

able to float in some kind of simple industrial environment. We have reason to believe that appropriate vocational training and social oversight of such children would in the end be far less costly than the present **laissez-faire** policy. Before many years it will probably become a matter of course to apply serial mental tests in the public schools to all pupils who are retarded or about to become retarded, or who give indications of unusual ability. The scientific management of special classes for atypical children in the public schools will be impossible until similar tests are multiplied indefinitely.

## THE PRESENT STATUS OF THE BINET AND SIMON TESTS OF THE INTELLIGENCE OF CHILDREN

BY F. KUHLMANN, *Faribault, Minnesota*

The writer used the 1908 series of the Binet and Simon tests in examining the inmates of the Minnesota School for Feeble-Minded and Colony for Epileptics, approximately 1,300 cases. In a later article the results of these examinations will be reported so far as they throw any light on the accuracy of the tests in determining the mental development of feeble-minded children. The object at present is to bring together the scattered results of others who have used and criticised the tests, and attempt an evaluation of these in the light of the combined results and of whatever the experience of the writer is able to add.

### A. The System of Tests as a Whole.

The tests are the first of their kind that have ever been offered for the purpose of determining the degrees of intelligence of children in terms of mental ages. They aim to and do accomplish much more than anything we have had heretofore. For this reason they have become at once widely popular. They have been used in many public schools throughout this country and abroad, and in a number of schools for defective children, reformatories and prisons for the practical purposes of grading intelligence. They have also been tried in an experimental way by various individuals for the purpose of testing their accuracy and to discover revisions where found to be needed. As a result of these combined circumstances and unusual activity we have already a considerable mass of data and criticisms that point the way to a rapid progress. There is, however, a sharp line to be drawn between two kinds of results from the use and study of the tests. These are (I) the actual degree of correlation found between the different tests of the system and the performance in

<sup>1</sup> With a few exceptions in the procedure in giving a test, they were Used exactly as given in my account of them in this Journal, Vol. XV, 1911. The reader is referred to this account for any information in regard to them that is assumed in this article.

them of normal children of the different chronological ages, and (2) generalizations and deductions as to the value of the tests, based largely on an **a priori** analysis of the nature' of the tests and on what we know or assume about the mental development of normal children. The latter can have but little value where they contradict the former, provided that the methods of determining the correlations are themselves free from criticism. The results and criticisms of the tests will therefore be considered under the following headings: (1) Statistics with normal children, and (2) general observations and criticisms.

I. Statistics with Normal Children. The point in question here is how closely the mental ages as determined by the tests agree with the chronological ages of normal children tested. The final proof of the degree of accuracy of the tests must be given by the degree of this correlation. For the 1908 series Binet and Simon tested 203 children of the schools who were up to grade, that is, were in the grades in which they should be according to their chronological ages. For 192 of these they give the results III the following table.

TABLE I.

Chronological Age	3	4	5	6	7	8	9	10	11	12	Total
Regular	3	9	13	5	7	16	11	14	13	2	93
Advanced 1 Year	3	2	6	8	7	5	9	2	..	..	42
Advanced 2 Years	1	..	..	..	1	..	..	..	..	..	2
Retarded 1 Year	4	4	4	6	3	1	2	9	5	5	43
Retarded 2 Years	1	..	1	1	..	..	3	2	4	..	12
Total	10	17	23	20	18	23	22	28	20	11	192

These figures give the number of children tested for THE chronological ages of three to twelve years. The term "regular" means children whose mental ages and chronological ages agreed. Likewise the terms "advanced" and "retarded" mean children whose mental ages were greater than or less than the chronological ages, respectively. 1 As is seen the correlation is perfect III

2 Le Development de L'Intelligence chez les Enfants. L'Annee Psychologique, 1908, P. 73.

3 In the totals the authors give 103, and 44 for 93 and 43, respectively, in the present table, apparently errors in adding.

93, nearly half, the cases, and there is a discrepancy of over a year in only 14 cases.

In giving the tests to other than French children a number of changes and adaptations of the original must be made, resulting from translation of verbal material used in the tests, and from incidental differences in the civilizations of different peoples. Further, the average normal intelligence of children of different nationalities might vary. Hence Goddard first determined the norms for these tests with 2,000 American school children, for 1,547 of which he gives tabular results. The following is one of his tables :

TABLE II.

Age	2	3	4	5	6	7	8	9	10	11	12	13	Total
4	..	1	2	2	3	..	..	..	..	..	..	..	8
5	..	2	4	8	40	40	16	4	..	..	..	..	144
6	..	1	..	3	29	48	69	9	1	..	..	..	160
7	..	..	1	2	8	15	114	50	4	3	..	..	197
8	..	..	..	2	2	1	87	86	16	12	3	..	209
9	..	..	..	..	..	..	27	54	56	58	4	2	201
10	..	..	..	..	..	..	15	24	19	124	27	8	222
11	..	..	..	..	..	..	4	13	25	50	60	12	166
12	..	..	..	..	..	..	4	10	13	42	36	39	144
13	..	..	..	..	..	..	1	5	6	30	19	21	89
14	..	..	..	..	..	..	..	1	1	6	5	4	20
15	..	..	..	..	..	..	..	..	3	..	1	2	6
Total	..	3	6	17	81	111	337	256	143	326	155	88	1547

The First Horizontal column gives the mental ages, and the vertical column on the left gives the chronological ages. The others give the number of children tested under each age. From these figures Goddard concludes that "To a person familiar with statistical methods the foregoing curve itself amounts to practically a mathematical demonstration of the accuracy of the tests. . . . We are forced to the conclusion that the questions that Professors Binet and Simon have selected are well graded, at least from the ages five to twelve, and that they fit the ages to which they are assigned."

In addition to this Terman and Childs give statistical results

Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence. Ped. Sem., 1911.

of the examination of 396 California public school children. 3  
They give the gross results in the following form :

TABLE III.

