness. These tests were naming coins, recalling a story read, making change, comprehending easy problem questions, repeating digits, defining in terms superior to use, copying a diamond, comparing remembered objects and arranging five weights. The first four of these tests are not in the point scale, and the value of the next two is obscured by the scoring system. The last three are included

in the five tests in Yerkes' scale that show the highest differential value for young subjects. The only test in the point scale showing a differential value for young children higher than 40% that is not included in Chotzen's list is that of constructing a sentence from three given words, and this test was not given a sufficiently large number of times by Chotzen. It is possible to conclude then that as a general rule the tests that most successfully diagnose mental defect in younger subjects, or most effectively differentiate the lower grades of feeble-mindedness, also show the greatest differentiation between the performances of 5 and 6 year subjects and 9 and 10 year subjects. The correspondence found between the results of the two studies of the diagnostic value of the tests and the results of comparing older and younger subjects are in agreement with the view that feeble-mindedness is a general slowing up of mental growth.

Again it is found that many of the tests that show the highest value in differentiating the higher grades of intelligence are tests that involve the use of language to a considerable extent. Yerkes' results afford the opportunity of studying the influence of language training on the tests. In table 31 (page 124) Yerkes gives the results of 196 children from 5 to 14 of non-English speaking parents. The present writer has computed the per cent. that the average number of points scored by the age groups 5 and 6, 7 and 8, 9 and 10, 11 and 12, and 13 and 14 was of the number of points possible to be scored. Subtracting these values from those given for similar groups of children of English speaking parents (shown in table 17) gives the influence of language training on each test in each group. These values are shown in table 18 in which the tests are arranged approximately in the order of the magnitude of the differences which they show between groups of different language training. A plus value indicates that the children of English speaking parents are ahead, a minus value that the children of non-English speaking parents are ahead.

The differences shown in table 18 vary from + 29% to - 9%, the median being + 4% ($Q=4.5\%$). There is then a general superiority of the results of children of English speaking parents.

TABLE 18.
Influence of Language Training.
Percentage Differences in Performance of Children of English and
Non-English Speaking Parents.

Test.	5&6	7&8	9&10	11&12	13&14
17. Defining abstract terms.....	0	0	+12	+18	+22
13. Using three words in a sentence.....	+1	+8	+22	+12	+8
14. Arranging five weights.....	-2	+29	+15	+4	+5
15. Detecting absurdities	0	+10	+14	+13	+13
12. Comprehending difficult questions.....	+6	+8	+7	+10	+19
20. Reconstructing dissected sentences.....	0	+1	+16	+8	+21
10. Comparing remembered objects.....	+4	+14	+11	+5	+2
18. Analogies	+1	+7	+10	+7	+10
9. Naming words	-1	+2	+5	+15	+13
6. Defining concrete terms.....	+3	+8	+9	+8	+5
16a. Length of letters.....	0	+9	+8	0	+15
1. Repeating sentences	+9	+8	+3	+6	-1
11. Counting backwards	-2	+10	+6	+2	+2
8. Indicating omissions in pictures.....	+11	+6	-1	+1	0
2. Describing pictures	+5	+1	+4	+1	+6
4. Comparing lines and weights.....	-5	+5	+10	+4	+2
3. Repeating digits	-5	-1	+4	0	+5
7. Aesthetic comparison	0	-3	+2	0	+2
5. Copying diamond and square.....	-6	-5	+3	-3	+2
16. Line suggestion	-4	-6	-9	+5	+2
19. Drawing designs from memory.....	-1	-1	-9	-3	+1

The correspondences shown in table 18 are of course more remarkable than the divergencies, but in the light of the high degree of correspondence, the fact of wide divergence would seem to indicate training. There are 19 differences higher than +10%, these differences being confined to but 10 tests. In three tests (nos. 8, 12 and 16a) differences higher than +10% occur in but one age group. In five of the tests (nos. 9, 10, 13, 14 and 20) these differences appear in two age groups. In two of the tests (nos. 15 and 17) these differences appear in three age groups. In five tests (nos. 12, 13, 14, 15 and 17) the average difference is +10% or higher. Although the evidence is thus directed against a few tests, it is probably impossible to say definitely on what tests the influence of language training is not present or on what tests the results are due to chance. The tests are arranged in table 18 in the approximate order of the magnitude of the differences found.

Two interpretations of the results are possible, the first, that the children of non-English speaking parents are under a serious handicap in some of the tests owing to deficient language training, the second, that these children have an inferior hereditary endowment, that they are less intelligent. The validity of the second interpretation may be examined by comparing the magnitude of the differences found between groups of younger and older children of English and non-English speaking parents with the magnitude of the values found for the tests in differentiating 13 and 14 year children from 9 and 10 year children, and for differentiating the latter group from 5 and 6 year children, or in other words by comparing the value of the tests as differential measures of growth with the supposed influence of language training.

As a measure of the amount of the influence of language training on younger subjects the differences found between the English and non-English speaking groups of 5 and 6 year children and 7 and 8 year children on each test were combined. The correlation (Spearman foot-rule) between these values and the magnitude of the differences between 5 and 6 year subjects and 9 and 10 year subjects is 0.11 ($p = 0.09$). As a measure of the amount of influence of language training on older subjects the differences found for the 11 and 12 year subjects and 13 and 14 year subjects were combined. The correlation between these values and the magnitude of the differences between 9 and 10 year subjects and 13 and 14 year subjects is 0.56 ($p = 0.09$). These values represent the relation between the differential value of the tests and the magnitude of the language differences at the extremes. The correlation between the differences found between the English and non-English speaking subjects of 9 and 10 and the differences between English speaking subjects of 7 and 8, and 11 and 12 is 0.35 ($p = 0.09$). In drawing inferences from these correlations it should be remembered that most of the tests that are the best differential measures of older children are passed by so few of the younger children that there is little opportunity of differentiating these children, or little opportunity for a language difference to appear. On the other

hand, some of the tests that show no differentiation between the performance of 9 and 10 year subjects and 13 and 14 year subjects fail to show a difference because they are too easy, and for the same reason these tests would fail to differentiate the language groups. These factors would tend to obscure the correlation between intelligence differences and language differences in the younger years, and to magnify this correlation for older subjects.

