Columbia University

Contributions to Education

Teachers College Series

SIGILLVM · COLLEGII · COLVMBIAE · NOVI · EBORACI

[redacted]

[redacted]

Non-Verbal Intelligence Tests for Use in China

By

HERMAN CHAN-EN LIU, PH.D.

Teachers College, Columbia University
Contributions to Education, No. 126

Published by
Teachers College, Columbia University
New York City
1922

Ss

[illegible]

TO MY FRIENDS
EMILIE BRETTHAUER
JAMES H. FRANKLIN
ANDREW MACLEISH

ACKNOWLEDGMENTS

Grateful thanks and appreciation are hereby expressed to Professor Edward Lee Thorndike, of Teachers College, Columbia University, under whose almost daily guidance and inspiration this study has been carried out; to Miss Margaret P. Rae, principal of New York Public School No. 108, for her ready coöperation and assistance in making the experiment successful; and to Professor William Anderson McCall, Professor Henry Alford Ruger, Miss Ella Woodyard, of Teachers College, and a host of others for much valuable aid.

HERMAN CHAN-EN LIU

CONTENTS

CHAPTER PAGE

I. INTRODUCTION 1
- A. The Problem 1
- B. Intelligence Examination in China 1
- C. The Development of Non-Verbal Tests in America 6

II. THE EXPERIMENT 12
- A. The Preliminary Plan 12
- B. Tests Used in the Experiment 18
- C. Method of Procedure 21

III. FORMATION OF A CRITERION 28
- A. Elements of a Criterion 28
- B. Test Scores Weighting 33
- C. Method of Selection of the Final Criterion . . 39

IV. SELECTION OF TEST ELEMENTS 43
- A. Selection of Test Elements by Correlation Method 43
- B. Selection of Tests by Rating 49
- C. Selection of Tests by Partial Correlation . . . 50
- D. Selection of Tests by Composite Method . . . 52
- E. Weighting by Regression Equation 55

V. RE-TESTING 58
- A. Procedure of Re-testing 58
- B. Statistical Study 61

VI. Alternative Forms and Standardization 63
- A. Alternative Forms 63
- B. Standardization 64
  - 1. Norms 65
  - 2. Scaling 66

CHAPTER PAGE

VII. THE CHINESE NON-VERBAL TESTS 68
A. The Nature of the Tests 68
B. Instructions to Examiner 69
C. Directions for Giving the Tests 70
D. Directions for Scoring the Tests 73
E. Treatment of Results 74
F. Caution 74

VIII. SUMMARY AND CONCLUSIONS 75

APPENDICES
A. Samples of Form A of the Chinese Non-Verbal Intelligence Examination 77
B. Samples of Records Kept 78

BIBLIOGRAPHY 83

INDEX OF TABLES

NUMBER PAGE

I. Distribution of Age and the Numerical Values Assigned . . 30

II. Age Distribution Showing the Slope and the Increase of Scores as Age Advances (Data from Boys Who Have Taken the Pintner Non-Language Tests) 35

III. Grade Distribution of Pressey Scale 36

IV. Weighting of the Scales According to Q 38

V. Data for School Criterion (10 Selected Pupils) 40

VI. Data for Calculation of the Final Criterion (10 Selected Pupils) 42

VII. Correlations of Individual Tests with Final Criterion by Sheppard's and Product-Moment Methods 45

VIII. Correlations of Individual Tests with Combination of Beta 4 and 6, by Sheppard's Formula 46

IX. Correlations of the Different Scales with the Final Criterion and the Inter-Correlations of the Individual Tests 47

X. Correlations between the Individual Tests and the Basic Tests (Pintner 2 and 3, and Beta 6). 49

XI. Ratings of the Individual Tests by Competent Judges . . . 51

XII. Individual Tests Rated *re* Application to Chinese 52

XIII. Correlations of the Individual Tests with the Final Criterion with the Elements of the Basic Tests Eliminated ($r_{12.3}$) . . 53

XIV. Combined Value of the Individual Tests, as Determined by Ratings and Partial *r* Method 54

XV. Data for Calculation of Regression Equation. 56

XVI. Distribution of Re-testing Scores by Grades 59

XVII. Distribution of Re-testing Scores by Ages 60

INDEX OF FIGURES

NUMBER PAGE

1. The Nine-Ring Puzzle 4
2. The Seven Mysterious Boards 4
3. Illustrations of the Seven Mysterious Boards 4
4. Illustrations of the Seven Mysterious Boards 4
5. Showing 27 Per Cent Overlapping of Grade III over Grade IV in the Scores of Pressey Primer Scale 37
6. Showing 21 Per Cent Overlapping of Grade III over Grade IV in the Scores of Myers Mental Measure 37
7. Showing 15.2 Per Cent Overlapping of Grade III over Grade IV in the Scores of Pintner Non-Language Tests 37
8. Showing 12 Per Cent Overlapping of Grade III over Grade IV in the Scores of Army Beta Examination 37
9. Showing 9.8 Per Cent Overlapping of Grade III over Grade IV in the Scores of Dearborn Group Tests of Intelligence 37

CHAPTER I

INTRODUCTION

A. THE PROBLEM

Psychological tests which have been applied in America with great success are now being experimented with in China. Progressive Chinese educators who are attempting to introduce the measurement movement into China, however, are confronted with the problem of procuring and selecting suitable test material. China, with its distinctive civilization and numerous dialects, presents a difficult field for the literal transcription of the American intelligence tests. This difficulty virtually prevents a widespread use in China of the language test, and makes necessary the construction of a non-language test. The present study is an attempt to develop a non-verbal scale, which, because of the elimination of language and schooling factors, may be used in China as an independent measure of general intelligence or as a supplement to a language test.

B. INTELLIGENCE EXAMINATIONS IN CHINA

The practice of setting intelligence examinations is not new in China. It is as old as our history, although the traditional methods have been crude and pseudo-psychological.

The earliest methods, which still prevail, are Kan Hsiang, physiognomy-reading, and Shan Ming, fortune-telling. Pseudo-psychologists in the guise of fortune-tellers and popular physiognomists are found everywhere. They are frequently consulted by uneducated parents as to the intelligence of their children, whose careers and destinies they foretell. The calculations of these pseudo-psychologists are said to be based upon the hour and date of birth, and physiognomic and anthropometric characteristics.

The system of competitive examinations, employed in China for centuries, was a sort of intelligence test. Its purpose was the selection of candidates for civil service. Scholars gathered at the examination halls, which were located in every district. There they were confined in little cells in which they composed classical essays on assigned subjects. Examinations were conducted and papers graded by high government officials. The results were announced with great ceremony, and the successful candidates honored with "Kung Ming,"—the equivalent of American academic degrees.

The practice was founded on the theory that only the intelligent and educated men should rule. No age or birth qualifications were required for participation in these examinations. Youngsters under twelve years of age, however, were sometimes released from the rigid, formal standards. In such cases the regular examination was often replaced by a series of "opposites or matching tests," in which the applicants were required to match assigned words and phrases. For instance, "East" would be expected to be matched with the word "West"; "above" with "below"; "mountains" with "oceans." The following is a typical "Dui Dzi," or opposites test:[1]

(*a*)	Chiang (將)	Fu 父	Djoh 作	Ma 馬
(*b*)	Wang 望	Dzi 子	Cheng 成	Lung 龍

The translation of matching phrase (*a*) with phrase (*b*) is as follows:[2]

(*a*)	Consider	Father	Being	Horse
(*b*)	Expect	Son	Becoming	Dragon

Of the old intelligence tests used in the schools of China, there were certain kinds called "Tien Dzih," that is, "completion tests." Some teachers occasionally employed these tests in judging the brightness of their pupils; others employed them as supplementary

[1] A story relates that a certain farmer carried his young son on his back to the examination hall. The examiner, upon the arrival of the youngster, was surprised at his presence and inquired of him how he had managed to come all the way from his distant home. The boy replied: "I came on my father's back." The boy's answer at once suggested to the examiner a topic for the opposites test, so he said: "Well, if you can match the phrase which I am about to give you, you are passed." The examiner then requested the boy to match "Consider Father Being Horse." The clever child, without a moment's hesitation, replied: "Expect Son Becoming Dragon." He had matched the assigned phrase so well, that he was given a pass without further examination.

[2] These are not strictly "opposites tests," as understood in America; but rather matching tests. They are comprehensive, requiring on the part of the examinee quick understanding and sound reasoning.

methods of teaching elementary composition. Problems in composition were often made by omitting a few words from a well-constructed sentence, necessitating the filling in of the blanks by the children.

A type of test similar to the puzzles used by Ruger is also quite common in China. The most famous of these puzzles is the "Kiu Lien Huan"—a nine-ring puzzle (see Fig. 1), consisting of nine connected copper rings mounted on a bar with a rod running through the center of the rings. The puzzle is how to get the rod out of the rings —a task which requires reasoning, and which seldom is solved by the trial-and-error method. The ring puzzle is used merely as a toy, not as a formal test, yet one often hears the remark, "Solve this puzzle and let us see how bright you are."

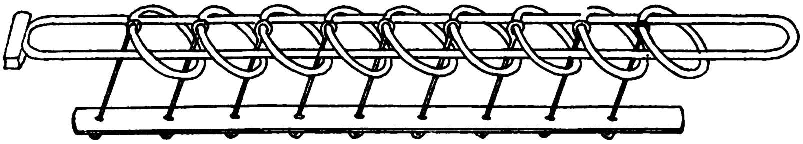


Fig. 1. The Nine-Ring Puzzle.

"Performance tests" also have been in use for centuries in China. The most noted one is "Yih Chih Tu," also called "Tsih Chiao Pan" (see Figs. 2, 3 and 4). Translated literally into English, it would be called "Increasing Wisdom Board," or the "Seven Mysterious Boards." It was called the "Increasing Wisdom Board" because playing with it was believed to increase one's wisdom. It was called the "Seven Mysterious Boards" because with the seven pieces of different shapes and sizes which made up the game, many forms of men, animals, birds, and inanimate objects could be constructed. The game was played by any number of persons, each with his own set of forms. The purpose was to see which person could construct the largest number of objects out of his seven pieces, the winner being considered the most intelligent person in the group.[1]

[1] It is said that the game originated in the ancient imperial palace among the women of the court, who, in the great amount of leisure time at their disposal, welcomed such sport. Later it became popular among the people.

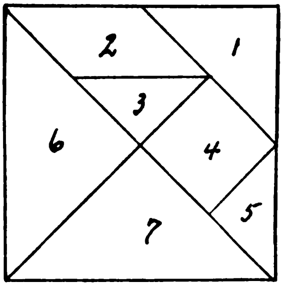


Fig. 2

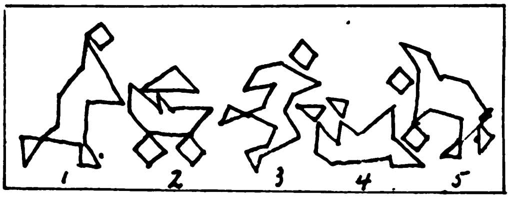


Fig. 3

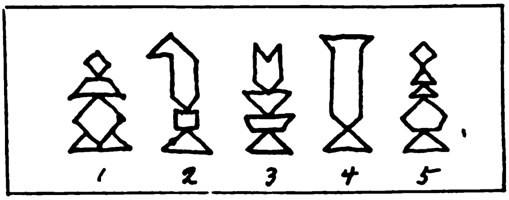


Fig. 4.

FIG. 2. The Seven Mysterious Boards.

FIG. 3. Illustrations of the Seven Mysterious Boards:
1 Man walking; 2 carriage; 3 man running; 4 and 5 two animals fighting.

FIG. 4. Illustrations of the Seven Mysterious Boards:
Candle sticks and different kinds of vessels.

The various tests which have been described here cannot be termed "intelligence tests" in the strictly psychological sense because they are not standardized. They are not extensively used as a measure of general intelligence, but rather as intellectual games. They do demonstrate, however, that the practice of intelligence examinations, although crude and pseudo-psychological, does exist and has existed in China for centuries. It is quite possible that some of these old methods and materials may prove useful in the construction of a genuine intelligence test for China.