No. Tested	29	83	26	29	43	49	33	44	35	17	6	2
Av. Chron. Age	4.75	5.5	6.37	7.5	8.5	9.5	10.5	11.46	12.33	13.42	14.58	15.2
Av. Men. Age	6.0	6.5	6.5	7.5	8.0	9.0	10.0	10.0	10.5	11	12	11.5

The first horizontal column gives the total number of children tested for the different chronological ages. In the second column the first figure, 4.75, gives the average chronological age of the 29 children, whose ages were between four and five. Likewise, the second figure, 5.5, is the average age of the 83 children, whose ages were between five and six, etc. The third column gives the average mental ages minus a half year in each case, which is subtracted from the authors' figures to make them more directly comparable with the others. Terman and Childs add a half year to the mental age of a child as determined by the tests, on the basis of the assumption that the chronological ages as given by Binet and Simon are all a half year smaller than they should be, since they seem not to have considered fractions of a year, but called all children between five and six, for example, five years old. They have also not followed Binet and Simons rule of adding a year to the mental age for every five tests a child passed beyond the age group in which he passed all or all but one. In place of this they added a half year for every three additional tests thus passed for the age groups III to VI, inclusive. For the seventh year a half year credit was given for four tests. From the ninth to the twelfth year, inclusive, three tests passed • counted again for a half year, and five tests passed for a whole year in the mental age. For these variations in the procedure corrections in their figures cannot be made from the data given. This complicates matters very much when we aim at a really accurate comparison of results. It is impossible to say with certainty in what direction Terman and Childs' procedure tended to vary their results from those of others. They conclude from  
5 A Tentative Revision and Extension of the Binet-Simon Measuring Scale of Intelligence. Jour. of Educat. Psychol., 1912.

their results that "the scale is far too easy at the lower end, while at the upper end it is too difficult."

Statistical results with the tests for normal children have been obtained also by Miss Johnston, who examined 200 school children of Sheffield, England; by Bobertag who adapted the tests for German school children and examined 435; and by Isabel Lawrence who tested 784 Minnesota school children, using only the tests that have to do with giving definitions of terms. In none of these, however, are figures given to show the degree of accuracy of the system of tests as a whole. Some of their figures will be considered when we come to discuss the individual tests. On the question of the degree of the accuracy of the tests as a whole as indicated by the degree of correlation given in statistics with normal children we are limited, therefore, to the results from the three sources given. Are the methods by which these figures have been obtained free from criticism, and what conclusions from them are justified? We may take up at this point such general observations as are concerned directly with these statistics themselves.

The authors themselves give results for only 192 children examined, and it has been objected that such a small number is inadequate. This inadequacy becomes obvious when we note that the total number of children examined for the different chronological ages ranges from ten to twenty-eight. The same criticism does not apply to Goddard's figures, excepting for the chronological age of four. Wallin apparently objects to these results on the grounds that Goddard's cases were not selected children, but undoubtedly included some mental defectives. The

6 Journ. Educat. Psychol., 1912, P. 70.

7 An English Version of M. Binet's Tests for the Measurement of Intelligence. Training School Record, London, 1910.

8 Ueber Intelligenzpruefungen (nach der Methode von Binet und Simon). Zeitschr. f. angew. Psychol., 1911.

9 A Study of the Binet Definition Tests. Psychol. Clinic, 1911.

10 Wallin, J. E. W.: The New Clinical Psychology and the Psycho-Clinicist. Journ. Educat. Psychol., 1911. P. 204. Terman, L. M.: The Binet-Simon Scale for Measuring Intelligence. Psychol. Clinic, 1911, P. 200.

same holds true of the results of Terman and Childs. He suggests selecting children who are up to grade in their school work as a method adequate for practical purposes. The present writer is inclined to add that some of the testing appears to have been done rather hurriedly. This is inferred from the statement that an examiner tested from twelve to thirty children a day, apparently during only the school hours. Concerning the time required to test a child carefully, I am on the whole in accord with Wallin's statement that "To examine five or six pupils in an hour at a given level in the scale means partial and perfunctory work, and will render the try-out essentially unscientific," with which statement Terman and Childs seem also to agree in noting that "tests carried through at the rate of 20 to 30 per day are sure to give unreliable and misleading results."<sup>12</sup> Another criticism would be that he seems not to have taken account of the exact chronological ages as much as should be demanded. If a child, for example, is called six years old until his next birthday it is obvious that the average age for large numbers called six years old will be about six and a half years. If in this procedure a scale of tests were adjusted so that the results would come out correctly according to the chronological ages given the tests would all be too difficult inasmuch as the}' would all fit higher chronological ages than indicated. But since the mental progress made from one year to another by a young child is much greater than for an older child, the errors in the scale of tests would be much greater for its lower than for its upper part, and would decrease proportionately to the rate of mental development. Moreover, when the actual chronological ages of a number of children who are called six years old range from six to seven it is evident that the mental ages determined by a scale of tests that is entirely correct must also range from six to seven for really average normal children. With an imperfect scale of tests and with a group of children varying from the average normal the degree of correlation between the results of the tests and the chronological ages given must therefore be considerably less than it

<sup>11</sup> Human Efficiency, Ped. Sem., 1911, P. 81.

<sup>12</sup> Journ. Educat. Psychol., 1912. Foot-note, P. 65.

Should be. This criticism is even more applicable to Binet and Simon's results, if they also have not taken account of fractions of a year in considering the chronological ages, since their number of cases is so much smaller. But one would hardly suppose that under this circumstance they would disregard this matter. This brings us to a comment on a part of the conclusion Goddard draws from his figures, namely, that the tests fit the ages to which they are assigned. The figures themselves show exceptions to this statement for several different ages. It will be noted in his table that five-year-old children are six years old mentally as often as the}' are five. Six-year-old children are seven years mentally oftener than they are six. The eight-year-old children are only seven mentally as often as they are eight, and those nine years old are mentally eight, nine, and ten with about equal frequency. For eleven-year-old children the mental age is ten nearly as often as it is eleven, and for the twelve-year-old the mental ages are ten, eleven and twelve with no conclusive difference. In other words, for six out of the nine chronological ages (excluding the chronological age of four because the number of cases here is inadequate) Coddard's conclusion does not quite hold. If we accepted Goddard's method of obtaining his norms as quite free from any criticism it would be true that for the chronological ages of five, six, eight, nine, eleven and twelve the tests give an error of a year in the mental ages as often as they give the correct mental age. Comparing the results of Goddard with those of Terman and Childs, it is seen that, even with considerable difference in procedure, the}' agree in showing the tests in age groups V and VI as too easy, and those of age groups XI and XII as too difficult. If we regard the above criticisms as essentially valid, it leaves the question we stated at the outset as to how closely the mental ages as determined by the tests agree with the chronological ages of normal children still largely an open one. However, there seems to be sufficient indication to warrant the claim that the tests on the whole give much more accurate results than we can obtain at present in any other way, except by close observation of the individual child for periods of many months or years. For all but the lowest part of the scale

an error of only a year in the mental age is very accurate compared with the judgment the teacher is usually able to give of her pupils. Normal children probably vary over about the range of a year from their average performance in a given chronological age.