It is not possible to demonstrate whether the differences are due to deficiency in language training or to deficiency in intelligence. The statement that the connection between language differences and intelligence differences becomes more intimate with increasing years, (based on the correlations 0.11, 0.35 and 0.56) is modified by the relations pointed out between these correlations and the difficulty of the tests, and after all this may only be another way of saying that the tests that most successfully differentiate the higher grades of mental defect or most successfully differentiate the growth of older subjects involve language training. The position that the differences are due to language training is therefore favored. If the results justify the position that children of non-English speaking parents have an inferior hereditary endowment, it would appear that this inferiority becomes more marked with increasing age. However it must be remembered that this inferiority is only being measured by comparatively few effective tests. The safest conclusion would probably admit the possibility of both factors.

The examination of the results of the five weight test should indicate what the nature of the conclusions concerning the individual tests should be. It is perhaps surprising to find the apparent influence of language training in the test of arranging five weights. This influence is most marked in the younger years in which there is the strongest reason to believe that the differences are due to language training rather than to intelligence. Binet considered this test important in differentiating morons from normals, and attached considerable importance to it because it presupposed no acquired knowledge and was absolutely independent of all instruction. The results of the present in-

vestigation show that it is useless in differentiating the higher grades of mental defect, but there is reason to suppose that the intellectual factors of comprehending a serial arrangement and making the logically necessary comparisons are correlated with intelligence. Chotzen's results shown the test valuable in differentiating the lower grades of mental defect, and Yerkes' results show a very rapid growth from 5 to 9. It is significant that the same range in which the test is most effective is the range in which the influence of language training appears. If this influence is actually language training, it would indicate that one of the most important factors in the test, and the one that probably gives it its strongest correlation with intelligence is simply that of understanding the instructions. The results of the two language groups would indicate that the non-English speaking subjects, even though they are able to make the sensory discrimination and to comprehend and execute a serial arrangement, fail the test because they do not understand the instructions.

If the conclusion that the test of arranging five weights is influenced by language training is not justified, then it follows that none of the tests are influenced by language training, for the five weight test shows as marked an influence of this factor as any other test. The fact that the apparent language differences on this test are not due to differences in intelligence is indicated by the low correlation between the language differences and growth differences in this region (0.11), and by the fact that two of the tests that are most effective in making a differentiation of growth in this region (nos. 5 and 11) show very slight language differences. It must be concluded then that the differences are either due to language training or to chance. If the differences on the five weight test are due to chance, all other differences are due to chance. The chance hypothesis would probably be overworked in accounting for the fact that 76% of the 105 differences found between English and non-English speaking subjects were less than 10%.

If the foregoing analysis of Yerkes' data is correct it follows that some of the tests are influenced to a considerable extent

by language training. In the solution of some tests the children of non-English speaking parents are under a serious handicap owing to deficient language training. In any event the factor can not be disregarded entirely, for if there is some truth in the hypothesis that the language differences are due to intelligence differences, and that the difference in intelligence would manifest itself in the long run, in an individual examination how is the experimenter to know whether the subject's failure is due to defective intelligence or defective training?

In the present investigation the writer by selecting groups of similar language training was able to keep the data bearing on the diagnostic value of the tests free from the influence of the language factor. The position that the Trenton data were free from this influence is strengthened by Yerkes' results, for it was found in the present investigation that if it were to be concluded that certain tests depended on language training, it would also be necessary to conclude that two of the tests (naming 60 words and using three words in a sentence) showed this influence in favor of the children of non-English speaking parents, or in other words that the training in two languages was a positive help in these two tests. The fact that Yerkes' results show the English speaking children ahead in these tests would indicate that the language differences found in the Trenton study were due to chance rather than to the positive influence of the language factor.

VIII. RESULTS OF OTHER INVESTIGATORS

It is beyond the scope of the present investigation to summarize all of the literature bearing on the correlation of various mental tests with intelligence. The results of other investigators bearing directly on the individual tests used have been mentioned in the detailed discussion of the tests. Simpson's investigation is reported here because the problem has many points in common with that of this investigation and the conclusions have a similar trend. The other investigations of Norsworthy, Terman (earlier study), Wallin, and Pyle are mentioned only on account of the correspondence in the method of group differentiation. Workers in the field will find a summary of the work bearing on the correlation between many of the standard tests and intelligence given under the heading "Dependence on intelligence" under each of the tests in Whipple's Manual (76). An analysis of the factors involved in many of the better known tests and the relation of these factors to intelligence is given in the second volume of Meumann's "Vorlesungen" (45), along with an evaluation of the results of various investigators who have applied the tests in this particular field. Lastly, for the real masterpieces in the creative portion of the field, the reader is referred to Binet's original articles which appeared from time to time in *L'Année Psychologique*, and which have recently been translated into English by Kite (39 and 40).

Simpson (58) gave 15 tests to two groups of adults who were taken to represent "the two extremes of 'general intelligence' as judged by the world", one group being composed of 17 professors and advanced students, and the other of 20 men who had never held any position demanding a high grade of intelligence. The tests used included two tests of perception (marking A's and marking geometrical forms), three tests of memory (memory of unrelated words, of passages, and recognition of forms previously seen), four of association (addition, easy opposites, learning pairs of words and forms, and com-

pleting words), three of selective thinking (hard opposites, mutilated prose, and absurdities), two of sensory discrimination (reproducing lengths and discriminating lengths), and one of motor control (scroll test).

Simpson found that "the tests reveal very marked differences in the two groups in language tests demanding selective thinking; marked but less difference in certain tests of memory; very decided differences in language tests demanding speed and accuracy in easy association; less difference in the more directly practiced and mechanical associations demanded in adding; in perception tests and in motor control the differences are somewhat less still; and in discrimination of lengths they are least of all." (page 55).

The method on which Simpson based his conclusions was that of comparing the per cent. of the poor group that surpassed the median of the good group, and the lowest four, two and one of the good group. From the results given in table II, (pages 30 and 33 of Simpson's monograph) the present writer has calculated the Maximum Diagnostic Value of each test. The list of tests in the order of the Maximum Diagnostic Value is as follows:

- 100% Test IV. Easy Opposites.
- 100% Test XII. Hard opposites.
- 94% Test XIV. Ebbinghaus' mutilated text.
- 94% Test V. Recognizing forms.
- 88% Test VII. Learning pairs.
- 85% Test VI. Memory for words.
- 78% Test VIII. Memory for passages.
- 75% Test XIII. Completing words.
- 73% Test III. Scroll test.
- 70% Test XI. Adding.
- 70% Test I. Marking A's.
- 69% Test II. Marking geometrical forms.
- 42% Test X. Estimating lengths.
- 21% Test IX. Reproducing lengths.

