

# A TENTATIVE REVISION AND EXTENSION OF THE BINET-SIMON MEASURING SCALE OF INTELLIGENCE.

## PART II. SUPPLEMENTARY TESTS.—CONTINUED.

LEWIS M. TERMAN AND H. G. CHILDS.

*Leland Stanford University.*

### 2. THE COMPLETION TEST.

Ebbinghaus characterizes this method as a real test of intelligence, which, he says, "consists in the elaboration of a whole into its worth and meaning by means of a manysided combination, correction, and completion of numerous kindred associations." Ebbinghaus finds a correlation between this test and intelligence; Wiersma, between capacity in the completion test and native ability; Krüger and Spearman, a correlation of 97 per cent. between capacity in the completion test and the hypothetical "central factor."

One objection to this test is the difficulty in standardizing it. To be suitable for this purpose a selection must have a close logical connection from beginning to end, and the degree of mutilation must be such that the performance is not the result of pure guess work or literary imagination, nor must it be so easy as to require no effort. By experimentation on about two hundred pupils the following form was devised as one which it was hoped would bring out quantitative and qualitative differences. While we used only one text, we would suggest that for a more extended test of the child's ability in this field two or three different tests, varying in sense and degree of mutilation, might well be employed. However, the use of a mutilated text of different degrees of mutilation, such as ours, partially obviates the need of employing a series of texts. The text used was typewritten, with bars (all of same length,  $\frac{3}{4}$ -inch) representing blanks. The three cross bars dividing the test form into four parts were not in the text used, but are inserted here for the convenience of the reader.

## COMPLETION FORM.

One \_\_\_\_\_ the eagle met \_\_\_\_\_ the other \_\_\_\_\_ to see who could \_\_\_\_\_ the highest. They agreed \_\_\_\_\_ he who \_\_\_\_\_ fly the \_\_\_\_\_ should \_\_\_\_\_ called the strongest. All started \_\_\_\_\_ the same \_\_\_\_\_ and flew \_\_\_\_\_ among the cl\_\_\_\_\_. One by \_\_\_\_\_ they \_\_\_\_\_ weary and re\_\_\_\_, but \_\_\_\_\_ eagle \_\_\_\_\_ upward and \_\_\_\_\_ un\_\_\_\_\_ he \_\_\_\_\_ a \_\_\_\_\_ speck \_\_\_\_\_ the heavens. | After \_\_\_\_\_ hours, when \_\_\_\_\_ sure \_\_\_\_\_ the other \_\_\_\_\_ had \_\_\_\_\_ up \_\_\_\_\_ contest, he decided \_\_\_\_\_ return to \_\_\_\_\_ . When he \_\_\_\_\_ back, \_\_\_\_\_ others \_\_\_\_\_ for him; and \_\_\_\_\_ he touched \_\_\_\_\_ , a sparrow \_\_\_\_\_ off \_\_\_\_\_ back, where \_\_\_\_\_ hidden, and \_\_\_\_\_ that \_\_\_\_\_ himself \_\_\_\_\_ strongest \_\_\_\_\_ . | "I \_\_\_\_\_ stronger the \_\_\_\_\_," said \_\_\_\_\_ , "for not \_\_\_\_\_ did I \_\_\_\_\_ high, but \_\_\_\_\_ he \_\_\_\_\_ downward, I came \_\_\_\_\_ hiding \_\_\_\_\_ up \_\_\_\_\_ little \_\_\_\_\_." After \_\_\_\_\_ boastful \_\_\_\_\_, the birds \_\_\_\_\_ heads \_\_\_\_\_ shame, \_\_\_\_\_ council \_\_\_\_\_ matter. | After \_\_\_\_\_ , they decided \_\_\_\_\_ was \_\_\_\_\_ for \_\_\_\_\_ only \_\_\_\_\_ he \_\_\_\_\_ high but \_\_\_\_\_ car\_\_\_\_\_ well.