2. General Observations and Criticisms. From the use and study of the tests there has resulted a miscellaneous group of observations and criticisms that are not a matter of statistics or based on statistical results. So far as these are not concerned with any one or few individual tests they will be considered next.

a. Lack of standardization. It has been pointed out that the authors have not given sufficient directions as to just how each test is to be given and how the results are to be interpreted in each case.<sup>13</sup> This is true in a large number of instances and in a variety of ways that cannot all be enumerated here. In my account of these tests<sup>14</sup> more specific directions have been added in some cases, but they are still far from complete. In response to this lack Wallin has also published an account of the tests in which an attempt is made to remedy this deficiency.<sup>15</sup> The authors in their 1911 revision of the tests have improved the 1908 series considerably in this respect.<sup>16</sup> The matter is of much importance, inasmuch as quite different results may often be obtained by only a slight variation in the procedure. It follows that in this refinement of the method the testing-out is yet to be done before the best ways are found with reference to these details.

A special and important instance of lack of standardization has appeared in the necessary adaptations from the French for other than French children. These adaptations have not always been equivalents of the original, and have in some instances included unnecessary changes. Substitution of sentences to be

<sup>13</sup> Wallin, J. E. W., *Ped. Sem.*, 1911, P. 78.

<sup>14</sup> See this Journal, 1911.

<sup>15</sup> A Practical Guide for the Administration of the Binet-Simon Scale for Measuring Intelligence. *Psychol. Clinic*, 1911.

<sup>16</sup> La mesure du developement de l'intelligence chez les jeunes enfants. *Bull. de la Societe Libre pour L'Etude Psychologique de L'Enfant*, 1911.

repeated for translations of the French, of words to be defined, of American coins for French in tests in which these are involved, of pictures used in some tests, and changes in the arrangement of words to be put in order to make a sentence, are illustrations.

b. Inequality of number of tests for different ages. In the 1908 series the number of tests for each age varies from three to eight. The rule given for determining the mental age from the results is to credit the child with the mental age of the highest age group of tests in which he passes all, or all but one, plus one year for every five tests he passes beyond this point. This complicates the scoring, especially when it is attempted to give the mental age in terms of fractions of a year. Thus, as Wallin notes, "If the subject passes age six by virtue of two failures in age seven he can obtain one and one-fifth year of credit for age seven ; i. e., one-fifth of a year more credit than if he were credited outright as having passed age seven."<sup>17</sup> It is obvious also that for those age groups in which there are only three or four tests an extra year of credit may be obtained by passing only two or three extra tests beyond the age group in which all but one are passed. This difficulty, however, is not in itself a serious matter, as it can be easily remedied. Terman and Childs suggest the plan of attributing a "unit value" to each individual test that is given by the fraction of one over the number of tests in the age group in which the test in question is found.<sup>18</sup> This assumes that a test has the greater value for determining the mental age the less the number of tests that are found in its age group, an assumption which the authors might be supposed to have made, since they left the number of tests unequal. There is no evidence that the assumption is correct, yet the plan is a considerable improvement over Binet and Simon's old procedure. In their 1911 revision of the tests the authors have reduced the number of tests to five for each age group, excepting for the four year group.

c. Communicability and coaching. It has been objected that whenever a group of children that associate with each other are examined the brighter ones who have already taken the

<sup>17</sup> *Ped. Sem.*, 1911, P. 80.

<sup>18</sup> *Psychol. Clinic*, 1911, P. 201.

tests may communicate them to others and coach them, which makes them tests on the ability to profit by such coaching rather than of native intelligence. To this Goddard has replied that a child cannot learn to do a thing if the task is beyond the natural abilities of his mental age, cannot retain what has been told him, and has found such coaching "practically without any influence upon any of his results." 20 The present writer has met some instances which tend to confirm Goddard's conclusion, but is not convinced that the rule has not too many exceptions to make the matter of possible coaching of some children by others already familiar with the tests a serious consideration for the upper part of the scale. A *priori* it seems quite plausible to suppose that a child might, for example, retain three words given him to use in a sentence he is to construct (Test X 3) and tell another so that the latter, with plenty of time, could think of such a sentence and thus be prepared for the test, even though both were considerably below the mental age of the group in which the test is found. The same might be said of the tests on definitions of abstract terms (Tests XI 4, XIII 3), of the test on "Words to put in order" to make a sentence (Test XI 5), of "Rhyming words" (Test XII 2), of "Drawing a cut in a twice folded piece of paper" (Test XIII 1), and possibly of "Drawing the figure of two juxtaposed triangles" (Test XIII 2). The question is an important one, since any test that can be communicated and be prepared for in any way can have only a temporary value, even outside of their application to children associated in groups, as in the schools and institutions. For in the long run such tests, used for such a purpose as testing a person's intelligence, are sure to become a matter of more or less common knowledge. We shall return to this question in an other connection.

d. The effect of training. The term "training" will be used here in the wide sense to include everything the child may acquire through the influence of his total environment. We will thus be concerned with a number of criticisms that have been worded differently, but which all amount essentially to the same

19 Wallin, J. E. W., in *Ped. Sem.*, 1911, P. 79.

20 *Ped. Sem.*, 1911, P. 233.

thing. This is that many of the tests are tests merely of what the child has acquired, has learned, and do not necessarily test his intelligence at all. This is true because what a given child has learned depends upon the opportunity for learning that his total environment has offered as well as upon his native intelligence. But since these opportunities vary so much in particular things from one child to another, his acquisitions may be no indication of his intelligence in any given case. Practically every writer commenting on the tests has made this criticism. The details of this discussion cannot be given here. In noting the different tests, however, that the critics have pointed out as affected by training it is seen that there is not very much agreement as to which tests are poor for this reason. Moreover, the authors admit that a number of the tests are thus affected, discuss the question at issue, and point out how such tests may still be used as tests of intelligence. Let us, therefore, turn first to the authors themselves.