The results of the absurdities tests were not used because they were not reliable.

In drawing conclusions concerning what sort of abilities were

connected with mental ability, Simpson used the intercorrelations of the tests as a basis of classifying them. The Ebbinghaus and hard opposites tests, for example, were classified under "selective thinking", and the easy opposites test under "association". The correlation between the first two tests was 85, so that they may be correctly classified under one heading. The easy opposites test correlated higher than these two tests of "selective thinking" (72 and 83) than with any of the other tests, so that it should properly be placed under this heading rather than under "association".

Grouping the tests together on a basis of their intercorrelations, Simpson figured the average correlation of each ability with all of the other tests. On the basis of the magnitude of these average correlations, Simpson concluded that "power of selective thinking is more intimately connected with, and more characteristic of, general mental ability than is any of the other abilities tested; that memory is next most highly correlated with general ability; the simpler forms of association next; perception next; motor control considerable less; and discrimination of lengths least of all." (page 67).

Simpson held that the tests were measures of mental capacity rather than measures of amount of training and education, because the correlation between the number of years of schooling and the rank in the eight tests that correlated most highly with the other tests was low (38). On the evidence of studies of retardation he held that "a small number of years schooling means inability to learn advanced and difficult language work, rather than lack of opportunity to learn it." (page 70). Further evidence that the tests of selective thinking were not measures of school training was derived from the fact that the subjects who were considered decidedly dull or stupid by their fellows did poorest in these tests. Simpson found further evidence in support of the proposition that language tests are fair tests of ability from the fact that the intelligence of primitive peoples may be measured by their language, and that feeble-minded children are deficient in acquiring higher forms of language.

A combined measure of general intelligence was taken for

each individual by adding his scores on the Ebbinghaus test, hard opposites, easy opposites, learning pairs and recognizing forms, the scores being compared in terms of the deviation from the median. The deviations of the good group varied from + 95 to + 21, while those of the poor group varied from 0 to - 127. There was a difference of 21 between the lowest member of the good group and the highest member of the poor group, this difference being 46.5% of the average deviation from the median (45.14). The combined score therefore differentiated the groups more completely than the score on any of the individual tests. Further evidence that the combined score on these five tests was a measure of "general intelligence" was obtained from the fact that the correlation between the ranking of the subjects of the good group according to these tests and their ranking according to the independent estimates of their intelligence by ten or more persons was .92. The correlation between the various tests and estimated intelligence varied from .96 (hard opposites) to -.20 (drawing lengths).

Norsworthy (48) gave twelve mental tests and four physical tests to 150 defective children and to large numbers of normal children in order to determine whether the mental defects of idiots were equalled by the bodily defects, whether idiots formed a separate species or not, and whether idiots showed a lack of mental capacity all around. The physical tests showed very slight differences between the groups, 26% of the idiots being above the median of the normals in the measurements of temperature, 44% in weight, 45% in height and 49% in the measurements of pulse. The median for the idiots in "intelligence tests" (part-whole, genus-species and opposites tests) was below the median of the normals 7 times the probable error. The median for idiots in memory tests (memory for dictated passages and related words) was below the median of normals 3.5 times the probable error. The median of idiots in "maturity tests" (reproducing a weight to a standard, memory for unrelated words, cancelling A's and cancelling a's and t's) was below the median of normals 2.7 times the probable error.

Norsworthy finds no evidence for the theory that idiots con-

stitute a separate species of individuals. The results show however that the differences that are found between idiots and normal children vary with the measurements used, which is of course another way of saying that the tests vary in their efficiency in diagnosing feeble-mindedness. Physical tests have little efficiency, and the controlled association tests (combined score) had twice the efficiency of the two memory tests.

Terman (64) examined seven of the brightest and seven of the dullest pupils in a school system made up of about 500 children, finding the bright boys superior in all the mental tests given, but below in the motor tests. No suggestions can be obtained concerning the relative efficiency of the tests in differentiating the groups on account of the small number of subjects.

Much information concerning the problem of what tests are diagnostic of intelligence may be gained from the correlation methods in which the standing of a group of subjects in a series of tests is compared with their rank order in intelligence as estimated by the school-masters, school-fellows, or other persons supposedly competent to diagnose mentality. The results of this method are of course no more accurate than the original independent rating of intelligence, and this rating is not absolutely reliable for the correlations between the ratings of one group by different observers are frequently low. Furthermore, there is danger that the individuals who make the ratings will stress some ability such as memory so that some tests will show a correlation with estimated intelligence somewhat higher than their probable true correlation with intelligence.¹

On the whole, these correlation methods are chiefly serviceable in determining the relationships between the various test abilities. Eventually, if the same tests are given to different groups,

¹ This point is well illustrated by Abelson's (1) results. Abelson instructed the teachers to estimate the "practical intelligence" of the children by considering in forming their opinion which of the children they would soonest trust on an errand requiring the sharpest intellect. The fact that the test that showed the highest correlation with intelligence was a test of memory for commissions shows that the teachers considered the mere retention of the instructions more important than the intellectual factors involved in the execution of the errand.

the evidence from different investigations will be very valuable in showing what tests are most diagnostic of intelligence. At present there are not enough investigations available. The work to date however would certainly support the two propositions that different tests vary in their effectiveness in diagnosing intelligence, and that a combined score of several tests is more effective than any single test.

Wallin (73) gave the Binet 1908 scale to a large group of epileptics, and compared their results with those of other investigators on normal and feeble-minded individuals. Certain tests proved to be especially difficult for epileptics just as certain tests in this investigation proved to be especially difficult for the retarded group. The results of the two investigations cannot be compared however, for Wallin referred the discrepancies between the performance of normals and epileptics to "inherent abnormalities in the mentation of the epileptics", and indeed there is no reason for believing that the tests that proved especially difficult for epileptics should also be the tests that are most highly diagnostic of feeble-mindedness.

Pyle (53) gave a series of class-room tests to groups of pupils classified as bright and dull on the basis of their school marks. He found that the completion, word-building, logical memory and controlled association tests were most valuable for the purpose of ascertaining the mental differences between the groups. Ability to do the cancellation test in some cases showed an inverse relation to the other tests. Ability in the ink-blot test showed an inverse relation with age, (the younger children doing better), and showed a negative relation to the other tests.