To \_\_\_\_\_ day \_\_\_\_\_ plumes \_\_\_\_\_ emblems \_\_\_\_\_ str\_\_\_\_\_ cour\_\_\_\_\_.

An examination of the text will reveal the fact that there are four progressive degrees of mutilation. These parts are separated by cross bars in the above reproduction. The first part is mutilated to the extent of  $33\frac{1}{3}\%$ ; the second  $45\%$ ; the third,  $54\%$ ; the fourth,  $66\frac{2}{3}\%$ . There are several advantages in the use of a graduated mutilation test instead of one of uniform difficulty, just as is the case in any other serial scale. It is equally available for children of several different ages and different degrees of ability. It also tends to prevent scattering of effort in spots over the entire paper. In nearly all cases the children filled in the blanks consecutively, with now and then an omission, so far as they were able in the time given. They got their bearings where the mutilation was slight, and this served to round out their later associations into a logical whole, even though the degree of mutilation was great.

*Source of Material.*—This test was given to all the pupils in the Palo Alto schools from the fourth to the eighth grades, inclusive. The method was as follows: E explained to the

class that the test to be set before them would be a story from which many words would be found omitted. They were directed to insert only *one* word in every blank space, or, where necessary, complete an unfinished word, and were asked to have, so far as possible, a well connected story when the filling in had been completed. The test papers were then distributed face down, and at a signal from E the papers were turned and work begun. At the end of fifteen minutes, work stopped and the papers were collected.

*Scoring.*—The usual method of marking such tests, by allowing the same credit for each word correctly inserted, would of course, not be reasonable in this text where the difficulty is not uniform. After the preliminary experimentation with the first graded completion test, the writers decided that to allow credit for each word correctly inserted in proportion to the degree of difficulty of the section in which it came would serve to bring out most clearly characteristic individual differences. The following method was therefore adopted. Each blank in Section I was credited 6 when correctly filled; Section II, 8; Section III, 10; Section IV, 13. The total score for a perfect performance would be 100.2, or approximately 100. One-half of the above credit was given where the words inserted made a well connected story, but were related only in a moderate degree to the thought that should have been given by the printed words. No credit was given for inserted words which did not make sense in their setting, or for words forming a connected story which was purely literary invention, having no connection with the thought given by the printed words. While individual words had to be frequently considered, yet on the whole the inserted words made sense, or the contrary, by series of words or by phrases. Credit was then given to the individual words in proportion as the phrase in which they were placed made sense in its entirety. Thought was considered rather than elegance of diction.

*Results.*—Table VII summarizes the results by years. It should be noted that the table shows for each year the grade reached or surpassed by two-thirds of those tested at that year, as well as median performances by years. Two samples of widely differing performance are also inserted.

TABLE VII.  
COMPLETION TEST.—SUMMARY

Age.	Number Tested.	Median Mark.	P. E.	Two-thirds Pass.
9	32	18.4	8.5	14.9
10	39	29.2	11.8	20.4
11	52	32.2	11.1	25.2
12	56	34.2	11.7	25.6
13	57	45.9	15.2	36.6
14	33	48.5	9.4	42.8

*Sample of an excellent performance by a fourteen-year-old S.  
Grade 100.*

One [day] the eagle met [with] the other [birds] to see who could [fly] the highest. They agreed [that] he who [could] fly the [highest] should [be] called the strongest. All started [at] the same [time] and flew [up] among the cl[ouds]. One by [one] they [grew] weary and re[turned], but the eagle [flew] upward and [upward] un[til] he [became] a [tiny] speck [in] the heavens. | After [two] hours, when [he was] sure [that] the other [birds] had [given] up [the] contest, he decided [to] return to [the earth]. When he [came] back, [the] others [were waiting] for him; and [as] he touched [the ground] a sparrow [flew] off [his] back, where [it had been] hidden, and [said] that [he] himself [was the] strongest [bird]. | “I [am] stronger [than] the [eagle].” said [the sparrow]. “for not [only] did I [fly as] high, but [when] he [flew on] downward, I came [from my] hiding [place and flew] up [a] little [way].” After [this] boastful [story] the birds [dropped their] heads [with] shame, [and they held] council [to decide the] matter. | After [having the council] they decided [that the eagle] was [the stronger bird], for [not] only [did] he [fly so] high, but [he] car[ried the sparrow as] well.