The authors do not give any clear account of their position on this question, and apparently do not point out all the tests that they would probably regard as seriously affected by training, and with what reservations each is to be used. Moreover, they do not definitely state all the assumptions that are implied in the various comments scattered throughout their several publications on the tests. This has confused the issue and has led to some un-called-for criticisms. In the first place, they admit that the tests do not all measure intelligence directly. They measure a complex, with the results depending on (1) intelligence, pure and simple; (2) acquisitions due to special training and teaching; (3) school acquisitions that appear at a certain age only; (4) acquisitions relative to language and vocabulary, due possibly to both school and home training. 21 Taking into account also comments made in connection with individual tests, we find the following: (1) When, in case of certain tests, the child passes no conclusion is to be drawn as to his native intelligence. For unusually favorable opportunities for learning the particular things in these tests may be the cause of the child's ability to pass. If,

21 *L'Annee Psychologique*, 1908, P. 80.

however', he fails in them it shows his lack of intelligence. In this class they give counting four pennies (Test V 4), copying a written phrase (Test VII 3), reading for two memories (Test VIII 1), in case of adults of thirty years or over, naming four common pieces of money (Test VII 8), naming four colors (Test VIII 3), naming the days of the week (Test IX 2), and naming the months of the year (Test X 1 . (2) When in certain tests the child fails to pass no conclusion is to be drawn, because unusually unfavorable opportunities to learn may be the cause of failure rather than lack of intelligence. But if he passes it shows a certain degree of intelligence, because this is involved in the acquisition in question. In this class are given reading for two memories (Test VIII 2), in case of children from eight to ten years, giving the date (Test IX 1), and naming the months of the year (Test X 1). These statements clearly involve certain admissions and assumptions, though they are not all definitely expressed. They admit (a) that environmental opportunities may be unusually favorable as regards acquiring the ability to pass certain tests, so that the results of these tests may not give any indication of the degree of intelligence; (b) that environmental opportunities may be unusually unfavorable, so that likewise the results of certain other tests do not give any indication of the degree of intelligence. They assume (c) that environmental opportunities are always at least adequate for the normal child to acquire the ability to do the things in some of the tests at the age indicated by the age group in which the tests are found. This is implied in the statement under "1" that if the child fails it shows lack of intelligence. They assume (d) that certain acquisitions must await the development of a necessary degree of intelligence that is involved in the acquisition, which development of intelligence cannot be accelerated materially by unusually favorable conditions. This is implied in the statement under "2." How does this affect the system of tests as a whole?

It will be seen that the authors name approximately a sixth of the tests as affected by training, and that whether or not these give any indication of the degree of intelligence depends on whether the child passes or fails in them in the different in-

stances. But it is difficult to see why others not named should not also be added to the list if some of those given belong there. If counting four pennies (Test V 4) belongs to class "1" why should not also counting thirteen pennies (Test VII 7)? If naming four common pieces of money belongs here (Test VII 8), why should not also naming nine pieces of money (Test X2) ? Also, we must add Test IX 5, since it is identical with Test VIII 1. Likewise for class "2", if a child fails to read (Test VIII 1) because of lack of opportunity to learn, why should he not also fail to write (Test VII 3, and VIII 5) for the same reason? If he fails in these for this reason, why should he not fail to know his own age (Test VI 5), to count, to name four or nine pieces of money, to count the value of stamps (Test VIII 2), and others that might be mentioned, for the same reason? It is noted that Test X 1 is given under both classes. In a word, there is no a priori reason why all the tests named for both classes and others besides might not as well belong to either one of the two classes alone. In the present writer's judgment the authors have done the same thing here as have their critics. The}' have picked out certain tests as affected by training because it seems plausible on the surface that they might be thus affected, instead of stopping with what empirical facts can show in regard to this question. But one of the results of the use of the tests and of the criticisms has been to show that this question cannot be decided in this way. It is one of the most important questions any test of intelligence has to deal with, and at the same time one of the most difficult to solve. The degree of validity attached to either of the implied assumptions of the authors as stated above is not yet determined. Yet in order that a test may be unaffected by training both assumptions must hold true for it. Meumann 22 makes an observation that applies here, to the effect that beyond the age of four there is no knowledge that may not be acquired through school or parental Straining and which may not vary with different children. We might

22 Der gegenwaertige Standt der Methodik der Intelligenzpruefungen mit besonderer Ruecksicht auf die Kinderpsychologie. Zeitschr. f. exper. Paedagogik, 1910.

note in general that differences in acquired knowledge and abilities of different children as due to differences in environmental opportunities undoubtedly increase rapidly with age, so that the problem of avoiding the effect of this in tests of intelligence varies accordingly. The contention has been repeatedly made that to test intelligence we must test mental functions directly (sensory discrimination, perception, memory, attention, etc.) as distinct from determining merely mental content. The development of these mental functions is supposed to be influenced but little by any differences in environmental opportunities. To the writer the supposition in itself seems a very plausible one. But in attempting to measure intelligence by testing these functions the following difficulties have been found. We cannot yet adequately isolate these functions in any tests so far devised, and in a given test we therefore do not always know what function we are testing most. Consequently we do not yet know what degree of correlation exists between any one or several of these functions and intelligence. Further, these functions cannot express themselves except through mental content any more than they can develop except through use. To test attention we must test attention to some particular thing, the same being true of every other function. Now the degree of perfection of a function as thus determined has been found to vary somewhat with the particular thing chosen. One thing is attended to better than another, one thing is remembered better than another, and a given function also seems to improve in a small measure at least with continued practice with a given mental content. In other words, even though we could completely isolate the mental functions for the purpose of determining their degree of development, and if there were a close correlation between them and intelligence, we would still not be entirely free from the question of the effect of training, but would only have materially reduced the degree of this effect. From these various considerations the writer is forced to the general conclusion that under the circumstances we can at present do no better than to keep these facts and criticisms in mind and rely on the actual empirical results of