It is only within the last few years that scientific psychological measurements have been inıroduced in China. The earliest known experimental work on the subject is that conducted by Dr.W .W. Creighton.[1] From 1915 to 1917, under the direction of Professor W. H. Pyle, of the University of Missouri, Dr. Creighton made a study of the mental and physical characteristics of Cantonese children. The subjects under examination numbered approximately five hundred, most of them ranging from ten to eighteen years in age, although twenty-five women were among those examined. The mental tests used in this experiment were those of rote memory, logical memory, substitution, analogies, and dot patterns. In conducting this experiment Dr. Creighton met with great difficulties as a result of the many dialects prevailing in this province. In his report he says: "In the mental measurements we were confronted at once by language difficulties."

In 1918 Professor G. D. Walcott[2] measured the intelligence of the students in the senior class in Tsinghua College, Peking, who averaged twenty-two years of age. Professor Walcott used the Stanford Revision of the Binet Scale, with the Scott Group Test as a check. The results of the experiment were not very satisfactory as, in addition to the insufficiency of the scale for persons of that age, the language difficulties were insurmountable.

Somewhat later, in the fall of 1920, the Nanking Government Teachers College tried psychological tests for the entrance examination. This is the first attempt made by Chinese educators to intro-

[1] Pyle, W. H.: "A Study of the Mental and Physical Characteristics of the Chinese," *School and Society*, Vol. VIII, No. 192 (August 31, 1918), pp. 264–69.

[2] Walcott, G. D. "The Intelligence of Chinese Students," *School and Society*, XI, 1920, pp. 474–80.

duce scientific intelligence tests into China. Two psychologists educated in America, Professors H. C. Chen and S. C. Liao,[1] devised five tests. The correlation of these tests and the average grades of the regular examination was .39.[2]

Psychological tests for entrance examinations were next taken up by the Peking Government Teachers College. The correlation between the tests and the average grades of the regular examination was practically zero (.000046).[3]

At the present time, Chinese progressive educators, especially those trained in America, are eager to introduce the use of psychological tests into China. Institutions, such as Nanking and Peking Teachers Colleges, as indicated above, have already started the movement. A few private and missionary schools have also adopted some form of tests. The Stanford-Binet Scale has been translated, although it is little used. Aside from these isolated experiments, however, very little has been done. Psychological tests remain virtually unknown. Here lies a great unexplored field of endeavor for the young Chinese schoolman trained in modern scientific method. He needs to understand, however, the difficulties which the use of the numerous dialects and the large percentage of illiteracy offer to the use of any language scale. Evidently a non-verbal test may hope to succeed where the language test is totally inadequate. The development of such a non-language scale is the purpose of the present study.

C. THE DEVELOPMENT OF NON-VERBAL TESTS IN AMERICA

Psychological tests may be roughly classified into two main groups: namely, language tests and non-language tests. The former includes those tests which require verbal response from the subject. The latter group of tests does not require such verbal response. The non-language tests, again, may be subdivided into a group of performance tests which require the doing of some task by means of

[1] *Journal of Educational Research*, Vol. III, No. 5 (May, 1921), p. 394.

[2] As this goes to press, the author has received a copy of *Mental Tests in China*, written by Professors H. C. Chen and C. S. Liao. It contains thirty-five different tests, twenty-four of which are translated from American tests.

[3] The data are found in the *Peking Teachers College Weekly*, No. 132 (September 11, 1921), p. 3, but the correlation was computed by the author by the product-moment method.

certain actual mechanical manipulations, and a group which require the subject to work with geometrical designs, figures or pictures, indicating the results of his thinking by making lines or pictures.

Non-language or non-verbal intelligence tests are the outgrowth of the intelligence measurement movement. Of recent development and used extensively only within the last two or three years, these non-verbal tests have shared the fame of the language tests. Among the tests devised by Alfred Binet, father of the movement for the measurement of intelligence, and published in his 1905–06 series are a number of tests which do not require verbal responses.[1] For example, in the visual coördinations test, the examiner moves a lighted match slowly before the subject's eyes and notes whether he follows the movement with the properly coördinated movements of the head and eyes. In the test known as "prehension provoked tactually," he places the small wooden cube in contact with the palm or the back of the subject's hand to determine whether he can execute properly coördinated movements of grasping. In the drawing test, he shows the subject two drawings, permits him to look at them for ten seconds, and then requires him to draw the views from memory. None of these tests expects a verbal response from the subject.

The scale devised by these French psychologists, Alfred Binet and T. Simon, was first translated and adapted for American use by Goddard.[2] Kuhlman [3] and Wallin [4] followed with further adaptations. The latest revisions of the scale are by Yerkes, Bridges, and Hardwick,[5] by Terman [6] and by Herring.[7] They all adhere

[1] The other non-language tests in the Binet-Simon 1905 series are tests numbered 3, 4, 5, 10, 12, 21, 22, 23, and 29. In the 1908 series the non-language tests are numbered 9, 10, 11, 12, 14, 16, 23, 24, 33, 54. For the complete account see A. Binet and T. Simon, "Le développement de l'intelligence chez les enfants," in *L'Année Psychologique*, 14, 1908, pp. 1–94; and A. Binet and T. Simon, "L'intelligence des imbeciles," in *L'Année psychologique*, 1909, pp. 1–147.

[2] Goddard, H. H.: "The Binet-Simon Measuring Scale of Intelligence, Revised," *Training School Bulletin*, Vol. VIII (1911), pp. 56-62.

[3] Kuhlman, F.: "A Revision of the Binet-Simon System for Measuring the Intelligence of Children," *Journal of Psycho-Asthenics*, monograph supplement, No. 1, p. 41.

[4] Wallin, J. E.: *Experimental Studies of Mental Defectives: A Critique of the Binet-Simon Tests.*

[5] Yerkes, R. M., Bridges, J. W. and Hardwick, P. S.: *A Point Scale for Measuring Mental Ability.*

[6] Terman, Lewis M.: *The Measurement of Intelligence.*

[7] Herring, John P.: *Significance of Certain Elements in Intelligence Examinations*, unpublished Ph.D. dissertation (Columbia University).

more or less closely to the original Binet Scale, and consequently some of their tests are non-verbal in nature.

In spite of the merits of the Binet-Simon Scale and its revisions, their chief deficiency lies in the large proportion of tasks requiring language responses. This criticism of the scale was vigorously presented by Ayres in 1911. He pointed out that the Binet tests predominantly reflect the child's ability fluently to use words, and do not reveal his ability to do acts. Thus, it gives "a warped and partial measure of his real degree of intelligence." [1]

✓ The language difficulty, inherent in the Binet-Simon Scale and its various revisions, became evident when the clinical psychologists attempted to apply it in various fields of practical work. They found that the Scale was utterly inadequate for the mental examination of non-English speaking people, speech defectives, the deaf, and those with language difficulties. Hence they introduced non-language tests which do not require language responses on the part of the child for adequate performance. Among those who first used the non-language test were Healy and Fernald.[2] In carrying out mental examinations at the Juvenile Psychopathic Institute of Chicago, they had been confronted with the problem of testing a cosmopolitan population. Some of the inmates were illiterate, and some, though educated in their own tongue, were unable to speak the English language. Since they represented most of the nationalities and languages of Europe no single test requiring language directions and responses could be adequate to measure them. In discussing their work, Healy and Fernald say: "The Binet-Simon Scale helps little where the language factor is a barrier, either on account of foreign parentage or insufficient schooling, or with uneducated deaf and dumb children." [3] They became convinced that language, as far as possible, should be eliminated from the mental examinations given to such subjects. They say: "In predicting the possible development of an individual under various conditions, it is most desirable to ascertain the mental ability quite apart from the individual's experience in formal training in our language, or indeed

[1] Ayres, L. P.: "The Binet-Simon Measuring Scale for Intelligence: Some Criticisms and Suggestions," *Psychological Clinic*, Vol. v (1911), pp. 187–96.

[2] Healy, W., and Fernald, G. M.: "Tests for Practical Mental Classifications," *Psychological Monographs*, Vol. 54, No. 2, pp. 4–5.

[3] *Ibid.*

any language. It often becomes necessary to classify mentally a subject who has had no education in English-speaking schools, or indeed who has had but little schooling of any kind."[1]

The work carried on at the Institute not only proved the inadequacy of the language tests, but demonstrated the practical value of the non-language tests. Healy and Fernald conclude as follows: "On one occasion we found ourselves able to demonstrate satisfactorily that a Gypsy boy of fifteen, quite innocent of schooling and knowledge of the three R's, had at least fair, if not good, native ability. And repeatedly a number of our tests have proved most serviceable in mentally classifying young, deaf and dumb children."[1]

Knox, in[2] his work among the immigrants at Ellis Island, found it impossible (even with the services of an interpreter) to use scales in which language responses were required. Faced with this language obstacle, and under the necessity to diagnose mental disease and mental deficiency among the immigrants, Knox devised a series of non-language tests, many of which are excellent and still widely used in psychological clinics.

Pintner and Patterson[3] also found the language scale "absolutely inadequate to test the mentality of deaf children." They experimented with the Binet-Simon Scale, but were confronted with numerous difficulties, such as lack of comprehension of certain tasks due to physical deficiency which in turn had made for lack in the environment of opportunity for forms of experience needed to acquire the proper test reaction. Consequently, they constructed a scale of performance tests which requires practically no instructions for the child other than natural gestures. Pintner and Patterson consider the non-language feature of the test as a *sine qua non* in the measurement of mentality in the deaf. As to the importance of the non-language tests, they say: "Here we have a group of individuals, completely shut off from hearing language, and for that reason laboring under a language difficulty that only in rare cases is surmounted to the extent of making them comparable in language

[1] *Ibid.*

[2] Knox, H. A.: "A Scale Based on the Work at Ellis Island for Establishing Mental Defects," *Journal of the American Medical Association*, Vol. LXII (March 7, 1914), pp. 741–47.

[3] Pintner, R. and Patterson, D. G.: "The Binet Scale and the Deaf Child," *Journal of Educational Psychology*, Vol. VI (1915), pp. 202 ff.

ability to ordinary hearing individuals. Any kind of tests involving reading or spoken language cannot be used as a test of their mentality. If we employ such tests for measuring the mentality of the deaf and use the standardization obtained from hearing children, we will not be measuring mentality but merely difference in language ability. There may be a greater percentage of feeble-mindedness among the deaf than among the hearing but the fact that a deaf child does not measure up to the language standard of a hearing child is no indication of mental deficiency." The performance tests lately have been used not only for the deaf but also for the non-English-speaking children, speech defectives, and children from different language environments.

The development of the non-language tests was greatly advanced, and their practical value definitely recognized, as a result of the United States Army psychological examinations.[1] In 1917, when the psychologists took up the personnel work in the Army they soon discovered that many of the men were handicapped by language difficulties. In order to permit the illiterates a real opportunity to show their ability, a non-language scale was constructed. Demonstration charts and pantomime were used to convey the instructions to the examinees. These methods require no language directions or responses. This scale, known as the Army Beta Examination, consisted of seven tests, including maze, cube analysis, X-O series, digit-symbol, number checking, pictorial completion, and geometrical construction. The scale was applied to 23,547 men. Its results correlate with the Army language examination Alpha to the amount of .80; with Stanford-Binet, .73; with the composite of Alpha, Beta, and Stanford-Binet, .91. This high correlation demonstrates the practicability of making non-language tests and the feasibility of their use where the language tests fail utterly. The unexpected efficiency of the Army Beta Examination thus demonstrated during the war, later brought about a mushroom growth of the non-verbal test material. Thorndike,[2] champion of the measurement movement, who had charge of much of the statistical work in the development of the Army tests, was first to utilize the data and experience gained from these tests. The Thorndike Non-Verbal

[1] Yoakum, C. S. and Yerkes, R. M.: *Army Mental Tests.*

[2] Thorndike, E. L.: "A Standard Group Examination of Intelligence Independent of Language," *Journal of Applied Psychology*, Vol. 3, No. 1 (March, 1919), pp. 13–32.