To [this] day [his] plumes [are worn as the] emblems [of] str[ength and] cour[age].

*Sample of poor result and badly scattered effort by a twelve-year-old S., who tested much below normal.  
Grade 17.9.*

One [day] the eagle met [crane] the other [was] to see who could [be] the highest. They agreed [that] he who [could] fly the [other] should [not] called the strongest. All started [at] the same [time] and flew [through] among the cl—, One by [one] they [went] weary and re[ach], but [the] eagle [went] upward and [crane] un[der] he [touch] a [little] speck [of] the heavens. | After [the] hours, when [the others] sure [and] the other [who] had [goi] up [to] contest, he decided [he] return to [back again]. When he [fell] back, [the] others [ran up] for him; and [when] he touched [the ground] a sparrow [fell] off [his] back, where [he was lying] hidden, and [hid] that [he] himself [was not] strongest [all]. | “I [though] stronger [than] the [others].” said [the stranger]. “for

not [as] did I [but to] high, but [when saw me] downward, I came [up from] hiding [and he look] up [and] little [while]." After [while] boastful [were] the birds [with great] heads \_\_\_\_\_ shame, [to join great] council [in the was] matter. After [while when they] they decided [about where there] was [a] \_\_\_\_\_ for \_\_\_\_\_ only \_\_\_\_\_ he \_\_\_\_\_ high but \_\_\_\_\_ car \_\_\_\_\_ well.

To \_\_\_\_\_ day \_\_\_\_\_ plumes \_\_\_\_\_ \_\_\_\_\_ \_\_\_\_\_ emblems \_\_\_\_\_ str \_\_\_\_\_ cour \_\_\_\_\_.

The results show clear differences in performance from year to year, except in the 12th year, which showed hardly any gain over the 11th. This test, therefore, fulfills the most important requirement for use in a measuring scale of mentality. We believe that it brings to light fundamental differences in the thought processes, and affording as it does a more extended performance than most of the short Binet tests, it seems to us of corresponding greater value than most of the latter. For clinical purposes, however, it will have to be determined how and to what extent an S's performance is influenced by taking the test with a group instead of individually.

### 3. BALL AND FIELD TEST OF PRACTICAL JUDGMENT.

Another test introduced for a try-out was the following: E draws a circle about three inches in diameter on a sheet of paper, leaving a small gap which is called a gate. S is then told that this circle represents a round field, that in this field a base-ball has been lost, that we have no idea what part of the field the ball is in, only we know that it is somewhere in the field. S is then asked to take a pencil and mark out a path showing what direction he would walk in hunting the ball. To make clearer what is wanted, E takes the pencil and marks the path as far as the gate, then gives the pencil to S. Before S begins, however, he is further informed that the field is covered with grass about six inches high, so that the ground cannot be seen clearly farther than ten feet on each side of him. E then adds "What direction will you go so as to be sure not to miss the ball?"

An examination of the results showed that four grades of performance could very readily be distinguished. The younger children gave no evidence of any rational plan of procedure, further than to fill up the entire field with marks. Their lines

were not parallel, were frequently crossed and re-crossed, and often were broken. By "broken" it is meant that the pencil was lifted and set down again at another part of the field. This general absence of plan was scored 1. Somewhat older children showed, as a rule, some degree of plan. If the lines were partly but not entirely parallel, only very slightly broken and seldom crossed, or if there was some other poorly adopted procedure, such as lines radiating from the center like spokes, the score 2 was given. If the lines were almost perfectly parallel (whether spiral or straight), not at all crossed or broken, the score was 3. Score 4 was given when in addition to the performance for score 3, S showed that he had taken into consideration the fact that the ground was visible only 10 feet on either side of him, and that the lines could be 20 feet apart, except around the border where the path had to approach within 10 feet of the fence. It was easy to ascertain by a question or two whether this had been taken into account. Figure III gives reproductions of typical reactions to this test.