Terman (65) has recently published a revision of the Binet scale, the selection of tests being based on an empirical verification of their validity. The method of demonstrating the validity of the individual tests was that of comparing each test with the scale as a whole. The subjects of each age were divided into three groups according to their "intelligence quotients" (IQ), and the tests that showed a higher per cent. passed in an inferior IQ group than in a superior IQ group were rejected. This method insures that each test is to some extent

coherent with the scale as a whole. The results of this method are best shown by the following quotation from Terman:

"When the tests were tried out in this way it was found that some of those which have been most criticized have in reality a high correlation with intelligence. Among those are naming the days of the week, giving the value of stamps, counting thirteen pennies, giving differences between president and king, finding rhymes, giving age, distinguishing right and left, and interpretation of pictures. Others having a high reliability are the vocabulary tests, arithmetical reasoning, giving differences, copying a diamond, giving date, repeating digits in reverse order, interpretation of fables, the dissected sentence test, naming sixty words, finding omissions in pictures, and recognizing absurdities." (Pages 76 and 77).

"Among the somewhat less satisfactory tests are the following: repeating digits (direct order), naming coins, distinguishing forenoon and afternoon, defining in terms of use, drawing designs from memory, and aesthetic comparison. Binet's "line suggestion" test correlated so little with intelligence that it had to be thrown out. The same was also true of two of the new tests which we had added to the series for try-outs." (Page 77).

"Tests showing a medium correlation with the scale as a whole include arranging weights, executing three commissions, naming colors, giving number of fingers, describing pictures, naming the months, making change, giving superior definitions, finding similarities, reading for memories, reversing hands of clock, defining abstract words, problems of fact, bow-knot, induction test, and comprehension questions." (Page 77).

From the standpoint of the desirability of comparing Terman's results on the individual tests with those of this investigation it is to be regretted that Terman's actual data are not yet available. However, he uses the agreement of each test with the scale as a whole as a criterion of the test's correlation with intelligence, and from his report it is possible to classify the tests into four grades of reliability, those showing a high correlation, a medium correlation, a less satisfactory correlation, and no correlation with intelligence, (the term "corrleation with

intelligence" being used interchangeably by Terman with "correlation with the scale as a whole").

It is not possible to compare all the results of this investigation with those of Terman on account of the differences of procedure. The agreement between Terman's procedure and that of the present investigation was very close on nine tests, and a direct comparison is possible. Terman found that the dissected sentence test showed a high reliability, and this test showed one of the highest diagnostic values found in the present investigation (71%). He also found that the tests of naming 60 words, giving rhymes with "day", "mill" and "spring", and naming the date showed a high correlation with intelligence, while the diagnostic values found in this investigation were not particularly high (40%, 32% and 29% respectively). The tests of naming the months and arranging five weights which Terman reported as showing a medium correlation with intelligence have diagnostic values of 43% and 18% respectively. Two tests that are classified as less satisfactory, the designs and 7 digits tests, show diagnostic values of 36% and 30% respectively. The line suggestion test which correlated so little with intelligence that it had to be thrown out showed one of the lowest diagnostic values found in this investigation (12%).

Five other tests used in this investigation may be compared with Terman's results although the procedure was somewhat different. Terman used five absurdity questions, three of them being the same as the present writer used. Terman reported this test as showing a high correlation with intelligence, and considered it "one of the most ingenious and serviceable tests in the scale", and "an invaluable test for the higher grades of mental deficiency", an opinion in keeping with the diagnostic value found in this study (53%). Terman used four grades of comprehension questions, and found that the test showed a medium correlation with intelligence. This test showed one of the highest diagnostic values in the present investigation (71%). The two studies used different words in the test of defining terms superior to use so that the results are but roughly com-

parable. Terman classified the test as having a medium correlation with intelligence, while the value found in this investigation was 51%. Two of the three words used in the abstract definitions test appear in the five used by Terman. The diagnostic value found was 51%, and Terman reported it as showing a medium correlation with intelligence. Terman used three problems from various facts, and found a medium correlation with intelligence. The two problems used in this study showed different diagnostic values (21% and 51%). It is not possible to compare Terman's results on the tests of describing and interpreting pictures with those of this investigation because the pictures used and the procedures in giving the tests were different.

The results of the two investigations, where comparison is possible, do not agree very closely. Tests that showed a high correlation with intelligence according to Terman showed diagnostic values of 71%, 53%, 40%, 32% and 29% in this study. The diagnostic values found for the tests classified as showing a medium correlation were 71%, 51%, 43% and 18%. The tests classified as less satisfactory showed values of 36% and 30%, while the test that was so unsatisfactory that it had to be eliminated showed a value of 12%.

The discrepancies between the results of the two investigations might possibly be explained by the difference of method. In this event the question arises as to which method is more reliable. A. S. Otis (49) points out that it is theoretically possible to have a coherent system of tests that are not tests of intelligence (tests of physical strength, for example), and that it is therefore necessary to have other criteria of the validity of the individual tests. This objection is of course largely theoretical, but it is possible that the tests in different portions of the scale vary in the degree in which they correlate with intelligence or depend on factors other than intelligence so that the criterion of coherency would not give results that were constant throughout the scale.

The 16 tests in years VIII and IX, for instance, include the tests of counting backwards from 20 to 0, naming coins, giving the date, making change, naming months and counting stamps

which were shown to depend on school training, the test of writing from dictation which Binet eliminated on account of school training, and the test of constructing a sentence from three given words, a test in which this factor was suspected. The large proportion of tests in this region (8 out of 16, or if alternatives are omitted 4 out of 12) that are dependent on training might account for Terman's finding that some of the tests "which have been most criticised have in reality a high correlation with intelligence." Logically the test of coherency would indicate dependence on training as much as on intelligence. In the light of the first study, the presumption would be that the test of coherency in the region of VIII and IX would show the tests of training to have an abnormally high validity. The training tests may be diagnostic of intelligence, but that is another matter. The objection against the test of coherency is that it fails to take account of variable factors.

In the absence of the actual data it is not possible to determine the cause of the discrepancies between Terman's results and those of this investigation. At least Terman may be considered as subscribing to the general thesis that the individual tests vary in the degree in which they correlate with intelligence or in their value in diagnosing intelligence.

