MENTAL TESTS AND THEIR USES

CHESTER E. KELLOGG

Professor of Psychology and Education, Acadia University, Nova Scotia

PERHAPS the earliest "mental test" on record is to be found in the Book of Judges. Gideon used a typical situation to secure samples of the behaviour tendencies, or "character", of his men:—

> Because of the Midianites the children of Israel made them the dens which are in the mountains, and caves, and strongholds. And so it was, when Israel had sown, that the Midianites came up and the Amalekites, and the children of the east, even they came up against them. And they encamped against them, and destroyed the increase of the earth, till thou come unto Gaza; and left no sustenance for Israel, neither sheep, nor ox, nor ass. For they came up with their cattle, and their tents, and they came as grasshoppers for multitude.....and they entered into the land to destroy it.....

At last Gideon, the son of Joash, in the village of Ophar, became convinced of his duty to lead a revolt, and built an altar upon which he sacrificed to Jehovah as a challenge to the enemy. After some difficulties with the townspeople, who were fearful of the consequences, the matter was reported to the enemy. The Midianites and Amalekites gathered in the valley of Jezreel. Gideon sent messengers to call the men of Manasseh, Asher, Zebulon, and Naphtali, and they "came to meet them".

Evidently Gideon was not over-confident of the quality of this army, coming from a long oppressed people. So he decided to win by strategy, with a few picked men. First the command was given:—

> Whosoever is fearful and afraid, let him return and depart early from Mount Gilead.

There returned of the people twenty and two thousand, and there remained ten thousand. From these ten thousand brave men Gideon wished to select the most cautious and wary:—

> So he brought down the people into the water; and the Lord said unto Gideon, Every one that lappeth of the water with his

> tongue, as a dog lappeth, him shalt thou set by himself; likewise every one that boweth down upon his knees to drink. And the number of them that lapped, putting their hand to their mouth, were three hundred men; but all the rest of the people bowed down upon their knees to drink water.

The three hundred could be trusted to obey orders. So, when they came into position about the enemy camp,

> The three companies blew the trumpets, and brake the pitchers, and held the lamps in their left hands and the trumpets in their right hands to blow withal; and they cried, The sword of the Lord and of Gideon. *And they stood every man in his place round about the camp*: and all the host ran, and cried, and fled.

Here was a mental test. Such tests are of practical value because people differ in ways which it is important for society to detect. One such variation is colour-blindness, which was first brought to the attention of scientists by John Dalton, the great chemist, in 1798. Dalton was a Quaker, but made no objection to wearing the scarlet gown of a Doctor of Laws, because—as he said—"To me its colour is that of nature, the colour of those green leaves." "In lecturing on optics", he tells us, "I got six ribands—blue, pink, lilac—and red, green, and brown,—which matched very well, and told the curious audience so. One gentleman came up immediately afterwards and told me he perfectly agreed with me. He had not remarked the difference by candlelight".

As the common forms of colour blindness, occurring in about 4% of men and somewhat over 1% of women, involve inability to discriminate between red and green, and as these colours are used extensively for signalling in railroad and marine service, tests for ready detection of the defect are of great value. The need has been met by Holgren's worsted tests and Nagel's card tests. Many other tests for sensory capacity and other specific abilities have been invented.[1]

In 1796, the astronomer in charge of the observatory at Greenwich, England, found himself compelled to discharge an assistant who had got into the habit of recording transits almost a second too late. For the benefit of those not familiar with the problem, a brief explanation will be given. With the telescope in a fixed position it is desired to know the exact instant of time when some star crosses the central hair. When the star is nearing the field, a seconds-pendulum is started, and the observer counts the beats, the exact second of starting having been recorded. If, then, the star appears in position *a* at, say, the 20th beat, and in position

1. See, e. g., Whipple's *Manual of Mental and Physical Tests.*

b at the 21st, the fraction is estimated and the total counted is added to the time recorded. The sources of error in this type of observation are now fairly well understood, as a result of extensive experimental studies of reaction-time, etc. But it has not been found possible to eliminate errors by selection of observers. Hence objective means of recording are now used as far as possible. The Greenwich case is important in the history of mental tests because its discussion and that of other problems in scientific measurement led to the development of the general theory of error, upon which rests the use of averages, mean variation, standard deviation, etc. The theory assumes that in the long run errors in one direction will balance those in the other, and that large errors will not occur so frequently as small ones, giving the law, first stated by Adrain, which is represented by the familiar probability integral and curve. Many physical and mental traits vary in accordance with this law, so that it is reasonable in various connections to speak of the average or normal man.