The ball and field test was not introduced until more than half the tests had been completed. In all, 113 children were given this problem, with the results shown in the following table.

TABLE VIII

Age.	Number of Cases	Score.			
		1	2	3	4
4	2	2	0	0	0
5	3	3	0	0	0
6	3	3	0	0	0
7	4	2	2	0	0
8	11	1	7	3	0
9	26	3	11	12	0
10	17	1	3	13	0
11	21	2	3	16	0
12	17	1	5	10	1
13	5	0	1	3	1
14	4	0	0	2	2

Score 2 may, therefore, be set provisionally as a standard performance for the eighth year, score 3 for the tenth year, and score 4 somewhere above the thirteenth year.

On the whole this is a test which seems worthy of a much more thorough trial than it has had in our series. It is easily and quickly given, attracts the interest of the S, and gives a performance which lends itself to standardization. A

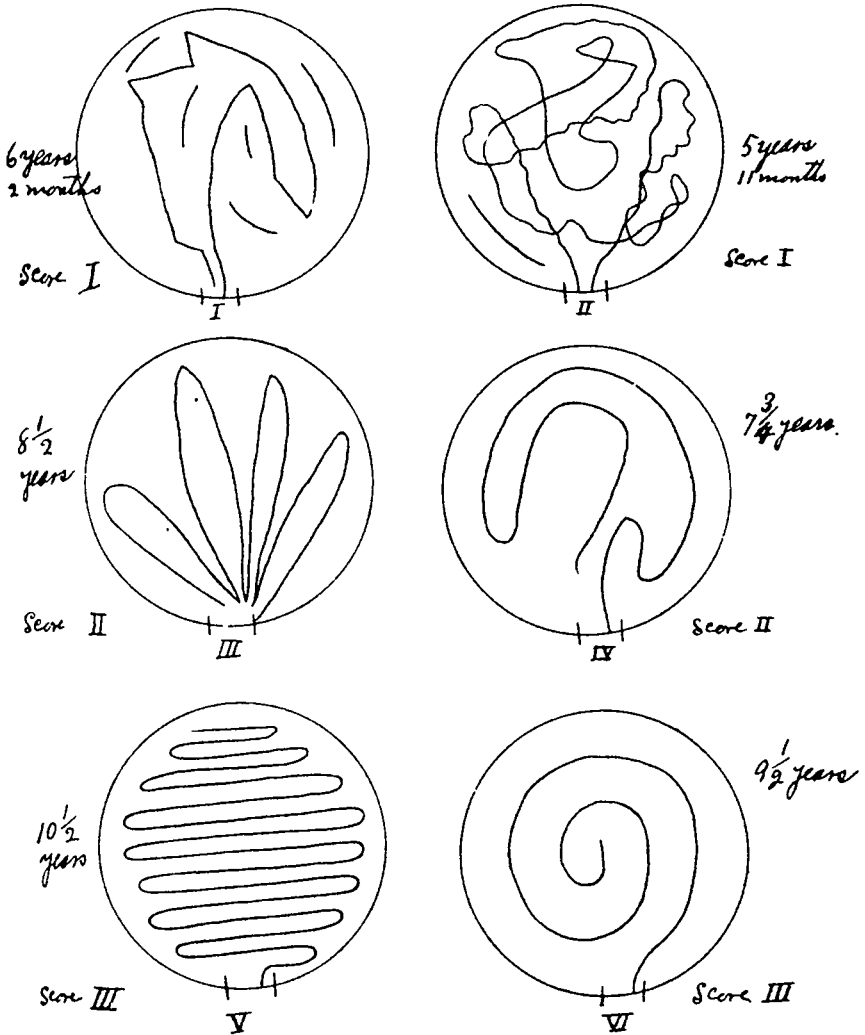


FIGURE III.