IX. CONCLUSIONS AND SUGGESTIONS

In all twenty-three Binet tests were used in this investigation, it being possible to draw conclusions concerning nineteen of them. The diagnostic values of these nineteen tests and their various sub-tests are shown in table 10. Ten supplementary tests were used, it being possible to draw conclusions concerning nine of them. The diagnostic values of these nine tests and their various sub-divisions are shown in table 16. Table 19 shows the various Binet tests and the supplementary tests arranged in the order of their diagnostic value as shown by this investigation.

The extreme divergency of the tests is clearly shown in Table 19. Three tests show diagnostic values higher than 70%, four higher than 60%, eight higher than 50% and twelve higher than 40%. Twelve tests show diagnostic values lower than 30%, five less than 20% and two less than 10%. The writer does not insist that all the values given in table 19 are absolutely final and definite. Indeed the influence of the one variable factor in the results, the personal equation, is so subtle that it can hardly be avoided. The experiment has however been reported in detail so that it can be repeated.¹

The reader may draw his own conclusions concerning the nature of the tests that are most diagnostic of intelligence, or the nature of the mental processes most intimately connected with intelligence. Inferences from the nature of the tests to the nature of intelligence are of course uncertain, for we know very little about the mental processes involved in the tests. The mere fact that a psychologist classifies a test as involving a certain process does not prove that that process is involved. In a general way it is perhaps interesting to note that there is

¹ The writer will gladly communicate further details of procedure that have not been reported to anyone wishing to repeat the experiment. All the data from the Princeton and Trenton experiments are on file at the Princeton laboratory, and are available for anyone who wishes to check up the writer's computations or to make further calculations.

TABLE 19.

List of Tests in the order of their Diagnostic Value.

- 74 Subtraction tests.
- 71* Comprehending difficult questions.
- 71* Reconstructing dissected sentences.

- 64 Healy cross-line tests.

- 53* Detecting absurdities in statements.
- 51* Defining in terms superior to use.
- 51* Defining abstract terms.
- 51* Solving problems from various facts (Problem b).

- 49 Balance test.
- 49 (42) Distinguishing between terms.
- 43* Enumerating the months.
- 40* Naming 60 words in three minutes.

- 36* Copying designs from memory.
- 35 Estimating lengths.
- 35 (30) Memory for commissions.
- 34 Puzzle tests. (Pooled score.)
- 33* Giving rhymes with "defender."
- 30* Repeating 7 digits.

- 29* Using three words in a sentence (2 ideas).
- 29* Giving the day and date.
- 28 Lifting the table asymmetrically balanced.
- 28 Influence of suggestion. (Suggestion by progressive lines.)
- 23* Repeating a sentence of 18 syllables.
- 23* Using three words in a sentence (1 idea).
- 21* Solving problems from various facts (Problem a).

- 19 Reproducing lengths. (Suggestion by progressive lines.)
- 18* Arranging five weights.
- 12* Resisting suggestion.

- 6* Interpreting pictures.
- + 2* Describing pictures.

Note: Tests marked with an asterisk (*) are in the Binet series.

much in common between Stern's (62) definition of intelligence as "general mental adaptability to new problems and conditions of life", Witmer's (74) definition of intelligence as "the ability of the individual to solve what for him is a new problem", and Pillsbury's (52) definition of reasoning as "the

application of any knowledge in a new way". Intelligence tests would seem to involve a new problem and a solution of the problem, or in other words, reasoning.

One logical implication of this study should be pointed out. Throughout the study the emphasis has been placed on the diagnostic value of the tests, or their merit in differentiating feeble-minded individuals from normal individuals. If the conception of diagnostic tests is carried to the extreme, the belief that certain tests could be found that are absolutely diagnostic of feeble-mindedness would imply that feeble-minded individuals constituted a separate species or a group of individuals who were in some respects completely different from normal individuals. Of course it has never definitely been shown that feeble-minded individuals do not constitute a separate species in some respects. Norsworthy's (48) results negate this view to some extent but not conclusively, for her results show feeble-minded individuals to be more distinct in some respects than in others. Logically of course there are no degrees of being a species, but there are degrees of accuracy of definition by which a species is specified, and Norsworthy in concluding that feeble-minded individuals did not constitute a separate species drew these conclusions on the basis of the tests used.

The view that feeble-mindedness is a general slowing up of mental development gets its chief impetus from the convention started by Binet and followed by others of defining normal development in terms of age, and this view of the intelligence of the feeble-minded if true would completely disprove the view that they were a separate species, unless intelligence changes in character in the course of its development. Theories of the correlation of intelligence with age have all been based on cross-section studies of different individuals at different ages, and the true nature of the development of intelligence will probably not be known until longitudinal studies of the same individuals through a number of years have been made. The results of this investigation show that as a general rule the tests that were most effective in diagnosing known differences

of intelligence also showed the greatest differentiation between subjects of different ages. The correlation found was not absolutely valid but in general supported the view that feeble-mindedness is a slowing up of mental development. These findings do not absolutely support this view for there may be changes in the character of intelligence with increasing age. The results of this investigation neither establish nor controvert the view that feeble-minded individuals in some respects constitute a separate species, a view the validity of which the belief in the possibility of discovering tests that are absolutely diagnostic of feeble-mindedness would necessarily imply. In the absence of definite experimental results the discussion of this point is largely speculative.

As a practical matter it is significant that the results of this investigation show that many of the tests that are diagnostic of the higher grades of mental defect involve the use of language. Six of the tests that show diagnostic values higher than 40%, abstract definitions, absurdities, comprehension, dissected sentences, 60 words and concrete definitions, stand first, fourth, fifth, sixth, ninth and tenth respectively in the list of tests in Yerkes' point scale arranged in the order of the magnitude of the differences found between English and non-English speaking children. The other six tests showing diagnostic values higher than 40% are not in Yerkes' scale. Two of the remaining four tests in the first ten of Yerkes' list (five weights and sentence test) show no diagnostic value in this investigation, and the other two (comparison and analogies) were not used. The character of two of the other tests showing diagnostic values higher than 40% (solving problems and distinguishing between terms) would indicate that they might be influenced by language training. Many of Simpson's (58) tests that most effectively differentiated his groups would also seem to involve language training.