Towards the end of the past century Francis Galton, perhaps because of his close relationship to Charles Darwin, became interested in the problem of the inheritance of mental ability, especially superior ability, and conducted researches which founded the study of eugenics. In tracing the inheritance of eminent ability Galton did not require any exact methods of measurement. He did, however, in the study of stature, etc., contribute much towards the development of exact statistical methods, and in order to study the power of visualization he sent out a *questionnaire* asking each of his correspondents to record the degree of his ability by reference to a carefully stated list of descriptions of imagery of varying degrees of excellence, the subject-matter being the breakfast table.[1] The method has since been widely imitated and extended to other sense fields.

Galton's imagery scale is an improvement on the familiar rating method, such as is used by teachers in grading the work of pupils as Excellent, Good, Fair, Poor, Unsatisfactory, or A, B, C, D, E. It is now generally admitted that such ratings, when made without objective standards, are highly inaccurate. We owe to Walter Dill Scott a method of rating, for personnel administration, which at least partially obviates this difficulty. In order to make a scale for any trait,—leadership, courage, intelligence, adaptability, mechanical ability, etc.— the names of a considerable number of persons known to the operator are set down. These are then rearranged as well as may be in order of excellence for the special

1. See Galton's *Inquiries into Human Faculty.*

trait in question. The names at the head and foot of the list, the one in the middle, and two intermediate between the latter and the two extremes, are set down as the scale. Persons to be rated in the special trait are thereafter compared with those listed as the scale. This method was used in rating officers in the U. S. Army in the recent war, and is now being used in many industries. Its success is of course conditional upon the ability of the operator to classify his original list accurately.

About 1900 the authorities of Paris, finding that the laggards in the public schools were becoming a serious problem, requested the great psychologist, Alfred Binet, to devise methods for detecting the feeble-minded, in order that they might discover to what extent the backwardness was caused by real inability to learn. The result was the well-known Binet-Simon Intelligence Scale. Binet's method rests upon the accepted fact that intelligence increases with age up to the adult level. Adult intellectual ability is not merely of a different sort (though there may be some real qualitative differences), but in general includes the ability of the earlier years. Binet did not work from any narrow definition of intelligence, but simply devised many little tests that would serve to take samples of the children's ability in various types of intellectual activity,—orientation, memory, constructive imagination, control of numbers, understanding of questions, of abstract terms, of pictures, etc. The tests were graded by careful experiment, and the best selected for each stage of development; e. g. in his final revision, 1911, the tests for age 3 are: –1. Points to nose, eyes, and mouth; 2. Repeats two digits; 3. Enumerates objects in a picture; 4. Gives family name; 5. Repeats a sentence of six syllables. For age 6: 1. Distinguishes between morning and afternoon; 2. Defines familiar words in terms of use; 3. Copies a diamond; 4. Counts thirteen pennies; 5. Distinguishes pictures of ugly and pretty faces, etc. The specific materials to be used, and the methods, are exactly prescribed. The results are stated in terms of mental age; i. e., if a child passes the tests which are passed by the average 8-year old child, his mental age is 8, etc. The scale has been repeatedly translated and revised for use in other countries.

The best adaptation in English is the Stanford revision, published by Lewis M. Terman in *The Measurement of Intelligence*, 1916. In addition to very extensive revision, Terman has extended the technique of the method by his study of intelligence quotients. The intelligence quotient (or I Q) is the ratio of mental age to chronological age. Terman has shown that the I Q tends to remain practically constant from year to year.

One case may be cited as an example:—

> H. S. boy, age 11; mental age 8-3; IQ approximately 75. At 8 years tested at 6. Parents highly educated, father a scholar. Brother and sister of very superior intelligence. Started to school at 7, but was withdrawn because of lack of progress. Started again at 8 and is now doing poor work in the second grade. Weakly and nervous. Painfully aware of his inability to learn. During the test keeps saying, "I tried anyway." "It's all I can do if I try my best, ain't it?" etc. Regarded defective by other children. Will probably never be able to do work beyond the fourth or fifth grade, and is not likely to develope above the 11-year level, if as high.

This means that it is possible to tell with fair accuracy how far above or below the average a person's intellect will be at maturity, by testing him in childhood. To the extent that vocational success depends upon intelligence, vocational guidance is thus possible at an early age, before vocational training need be begun.