few tests of this general type are much needed to round out the Binet scale, but unfortunately unobjectionable tests of this kind are hard to devise. The problem set may be too easy or too difficult, it may permit of a mechanical solution, or it may give too large a play to the element of chance and thus approximate the puzzle. The ball and field test meets these requirements and offers striking development in performance with increasing age. For example, plans II, IV and VI of the illustrations above reproduced are the performances of three sisters, aged 5-11, 7-9 and 9-6, respectively.

A somewhat more difficult form of this same test is now being used with older children and adults.

#### 4. THE VOCABULARY TEST.

The experiments of Kirkpatrick and Whipple show that it is possible to ascertain with serviceable accuracy the size of an individual's vocabulary by means of a test of 100 words properly selected. Considering the facility with which such a test may be applied, the relative ease of establishing age norms, and the suggestiveness for clinical purposes of the quality as well as the quantity of definitions given, it would seem that some kind of vocabulary test ought to be devised for use in each year. Its application requires only from five to fifteen minutes, and its results are much more significant than the scattering-definition tests used by Binet in years 6, 9 and 13 of his scale. If the vocabulary test is given, the Binet definition tests may be omitted.

It has seemed to the writers that a vocabulary test to be used upon children and as a part of a measuring scale of intelligence should be based in so far as possible upon the mass of familiar words employed as the vehicles of the most common and therefore presumably the most fundamental concepts in general use. The more it is desired to measure intelligence rather than the effects of training, the more necessary it becomes to make the selection from a representative list of fundamental terms instead of from a complete list of language units. The more complete the list from which the selection is made, the more it tends to reveal accidents of training rather than real intelligence. For our present purpose, therefore, the complete or un-



abridged dictionary is not to be considered. Kirkpatrick's test is based upon Webster's Abridged Dictionary, containing 28,000 words. Even this has a very large proportion of words the interpretation of which depends upon the technical training, as is indicated by the fact that Kirkpatrick's test gives even for college students an average vocabulary index of only about 70%. In other words, more than one-fourth of the words are so technical that highly selected young men and women of this extended academic training can not define them. Casting about for a fairer test of intelligence the writers, after some preliminary trials, have arranged a list of 100 words selected from Laird and Lee's Vest-Pocket Webster Dictionary, 1904 edition. This dictionary meets the above specified requirements more nearly perhaps than any other, containing as it does about 18,000<sup>1</sup> of the most basic words of the language. Taking the last word of every sixth column gave the following representative list of 100 words used in our test:

## VOCABULARY TEST.

afloat	frustrate	philanthropy	achromatic
apish	gelatinous	plumbing	amberggris
artless	gown	pork	cameo
avarice	guitar	priceless	casuistry
haste	harpy	promontory	complot
bewail	herd	puddle	declivity
bonfire	hysterics	quake	exaltation
brunette	impolite	ramble	fen
charter	insure	reception	hookah
civilly	irony	regard	incrustation
coinage	juggler	reposing	infuse
conscientious	lecturer	roarer	laity
copper	lotus	rule	limpet
crunch	majesty	scorch	ocher
curse	Mars	shrewd	paleology
depredation	mellow	skill	pärterre
dilapidated	milksoap	snip	perfunctory
disproportionate	misuse	southern	piscatorial
drabble	mosaic	sportive	precipitancy
easterly	muzzle	stave	retroactive
embody	nerve	straw	sapient
envelope	noticeable	swaddle	selectman
eye-lash	orange	tap	shagreen
flaunt	outward	tolerate	sudorific
forfeit	peculiarity	treasury	theosophy

<sup>1</sup>The publishers incorrectly state the number as 30,000.