*a* test, the results showing how closely it actually does correlate with known degrees of intelligence, in judging its value and in deciding on the need and nature of a change or substitution,

e. Some of the tests are too mechanical. Decroly and Degand, chiefly, have advanced the criticism that a number of the tests are too mechanical. This means that the tasks involved in them can be performed in a semi or entirely automatic manner without intelligence taking any part. Counting pennies, naming the days of the week, and the months of the year are illustrations. In its further analysis their criticism really reduces itself to the just preceding, the effect of training. For the performance of any of these tasks was not automatic from the beginning with the child. They are not inherited reflexes, but had to be acquired through the combined influence of intelligence and training. If then any of the tests are too mechanical it can mean simply, so far as this criticism alone is concerned, that they are placed too high in the scale. They should be placed at the point where the child has not yet learned to do the tasks involved automatically. We thus see that the last three criticisms considered are concerned with essentially the same question.

f. Some of the tests are wrongly placed in the scale. The several revisions of the system of tests that have been offered so far have been concerned mostly with shifting various individual tests up or down the scale because they were regarded as too easy or too difficult for the age groups in which they were originally placed. Since the correct rating of each individual test is the fundamental thing in the whole system, this question is of the first importance. If the tests thus affected are otherwise good, however, the defect is much more easily remedied than is the defect that comes from varying degrees of the effect of training with different children. The second part of this article will be concerned with the question of what individual tests are subject to this criticism. We will note only a few general conclusions at

For the main criticisms and discussion on this question of the effect of training on the tests see especially the following. Binet and Simon, in *L'Année Psychologique*, 1908; Decroly and Degand in *Archives de Psychol.*, 1910; Bobertag in *Zeitschr. f. angew. Psychol.* Vol. V.

this point. Wallin<sup>24</sup> thinks that the "aggregate difficulty of the tests for a given age may be greater than that for a higher," and that the upper part of the scale is especially defective. Miss Johnston<sup>25</sup> finds the tests for 'the age groups of six and seven too easy, and those of the age groups from ten on too difficult. Terman and Childs' conclusion was already given above, namely, that "the scale is far too easy at the lower end, while at the upper end it is too difficult." The same conclusion seems also indicated by the results of twenty-four children tested by Alice Descoedres.<sup>26</sup> On the other hand, this does not agree with Goddard's revision of the scale, and agrees only in part with the revision of Binet and Simon, though Goddard's own figures tend again to verify it.

g. Defects of omission. It has been objected that the tests are of the intelligence chiefly, but the intelligence is only one of a number of functions of the mind. We should have tests of all the other mental functions as well.<sup>27</sup> It is his total mental development that we are interested in, not its development along one particular line alone. To carry this suggestion out in detail would mean that we should have tests that would give us not merely the mental age in general, but the age development of sensory discrimination, of motor co-ordinations, of memory, of perception, of attention, of the feelings and emotions, etc. Any adequate discussion of this question would involve a precise definition of the term "intelligence," which the writer is not prepared to give. But there are several things that may be noted without attempting such a definition. The authors use the term "intelligence" loosely, and their tests are not merely or chiefly, even, of intelligence in the narrower sense of the term as used in current psychology. It would be difficult to say just what they do test in each individual case, if an analysis of the mental processes in-

<sup>24</sup> Ped Sem., 1911, P. 78.

<sup>25</sup> Journ. Exp. Ped. and Training College Record, 1911.

<sup>26</sup> Archives de Psychol., 1911.

<sup>27</sup> See Huey, in Journ. Psycho-Asthenics, 1910; Wallin, in Ped. Sem., 1911, and in Journ. Educat. Psychol., 1911; and Pyle, in Journ. Educat. Psychol., 1912.

involved in the child's mind is asked for. They are tests that have been found in an empirical way to give results which show the general mental development of a child. The question as to what mental processes are involved in the tasks the child has to perform is ignored in most cases. Clearly the tests do not aim at a systematic determination of the development of any particular mental function. Further, some of the mental functions are hardly involved in any of the tests, the general motor co-ordinations, and the feelings and emotions for example. It is to this latter fact that Huey and Wallin object chiefly. It is in the general form given above that Pyle states the criticism. The latter seems to imply that we would arrive at a better solution of the practical problem of finding tests that will accurately determine the general mental development by setting out to devise tests that will systematically test the different mental functions. For the reasons already given above concerning the testing of mental functions, the writer doubts very much that this could be done at present. It would, of course, be highly desirable to do so as a means of determining special mental traits aside from the task of determining general mental development. But that is another matter, and it is not a defect of the tests that they do not accomplish what they do not aim to do.

A criticism of more importance is Huey's contention that mental development goes on to adult age and beyond, while the tests stop at the chronological age of thirteen. It is just at this point, in fact, that many of the sexual and social instincts, for example, begin a period of rapid development. Many of the conflicts between the individual and the laws and customs of society occur because of the combined influence of a previous slight defect in development and what appears in the total mental development after thirteen. Thus it happens that the tests fail to cover just that part of the field where accurate determinations are at present needed perhaps most of all, at and around the border-line between the normal and the defective for the higher chronological ages.