By the Binet-Simon method and its revisions, examinations have to be conducted individually, the responses called for being mostly oral. For a complete examination it is required that the subject pass all the tests of some one year. Testing must then be continued up the scale until a year is reached in which there are no successes. This can rarely be done in less than half an hour, and often requires an hour or more. Though no special psychological knowledge is required for routine work, the tests cannot be conducted or scored properly without considerable practice, so the method is too expensive for very general use. As a result, group tests have been devised. In such trials, tests are somehow presented to a whole group of persons at once, either orally or in writing, or both, and their responses recorded in a form that will permit later scoring. Tests of performance, as, e. g., picture puzzles, may be conducted by the group method, but group tests are usually by pencil-and-paper. Complete written answers to questions are very difficult to score fairly. Also, facility in writing influences the scores attained. This difficulty was first met by Arthur S. Otis, who prepared tests requiring the underlining of the proper one of several words or diagrams, the insertion of numbers or letters in a space provided, etc., all responses calling for very little pencil work, and readily scored by means of a transparent stencil.[1] Otis's technique and a large part of his actual test materials were incorporated in the tests prepared in 1917 by members of the American

1. Cf. *The Otis Intelligence Examinations.*

Psychological Association for use in classifying men drafted for the U. S. Army.[1]

Many other group tests have since been completed for use in the public schools, in colleges, and in industry. These can be easily conducted and scored by any fairly intelligent person. The only requirement is willingness to exercise care in reading directions, keeping to the prescribed time-limits, etc. Extensive use of mental tests is now easy, and a good beginning has been made.

Mental tests are not, however, coming into general use without opposition. Every psychological examiner has met people who profess the ability to judge the intelligence of others without any assistance from tests. There is, of course, an element of truth in this claim. We may grant that it is quite possible for most of us to classify as intelligent or dull or just ordinary those with whom we are well acquainted. Indeed, as will be shown presently, mental tests could not be devised if this were not true. But even here it is easy to make mistakes. Teachers often over-estimate the intelligence of laggards, comparing them with others in their classes rather than with the classes in which they would be if normal. Last winter, a girl examined by the writer was found to have a mental age of 7 years 4 months. The teachers in the special school which she was attending thought her not much below average, as she was doing fairly well with first-grade work. But she was fourteen years old, so practically an imbecile. They had compared her achievements with those of the normal first-grade child, only half her age. Even when such factors do not enter, only a very rough placing of the individual upon the scale of intelligence can be expected. Accurate measurement of general intelligence can result only from the use of tests, just as—in school and college—examinations are required to assess the results of study.

We have admitted that it is possible to estimate roughly the intelligence of one's acquaintances. Some persons profess much more than this, namely, the ability to detect intelligence and to read character in general from the physical appearance, without previous acquaintance. To a certain extent, even this claim is justified. At any rate, we commonly respond favourably or otherwise to the appearance of those we meet, and while many, probably most of us, find it necessary frequently to revise such first impressions, a certain minority are more fortunate, i. e., if we may accept their own statements at face value. Some of them have set down systems based upon complexion, shape of the features, etc. But the systems have not as yet been developed by scientific methods,

1. Cf. *Army Mental Tests*, by Clarence S. Yoakum and Robert M. Yerkes.

and do not succeed when tried by others. It is evident that the authors do not actually know the basis for their own judgments. As Woodworth says, "No good judge of character really goes by the shape of the face; he goes by little behavior signs which he has not analyzed out, and therefore cannot explain to another person." [1]

It is the accepted custom, when filling positions calling for any special skill or responsibility, to require letters of application, photographs, and letters of recommendation, and to interview personally those applicants thought to be most promising. But the ability to write a good letter of application may not necessarily imply the type of ability desired, and a letter of recommendation is not of much value unless the writer's standards are known or some fairly objective method of rating is used so that one recommendation can be compared with another. Rudolph Pintner has shown that ability cannot be accurately judged from photographs.[2] Photographs do not offer those 'behavior signs' which seem to be essential in judging personality. Final selection usually depends upon a personal interview in which the questions asked may or may not throw real light upon the candidate's ability and character. With experience, any employment manager tends to use a stock set of questions. Whether these are significant or not cannot be told unless the replies are recorded and checked up with degrees of success in employment.