As already stated, all the vocabulary tests were given individually. The child looked at the word, heard it pronounced, and then gave its meaning orally. E did not write down the definitions (only in case of exceptional or peculiar answers), but scored them as given. We believe, however, that a *qualitative* analysis of the verbatim definitions of children of different ages and mentality would be of great value.

In scoring, full credit was allowed for one correct meaning given (regardless of whether that meaning was the most common one), and half credit for a definition which was partially correct. No value was attached to the logical form employed in the definition, since the test is meant to explore the range of ideas rather than the evolution of thought forms. When it was clear that the child had one correct meaning for a word he was given full credit for it, however poorly the definition may have been stated. The difficulty comes in deciding when a meaning is "correct," since definitions may be of all grades of excellence. Individual differences in E's will inevitably appear here, and in order to minimize them it may be necessary ultimately to indicate definitely for each word what definitions are acceptable, what deserve half credit, what none.

Whatever standard of excellence is accepted it is bound to be more or less arbitrary. Almost every word has several different usages, to say nothing of the innumerable delicate nuances of meaning due to differences in contextual setting. Besides, our own stores of meanings are never garnered as unchanging units of intellectual riches, but instead are constantly undergoing transformation, substitution, amplification and revision. To ascertain that one S has half the vocabulary that another has, as measured by this test, may be of value in estimating their difference in ability, but it does not by any means fully represent the extent to which they differ in language mastery.

On the whole, leniency is to be commended in judging the definition for the reason that the child's power of expression runs farther behind his understanding than is true of adults, and also because for the young S the word has probably a relatively less unitary existence. We are all at a disadvantage in defining isolated words, the child doubly so. To give an idea of the standard of perfection employed in the tests

the following illustration will be of service: Afloat, "a ship floats on the water (full credit); civilly, "its when you treat a person nice" (full credit); hysterics, "you act funny or crazy" (full credit); majesty, "what you say in speaking to a king" (full credit); copper, "something you make money out of" (full credit); sportive, "to like sports" (half credit); pork, "meat" (half credit only unless kind of meat can be specified). It is seen from this that a very liberal standard has been employed. Questioning for the sake of drawing out meanings was not resorted to except in rare instances to overcome the child's timidity. Usually the responses were quickly made and not infrequently the most absurdly incorrect definitions were struck off with amazing confidence.

The vocabulary list was not devised until a large part of the testing had been completed, hence we have data for only 161 cases, ranging from 5 to 13 years of age. The following table shows the main results:

TABLE IX.

*The Growth of Vocabulary as Indicated by the New Vocabulary Scale Applied Individually and Orally.<sup>1</sup>*

Age.	Median Age.	Median Test Age.	Number of Pupils	Median Vocabulary.	Vocabulary Reached by Two-thirds.
6	6.5	7.5	5	2500	2300
7	7.5	8.3	14	2600	2300
8	8.5	9+	28	3000	3000
9	9.5	10—	35	5000	4000
10	10.58	10+	24	6000	4500
11	11.5	11	29	6100	5500
12	12.4	11.5	19	7700	6500
13	13	11.5	7	8800	7400

<sup>1</sup>It is not at all certain, of course, how written and mass tests would compare with these results, though presumably they would give a somewhat lower index.

With the exception of year 6, where the cases are too few to afford even a tentative norm, it is seen that the vocabulary index shows a fairly steady growth. For further experimentation the point reached or passed by two-thirds of the children of any year will be tentatively assumed as the test norm for that year. Our test, as may be seen by comparison, gives a larger percentage index than Kirkpatrick's, because it is composed of words on the whole somewhat more common; but it necessarily gives on the average a smaller absolute vocabulary since it is based on 18,000 instead of 28,000 words.

(Concluded in the May number.)