## B. The Individual Tests.

1. Statistics on Individual Tests. It is not implied or as-

sumed in these tests that each individual one will always give correct results with normal children, but only that it will do so in a certain fixed percentage of cases. It is assumed, let us say, that seventy-five per cent., at least, of normal children of the age indicated by the age group in which a test is placed will pass that test. The authors do not give any such exact percentage on the basis of which each individual test is rated. Goddard chooses seventy-five per cent, and Terman and Childs, sixty-six per cent, as a basis. Under this condition an error in determining the mental age may accidentally occur, but the chances for such an error are decreased, of course, with the increase in the number of tests in the age group. The figures given above showed with what frequency such an error occurred, assuming that the methods of obtaining these figures are themselves faultless. We are now to consider similar results for the different individual tests, to determine the accuracy of the assumption that each test will give correct results in the given fixed percentage of cases. Complete statistics to cover this point are not yet at hand, since no one has tested a sufficient number of normal children for the lower and upper limits of the scale. Further, there are a number of other considerations which must be taken into account, which make the comparison of results of different authors difficult. The chief ones of these are (a) that the exact chronological ages of the children tested have not been given, fractions of a year not being considered, with the exception of Bobertag's results given in the next table, (b) The number of children for any individual test is often very small. From these two facts combined, there is no way of knowing what the exact average ages of the children were whose ages are given as five, six, seven, for example. With large numbers for each of these ages we might reasonably assume that the chance distribution would make the average ages five and a half, six and a half, and seven and a half, but this can obviously not be assumed for such small numbers. In any given case children called six years old may have been practically seven years old. Consequently we would expect from this alone that a frequent variation of a year in the mental ages as determined by the tests in the hands of different authors

with different groups of children would occur. A discrepancy of a year, therefore, can be of no great significance as regards the accuracy of a test. (c) Different authors have not used the same adaptations required for some of the tests for other than French children, and there is also considerable occasion for other variations in the procedure of giving a test and in the interpretation of the child's responses for a number of tests in which this procedure has not yet been sufficiently standardized. We do not yet know definitely what differences these variations in the procedure may produce in the results. The following table shows the lack of agreement of the different authors as to the accuracy of the individual tests.

TABLE IV

	V	VI	VII							
	1 2 3 4	1 2 3 4 5 6 7	1	2	3	4	5	6	7	8
Goddard	c c c c	c D c c c c c	c	c	c	c	c	c	c	c
Terman & Childs	e e e e	c d E E E e e	d	c	c	d	D	e	d	
Bobertag	c	d d D	c	c	d	c	e			
Johnston			d	e	d	c	c			
Binet & Simon	c c c c	d c c d c	d	e	e	c	e			

  

	VIII	IX	X	XI	XII
	1 2 3 4 5 6	1 2 3 4 5 6	1 2 3 4	1 2 3 4 5	1 2 3 4
Goddard	d c c c d e	c e d d c c	e c c c	c d c d c	c e d
Terman & Childs	d D E d e d	c E D D D D	e c c D	D e c D D	D E D
Bobertag	c c c d e c	d	c d d	D d d	d
Johnston	E E d E	d d d d	d	e d D D D	d D D
Binet & Simon	e e c c	e c c d	e e c c e	d d d	D D D

The first horizontal columns give the test numbers, the Roman numerals giving the age groups. The Arabic numbering of the individual tests corresponds to the numbering of the tests in my account of them in this Journal, 1911. A "c" means that a test thus marked is correctly placed in the system according to the percentage of children that passed it, as tested by the author in-

dicated on the left. A small "d" means that the test is too difficult, and should be moved up one year in the scale. A small "e" means that the test is too easy, and should be moved down one year. A capital "D" or "E" means that the corresponding test should be moved two or three years up, or down, respectively. For the blank spaces no figures are given by the authors on the corresponding tests. For the first three these ratings of the tests are based strictly on the figures given by the authors, and do not in all instances agree with the rating the authors give themselves, as they seem to have taken into account other data besides their own statistics on a test. The ratings given for Binet and Simon are taken from their 1911 revision of the scale. This revision is based only in part on the statistics they give. 28 A test is regarded as belonging in the lowest age group in which seventy-five per cent, or more of the children of the corresponding chronological age pass. If, for example, sixty per cent, of five-year-old children pass Test VI I, and seventy-five per cent, or more of six-year-old children pass it the test is regarded as correctly placed in age group VI. With this procedure it happens with a number of tests that wrong results are obtained with the majority of children of a given age, since if between fifty and seventy-five per cent, of the children of a given chronological age pass a test that test is placed in an age group higher than this chronological age. A glance at this table now will show the following. (a) There are but few tests on which all authors agree, nearly all suffering a shift of a year up or down the scale. (b) The greatest disagreement is for tests in the upper part of the scale. (c) The results of Terman and Childs indicate much more frequent and greater errors than do those of any of the others. (d) The results of all taken together show that 2\ of the 44 tests considered should not be shifted by more than one year. Twenty-three should be shifted by two or three years according to one author or another. (e) Excluding the results of Terman and Childs, only 6 out of 35 should be shifted by more than a year, up to the group of XI.

26 Aliss Johnston's original article was not available to the writer. The ratings given here are taken from a table in the Journal of Educational Psychology, 1912. P. 104-5.

These are VI 2, 7, VIII 2, 3, 6, X 4. (f) For 15 of the tests the ratings in each range from too easy to too difficult. These are VI 3, 5, 6, 7, VII 1, 4, 5, VIII 2, 6, IX 1, 2, X 2, XI 1, 2, XII 2. (g) There is substantial agreement on the following: VI 2, VIII 1, IX 6, XI 4, 5, XII 3 are too difficult. VIII 3, IX 2, X 1 are too easy. What general conclusions can we draw from these results? From what has been noted already, we may say in the first place that the very frequent shift of one year only as indicated by these results may be a consequence of the general procedure in not taking account of fractions of a year in the chronological ages and the small number of children for each chronological age, instead of showing that the tests are too easy or too difficult. The results are what we would expect from the procedure. We do not know what Binet and Simon themselves did in regard to this question. If their test ages are all a half year less than they should be, from not having taken account of fractions of a year in the ages of the children tested, and the tests were all correctly placed, the results of these other authors would sometimes show a test too easy by a year and sometimes too difficult by a year. In the rough this is the case, there being 65 d's and 38 e's in the above table. If, on the other hand, the chronological ages of the children tested by Binet and Simon were all just as given, namely, five, six, seven, etc., years instead of five and a half, six and a half, seven and a half, etc., the results of the other authors would frequently show the tests as too easy rather than as too difficult, which, as is seen, is not the case. In this connection it is worth noting further that Bobertag's results show twelve tests too difficult and only two too easy. Bobertag's children were all of a chronological age within two months of the test ages, being an exception to the procedure of the others, and making his children presumably about a half year younger for each age given than were those of the other authors. The second general conclusion indicated by this table is that the smaller differences in the procedure in giving a test and in interpreting the children's responses as followed by different examiners may make a large difference in the results obtained, and that better standardization of the tests in this respect is one of their chief needs. We