Many of the tests that most effectively differentiate the higher grades of mental defect involve the use of language. In view of the probable close connection between intelligence and reason-

ing ability, it is not surprising to find a close connection between reasoning and the common vehicle of its expression, language. Of course it is possible to have reasoning in action as well as in thought but the fact remains that we have very few tests of reasoning in action.

The dearth of intelligence tests that are independent of the language factor indicates the magnitude of the problem that faces American investigators who wish to test the intelligence of individuals who have not had adequate training in English, immigrants, children of non-English speaking parents etc. It is possible to find many tests for older persons that do not involve language but a great many of these tests do not involve intelligence. With younger children and lower grade cases, the influence of language may appear in the instructions rather than in the test itself, as in the case of the five weight test. The problem of testing individuals without adequate training in English has considerable practical importance in this country with its cosmopolitan population. The solution of the problem will most certainly involve many careful and skilful researches.

To one who has followed the analysis of the various intelligence tests through these pages, there is apparently a hopeless confusion in the field. Some tests involve the influence of the personal equation to a marked degree, others depend on school training, some depend on linguistic training, and far too many depend too little on intelligence. The only hopeful aspect of the situation is that it is possible to place the whole field of mental tests on an absolutely empirical basis.

The personal equation is a difficulty but not an insurmountable difficulty and the presence or absence of this factor may be empirically demonstrated. The influence of scholastic training may be determined by comparing groups of similar ages but different training. The influence of sex differences may be determined by comparing the results of boys and girls of the same age and with the same training. The dependence of the tests on language may be determined by comparing the results of groups of English and non-English speaking children with

the same environmental opportunity. And lastly, the reliability of a test as a measure of intelligence may be determined by giving it to groups of the same age and opportunity, but of known differences of intelligence. There is no room for a priori objections or inferences. Every factor may be empirically demonstrated. It is possible to construct scales for measuring intelligence without knowing exactly what intelligence is. A person may determine the presence or effectiveness of intelligence without knowing its nature.

In view of the fact that the merit of tests and systems of tests may be empirically demonstrated, it is legitimate to demand that the investigator who proposes a new scale or another revision of existing scales should offer a demonstration of the reliability of his method. If he has a scale for measuring the higher grades of defect, let him show that it will actually diagnose these conditions. If he has a scale for differentiating adolescents from adults, let him demonstrate that the scale will actually make this differentiation.

It is surprising that up to this time very few complete demonstrations have been made of the reliability of measuring scales of intelligence. Binet's only experimental verification of his scale consisted in showing that the distribution of the children testing "at age" and below and above age was normal.² As a matter of fact there was practically no experimental verification of the scale, and its validity rests on Binet's merit as a psychological observer. The fact that some of the tests are worthless merely proves that Binet was occasionally mistaken in his opinion, and the fact that so many of the tests are extremely valuable is a lasting tribute to his experimental genius. Goddard's statistical verification of the Binet scale has been shown to be faulty by Ayres (2), Schmitt (56), Thorndike (69), Yerkes (82), and others.

Yerkes evidently revised the Binet scale on a point basis without experimental verification of his opinions. He threw out some tests because he thought they depended on school train-

² The totals of the table demonstrating this fact do not add up correctly. *Année Psychologique*, 1908, Vol. 14, page 73.

ing, and weighted the tests and their various parts on an entirely arbitrary basis. Of course Yerkes guessed very shrewdly, for only two of the tests in the scale involve school training, as a general rule the most valuable tests receive the most weight, and only three or four of the tests are worthless on account of faulty weighting. So far as the writer knows, Terman (65) is the only investigator who has offered an empirical verification of the individual tests along with the publication of the system of tests. Terman's method of demonstrating the validity of the individual tests has been discussed in the previous chapter.

In affording an experimental verification of his scale the experimenter should in every case make his demonstration from the individual tests, and not from the total score of the whole system, for the only source of progress in the perfection of systems of tests lies in perfecting the individual tests. The system as it stands may be fairly effective, but the analysis of the individual tests will usually show that it could be more effective.

Every study of systems of tests will probably show that the total score has greater reliability than any of the individual parts. The twenty-three Binet tests used in this investigation scored together in the form of "mental ages" had a maximum diagnostic value of 69%, which is more than double the average diagnostic value of the individual tests. Yet if the total score were computed from 5 of the most effective tests, the diagnostic value was 83%. The same has been indicated time and again. The pooled score always shows a higher correlation with intelligence than the individual tests. The whole is more reliable than the parts, yet the effectiveness of the whole is raised by increasing the reliability of the parts.

To one who has followed the analysis of the Binet tests through these pages, it is perhaps surprising that the scale works. Yet it does work within certain limits. It will indicate pronounced defect in children over certain ages, and as the writer has shown in a previous article (14), it will diagnose the finer shades of intelligence from 7 to 11 as expressed by

the teachers' judgments. The merit of the scale as it stands undoubtedly rests on two principles. In the first place, the tests are arranged in the approximate order of their increasing difficulty (by the criterion of 75% passed) so that the experimenter can find tests within the subject's range of ability by exploring from the "basal age" upwards. In the second place, the experimenter in exploring this ability gives the subject a number of tests. The merit of the scale rests then on the principles of having a number of tests, and of having those tests within the range of the subject,—neither too far above or too far below his ability. There is nothing essentially new in these principles. It is a matter of common observation that the user of the shot-gun frequently chooses the "spread" instead of the "choke" barrel, and that the user of the rifle invariably adjusts his sights to suit the range.

In regard to the principle of having a number of tests, Binet is quite frank. His explanation of the reason for measuring intelligence by groups of tests follows. "Obviously it rests upon the principle that a particular test isolated from the rest is of little value, that it is open to errors of every sort, especially if it is rapid and is applied to school children; that which gives a demonstrative force is a group of tests, a collection which preserves the average physiognomy. This may seem to be a truth so trivial as to be scarcely worth the trouble of expressing it. On the contrary it is a profound truth, and good sense is so far from being sufficient to divine this so called triviality, that up to the present it has been constantly disregarded. One test signifies nothing, let us emphatically repeat, but five or six tests signify something. And that is so true that one might almost say, 'It matters very little what the tests are so long as they are numerous.'" (Kite's (39) translation pg. 329).