Examination follows the general nature of the Army Beta, but eliminates one weakness by providing ten alternative forms of the examination instead of the single form, thus reducing the error in measurement caused by unfair tutoring. Such alternative forms widen the field of usefulness of tests in many ways, permitting a study of the growth of intelligence by repeated testing, comparison of groups and individuals and increased reliability in the determination of the intelligence of groups and individuals.

Pintner,[1] who with Patterson constructed the Performance Scale, has also, since the war, devised a non-language group intelligence test. He realized that his Performance Scale, although it required no language response, was still clumsy and not convenient for application to a group. Consequently, in his later scale he devised a set of six non-language tests for group use. When compared with the results obtained from the Binet-Simon Scale the correlation was found to be .66. He recommends that such tests be used in mental survey work for school children and adults, particularly in communities containing a large foreign or illiterate element.

In addition to the non-verbal tests which have already been discussed, there are many others available. Among the more well-known scales are Myers' Mental Measure, Pressey's Primer Scale, Kingsbury's Primary Group Intelligence Tests, and Dearborn's Group Intelligence Tests. All these tests have been widely employed, with varying degrees of success, by psychologists.

The rapidity of the development of non-language tests has been phenomenal and indicates that it meets an important need. In the Binet-Simon Scale, there were only a few tests which required no language responses. Then followed the performance scales, developed by Healy, Knox, Pintner and others, in which language responses are completely eliminated. The Army Beta Examination, with its wide application among the millions of soldiers, demonstrated its practical value for intelligence measurement and for group use. Others have succeeded in advancing the non-verbal tests beyond the experimental stage. These tests are now applied to individuals and groups, both as an independent measure and as a supplement to language tests, with confidence that the results are trustworthy and fairly adequate.

[1] Pintner, R.: *The Mental Survey.*

CHAPTER II

THE EXPERIMENT

A. PRELIMINARY PLAN

In drafting a preliminary plan for the experiment it was first decided to devise a large number of tests and to try them out on Chinese children in America. Since the purpose of the experiment was to develop a non-verbal intelligence scale for use in China, it appeared essential that the subjects be Chinese. Ten non-verbal tests were consequently constructed and mimeographed for trials. Fifty-one persons were examined with these tests, after which the examinations were discontinued as impracticable. The reasons for the disuse of the examinations were threefold: first, the tests were constructed by the subjective method instead of by the objective or scientific method; second, the tests were mimeographed instead of being printed, causing the test material to be in many instances indistinct and difficult of recognition; third, the scarcity of Chinese subjects, and the difficulty of dealing with the few which were available. Three months' time and much labor had been expended, and naturally the results were discouraging. An important fact, however, was revealed by these examinations; namely, the children of naturalized Chinese and of Chinese long resident in this country had been affected by their American environment and training, so that they were more American than Chinese. Tests which were applicable to American-Chinese children would be quite irrelevant if applied to children in China. As a result of these findings, the mimeographed tests were abandoned and thought was turned to the formulation of a new plan.

A careful study was then made of all the available intelligence tests, especially the non-verbal forms. The new plan under consideration was to select the best elements in the American non-verbal tests and to attempt to develop them into a non-verbal scale for use in China. At the time (1920), there were available the following non-verbal and semi-non-verbal tests:

1. Army Beta Examination

2. Dearborn Group Tests of Intelligence, Series I
3. Haggerty Intelligence Examination Delta 2
4. Holley Picture Completion Test for Primary Grades
5. Myers Mental Measure
6. National Intelligence Tests
7. Otis Group Intelligence Examination
8. Pintner's Mental Survey Tests
9. Pressey Primer Scale
10. Trabue Mentimeters

The question arose whether all of these tests, or whether any of them, could be used in the experiment. Before coming to a decision, it was necessary to formulate definitely the principles to be embodied in the proposed scale for use in China. After considerable study, the following principles were adopted as criteria:

1. Tests should involve no language responses from the subjects.
2. Test materials should be drawn from social environment common to all peoples.
3. Test material should exclude, as much as possible, school training.
4. Test material should be of interest to all types of subjects.
5. Tests should be valid as a measure of intelligence.
6. Tests should be reliable.
7. Objective methods should be employed in both giving and scoring of tests.
8. Tests should measure a wide range of intelligence.
9. Tests should indicate mental growth.
10. Tests should be adapted for group use.
11. Time for testing and scoring should be reasonably short.
12. Instructions for testing and scoring should be simplified for use by teachers and others who are not specialists in measurements.
13. Tests should have alternative forms as a preventive against the vicious effect of coaching.
14. Test material should be inexpensive, easy to handle, of small bulk, and easily kept in order.

1. *Tests should involve no language responses from the subject.*

General intelligence signifies a group of related inborn capacities for adapting one's self to specific situations in life. Inborn capacities, however, are never measured directly but are always inferred from the ability displayed. Language use is one of these abilities which ordinarily is a good index to intelligence, but it has its limitations. It cannot be employed, for instance, as a medium to measure intelligence when the language varies among the subjects under examination sufficiently to make understanding or executing the tasks difficult, slow, or impossible. Such a condition exists in China. The languages spoken in various sections differing widely, people from Peking do not understand the dialect of Canton, and the Shanghai dialect is different from that of Hankow. This diversity of dialects is not only characteristic of the provinces but exists in local districts of the same province. The written language, it is true, is identical throughout China, but comparatively few can read, 90 per cent of the Chinese people being illiterate. Under these conditions, a non-verbal test for use in China would have great superiority over any existing language test.

2. *Test material should be drawn from social environment common to all peoples.*

It is a well-known fact that social environment affects the development of intelligence. Edison, born and raised in the wilds of Thibet, would doubtless never have developed into the particular kind of a mechanical genius he now is. To measure a Thibet-born Edison by the standards used in examining an American-born Edison, would manifestly be inaccurate and unfair. The uncivilized Miaotze boy in Yunnan could not be expected to answer questions on automobiles or airplanes; and the New York boy, raised in the Bronx, could not be expected to answer intelligently questions on rice growing. There should be common grounds; the test material should be drawn from an experience common to all. Tests should measure capacity, and this can be accomplished by measuring only those traits possible of development by all subjects. Tests, based on such a principle, could be employed over all of China.

3. *Test material should exclude, as far as possible, school training.*

As ninety per cent of the Chinese people are illiterate, test material which requires school training must prove inadequate. Culture and school training are both acquired, not innate. They vary in different persons according to the environment to which they have been subjected. The boy ignorant of mathematics could not be expected to solve problems in algebra as well as the son of an instructor in mathematics. In order to compare the native ability of children, therefore, the products of school training should be excluded from the test material.

4. *Test material should be of interest to all types of subjects.*

Interest in the tests is essential to proper reaction; therefore, a good test should arouse interest in the subjects of widely differing mentality and type of intellect. Unless this is accomplished, the results of the test will not indicate the actual intelligence. Errors have been made in drawing conclusions as to the intelligence of the individuals in a group, when these individuals have had interests different from those called out by the test. For instance, a mechanical test given to a co-educational class usually results in a higher score for the boys than for the girls. The scores in this case do not prove that the boys are more intelligent than the girls; they probably indicate rather the difference in degree of interest in the subject between the boys and girls. It is, therefore, evident that the tests to be adequate must be of common interest to the entire group.

5. *Tests should be valid as a measure of intelligence.*

A test is valid when it actually measures the trait which it professes to measure. A valid test, therefore, implies actual, consistent measurement. Whether a test is valid or not is determined by the correlation of test scores and the elements of the intelligence, as objectively known by other means. The checks on validity most often used are school marks and progress, and estimates by teachers and associates. In applying this principle, the reliability of such checks should be investigated.

6. *Tests should be reliable.*

Reliability in a test indicates the obtaining of similar results from two or more testings of the same subjects under the same conditions. Perfect reliability implies identical results from two or more testings under identical conditions, and is, therefore, never completely attained; but competent authorities agree that the coefficient of reliability should be .90 or higher for a group of equal age.

7. *Objective methods should be employed in both giving and scoring of tests.*

Objectivity is attained when the methods and procedure of testing and scoring are uniform and independent of personal opinion so that the results may be verified by other testers. That is to say, methods of testing and scoring should be identical at all times for all testers. The personal equation of the teacher should be eliminated as far as possible. The results of the testing should endure verification in all cases where the same tests are applied to the same subjects, using the same methods under similar conditions.

8. *Tests should measure a wide range of intelligence.*

The term "general intelligence" means the combination of many mental traits. It varies in amount in individuals from practically zero in the lowest grade of idiots to that large quantity, at present unmeasured, of the world's greatest genius. Its distribution, according to the best available estimates, approximates a bell-shaped curve; that is to say, there are few of genius level, a large number of ordinary or average people, and comparatively few idiots. An intelligence test, to be entirely satisfactory, should be easy enough for all except the hopeless idiots to make some score and sufficiently difficult for a person of great genius not to make a perfect score. On the other hand, the scores should be distributed continuously and around one mode. Furthermore, the tests should measure a large number of unlike or differentiating traits. The ideal way would be to measure every trait that contributes to intelligence and to give each trait a weighting proportional to its contribution to the total intelligence. This is impossible in our present state of knowledge, but an intelligence test should measure as many differentiating traits as possible.

9. *Tests should indicate mental growth.*

Intelligence develops along with the advance of chronological age up to a point believed to be somewhere near the end of the adolescent period. As the child grows older, his native endowment unfolds. So a normal ten-year-old child should be able to do more than the eight-year-old child and a normal eight-year-old child should know more than a six-year-old child. The intelligence test should reveal the different stages of development by improved scores with each increase in chronological age. This mental index is known as mental age. The mental age divided by the chronological age gives what is known as the intelligence quotient.

10. *Tests should be adapted for group use.*

Group testing enables the examiner to test many persons at a time and therefore makes possible the testing of many more children with the same expenditure of time, labor and money, than can be achieved by testing them singly. The success of the group-testing method was shown in the United States Army during the war. To test two million soldiers individually in so short a time was totally impossible, but by means of group tests the men were speedily sorted and classified. Group testing may not give such an accurate diagnosis as does individual testing, but it is generally satisfactory and can be supplemented by individual tests in exceptional cases. For general use in China, the tests must be adapted for group use.

11. *Time for testing and scoring should be reasonably short.*

Time for testing should be long enough for the average subject to give response without hurry, but it should be reasonably short so as not to cause fatigue in the subjects nor to entail such administrative inconvenience as to prevent its use. If two scales, for instance, give the same result, and one takes thirty minutes to give, while the other takes two hours, the former is certainly preferable to the latter. As to scoring, the test should be constructed so that it may be accurately, uniformly, and rapidly scored with little dependence upon the judgment of the persons doing it. Mechanical scoring devices should be employed whenever advisable.

12. *Instructions for testing and scoring should be simplified for use by teachers and others who are not specialists in measurement.*

There are not many psychologists in China. Most of the measurement work probably will be done by the ambitious teachers and others who are not specialists in measurements. To facilitate the work, it is absolutely necessary that the instructions for both testing and scoring should be simplified so that they can be followed easily. The instructions should be clear, concise, and adequate, but must be brief, consistent, and uniform for all who are to be testers.

Whenever possible, instructions should employ a preliminary demonstration test in order that the subjects may understand clearly what they are expected to do.

13. *Tests should have alternative forms as a preventive against the vicious effect of coaching.*

The one-form scale has at least two defects. First, if the tests are to be used as a basis for promotion in education or business, ambitious parents will be likely to purchase the material and coach their children with the object of increasing their scores. Second, the one-form scale cannot be used in retesting for a study of mental growth. Therefore, alternative forms should be prepared. They should have the same value, however, as the original form, and measure the same traits.

14. *Test material should be inexpensive, easy to handle, of small bulk and easily kept in order.*

As communication is inconvenient in some parts of China and the merit of intelligence measurement is not as yet widely demonstrated there, it is important that every advantage be taken to facilitate the use of the tests. They should, therefore, be easy to handle; they should not be bulky nor contain apparatus which is difficult to keep in order; and the cost of the test material should be small.