are forced to this conclusion except for two other possibilities: (a) The average intelligence of the children tested by one author may have differed from that of the children tested by another author. This is not likely from general considerations, and is also not indicated by the details of the results. It is not likely, for instance, that the younger children tested by Terman and Childs were more intelligent than the younger children tested by others, while the older children tested by them were not. (b) The tests may be much affected by home and school training, and the acquisitions from these sources may have varied from the children of one author to those of another. But again this does not seem likely. On this supposition the results for the two different groups of American children should be more alike than for two groups of children of different nationalities. But the opposite is the case. This lack of agreement of the different authors is very probably due to the examiners, not to the children tested or to the tests. The third conclusion is that several of the tests are seriously misplaced in the system, the results of all substantially agreeing that these are too easy or too difficult.

2. General Observations on Individual Tests. The several authors quoted above have not always followed their own figures in concluding that a test is too easy or too difficult, but apparently have taken other observations into account. These other observations are usually not given, but we may note the instances in which the statistics have not been the sole criterion for passing on the accuracy or inaccuracy of a test. Goddard's figures show that tests VIII 3, (6), IX 2, XI, XII 2, (4) are too easy; that VI 2, VIII 1, 5, IX 3, (4), XI (2, 4), XII 3 are too difficult. Those enclosed in parenthesis he does not name as wrongly placed, and he adds X 2 as too easy.<sup>29</sup> In his revision of the scale so he does not follow either his figures or his recommendations in a few instances. Here he transfers VII 5 to VIII, retains IX 3 in IX, drops X 4b (second series of questions in X 4) for a new test, and retains XII 3 in XII. Bobertag dismisses a number of the tests as poor for various reasons without giving

<sup>29</sup> Ped. Sem., 1911.

<sup>80</sup> Training School, 1911.

his figures that show how the tests actually worked in practice. Some others he speaks of as good or poor apparently without reference to his figures, and concedes at the outset that he does not regard his statistical results as having any great value, because further careful testing might give essentially different results. He had about forty children for each age tested. Thus VII 2 is mentioned as too easy in contradiction to his figures. IX 2 is regarded as worthless, although 75 per cent, of eight-year-old children pass and 97 per cent, of nine-year-old children pass, a good difference for two consecutive ages. IX 6 is mentioned as a particularly good test, while for nine years 60 per cent, pass and for ten years only 78 per cent, pass, a smaller difference between the two consecutive ages than in IX 2. Binet gives only a part of the figures that were used in making the 1911 revision of the scale.<sup>31</sup> Those that are given show some striking differences, as regards the proper place of some tests, from the places given these tests in the revised scale.

We may next summarize briefly the more important comments that have been made on different individual tests which are not based chiefly or at all on any statistical data. This has already been touched upon above in pointing out some tests as illustrations that came under general criticisms of the system as a whole. The most extensive criticisms of this sort come from Decroly and Degand<sup>32</sup> who tested only forty-five children with the 1908 series, but who have made a considerable study of the general problem of measurement of intelligence. They make the following observations :

Too easy—III 1, IV 2, VI 3, 4, 5, VII 3, 6, IX 6, X 4, XI 1, 3.

Too difficult—XIII 1, 2, 3.

Too mechanical—V 4, VII 7, IX 2, X 1.

Affected by training—III 5, IV 1, VI 6, VII 2, 3, 8, VIII 1, 1, 2, 4, 5.

Too dependent on memory—III 2, 3, IV 3, VI 2, VII 5, **XII**

Bobertag's general observations on individual tests indicate

<sup>31</sup> L'Annee Psychologique, 1911.

<sup>32</sup> Archiv de Psychologic 1910.

similar criticisms, but, unfortunately for harmony, they do not affect the same individual tests. Omitting tests regarded as too easy or too difficult, which have already been considered for Bobertag's results, he points out the following:

Affected by training—V 4, VI 6, VII 2, 3, 7, VIII 2, 5, IX 1, X 1, 2.

Too mechanical—VI 5, IX 2.

Affected by chance—VI 7.

Affected by interpretation—VI 4.

Good tests—VI 1, 3, VII 1, 4, 8, VIII 3, 4, IX 6.

In revising the 1908 series Binet and Simon drop two classes of tests. These are:

Too mechanical—VI 6, VII 2, IX 2.

Affected by training—VII 3, VIII 1, 5, 88

A comparison of one author with another on this class of observations shows that there is but little agreement between them as to what criticism applies to what tests. This is particularly true of the criticism that training affects some of the tests. The reader will find it interesting, further, to note the tests any given author has designated as affected by training, and then go over the whole list in the system of tests and pick out additional ones that in his own judgment might logically be added on the basis of those the author in question gives. His disagreement with the given author will illustrate why the authors themselves disagree. As the writer has stated above and elsewhere, estimations of the value of different tests based on this sort of observations can in themselves have but little value. The concrete results, the relative number of normal children of different chronological ages, who pass or fail in a test can alone be the basis for a final decision on the value of a test. It is apparent by this time that even such results may from a variety of causes be misleading. Criticisms coming from other sources are important in pointing the way to questions that need figures to decide them, but beyond this should receive little consideration. It is partic-

<sup>89</sup> The term "Mechanical" is used throughout here as meaning that the test so described may be passed quite independently of the intelligence. The authors quoted do not all use just this term.

ularly unfortunate that in the use of the Binet and Simon tests apparently everyone has disregarded statistical results at times, and has even broken away from his own figures to criticise a test as poor or misplaced in the system.

#### C. Summary of Conclusions.

1. The procedure in obtaining statistics to show the degree of correlation between the chronological ages of normal children and their mental ages as determined by the tests has not taken sufficient account of the exact chronological ages of the children tested, nor adequately eliminated from the supposedly normal children tested those that were below normal or precocious, and has probably lacked necessary uniformity in ways of giving the tests and in interpreting the child's responses. We cannot conclude with certainty from these statistics that any given test is on the whole too easy or too difficult by a year if the discrepancy is not greater than this.