Inasmuch as no article on mental tests is complete without a suggested revision, the writer will indicate his opinion on the nature of such a revision. If the presence of so many variable factors in mental tests means that it is impossible to have quantitative measuring scales of intelligence, then of course

the development of mental tests is merely the development of more diagnostic tests with studies of the types of performance that are qualitatively diagnostic. However, the assumption underlying a qualitative diagnosis is that owing to certain facts the experimenter expects the subject to do differently. The certain facts that the experimenter takes into consideration represent merely the performance of other similar children. In other words the assumption underlying a qualitative diagnosis is the assumption of norms of performance, and if we can have norms of performance on individual tests, we can have norms on groups of tests with the probability that the group of tests will be more reliable than its components. The assumption that norms of performance are possible rests on the assumption of equality of opportunity. The following remarks are made on the assumption that it is possible to classify individuals as having had equal opportunity.

One of the first principles to be recognized in considering what the nature of an adequate measuring scale would be is that the individual tests should be scored on the point or partial credit system rather than on the all or none system, for as Yerkes points out, the scoring of tests by points brings out the full value of the testing material and minimizes the influence of the personal equation. It might be objected that some tests can not be scored on a basis of partial credits, for they are all or none tests, they are either passed or failed. The months test and the counting stamps tests are examples of all or none tests. It is probably true that no test should stand entirely by itself, as there is always a danger that the subject has been told about some individual tests. If the tests of naming the months, days of the week, giving the date etc. are diagnostic tests, their efficiency would not be decreased by combining them into one test and weighting them according to their relative difficulty, or if the ability in the counting stamps test is diagnostic, a person would stand a better chance of measuring this ability by giving other problems than that of counting three 2 cent stamps and three one cent stamps (counting two 2's and two 1's, three 3's and three 1's, for example). It has been shown

that there are dangers in weighting parts of tests according to the point system, but the parts may be weighted empirically according to their difficulty.

Another principle that should be recognized in considering the nature of an adequate measuring scale is that a number of tests should be given well within the range of ability of the subjects tested, for the best differential measure of a number of groups is one well within the ability of the groups. This was brought out in the studies of the personal equation, grade correlations, sex differences, and diagnostic value of the tests. It is an obvious principle, but one that needs recognition. An example from Yerkes' point scale will make this clear.

The examination of Fig. 5 in which the growth of each test in the point scale with age is shown graphically shows that for the most part the tests are either one thing or another—they are either tests for young children or for older children. Tests 4, 5, 7, 8, 10, 11 and 14 are useful for indicating growth from 5 to 9, but are practically useless beyond that point for the abilities are almost completely developed at 9. The 9 year subjects scored over 75% of the possible number of points on these seven tests, and there is less than 20% improvement manifested above 9. At the other extreme are tests 12, 15, 16a, 17, 18, 19 and 20 that are apparently valuable for indicating growth above 9, but are practically useless below 9. Of the remaining 7 tests, 1, 2, and 3 are useless for all ages, tests 6 and 16 have doubtful significance, and tests 9 and 13 alone have value for differentiating the intermediate growth from 8 to 12.

From the data in table 30, page 123, the present writer has calculated norms for each year for four point scales, (1) the original point scale without test 16a, (2) the same scale eliminating tests 1, 2, 3 and 6 that have an error in the scoring, (3) a scale for younger children consisting of tests 4, 5, 7, 8, 9, 10, 11, 13 and 14, and (4) a scale for older children consisting of tests 9, 12, 15, 16a, 17, 19 and 20. The norms are calculated on the basis of the per cent. that the number of points scored is of the number possible to score, that number being 100 points for the first scale, 72 points for the second scale, 34 for the third and 36 for the fourth. These norms are given in table 20.

TABLE 20.
Norms for Four Point Scales.

Chronological ages	5	6	7	8	9	10	11	12	13	14
Scale 1.....	22	30	36	42	56	62	65	77	79	81
Scale 2.....	12	20	27	35	51	59	62	75	79	82
Scale 3.....	22	35	45	57	76	82	85	93	93	94
Scale 4.....	3	6	10	14	31	38	43	63	70	73

The effect of throwing out tests 1, 2, 3 and 6 is to lower the norms of the younger children without changing those of the older children to any extent, so that the subjects may express their ability within a range of 70 points instead of 59 points. In other words the scale is more effective, for the greater the range in which differences may be expressed, the greater the possibility of differentiation. In the same way scales 3 and 4 are more effective than scale 2. Scale 2 gives a range of 39 points between 5 and 9, while scale 3 gives a range of 54 points. Scale 2 gives a range of 31 points from 9 to 14 while scale 4 gives a range of 42 points. Scale 1 has a range of 59 points, 58% of the increase being scored by children from 5 to 9 and 42% by children from 9 to 14. Scale 2 has a range of 70 points, 56% being scored by the younger group and 44% by the older group. On the other hand, 75% of the 72 points in scale 3 are scored by the younger group (5 to 9), and 60% of the 70 points in scale 4 are scored by the older group (9 to 14).

The foregoing demonstration does not mean that Yerkes' scale as it stands would be more accurate if broken up into two parts, for the accuracy is in some measure dependent on having a sufficiently large number of tests. It is true nevertheless that the use of tests above or below the ability of any group lessens the possibility of differentiating that group. The accuracy of any point scale system of testing intelligence increases with the number of parts, in the degree in which it tends away from universality. In a recent communication to the "Symposium" in the *Journal of Educational Psychology* (63) Yerkes stated that he had abandoned his plan for a universal point scale, and suggested three age scales, from birth to 4 years, from 4 to 12, and from 12 to maturity or 16. Ultimately, he will probably have to split the scale from 4 to 12 into two scales at least.

Aside from the fact that the difficult tests decrease the possibility of differentiating the younger children, and the easy tests decrease the possibility of differentiating the older children, it is a waste of the experimenter's time to ask a young child to perform tests obviously beyond his ability, or to ask an older child to do tests way beneath his ability. But the experimenter never knows the range of a child's ability till he starts to test. What is needed then to avoid waste of time and to provide for the rapid exploration of a subject's possibilities of accomplishment is a series of over-lapping point scales in which a

certain number of tests may be found within the range of every child.