B. TESTS USED IN THE PRESENT EXPERIMENT

In consideration of the above adopted principles, a selection was made from the ten non-verbal tests listed on pages 12 and 13 of the following tests, to be used in the experiment:

1. Myers Mental Measure
2. Pintner's Non-language Tests
3. Pressey Primer Scale
4. Army Beta Examination
5. Dearborn Group Examination, Series I
   General Examination 1
   General Examination 2
   General Examination 3

A brief description of each of these tests follows:

1. *Myers Mental Measure.*[1]

[1] *School and Society*, Vol. 10, pp. 353-60 (1919).

The Myers Mental Measure was devised by Carolyne E. Myers and Garry C. Myers for school use, the Measure being based upon the Army Beta tests. Mr. and Mrs. Myers were interested in the classification of children as early as possible on the basis of intelligence in order that children of marked ability might be selected for rapid advancement, and that those of very low grade intelligence might early be segregated. To do this, they devised a scale, universal in nature, with the hope that it could be applied to school children of all ages and given in 15 or 20 minutes to a large number of individuals.

The scale consists of four tests, all of which are pictures. The first test is called a directions test. It requires the child to obey certain directions, such as to draw a line or make a mark in a particular way. It furthermore needs no preliminary demonstration other than a brief pantomime with very little spoken instruction. The second test is a picture-completion test consisting of pictures of familiar objects or situations, with one important element missing which the subject must supply. The third is a learning test which requires the subject to make substitution of proper symbols for other symbols, while the fourth is a common element test in which the subject is asked to mark the pictures which are similar in some way.

Mr. and Mrs. Myers used the Stanford-Binet Scale as a check upon their own scale. Omitting test 3, which gives practically zero correlation, the total of tests 1, 2 and 4 correlates about .80 with Stanford-Binet.

2. *Pintner's Non-Language Tests.*[1]

Pintner's Non-Language Tests were devised by Professor Rudolph Pintner with the purpose of measuring the general intelligence of the deaf, illiterate, and non-English speaking. A knowledge of English is not needed either to understand the directions or to make responses. The scale consists of six tests which have been arranged for group testing, suitable for children and for adults. The first is the imitation test which is essentially the same as the Knox Test. The second and third are "easy learning" and "hard learning" tests respectively. The task in the next one is a "drawing completion" test, which is an abbreviated form of the larger drawing test devised

[1] Pintner, R.: "A Non-language Group Intelligence Test," *Journal of Applied Psychology*, Vol. III, No. 3 (September, 1919).

by the same author. The fifth is the "reversed-drawing" test, which requires the subject to draw the reversal, or counterpart of a drawing given. The last test is "picture-reconstruction," involving the rearranging of picture sections with the object of completing the entire picture. All the correlations between each test and the total score were found positive and fairly high; and the correlation between the I Q on the Stanford-Binet and the percentile rank on the Pintner's Non-Language Tests was .66.

3. *Pressey Primer Scale.*[1]

Pressey Primer Scale is known as the "crossing-out" test. As the authors describe, "each test asks of the subject that by crossing out some one thing, he eliminate a wrong, irrelevant, or extreme element in a situation." The scale was devised for the use of the first three grades. In the first test, the subjects are required to cross out an unnecessary dot in each of several groups of dots. The second involves the crossing out of the most discordant, or dissimilar object from a group of three objects; the third, for the crossing out of the superfluous block in each square, after the other blocks have been fitted into four patterns at the top of the page; and the fourth test provides for the crossing out, in each picture, of the absurd part.

4. *Army Beta Examination.*[2]

The Beta Examination was introduced primarily for the group testing in the Army during the World War of those illiterate in English. Instructions were given in the form of four demonstrations at the beginning of each test with gestures and pantomimes. The original or trial series consisted of fifteen tests, but after an extensive trial, seven tests were finally retained. These tests are known as maze, cube analysis, X-O series, digit-symbol, number checking, pictorial completion, and geometric construction. The maze test, devised by C. R. Brown, was retained from the preliminary trials because it could be successfully demonstrated, gives few zero scores and correlates fairly well with the total scores of army Alpha and

[1] Pressey, S. L. and Pressey, L. W.: "Cross-out Tests," *Journal of Applied Psychology*, Vol. III (1919), pp. 143–150.

[2] See Yerkes and Yoakum, *Army Mental Tests;* also *Memoirs of the National Academy of Science*, Vol. XV.

Beta. The cube analysis test was originally devised by Edwards at Camp Lee, to take the place of the usual form of test for arithmetical reasoning. Test 3 (X-O series) was an attempt to provide the equivalent of test 8 of Alpha. It proved to be an easy and effective way to indicate the institutional feeble-minded group. The digit-symbol test was modeled after the well-known substitution test which had been used in various forms by Woodworth, Pintner, Whipple, and others. Number checking was devised by Thorndike, and found satisfactory on all counts. The pictorial completion test was devised by Kelley and patterned originally after the Binet mutilated pictures. The last test, geometrical construction, was patterned after the various form-board tests. It was found particularly good in picking out the higher levels of ability. The product-moment coefficient of correlation between the Beta Examination weighted score and Stanford-Binet mental age was reported to be .731 ± .012.

5. *The Dearborn Group Tests of Intelligence, Series I.*[1]

The Dearborn Group Tests of Intelligence, Series I, were devised and standardized by Professor W. F. Dearborn, of Harvard. They are not linguistic, and consist of three parts (known as General Examinations 1, 2, and 3) for use in the first three grades. General Examination 1 contains a "directions test," a "clock test" and a "circus" test. General Examination 2 consists of seven "games" which, in order, are "color blocks," "substitution," "ladders," "picturemaking," "picture recognition," and "dominoes." General Examination 3 consists of "picture completion," "map of town," "ruler," and "number form puzzles." A correlation of .87 of the Stanford-Binet Scale with Dearborn tests has been reported.[2]

C. METHOD OF PROCEDURE

The present testing was carried on in Public School No. 108, situated in the section of New York City which is populated and inhabited by immigrants. This school has only the kindergarten and the first four grades. Each grade is divided into two sections, so there are altogether nine sections in the school. During the fall of 1920

[1] Dearborn, W. F.: *The Dearborn Group Tests of Intelligence.*

[2] *Journal of Educational Research,* VOL. III, No. 4, p. 308 (April, 1921).

when the experiment was started, the enrollment was about 1,000. After a preliminary trial, it was found impossible to test all the pupils, as those who were in the kindergarten and the first grade could not follow the directions of the tests satisfactorily. They were eliminated from the testing and only the children from grades 2B to 4B were tested. In these grades, there were 185 boys and 216 girls, a total of 401. The distribution of the children according to nationalities was as follows:

Nationality	Number	Per Cent
Italian	362	90.27
Chinese	21	5.24
Jewish	14	3.49
Jewish-Italian	2	.50
Chinese-Jewish	1	.25
Spanish-Italian	1	25

Only a few of these children were Chinese; more than 90 per cent were Italians. However, since the purpose of the experiment was to select the best non-verbal tests, and since special forms for use in China would have to be made later and no norms were expected to result from the testing, the nativity of the subjects was wholly immaterial. Prior to this experiment, the school had never used any standardized psychological or educational tests. The principal and the teachers were all deeply interested in the experiment and offered every possible assistance to make it a success. The writer took advantage of this unusually excellent opportunity to visit the school frequently and make friends with both teachers and pupils. In consequence, when he was ready to test the children, although a foreigner, he was no longer a stranger to the school population.

All the testing was done in a large classroom equipped with desks, blackboard and comfortable chairs. Twenty-eight pupils were brought to this room, to be tested, at one time. The pupils were seated apart from each other, so the possibility of copying was reduced.

Before giving a test to the children, the writer familiarized himself with the instructions by trying them with other children. All the examinations were conducted by the writer himself with the assistance of the principal, Miss Rae, and a college trained teacher. Pains were taken to maintain uniformity both of the procedure in the testing and the environment in the room. The order at testing was always from the younger ones, then to the older ones; that is, from Grade 2B to 4B. The testing time was from 10 A. M. to 12 M. and from 1 P. M. to 3 P. M. Every effort was made to make the testing informal and pleasant yet stimulating and searching.

The scales were given on the following dates:

No. 1 Nov. 24–26, 1920—Myers Mental Measure
No. 2 Dec. 4–6, 1920—Pintner's Non-language Tests
No. 3 Dec. 14–16, 1920—Pressey Primer Scale
No. 4 Dec. 20–22, 1920—Army Beta Examination
No. 5 Jan. 4–8, 1921—Dearborn Group Tests of Intelligence

In giving the scales, all the original directions were followed literally, except in the cases of the Dearborn Group Tests of Intelligence and the Army Beta Examination, both of which were modified to meet the peculiar needs. The altering of directions for the Dearborn Tests of Intelligence was very slight. The only change made was in the "clock" test of General Examination 1. The original direction calls for the subjects to draw in the clock hands, indicating the time when school begins in the morning, when school begins in the afternoon, and when school closed in the afternoon. In the school where the testing was done, starting and closing time is different for different children. As it was therefore confusing for the children to answer these questions, the following directions were substituted: "In the first clock draw in the hands so as to show what time school assembly begins in the morning. Draw the hands in the next clock, to show what time school recess begins for lunch. In the third clock show what time school begins in the afternoon." As suggested by Dearborn, the tests were given in two periods, but with one day interval between them.

In the case of the Army Beta Examination, the procedure was considerably modified. The original directions call for a blackboard frame consisting of eight fitted sections, a blackboard chart in a continuous roll 27 feet long, cardboard pieces for Test 7, and patterns for constructing Test 2. It was impossible to get the original apparatus, so the school blackboard, self-made cardboard pieces, and real wood cubes were substituted. Furthermore, according to the original form, it was necessary to have an examiner, a demonstrator, and a number of orderlies. The demonstrator was charged with the single task of doing before the group just what the group was later to do with the examination blanks. The use of a special demonstrator, as provided for in the original tests, was considered both superfluous and cumbersome. The examiner also performed the duty of the demonstrator. As in other scales, he was to give the directions as well as demonstrate to the class the preliminary test. The adapted directions were as follows:

Directions for Test 1

As soon as the pupils have been properly seated, and examination blanks distributed, the examiner says, "Here are some papers. You must not open them or turn them over until you are told to."

Holding up the Beta blank, the examiner continues: "In the place where it says name, print your name very clearly. Remember, print your full name. If you are Mary Jones, print Mary Jones; if you are John Smith, print John Smith. Right after your name, in the place where it says rank, write your grade. Do you know in which grade you are? That is fine. Write down your grade very clearly, so that I can read it. Look over your paper again and show me whether all of you have written your name and grade very clearly."

Before the examination begins, each paper should be inspected by the assistants in order to make sure that the name and grade are clearly written. Then the Examiner remarks, "Attention! Watch what I do on the blackboard. I am going to do here what you are going to do on your papers. Ask no questions. Wait till I say, 'Go ahead.' Now is everybody ready? Turn your paper over. This is Test 1, here (pointing to the page of record blank). Have you found it?" After all have found the page, the Examiner continues, "Don't make any marks till I say 'Go ahead.' What I want you to do is to draw a line which shall pass through the pictures from left to right without touching any line. Now watch me work on the blackboard." After touching both arrows, the Examiner traces through the first maze with chalk, slowly and purposely makes one mistake by going into the blind alley at upper left-hand corner of the maze and asks the class, "Is this correct?" After the class answers "No," the examiner places his hand back to the place where he may start right again, and traces through the rest of the maze, indicating an attempt at haste and hesitating only at ambiguous points. After this is done, he says, "Everybody ready! All right. Go ahead. Hurry up." At the end of two minutes, the examiner says, "Stop! Turn over the page to Test 2."

Test 2

The examiner then continues, "This is Test 2 here. Look!" After everyone has found the page he says, "I want you to count the cubes and write the number in the little square below the picture. Now watch me work on these blocks." The order of procedure is as follows:

a The examiner points to the three-cube model on the blackboard, making a rotary movement of the pointer to embrace the entire picture.

b With similar motion he points to the three-cube wood model on the desk.

c The examiner next points to picture on blackboard and asks the class, "How much?"

d The examiner turns to cube model and counts aloud, putting up the fingers while so doing, and encouraging the class to count with him.

e The examiner taps each cube on the blackboard and asks the class, "How much?"

f After the class answers correctly, the examiner counts the cubes on blackboard silently and writes proper figures in proper places. (The rest is the same as the original directions).