Mental tests are valid measures because they are standardized. The tests for the Binet-Simon Scale and its revisions have been selected by a long process of trial and error, in order that when a large number of unselected children of any given age are tested the mental age scores resulting may cluster about the given chronological age. In general, a test is considered a satisfactory measure if those persons already known to rank high in the trait in question score high in the test, while those ranking low score low. The more consistent are the results, the better the test. We are thus, as remarked above, dependent for the standardization of any test upon judgments which are based upon long experience, for no other criterion is available.

The statistical method used for this purpose is some one of the forms of correlation. As an illustration which may be of some service to the reader, one of the simplest forms will be stated. In the theory of chance, certainty is represented by 1, complete uncertainty by O. In developing a formula for correlation, the same idea is used. Perfect correspondence of one trait with another, or test result with trait measured, is to be represented by the

1. *Psychology*, p. 446.

2. See *The Psychological Review* for 1918.

value 1 from the formula; entire lack of relationship by O; also, inverse relationship (as might occur when errors instead of successes in a test are counted) will run from 0 to –1, the latter denoting perfect negative correlation. Now, suppose we plot cases on a diagram to indicate the corresponding values in two traits, or in a test and a trait. Let O –X represent the values of one variable, O –Y those of the other. Then it is evident that if the relationship is positive and very close, the cases will cluster about the diagonal from O. If the relation is negative, the cases high in trait Y being low in trait X, and *vice versa*, the cases will cluster about a line XY; while if there is no relationship, the cases will be scattered over the diagram. Now draw the lines MY and MX, so that an equal number of Y's will fall above and below MY, and an equal number of X's to right and left of MX. Then if the relation is positive, most of the cases will fall in the squares marked + + and — —, i. e., with like signs, and few in those marked with unlike signs. It is evident that the smaller the percentage of unlike signs, the higher the correlation. We desire a formula that will give 1 when the percentage of unlike signs is 0, 0 when the percentage is 50, and –1 when the percentage is 100. Such a formula is the following:—

R (the correlation or relationship) = $\cos \frac{U}{U+L} \pi$, since $\cos o° = 1$, $\cos 90° = 0$, and $\cos \pi$, or $180° = -1$. This formula is not exact, but is useful to indicate the trend of the results as a preliminary to more thorough treatment. Only tests giving a high correlation are worth considering for practical use.

Last spring Dr. G. B. Cutten and the writer conducted tests in the public schools of Wolfville, N. S. The group tests used were Otis's Primary Intelligence Examination, Form A, in the Kindergarten and Grades I to III; the National Intelligence Tests, Scale A, Form 1, in Grades IV to VI; and Terman's Group Test in Grades VII to XI. The mental ages were determined from the published forms, and IQ'S calculated. The IQ's were tabulated in accordance with Terman's classification:

IQ	Classification
Above 140	"Near" genius or genius
120-140	Very superior intelligence
110-120	Superior intelligence
90-110	Normal, or average, intelligence
80-90	Dullness, rarely classifiable as feeble-mindedness
70-80	Border-line deficiency, sometimes classifiable as fullness, often as feeble-mindedness.
Below 70	Definite feeble-mindedness

The tabulation follows:—

Grade	"Near" Genius 140–	Very Superior 120–139	Superior 110–119	Normal 90–109	Dull Normal 80–89	Border Line 70–79	Mentally Deficient –69	Total
XI	..	..	2	11	2	..	..	15
X	..	3	6	9	..	..	..	18
IX	..	..	2	11	6	2	1	22
VIII	..	9	8	8	4	..	..	29
VII	1	9	5	8	1	..	..	24
VI	..	5	4	6	5	..	1	21
V	1	3	3	20	4	6	2	39
IV	1	4	2	13	6	5	2	33
III	..	6	2	24	3	3	3	41
II	3	11	6	16	4	1	2	43
I	..	5	5	15	4	7	7	43
Totals	6	55	45	141	39	24	18	328

(Tests in Kindergarten were not satisfactory and are not listed)

The results were checked by individual examination of a considerable number, of both superior and inferior, with the Stanford-Binet. No very great changes resulted, so the above table may be considered as substantially correct. If these results are found to be representative of the Province as a whole, they are of very great significance, for they indicate 6% or more of mental deficiency, as against 1 to 2% stated by most authorities, and, on the other hand, a marked excess of highly superior ability. A few cases may be cited to illustrate the situation.