This of course is just another way of describing the Binet scale. The Binet system in which the experimenter finds the basal age and then gives more and more advanced tests until the subject fails several in succession enables the experimenter to give the subject a number of tests within the range of his ability. The four Princeton experimenters using the narrow range method of testing actually averaged about 19 tests for each individual, a number which is remarkably close to the 20 tests prescribed by Yerkes. The proof that this system is reliable rests on the fact that the Binet system has been found to be reliable within certain ranges. It breaks down at the extremes or in those regions below which or above which there are no more effective tests by which the subjects may be differentiated. The chief limit is the number of tests.

So far then our analysis of the nature of an adequate measuring scale has led to the conclusion that this scale must be a point scale on the Binet age scale basis, or in other words that both scales have desirable features. Yerkes' point scale is superior in that it has the partial credit rather than the all or none system of scoring, and the Binet age scale method is superior in that it has greater adaptibility so that the experimenter can find more tests within the subjects' range of ability. The advantage of having post-experimental norms is not peculiar to the point scale for it is just as easy to compute the average "mental age" of groups of non-selected children of different age, sex, nationality, sociological status, etc., as it is to compute the average number of points scored. From a practical standpoint, both scales are probably satisfactory within certain limits. Both scales have a large number of users and have yielded valuable results.

The chief objection against the Binet type of scale is that it is a closed system of tests and admits of no improvement. The moment an experimenter changes a test or the position of a test in the various Binet scales, all the norms of past experimental work are useless. The study of the individual Binet

tests in this investigation certainly shows that the scale can be improved, yet its structure precludes the possibility of improvement without discarding all the norms of previous experimental work. In admitting that the individual tests in the Stanford revision vary in the degree with which they correlate with intelligence (see page 231), Terman admits that the scale can be improved, yet it is impossible to improve it without changing the "intelligence quotients" so carefully worked out on an empirical basis. The chief objection to Yerkes' scale is its lack of adaptability, the fact that its structure places a limit on the number of tests that can be found within the range of ability of the children of the ages for which it is designed to test.

The need of a scale that has an elastic structure and one that may be improved upon indefinitely brings us back to Huey's (36) original conception of a point scale, which consisted of "the per cent. of intelligence obtained by adding together all the points earned, multiplying by 100, and dividing by the sum of the points allotted to the tests actually given and counted as given", or in other words the per cent. that the number of points scored is of the possible number of points. On this basis the norms of performance are not given for any whole system of tests, but for the individual tests. Each experimenter must select his tests to meet his problem, and must compute his norms from the published norms on individual tests.

Such a system has the obvious advantages of adaptability and improvability. If the problem were that of differentiating younger children, the experimenter would select tests that others have found useful in this respect. Under present conditions there is no measuring scale for younger children in which an experimenter can use such an historically valuable test as the form board. If the problem were the diagnosis of the higher grades of defect, the experimenter would select tests useful for making this differentiation. Under such conditions each clinic would probably have its own group of tests, and the only basis of comparability of results would be the individual tests. Huey's system is really nothing more than a percentage system of scoring the results of tests, and the advantage of

the method lies in the fact that a system of tests is more reliable than the individual tests—that a person is more apt to hit a bird with the “spread” barrel than he is with the “choke” barrel.

On this system scales could be constructed on almost any basis, the only limit being the number of standards to which the percentages could be referred. The writer has in mind two types of scales which should prove immediately useful, one an age scale, and the other a scale for feeble-mindedness.

A convenient form of age scale would be a series of tests arranged according to increasing difficulty, the arrangement being made according to any arbitrary criterion such as a score of 50% or 75% at certain ages or combinations of ages. Each test would consist of a number of parts empirically weighted. It would probably be convenient to weigh all tests on some arbitrarily selected scale such as 10 which has the advantages of the decimal system. It would be unnecessary to weigh tests differently for eventually it should be possible to have nothing but valuable tests in the scale. The tests could conveniently be placed in groups according to equality of difficulty, and the score computed from any number of successive groups of tests. The only limit to the number of tests in a group would be the number of tests that could be found of approximately the same difficulty, and any number of groups could be combined.

For purposes of illustration the writer will take five tests for a group and make every successive twenty tests a successive scale. The form of such a scale would be

A ₁	B ₁	C ₁	D ₁	E ₁	F ₁N ₁
2	2	2	2	2	2	2
3	3	3	3	3	3	3
4	4	4	4	4	4	4
5	5	5	5	5	5	5

Group ABCD would constitute one point scale, group BCDE another point scale, etc. If the norms on the individual tests are accurate the per cent. that an individual scores on scales ABCD, BCDE, CDEF, etc. should refer to the same age, only of course the scale nearest the subject's range of ability would

give the most accurate estimate. A preliminary series of tests such as A_I, B_I, C_I, D_I, etc. could be given to ascertain the subject's probable range. Indeed a preliminary scale that would roughly place the subject near his proper level could probably be arranged for class-room testing. In this way a whole school system could be tested, selecting the doubtful cases on a class-room scale, and giving these cases more and more detailed examinations, the number of tests that can be found being the only limit. Any test that is found to be inaccurate could immediately be thrown out and another substituted, it being a simple matter to recalculate norms.

A series of scales for feeble-mindedness could be arranged without reference to age norms. The scales could be for the three groups, idiots, imbeciles and morons, limiting the groups by some such arbitrary criteria as Binet proposed, viz. that the idiot is one who never acquires spoken language, and the imbecile one who never acquires written language. Illiteracy could be experimentally determined from imbecility by researches on groups of individuals who after long training have not been able to acquire written language. The differentiation of the moron from the normal individual requires further experimental work. Within each group the individuals could scale over a range of 100%, and further study would probably indicate the border-line for the sub-divisions "low", "middle" and "high". These scales could be improved and elaborated as more diagnostic tests were discovered. Of course the scales would not necessarily be limited to the groups for which they were designed. The imbecile could be given the scale for morons, just as well as either one could be given the age scale for normal individuals or as normal individuals could be given the feeble-minded scales. It is the writer's opinion that if scales for feeble-mindedness could be developed that were not standardized on an age basis and the results were then studied in the light of their age correlations, the results would be illuminating.

The chief advantage of systems of tests modelled according to Huey's plan is that they admit of improvement. It is probably impossible for any one person to perfect a final quantitative

scale. It would seem better to adopt a skeleton plan which would allow a more perfect scale to evolve. Such a plan would encourage researches on individual tests both in the public schools and in institutions for the feeble-minded, and it is only by co-ordinating such researches that we will ever be able to solve the problem of the nature of intelligence or the corresponding problem, the nature of feeble-mindedness.