After the demonstration is completed, the examiner says, "Everybody ready! All right. Go ahead. Hurry up," and at the end of 2½ minutes he says, "Stop! Look at me and don't turn the page."

Test 3, x-o Series

"This is Test 3 here. Look." After everyone has found the page, he says, "I want you to draw in X or O in the proper squares which are empty. Now watch me work on the blackboard." The examiner first points to the blank rectangles at the end, then traces each "O" in chart, then traces outline of "O's" in remaining spaces and draws them in. Then he traces first "X" in next sample, moves to next "X" by tracing the arc of an imaginary semicircle joining the two, and in the same manner traces each "X," moving an arc to the next. He then traces outlines of "X's" in the proper blank spaces, moving over imaginary arc in each case, and asks the class what should be drawn in. The examiner follows the answers of the class and fills in remaining problems very slowly. After the demonstration is finished, the examiner says, "All right. Go ahead. Hurry up!" At the end of 1¾ minutes he says, "Stop! Turn the page to Test 4."

Test 4. Digit-Symbol

"This is Test 4 here. Look." After everyone has found the paper—"I want you to study each number and memorize the symbol which represents it. Put in the right symbol under the right number." The examiner touches the number in first sample with index finger of right hand; holding finger there, finds with index finger of left hand the corresponding number in key; drops index finger of left hand to symbol for the number found; holding left hand in this position writes appropriate symbol in the lower half of the sample. Similar with the other sample. But for the last three samples the class is asked to give the correct symbols. At end of the demonstration, the examiner says, "All right. Go ahead. Hurry up!" At the end of 2 minutes the examiner says, "Stop! But don't turn the page."

Test 5. Number Checking

"This is Test 5 here. Look." After everyone has found the page—"I want you to find out whether the two numbers are the same. If they are the same, write 'X' on the dotted line between them; if they are not the same, write 'O' on the dotted line between them. Now watch me do this on the blackboard." In this demonstration the examiner must get "Yes" or "No" responses from the class. If the wrong response is volunteered by the group, the examiner points to digits again and gives right response, "Yes" or "No" as the case may be. The examiner points to the first digit of first number in left column, then to second digit first number in left column and second first number in right column. He says "Yes" to the class and marks an "X" on the dotted line between the number. The examiner does the same for the second line of figures, but here he indicates by "O." In the last three samples, the class is asked to answer "Yes" or "No." After the demonstration is over, the examiner points to page and says, "All right. Go ahead. Hurry up!" At the end of 3 minutes, he says, "Stop! Turn over the page to Test 6."

Test 6. Pictorial Completion

"This is Test 6 here. Look! A lot of pictures." After everyone has found the page—"Every picture has something gone. I want you to fix it. Now look at the pictures on the blackboard." The examiner points to the picture of the hand, then to the place where the finger is missing and asks the class, "What is gone?" After the class has given a correct answer, he says, "That's right. The finger is gone." Then he draws in the finger. Similarly with the other samples. When the demonstration is finished, the examiner says, "Fix all the pictures on the whole page. All right. Go ahead. Hurry up!" At the end of 3 minutes, the examiner says, "Stop! but don't turn over the page."

Test 7. Geometrical Construction

"This is Test 7 here. Look." After everyone has found the page—"Here are blocks. Imagine that you could fill them in this square, and then draw in the intersecting lines in this square. Now watch me." The examiner points to the first figure on blackboard. He then takes the two pieces of cardboard, fits them on the similar drawing on blackboard to show that they correspond and puts them together on the square on blackboard to show that they fill it. Then, after running his finger over the line of intersections of the parts, he removes the pieces and solution in the square on the blackboard. Similarly for the other samples. At the end of the demonstration, the examiner holds up the blanks, points to each square on the page and says, "All right. Go ahead. Hurry up!" At the end of 2½ minutes he says, "Everybody stop!" Papers are then collected by monitors immediately.

While the children were doing the tests, a general impression of their attitude was recorded. As a whole they showed fine spirit, worked enthusiastically, and seemed to enjoy the work. Judged by their manner, they seemed especially interested in the Pressey

Absurdity Test and in all the Pictorial Completion Tests. But in the case of the Dearborn Tests, the majority of the children were bewildered by lack of clearness in the directions and showed signs of fatigue due to the over-long time required.

Practically all the tests were scored by the writer himself, with great care. The Dearborn Group Tests of Intelligence were found to be the most difficult to score. It took on an average of fifteen minutes to score a child's paper containing the three examinations. The amount of time required in scoring the Dearborn Tests seemed greatly out of proportion to the results obtained.

CHAPTER III

FORMATION OF A CRITERION

The chief object of the present study is to select the best group of tests from the five intelligence examinations which were given to the children in New York City Public School No. 108, with the view to modify them for use in China. In order to do so, it is necessary to have a definite, constant criterion with which to compare tests. This criterion should be made up from as many factors as possible that are known to be indices of the constituents and development of general intelligence. These factors must be reliable indicators if the criterion, which is depended upon to determine the value of the selected tests, is to be trustworthy. In this chapter, an account of the selection of the best elements to be included in the criterion with which to compare the tests is given.

The elements of the criterion adopted are: (1) age, (2) school marks, (3) school progress, (4) teachers' estimates of intelligence, and (5) composite test scores of (*a*) Dearborn Group Tests of Intelligence, (*b*) Pintner Non-language Tests, (*c*) Army Beta Examination, (*d*) Myers Mental Measure, (*e*) and Pressey Primer Scale. Certain weights, to be described later, are given for age, school marks, school progress, and teachers' estimates of intelligence; and to each of the mental test scores, and the total combined into one rating called Final Criterion.

A. ELEMENTS OF THE CRITERION

The elements which may be included conveniently in a criterion for pupils' intelligence are age, teachers' estimates, school marks, school progress and test scores. All of these measure general intelligence in different ways, though their values are not equal. Some or all of them should be used in combination and given weight in reference to their special significance in showing the presence of intellectual ability.

1. *Chronological Age.*

The chronological ages of the children were copied directly from the school record, in order that they might be accurately known. As the administering of the intelligence scales extended from November 24, 1920, to January 8, 1921, the median date was taken as a standard to calculate the ages, that is, December 17, 1920. All the ages, therefore, shown on the record book date from the birth of the individual up to December 17, 1920.

The dependence of intelligence upon age in adults is a theoretical problem, but gradual mental growth in children is accepted by all psychologists as beyond dispute. Binet, Terman, Thorndike, and others have all found that the general intelligence of a child gradually develops as his age advances until he reaches maturity. Kelley [1] and Fretwell,[2] according to their experimental studies, find that there is a negative correlation between achievement—an indication that with pupils in the same grade the younger are the brighter ones. Since all the subjects in this study are below fifteen years of age, it is evident that age should be considered in the making of the criterion for the selection of the tests, but that the young child in an advanced grade should be given a bonus, and the older child in the early grade a demerit in utilizing age as a criterion of intelligence. Therefore an age distribution table was prepared and a numerical value was assigned to different ages in different sections of the grades. As the ages for the sexes were different, so the values assigned were also different. For instance, in section B of the second grade boys, 9 was assigned to 7 yrs. 6 mo.; 8 to 8 years; 7 to 8 yrs. 6 mo; 6 to 9 years; 5 to 9 yrs. 6 mo.; 4 to 10 years; and 3 to above 10 years. For a complete record, see Table I.

2. *Teachers' Estimates.*

A teacher associates with children daily. She knows in a general way a pupil's strong points as well as his weak ones. Her estimate of his general intelligence should be accurate in some particulars if she clearly understood that her rating was to be based upon native intelligence and not school achievement. Fretwell found that the correlation of the composite of teachers' judgments of pupils with the composite of eleven tests was .66. Kelley found that "the cor-

[1] Kelley, T. L.: *Educational Guidance.*

[2] Fretwell, E. K.: *A Study in Educational Prognosis.*

TABLE I

Distribution of Ages and the Numerical Value Assigned to Each Age

BOYS

Grade	2B		3A		3B		4A		4B	
Age	No.	Value	No.	Value	No.	Value	No.	Value	No.	Value
Yr. Mo.										
7.0	..	10	..	11	..	12	..	13	..	14
7.6	2	9	..	10	..	11	..	12	..	13
8.0	17	8	13	9	..	10	1	11	..	12
8.6	3	7	9	8	..	9	..	10	..	11
9.0	5	6	16	7	20	8	8	9	2	10
9.6	2	5	2	6	7	7	1	8	..	9
10.0	2	4	6	5	6	6	15	7	13	8
10.6	..	3	..	4	1	5	1	6	5	7
11.0	2	3	1	3	3	4	4	5	7	6
11.6	..	3	..	3	..	3	..	4	2	5
12.0	..	..	..	3	..	3	2	3	5	4
12.6	..	..	..	..	..	3	..	3	..	3
13.0	..	..	..	..	..	..	..	3	1	3
13.6	..	..	..	..	..	..	..	..	1	3
14.0	..	..	..	..	..	..	..	..	1	3
14.6	..	..	..	..	..	..	..	..	..	3
15.0	.	..	..	..	..	..	..	..	1	3

GIRLS

Grade	2B		3A		3B		4A		4B	
Age	No.	Value	No.	Value	No.	Value	No.	Value	No.	Value
Yr. Mo.										
7.0	..	10	..	11	..	12	..	13	..	14
7.6	..	9	1	10	..	11	..	12	..	13
8.0	30	8	10	9	5	10	..	11	..	12
8.6	1	7	7	8	4	9	1	10	..	11
9.0	17	6	18	7	15	8	11	9	3	10
9.6	2	5	2	6	4	7	10	8	2	9
10.0	2	4	6	5	6	6	18	7	19	8
10.6	..	3	..	4	6	5	3	6	3	7
11.0	..	..	..	3	1	4	2	5	9	6
11.6	..	..	..	..	1	3	..	4	..	5
12.0	..	..	..	..	1	3	..	3	3	4
12.6	..	..	..	..	..	3	1	3	..	3
13.0	..	..	..	..	..	..	1	3	4	3

relation between class standing and the regression equation combination of the estimates of traits by teachers" was .76. He remarked, as a result of his investigation, "With such a high correlation, a division of pupils into classes by means of teachers' estimates would be highly reliable." [1] Teachers' estimates are not enough, however, because a teacher may overemphasize some factors and neglect others. Terman found that teachers frequently err in estimating general intelligence because they neglect to consider age and emotional differences. Whipple found that teachers estimate the dull children too high and the bright children too low.

In this study, teachers were requested to separate their children into five classes A, B, C, D, and E on the assumption that "Intelligence is a general capacity of an individual consciously to adjust his thinking to new requirements: it is general mental adaptability to new problems and conditions of life."[2] They were asked to rate few as A's or E's, comparatively more as B and D, and a larger number as C. The teachers were warned not to grade the intelligence of their children by their school achievement and deportment but by their general abilities or brightness shown both in their academic work and extra-curricular activities. In order to be fair to the children, the teachers were requested to grade their children independently three times, November 24, 1920, December 16, 1920, and January 11, 1921. The dates were sufficiently far apart so that the teachers scarcely remembered their previous marks. An aggregate of the three estimates was taken as the estimate of the teacher for the general intelligence of the child.

In order to make possible statistical treatment, the letters A, B, C, D and E, given by the teachers, were transmuted into numerals. They are shown as follows:

Teachers' Estimate	Numerical Value Assigned
A	4
B	3
C	2
D	1
E	0

[1] Kelley, T. L.: *Educational Guidance*, p. 16.

[2] Stern, W.: "The Psychological Methods of Testing Intelligence," translated by G. M. Whipple, *Educational Psychology Monographs*, No. 13, p. 3.