CASE 1. Age 6-11, Mental age 9-6, IQ 138. Doing very good work in first grade. With a little coaching, could be promoted to the third grade instead of the second.

CASE 2. Age 10-1; M. A. 6-3; IQ 62. In first grade, doing almost nothing. Is of degenerate stock.

CASE 3. Age 11-10; M. A. 6-7; IQ 56. Doing rather poor work in second grade.

CASE 4. Age 15-1; M. A. 9-6; IQ 63. Very poor work in fifth grade. Barely capable of promotion to the fourth grade.

CASE 5. Age 8-9; M. A. 11-7; IQ 132. Excellent work in second grade. Mentally qualified for the fifth grade. Could very easily skip the third grade.

CASE 6. Age 7-7; M. A. 11-6m IQ 152. Very good work in second grade. Should certainly be given special opportunity for advancement.

CASE 7. Age 9-4; M. A. 15-4; IQ 164. Near head of fourth grade. Could easily do work of eighth grade.

Certainly it is important that children who are definitely feeble-minded, and can therefore never go beyond the level of common unskilled labour, should not compete in the same class-room with those who will easily make brilliant records in college and in the professions. Special classes for the deficient have been established in many large communities, offering them a reduced amount of the usual studies, in order that their helplessness may not be increased by discouragement over failure for which they are in no way responsible, and also giving them such vocational training as is possible.

The superior children, though far more important to society, have usually been allowed to drift along in the ordinary curriculum. The subject-matter of the traditional curriculum is indeed suited to them, and—in the upper grades—to them only. But their progress in such subjects as Latin, algebra, and geometry has been delayed by the presence in the same classes of a majority of average children to whom these subjects mean little or nothing. This majority does not from the old curriculum get the broad training it should have for citizenship and business life. Enrichment of the curriculum would probably be better for the superior children than extremely rapid promotion. Opportunity classes for such children are not now unknown. For example, such a class is to be established by the Boston, Mass. Y. M. C. A. Our modern graded schools are much better off in physical equipment, offer many social advantages, and probably on the whole more efficient teaching, but have lost the flexibility characteristic of the old district school. The awkward administrative problems resulting from failure in one or two subjects are painfully familiar to all educational administrators, and in various places efforts have been made to effect a compromise between the new and the old type of school. In all schools large enough to warrant subdivision of the grades, the problem can be reduced to very small proportions by selecting pupils for the divisions by means of mental tests, and adjusting the curriculum requirements accordingly. In the smaller towns a special class for the superior, and one for the inferior, could be established, to be conducted as one-room schools.

Such adjustments do not, of course, offer any permanent solution of the problem of the feeble-minded. How serious this problem is, is quite evident from the reports of the Superintendents of Ne-

glected and Delinquent Children for the various Provinces, and the recent Mental Hygiene Survey of Nova Scotia. Such families as the Jukes, the Kallikaks, and the Hill Folk have been made familiar to students of social problems. Readers of this journal will recall a recent interesting study of such a family.[1] But it is not so well-known how commonly they occur. A survey conducted last year for the public schools of Wolfville by Miss Edith M. White, In-Instructor in Social Service at Acadia University, revealed the presence of a number of families of the same low level within the school district. These have intermarried frequently, and account for a large part of the feeble-minded in the district. Some of these matings have been traced back two or three generations. For obvious reasons it is not permissible to give details, but one inferior individual, no longer living, is known to have over ninety descendents now living within the school district of Wolfville. Few of these can be considered of any value to the community. Many are known to be definitely feeble-minded. A large part of the business of the local courts and of the Children's Aid Society is due to them. The known matings, leaving out the frequent cases of illegitimacy, include twelve with other degenerate stock. As no attempt is being made to segregate the feeble-minded, the prospects for matching the cases already famous are at least fair.

Adoption of intelligence tests as part of the routine of the public school system would be of great value. The group tests could easily be conducted by the school principals. Individual examinations should not be attempted by anyone not thoroughly familiar with the methods; but in many communities a physician interested in mental hygiene might be found who would be willing to make the effort necessary to acquire skill in using tests. The results would not only be directly serviceable in the administration of the schools, but would quickly indicate the status of the population at large. General testing of adults would not be practicable, but, since the IQ remains practically constant into adult life, the school children could be considered a representative sample. Public policy could then be formulated in accordance with the facts shown.

1. Maud A. Merrill, *Feeble-Mindedness and Crime.* Dalhousie Review, Vol. 1. No. 4.