2. The tests for the upper part of the scale give the greatest irregularity in the results obtained by different examiners, and are on the whole probably too difficult. For certain fields of work great accuracy is especially needed in this part of the scale. It needs to be corrected, supplemented and extended at the upper end so as to give us a more reliable means of distinguishing between the normal and nearly normal for these higher chronological ages.

3. The scale as it stands may undoubtedly give frequent errors of a year in the mental ages, and more or less occasionally an error of two years. This degree of accuracy is greater than we can at present obtain in any other way for all but the lower part of the scale, except by prolonged careful observation of the individual child by a skilled observer.

4. A number of the individual tests have been shown, by the substantial agreement of the results of different writers, to be too easy or too difficult for the part of the scale in which they are placed.

5. One of the most immediate needs of the scale is a more thoroughgoing standardization of the tests, both as regards how each individual test is to be given, and how the results are to be

interpreted. This lack of standardization has brought in the personal and varying factor of the examiner, and is probably very largely responsible for the different results obtained by different writers.

6. The question of the effect of training on the value of a test of intelligence is among the most important. No test can probably ever be entirely free from such effect, but there are grounds for believing that it can be eliminated sufficiently for all practical purposes of accurate testing. We cannot determine from a priori considerations alone the degree in which any given test is thus affected.

#### D. Literature.

- AYRES, L. P.: The Binet-Simon Measuring Scale for Intelligence: Some Criticisms and Suggestions. *Psychol. Clinic*, 1911.
- "J. C. B.": Recent Literature on the Binet Tests. *Journ. Educat. Psychol*, 1912.
- BINET A.: A propos de la mesure de l'intelligence. *L'Annee Psychologique*, 1904.
- BINET A. et SIMON T.: Sur la necessite d'etablir un diagnostic scientifique des etats inferieure de l'intelligence. *Ibid.*
- :Methodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *Ibid.*
- :Application de methodes nouvelles au diagnostic du niveau intellectuel chez des enfants normaux et anormaux d'hospice et d'ecole primaire. *Ibid.*
- :Le development de l'intelligence chez les enfants. *Ibid*, 1908.
- :L'intelligence des imbeciles. *Ibid*, 1909.
- :La mesure du development de l'intelligence chez les jeunes enfants. *Bul. de la Societe Libre pour l' Etude Psychologique de l'enfant*, 1911.
- BINET, A.: Nouvelle recherches sur la mesure du niveau intellectuel chez les enfants d'ecole. *L'Annee Psychologique*, 1907.
- BOBERTAG, O.: Ueber Intelligenzpruefungen (nach der Methode von Binet und Simon). *Zeitschr. f. angew. Psychol.*, 1911.
- DECROLY, O. et DEGAND, J.: Les tests de Binet et Simon pour la mesure de l'intelligence. *Archiv de Psychol*, 1907.
- :La mesure de l'intelligence chez des enfants normaux. *Ibid*, 1910.
- DESCOEUDRES, A.: Les tests de Binet et Simon et leur valeur scolaire. *Ibid*, 1911.
- :Exploration de quelques tests d'intelligence chez des enfants anormaux et arrieres. *Ibid.*
- GODDARD, H. H.: The Grading of Backward Children. The De Sanctis

- Tests and the Binet and Simon Tests of Intellectual Capacity. *Training School*, 1908.
- :Binet's Measuring Scale for Intelligence. *Ibid*, 1910.
- :Four Hundred Feeble-Minded Children Classified by the Binet Method. *Journ. Psycho-Asthenics*, 1910.
- : Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence. *Ped. Sem.*, 1911.. (Also brief account in *Training School*, 1911, and in *Proc. V. E. A.*, 1911.)
- HUEY, E. B.: A Syllabus for the Clinical Examination of Children. Lincoln (Illinois) State School and Colony. *School Print*, 1910.
- :Retardation and the Mental Examination of Retarded Children. *Journ. Psycho-Asthenics*, 1910.
- The Binet Scale for Measuring Intelligence and Retardation. *Journ. Educat. Psychol*, 1910.
- :Backward and Feeble-Minded Children. A Clinical Study in the Psychology of Defectives, with a Syllabus for the Clinical Examination and Testing of Children, **Baltimore, 1912.**
- JOHNSTON, K. L.: An English Version of M. Binet's Tests for the Measurement of Intelligence. *Training School Record*, London, 1910.
- :M. Binet's Method for the Measurement of Intelligence—Some Results. *Journ. Exp. Pedagog. and Training College Record*, 1911.
- KUHLMANN, F.: Binet and Simon's System for Measuring the Intelligence of Children. *Journ. Psycho-Asthenics*, 1911.
- :Dr. Ayres' Criticism of the Binet and Simon System for Measuring the Intelligence of Children—A Reply. *Ibid.*
- LAWRENCE, I.: A Study of the Binet Definition Tests. *Psychol. Clinic*, 1911.
- MEUMANN, E.: Der gegewaertige Standt der Methodik der Intelligenzpruefungen (mit besonderer Ruecksicht auf die Kinderpsychologie). *Zeitschr. f. exper., Padagogik*, 1910.
- PYLE, W. H.: A Suggestion for the Improvement and Extension of Mental Tests. *Journ. Educat. Psychol*, 1912.
- TERMAN, L. M.: The Binet-Simon Scale for Measuring Intelligence. *Psychol. Clinic*. 1911.
- TERMAN, L. M. and CHILDS, H. G.: A Tentative Revision of the Binet-Simon Measuring Scale of Intelligence. *Journ. Educat. Psychol*, Feb., March, April, and May, 1912.
- WALLIN, J. E. W.: The New Clinical Psychology and the Psycho-Clinicist. *Ibid*, 1911.
- :Human Efficiency. *Ped. Sem.*, 1911.
- :A Practical Guide for the Administration of the Binet-Simon Scale for Measuring Intelligence. *Psychol. Clinic*, 1911.
- WHIPPLE, G. W.: *Manual of Mental and Physical Tests*. Baltimore, 1910.