3. *School Marks.*

School marks have been the most universal method used for grading pupils. In the past, it has been the only method of judging the ability of the children recorded in school reports. While it is true that teachers often do not agree with each other, yet school marks are a fair measure of mental ability. Fretwell found the correlation between school marks and a group of tests as high as .57.[1] McCall says, "Teachers' marks are important because they are now and will continue for some time to be the most universal method of rating pupils. In fact, they may continue forever to be the criterion for classification because teachers will soon be familiar with the simple mysteries of scientific measurement." [2]

In the school in which this experiment was carried on, there were weekly, monthly, and term examinations. The teachers mark the children by letters. The marks used in this study are the average school marks of the children in the fall term of 1920–21. For the convenience of statistical study, the school marks were turned into figures as follows:

School Marks	Numerical Value Assigned
A	10
B	8
C	7
D	5
E	3

4. *School Progress.*

By school progress is meant the progress which a child has made in the school, that is, his present class standing. The very reason that one could be promoted to a certain grade and maintain his standing there shows that he must have the mental ability to handle the subjects. When a pupil fails to make satisfactory progress in his school work, he is ordinarily retarded or eliminated. It is clear, therefore, that advance in grade usually indicates development of intelligence, although there may be exceptions. Sometimes the school permits a pupil to move up a grade or class even though he has not done the work below, because the parents of the child insist upon it; or because the teacher wants to get rid of the backward child; or

[1] Fretwell, E. K.: *A Study in Educational Prognosis*, p. 17.

[2] McCall, W. A.: *How to Measure in Education.*

because the school must make room for younger pupils. However, the majority of the pupils are promoted because their mental ability permits the expected scholastic attainment, and therefore the grade reached should be utilized in building up the criterion for the selection of tests.

In Public School No. 108, classes are divided into A and B sections, and pupils are promoted by sections twice a year, A being the lower section. The following numerical values were assigned to the different sections of the different grades:

School Progress	Numerical Value Assigned
Grade IIB	0
Grade IIIA	5
Grade IIIB	10
Grade IVA	15
Grade IVB	20

5. *Test Scores.*

All tests given are standardized. They all are claimed to correlate highly with general intelligence. The correlation between Myers Mental Measure and Stanford-Binet Scale was reported to be .80; between Pressey Primer Scale and Stanford-Binet Scale, .60; between the Army Beta Examination and Stanford-Binet Scale, .73; and between Dearborn Group Tests of Intelligence and Stanford-Binet Scale, .87. It is safe to assume that if the individual scales are so valid as a measure of intelligence, a combination of the test scores of all these scales would result in an excellent measurement of general intelligence. Based upon this assumption, the combined test scores were included in the final criterion.

B. TEST SCORES WEIGHTING

In order to study the combined value of all the scales given to the children, it was necessary to have a composite of all the test scores. This could be done by summing all the raw test scores of the different scales. But the merits and variabilities of the scores of the different scales are different. To sum the raw scores is, therefore, unfair. The problem then to be next considered was how to weight the different scales properly.

It was important to know the merits of the different scales, when a weight was attached to each. One of the simplest methods for finding them was to prepare an age or grade distribution table and inspect the slope shown. This was based on the assumption that a child, as he advanced in age and grade, should make a higher score in an intelligence test. This gradual increase of scores in proportion to the advance in age and grade permits the appearance of a slope on the distribution table. When scores for a given age are near together and on the whole greater for each increased age, which is shown graphically by their clustering about the slope line, the more valuable is the scale.

Based on the above assumption, age and grade distribution tables of each scale were prepared for both sexes. An inspection of the tables shows the existence of some slope in all the scales; Pintner Non-language Tests, Army Beta Examination, and Dearborn Group Test, however, seemed better than the rest. All of the tables could not be shown here, but for illustration, the distribution of Pintner's Non-language Tests is given in Table II. Attention is called to the slope and the gradual increase of scores as expressed by the medians, from 50.83 for age 8 to 106.5 for age 11.

Another rough method for finding the merits of the different scales is to compute the extent of overlapping of the two groups of scores. The assumption is this: the less overlapping in different grades the tests show the better measures of intelligence they are. For instance a good scale should show the differences in mental traits between the child in Grade III and the child in Grade IV. The more the scale can indicate the difference, the more reliable the scale. Such overlapping can be computed by comparing the two total distributions of the test scores by stating the variabilities of the two groups and their central tendencies. The method used in this study, however, is a shortened one, based on the following formula:[1]

$$\text{Per cent overlapping of } A \text{ over } B = \frac{A\,(\text{No. of cases}) > \text{median in } B}{N\,A}$$

To illustrate the method, the data in Table III are taken. In this case A in the formula means Grade III and B Grade IV. The median for Grade IV is 67.5, which falls midway in the step 65–70.

[1] See Thorndike, E. L.: *Mental and Social Measurements*, p. 128 ff.

TABLE II

AGE DISTRIBUTION SHOWING THE SLOPE AND THE INCREASE OF SCORES AS AGE ADVANCES; DATA FROM BOYS WHO HAVE TAKEN THE PINTNER NON-LANGUAGE TESTS

Age	7	8	9	10	11	12	13	14	15
Score									
130-135			/						
125-130					/				
120-125					//				
115-120				//	/	/			
110-115		/		//	///	/			
105-110			////	//	/		/		
100-105			/	/		///	/		
95-100			////	////					
90- 95		/	//	///	//				
85- 90			/	/	/				
80- 85		//	////	/////					
75- 80		//	/////	//	/				
70- 75		/	/////	//					
65- 70		//	////	////	/				
60- 65		//	///	//					
55- 60		////	/////	/					/
50- 55		///	//	///		/			
45- 50		//	///	/	/				
40- 45		/	///						
35- 40		///	/	/					
30- 35		//	///					/	
25- 30			///						
20- 25		///	//						
15- 20	/	//	/		/				
10- 15	/	//	//	/					
5- 10		/							
0- 5		/	/						
Number of Cases	2	35	60	37	15	6	2	1	1
Median		50.83	66.25	81.50	106.5				
Quartile		10.50	17.10	13.45	16.25				

TABLE III

GRADE DISTRIBUTIONS OF THE PRESSEY SCALE

SCORE	GRADE III	GRADE IV
85–90	. .	9
80–85	5	8
75–80	9	26
70–75	18	19
65–70	17	17
60–65	23	16
55–60	20	20
50–55	10	10
45–50	13	8
40–45	15	2
35–40	7	2
30–35	3	2
25–30	4	. .
20–25	3	1
15–20	1	. .
10–15	. .	1
5–10	1	. .
0– 0	. .	. .
Number of Cases	149	141
Median	59.375	67.5
Quartile	10.6	9.6

The number of Grade III pupils who equal or exceed this score is therefore $\frac{17}{2} + 18 + 9 + 5$ or 40.5, which is 27 per cent of the number in the third grade, 149. The per cent of overlapping of the third and fourth grades is, therefore, 27 per cent. It is illustrated by Fig. 5.

By this method the per cents of overlapping were computed for Grade III and Grade IV in all the scales. The results were as follows (for illustrations, see Figs. 6, 7, 8 and 9):

VALUE NUMBER	SCALE	PER CENT OVERLAPPING OF GRADE III OVER GRADE IV
1	Dearborn	9.8
2	Army Beta	12.0
3	Pintner	15.2
4	Myers	21.0
5	Pressey	27.0

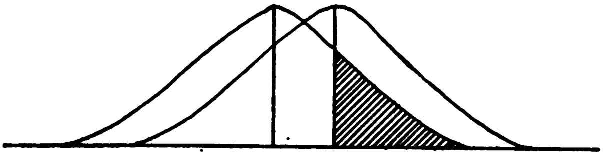


Fig. 5. Showing 27 percent overlapping of Grade III over Grade IV in the scores of Pressey Primer Scale.

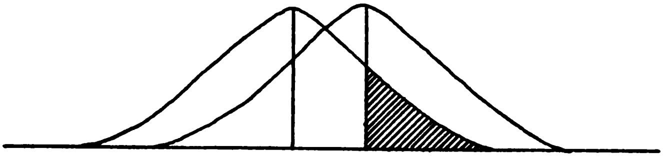


Fig. 6. Showing 21 percent overlapping of Grade III over Grade IV in the scores of Myers Mental Measure.

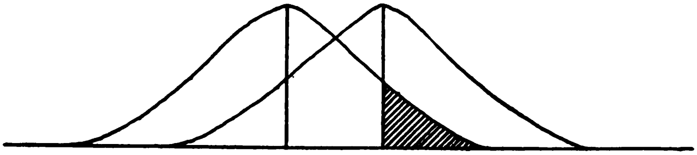


Fig. 7. Showing 15.2 percent overlapping of Grade III over Grade IV in the scores of Pintner's Non-language Tests.

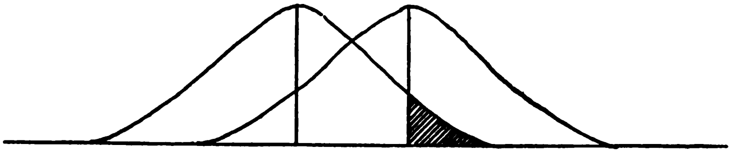


Fig. 8. Showing 12. percent overlapping of Grade III over Grade IV in the scores of Army Beta Examination.

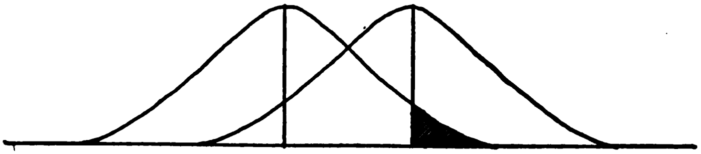


Fig. 9. Showing 9.8 percent overlapping of Grade III over Grade IV in the scores of Dearborn Group Tests of Intelligence.

The Dearborn Group Tests of Intelligence, the Army Beta Examination, and the Pintner Non-language Tests, which were found better than the others according to the slope method, also stand high here. However, tests should be ultimately weighted according to the variabilities of their scores; the range and deviations from the averages should be taken into consideration. The measure of variability used in this study is Q or quartile-deviation. Q is that distance on the base line of the normal curve which includes roughly half of the measure, when laid off on each side of the average. It is computed by $Q = \frac{Q_3 - Q_1}{2}$. That is, Q = half of the distance between the 75 percentile and 25 percentile.

Q was computed for ages 8, 9, 10 of both boys and girls as shown in Table IV. The sum of these Q's in the different scales is 62.1 for Pressey Primer Scale, 104.1 for Pintner Non-language Tests, 29.8 for Myers Mental Measure, 69.3 for the Army Beta Examination, and 150 for the Dearborn Group Tests of Intelligence. These

TABLE IV

WEIGHTING OF THE SCALES ACCORDING TO Q

Age	Pressey	Pintner	Myers	Beta	Dearborn
BOYS					
8	11.5	10.5	5.4	13.3	39.0
9	8.6	17.1	5.1	11.5	23.0
10	11.7	13.5	5.6	11.6	23.0
GIRLS					
8	14.8	19.2	4.0	9.8	26.0
9	8.6	28.8	5.2	10.2	16.0
10	6.9	15.0	4.5	12.9	23.0
Total	62.1	104.1	29.8	69.3	150.0
or					
Abbrev. Total . . .	6	10	3	7	15
Multiplier	1	1	2	2	1
Resulting Weight .	6	10	6	14	15

numbers were then reduced, for convenience, to 6, 10, 3, 7, 15 respectively for five different scales. These values of the Q's show that, if the raw scores of the different scales were summed up just as they appear, the Dearborn Scale with its Q of 15 would have five times as great weight as the Myers Scale with its Q of 3; it would give Army Beta Scale almost the same weight as the Pressey Scale, and these weights did not appear to correspond with the real value of the tests. After several trial weightings, it was finally decided to multiply the Myers Scale scores and the Army Beta scores by 2, and the other scores by 1. The results showed that they were thus weighted fairly as their value corresponded roughly with the results previously found by the overlapping method. Army Beta Examination was found to be one of the best scales, and it should have at least as much weight as the Dearborn Scale. Although the Myers Scale was not considered one of the best, it was fair to assume that it should carry weight equal to the Pressey Scale. The following table indicates the weights:

Scale	Value No.	Per Cent Overlapping	Weighting
Dearborn......	1	9.8	15
Army Beta.....	2	12.0	14
Pintner	3	15.2	10
Myers	4	21.7	6
Pressey	5	27.0	6

After the raw scores of the different scales were weighted, they were summed up to get a composite score for each individual.

C. METHOD OF SELECTION OF THE FINAL CRITERION

After consideration of the various facts known about the subjects, and inspection of their correlations with the composite test score, the following composite (termed school criterion) of age, school marks, teachers' estimates and school progress was tried. As previously explained (page 29) numerical values were assigned to different ages, so that a young child in an advanced grade receives more credit than an older child in the same grade (see Table I). Likewise, teachers' estimates of intelligence and school marks (see pages 29, 31–32), both of which were registered in letters, were transmuted into numbers. Numerical values were also assigned to the grades reached (see pages 32, 33). To illustrate the procedure, ten cases are shown in Table V. Pupil A received a credit of 9 for her age, 10 for her school marks, 12 for teachers' estimates of her intelligence and 20 for her school progress. Similarly, pupil J received a credit of 3 for his age, 5 for his school marks, 4 for his teachers' estimates of his intelligence and 0 for his school progress. In the same way, credits were assigned to all the elements of the school criterion for each pupil.

This seemed a reasonable weighting of the facts. Their correlations with the composite test score were .71 for the boys and .91 for the girls, with an average of .81. Since we may assume that the composite average of all the intelligence tests is a fairly true measure of intelligence, these high correlations are evidence that the school criterion is reasonable.

TABLE V

DATA FOR SCHOOL CRITERION (10 SELECTED PUPILS)

Grade		IV B (Girls)					II B (Boys)				
Pupil		A	B	C	D	E	F	G	H	I	J
Age	Chron. Age	9 yrs 6 mo.	10 yrs 3 mo.	9 yrs 6 mo.	10 yrs	10 yrs	8 yrs 5 mo.	9 yrs 6 mo.	8 yrs 6 mo.	8 yrs	7 yrs
	Credit	9	8	9	8	8	8	5	7	8	3
School Marks	Marks	A	B+	A	B	B+	C	C	C	B	C
	Credit	10	8	10	7	8	5	5	5	7	5
Teachers' Estimates	Marks	A A A	A A A	A A A	B+ B+ B	A+ A+ A	C+ C+ C	C+ C+ C	D+ E+ D	B+ C+ C	C+ E+ C
	Credit	12	12	12	9	12	6	6	2	8	4
School Progress	Grade	$4B_1$	$4B_1$	$4B_1$	$4B_2$	$4B_2$	$2B_1$	$2B_2$	$2B_2$	$2B_2$	$2B_2$
	Credit	20	20	20	20	20	0	0	0	0	0
Total		51	48	51	44	48	19	16	14	23	12

However, a combination of the composite test score and the school criterion might be still more useful. So, we combine them into what we have called the Final Criterion. The S.D. for the school criterion is 8 and that for the test score 12. It seems desirable to give each equal weight; therefore, the raw score of school criterion were multiplied by 3 and the composite test score by 2. This may be expressed by an equation as follows:

Final Criterion = (3 × School Criterion) + (2 × Weighted Test Scores)

that is,

Final Criterion = [3×(Age+School Marks+Teachers' Estimates +School Progress)] + [2× (Dearborn + Pressey + 2 × Army Beta + 2 × Myers)]

For further explanation, see Table VI, which contains data for ten pupils. Similarly, the final criterion was calculated for all the pupils.

The Final Criterion is the standard used to select the best test elements from the five intelligence scales for development into a valid and reliable measure of intelligence for use in China. The success of the work is, therefore, largely dependent upon the validity and reliability of the criterion. Now the questions arise: Is the final criterion valid and reliable? Are not the elements which made up the final criterion repeating themselves? Is it right to include the scores of tests in the final criterion and then use the combination with the tests elements? It is admitted here that the criterion elements do overlap and there is no line of demarcation to differentiate them. For instance, when a teacher estimates the general intelligence of a child she considers his age and school achievement; and school progress involves many factors such as age, school marks and teachers' estimates. But there is no doubt that every element measures something which is somewhat different from that which the other elements measure, that no two of them measure exactly the same traits. Furthermore, all the criterion elements, as explained before, are in some degree measures of general intelligence and each of the five scales has been reported to be a reliable intelligence test. A combination of all these factors certainly should make the final criterion reliable. Finally it must be kept in mind that the purpose of this study is to select the best test elements from the five scales; and the criterion required is simply a definite constant standard. It really makes little difference whether the criterion elements to some extent overlap in their functions, for the final criterion will be applied uniformly to the tests elements.

TABLE VI

DATA FOR TEN SELECTED PUPILS FOR CALCULATION OF THE FINAL CRITERION

Grade.	IV B (Girls)					II B (Boys)				
Pupils.	A	B	C	D	E	F	G	H	I	J
School Criterion										
Age	9	8	9	8	8	8	5	7	8	3
School marks	10	8	10	7	8	5	5	5	7	5
Teachers' estimates	12	12	12	9	12	6	6	2	8	4
School progress	20	20	20	20	20	0	0	0	0	0
School Criterion Total	51	48	51	44	48	19	16	14	23	12
3 × School Criterion Total	153	144	153	132	144	57	48	42	69	36
Test Score										
Pressey	65	59	78	69	78	45	31	61	40	24
Pintner	130	91	130	121	104	5	4	51	9	15
Myers	19	24	28	15	28	0	0	0	6	10
Army Beta	90	73	108	82	60	16	39	42	28	15
Dearborn	197	169	210	205	220	14	67	138	70	87
Test Total Pressey+ . . . . Pintner+ . . . . 2× Myer+ . . . . 2× Beta+ . . . . Dearborn	610	513	690	589	578	96	180	334	187	176
2× Test Total (abbreviated)	122	102	138	120	116	20	36	66	38	36
Final Criterion [(3 × School Criterion Total) + (2 × Test Total)]	275	246	291	252	260	77	84	108	107	72

CHAPTER IV

SELECTION OF TEST ELEMENTS

A. SELECTION OF TEST ELEMENTS BY CORRELATION METHOD

The ultimate aim of this.study is to select the best single tests from the five intelligence scales, with the hope that they may constitute a non-verbal intelligence scale for use in China. Chapter III has discussed the "final criterion." The present task is to utilize it as a basis for selection. For this purpose the correlations of every single test of the five scales with the final criterion have been worked out. It is assumed that any test element which correlates highly with the final criterion is good. This, however, does not mean that all the tests which correlate highly with the final criterion should be adopted in the Chinese Scale. A high correlation between two tests may be because they measure the same traits; and the correlations so obtained are simply self-correlations. A good intelligence scale should measure a combination of different traits, so the test elements in the scale should measure as many different mental traits as possible. Consequently, the ultimate object should be to select those test elements which individually correlate highly with the final criterion but which correlate but little with each other. The writer has adopted $r = .80$ as a standard. It is aimed to discover a group of test elements from the five scales, which, combined together, will give a correlation above .80 with the final criterion.

Scattergrams were prepared charting every single test element against the final criterion, an inspection of which showed the following to have high correlations.

Pressey Primer Scale, test 4
Pintner Non-language Tests, tests 2 and 3
Myers Mental Measure, test 2
Army Beta Examination, tests 4, 5 and 6
Dearborn Group Tests for Intelligence, Series I:
  General Examination 1, test 17
  General Examination 2, test 4
  General Examination 3, test 1

After the scattergrams were inspected, the next step was to determine roughly the correlations of all the tests. The formula used is Sheppard's, $r = \cos \pi U$ where U is the "percentage of unliked signed pairs,"[1] and

$$U = \frac{u + \left(\dfrac{\dfrac{u}{u+l} + \dfrac{1}{2}}{2} \right) d}{n}$$

n = the number of cases
l = the number of + + and − − pairs
u = the number of + − pairs
d = the number of oo, o+ and o− pairs

All the correlations which, by this method, were found to be above .60 were computed also by the product-moment method. Table VII shows the results as found.

Among these tests, the two types which appear the most promising are the completion tests and learning tests. Other workers in this field find similar results. Each of these was used by the makers of three of the five scales tried out and was included in their final forms because of its value as an independent measure of intelligence. Consequently, these two types of tests have been made the core of the proposed Chinese scale. The other elements to be chosen should not correlate highly with these two combined, since any other test which does correlate highly with them probably measures the same traits and, consequently, would add little to the measurement. The learning and completion tests selected were those from Army Beta (tests 4 and 6) rather than from the others, because this scale has had a wider use and more searching criticism than any of the others. With these as a basic group, the correlations with every other test in all the five scales were made. Table VIII shows the results.

However, all the completion and learning tests show high correlations with the final criterion and certain of the correlations of the tests against Beta 4 and 6 combined give promise. But to explore further to see whether a better basal combination could

[1] Thorndike, E. L.: *An Introduction to the Theory of Mental and Social Measurements*, pp. 170–71.

TABLE VII

CORRELATIONS OF INDIVIDUAL TESTS WITH FINAL CRITERION BY SHEPPARD'S PRODUCT-MOMENT METHODS

Tests	Number of Cases	Column I r (Sheppard)	Column II r (Product-Moment)
Pressey 1	233	.51	..
Pressey 2	230	.28	..
Pressey 3	230	.42	..
Pressey 4	216	.61	.54
Pintner 1	234	.59	..
Pintner 2	235	.90	.69
Pintner 3	235	.84	.62
Pintner 4	234	.48	..
Pintner 5	234	.59	..
Pintner 6	235	.59	..
Myers 1	235	.63	.51
Myers 2	234	.66	.47
Myers 3	235	.68	.50
Myers 4	231	.66	.49
Army 1	229	.54	..
Army 2	234	.39	..
Army 3	234	.30	..
Army 4	233	.61	.44
Army 5	234	.68	.52
Army 6	234	.75	.65
Army 7	233	.45	..
Dearborn I 7	234	.66	.45
Dearborn I 8	232	.51	..
Dearborn I 9	235	.40	..
Dearborn I 10	234	.63	.52
Dearborn I 11	235	.66	.50
Dearborn I 12	235	.42	..
Dearborn I 15	235	.36	..
Dearborn I 16	212	.72	.56
Dearborn I 17	235	.51	..
Dearborn II 1	235	.39	..
Dearborn II 2	235	.82	.58
Dearborn II 3	234	.36	..
Dearborn II 4	234	.66	.46
Dearborn II 5	234	.45	..
Dearborn II 6	234	.48	..
Dearborn II 7	234	.75	.56
Dearborn III 1	235	.68	.58
Dearborn III 2	..	..	..
Dearborn III 3	227	.64	.43
Dearborn III 4	233	.56	..

TABLE VIII

CORRELATIONS OF INDIVIDUAL TESTS WITH COMBINATION OF BETA 4 AND 6 BY SHEPPARD'S FORMULA

TESTS	NO. OF CASES	*r* (SHEPPARD)
Pressey 1	346	.45
Pressey 2	338	.51
Pressey 3	341	.45
Pressey 4	344	.61
Pintner 1	312	.48
Pintner 2	313	.42
Pintner 3	313	.61
Pintner 4	313	.51
Pintner 5	312	.48
Pintner 6	312	.34
Myers 1	297	.45
Myers 2	324	.61
Myers 3	319	.51
Myers 4	292	.56
Army 1	371	.66
Army 2	370	.42
Army 3	370	.51
Army 5	368	.58
Army 7	374	.33
Dearborn I 7	334	.45
Dearborn I 8	324	.36
Dearborn I 9	335	.19
Dearborn I 10	333	.36
Dearborn I 11	331	.33
Dearborn I 12	336	.19
Dearborn I 15	329	.31
Dearborn I 16	331	.56
Dearborn I 17	297	.48
Dearborn II 1	341	.45
Dearborn II 2	342	.51
Dearborn II 3	340	.31
Dearborn II 4	340	.37
Dearborn II 5	343	.34
Dearborn II 6	303	.28
Dearborn II 7	305	.37
Dearborn III 1	266	.48
Dearborn III 3	341	.45
Dearborn III 4	303	.51

be made, correlations were worked out between the criterion and various other combinations of tests. The results are shown in Table IX.

TABLE IX

Correlations of the Different Scales with the Final Criterion and the Inter-Correlations of the Individual Tests

Correlation Between	No. of Cases	r (Pearson)
Final Criterion and Pressey	237	.58
Final Criterion and Pintner	235	.78
Final Criterion and Myers	235	.65
Final Criterion and Army	235	.75
Final Criterion and Dearborn	235	.80
Final Criterion and Dearborn I	235	.69
Final Criterion and Dearborn II	235	.76
Final Criterion and Dearborn III	235	.67
Final Criterion and Dearborn I, 1–6	230	.20
Final Criterion and Dearborn I, 7–15	234	.63
Final Criterion and Pintner 2+3	236	.73
Final Criterion and Army 3, 4, 5, 6	232	.714
Final Criterion and Army, 4+6	234	.711
Final Criterion and Army 3, 4, 5, 6+ Pressey, 2, 4	234	.696
Final Criterion and Pressey, 2, 4	233	.47
Final Criterion and Army 4, 6+ Pressey 2, 4	235	.56
Final Criterion and Pintner 2, 3, Army 6	233	.815
Army 3, 4, 5, 6, and Pressey 2, 4	235	.38
Pressey 4 and Army 5	..	.49
Pintner 2 and Pintner 3	337	.73
Pintner 2 and Army 6	313	.28
Pintner 3 and Army 6	313	.37
Pintner 3 and Dearborn I	313	.614
Pintner 3 and Dearborn II	315	.551
Pintner 3 and Dearborn III	316	.57

Here are shown significant results, establishing the fact that a better combination than Army Beta 4 and 6 is Pintner's 2 and 3 and Army Beta 6, its correlation with final criterion being .815. They are, however, still learning and completion tests. Tests 2 and 3 of Pintner's scale both correlate low with test 6 of the Army Beta (.28 and .37). These two types of tests really measure different traits. Pintner 2 and 3 are both included rather than either one alone because they really form a single test.[1]

These three tests finally selected were now termed "The Basic Tests." They take only ten minutes to perform. The other tests to be included with these must be of different type. This could be found out by correlating the individual tests with the basic tests. The correlations between the basic tests and the individual tests were computed and compared with their correlations with the final criterion, as shown in Table X. The results indicated that the other tests were fairly good as independent measures because their corre-

[1] A second significant correlation shown in Table IX is that of the final criterion and the entire Dearborn test (.80). However, this should not be interpreted as proof that the Dearborn Group Tests are the best of the five scales. They take more than two hours to finish, and consequently the high correlation may be due to practice effect. Any test if prolonged might result in a fairly high correlation. No single test in the Dearborn battery, however, correlates higher than .58 (see Table VI) with the final criterion.

It is worth noting (see Table VIII) that when tests 1 and 6 are eliminated from Dearborn Group Examination I, little change in the total correlation is made; also that Group Examinations I, II and III each has almost the same value as the other, the correlations being .69, .76, .67 respectively. Each part of the Dearborn Scale, when used as a single measure of intelligence, is better than the Pressey Scale and just as good as the Myers Scale. Each of the three parts of the Dearborn Scale also correlates fairly high with the Pintner Scale, which also indicates the value of each part as a measure of intelligence.

As a whole test, the Pressey Primer Scale seems to be the poorest of the five scales used in this experiment. Its correlation with the final criterion is only .58. Tests 2 and 4 were found better than the other two tests, but their correlation with the final criterion was only .47. The correlations were not raised when combined with Tests 4 and 6 or Tests 3, 4, 5 and 6 of the Army Beta Examination.

According to this investigation, Myers Mental Measure was better than the Pressey Primer Scale, but it was inferior to the other three scales. The individual tests, however, all showed fairly high correlations with the final criterion.

Army Beta Examination as a whole had a correlation of .75 with the final criterion, which was good. When only the combined scores of Tests 3, 4, 5 and 6 were correlated with the final criterion, the result was $r = .714$; and the correlation between Tests 4 and 6 alone and the final criterion gave just as good result (.711). This proved that Tests 4 and 6 were the best test elements for our purpose in the Army Beta Examination. The conclusion was further confirmed when these two tests, combined with tests from other scales, failed to raise the correlation higher than .711 (see Table IX).

Other things as well as the correlation being taken into account, Pintner's Non-language Scale seemed to give the best measure of intelligence, because (*a*) it correlated highly (.78) with final criterion, (*b*) it did not take a long time to give, and (*c*) it was easy to score. The individual tests also correlated highly with the final criterion. Pintner Tests 2 and 3 with Test 6 of the Army Beta stood highest among all the individual tests in the five scales.

lations with the basic tests were in general lower than with the criterion. Test 4 of the Pressey scale and Test 7 of Dearborn Examination I were the best, as their correlations both were below .30.

TABLE X

CORRELATIONS BETWEEN THE INDIVIDUAL TESTS AND THE BASIC TESTS (PINTNER 2, 3 AND BETA 6)

NAME OF TEST	NO. OF CASES	r (PEARSON)
Dearborn I— 7	287	.26
Dearborn I—10	291	.35
Dearborn I—11	289	.47
Dearborn II— 2	293	.47
Dearborn II— 4	294	.40
Dearborn II— 7	295	.46
Dearborn III— 1	293	.44
Dearborn III— 3	293	.36
Army 4	313	.58
Army 5	309	.38
Myers 1	253	.51
Myers 2	278	.39
Myers 3	280	.34
Myers 4	263	.73
Pressey 4	291	.25

However, as there are other factors to be considered in the selection of the tests besides the correlations, the tests to be combined with the basic tests were not finally selected pending further investigation.

B. SELECTION OF TESTS BY RATING

The rating method is not so accurate as the correlation method, but when the results of the latter are known the former can be wisely used to help in the selection of tests. Sometimes the judgments of specialists are as valuable as objective computation.

On October 20, 1921, the members of the psychology seminar at Teachers College, who are instructors and graduate students in the field of measurement, were asked to rate the individual tests in the different scales. A copy of the test material was distributed to each member and the instructions for administering the tests were read to them. They were then asked to rate two characteristics of each individual test as follows:

a Can many alternative forms be prepared for the test? Assign a value of 0 to 10 or more for alternative forms, 0 value for no alternative forms, and the others in proportion.

b Is success in doing the test due to verbal instruction? Assign a value of 10 if the success in doing the test is entirely independent from verbal instruction, a value of 5 if the success is fairly due to the verbal instruction, 0 value if the success is entirely dependent upon the verbal instruction, and other values in proportion.

The results of the rating are shown in Table XI.

It was assumed that the instructors and the writer, being familiar with tests and their making, would be better judges than the members of the class, and therefore their judgments were weighted four times as heavily as those of the students.

A question of prime importance in the case of any test is whether or not it is applicable to Chinese. Consequently, ten Chinese advanced graduate students of education were asked to rate the tests in the same way as the seminar students. The test material was given to each of the judges and the instructions for giving the tests were read and explained to them. They were asked, "Is this test applicable to Chinese? Assign a value of 10 if it can be applied to Chinese very easily, a value of 5 if it can be applied with some difficulty and 0 value if it cannot be applied to Chinese at all." The results of these ratings were not so satisfactory as anticipated. The writer finally assumed the responsibility, although he was guided by the ratings of other Chinese judges, to rate the individual tests. The results are shown in Table XII (page 52).

Both of the ratings of the two groups of judges indicate different values for the different tests. The three best tests according to this investigation were Tests 4 and 5 of the Army Beta Examination and Test 4 of the Pressey Primer Scale. This finding was still not cosidered final and a further investigation was made.

C. SELECTION OF TESTS BY PARTIAL CORRELATION

To be certain that the other tests to be included in the Chinese scale should be different in their nature from the basic tests, all the completion and learning elements should be eliminated from the

TABLE XI

RATINGS OF THE INDIVIDUAL TESTS BY COMPETENT JUDGES

A = Can many alternative forms be prepared?
V = Is success due to verbal instruction?

Judges........		HL	BC	AG	RP	LH	(*a*) Av.	*A*	*B*	*C*	*D*	*E*	*F*	*G*	*H*	*I*	*J*	(*b*) Av.	[2×(*a*)+(*b*)]÷3
Army 4	*A*	10	10	10	10	10	10.0	10	10	10	8	10	10	8	10	10	5	9.1	9.7
	V	9	7	8	8	8	8.0	5	10	9	8	8	5	10	10	8	5	7.8	7.6
Army 5	*A*	10	10	10	10	10	10.0	10	10	10	7	10	10	10	10	10	10	9.7	9.9
	V	8	4	8	4	5	5.8	3	7	6	10	5	2	8	8	5	7	6.1	5.9
Dear. I—7	*A*	3	4	7	1	6	4.2	10	10	8	8	0	10	9	5	5	2	6.5	4.9
	V	2	0	2	2	0	1.2	1	0	0	5	0	0	7	0	0	5	1.8	1.4
Dear. I—10	*A*	4	7	3	4	10	5.6	0	10	6	0	6	10	8	3	5	0	4.8	5.3
	V	2	1	1	3	2	1.8	0	0	3	2	1	0	0	0	2	2	1.0	1.5
Dear. I—11	*A*	2	1	6	1	2	2.4	10	10	2	0	3	4	0	3	1	3	3.6	2.8
	V	1	0	0	1	1	.6	0	0	0	0	0	0	8	0	1	0	.9	.7
Dear. II—2	*A*	8	2	6	8	10	6.8	10	10	10	2	10	10	2	3	5	3	6.5	6.7
	V	4	1	0	2	5	2.4	0	0	0	1	0	0	9	0	5	3	1.8	2.2
Dear. II—4	*A*	10	10	8	7	10	9.0	10	10	10	3	8	10	2	3	10	9	7.5	8.5
	V	2	0	1	2	3	1.6	0	0	0	5	0	0	10	0	5	0	2.0	1.7
Dear. II—7	*A*	4	2	5	5	10	5.2	10	10	9	0	8	2	2	2	10	3	5.6	5.3
	V	3	2	0	2	2	1.8	0	0	0	1	1	0	10	0	3	0	1.5	1.7
Dear. III—1	*A*	10	2	8	8	10	7.6	6	10	9	0	10	10	0	5	10	5	6.5	7.2
	V	4	5	3	2	1	3.0	3	0	1	7	3	0	10	5	2	3	3.4	3.1
Dear. III—3	*A*	2	1	4	2	3	2.4	10	1	5	0	5	4	0	2	10	2	3.9	2.9
	V	2	3	0	1	0	1.4	0	0	0	0	0	0	10	0	0	0	1.0	1.3
Myers 1	*A*	10	5	9	9	10	8.6	10	10	10	0	10	10	0	10	10	8	7.8	8.3
	V	1	0	0	1	1	.6	0	0	0	6	0	0	8	0	0	0	1.4	.8
Myers 2	*A*	10	5	8	9	10	8.4	10	10	10	5	10	10	10	10	10	5	9.0	8.6
	V	8	4	8	9	7	7.2	5	0	5	10	4	0	7	3	5	6	4.5	6.3
Myers 3	*A*	10	4	9	7	10	8.0	10	10	10	4	10	10	10	10	10	6	9.0	8.3
	V	8	3	2	3	7	4.6	0	0	1	10	1	0	4	0	5	2	2.3	3.5
Myers 4	*A*	10	4	8	7	10	7.8	10	10	10	6	10	10	2	10	10	6	8.4	8.0
	V	8	1	2	2	6	3.8	0	0	0	10	0	0	7	0	3	3	2.3	3.3
Pressey 4	*A*	10	10	8	9	10	9.4	10	10	10	8	10	10	5	4	10	5	7.8	8.9
	V	8	4	7	6	7	6.4	5	5	4	10	3	3	3	5	7	2	4.7	5.8

TABLE XII

Individual Tests Rated re Application to Chinese

Tests	Application
Army Beta 4	9.60
Army Beta 5	8.03
Dearborn I—7	6.00
Dearborn I—10	7.70
Dearborn I—11	7.30
Dearborn II—2	8.46
Dearborn II—4	8.16
Dearborn II—7	7.00
Dearborn III—1	9.76
Dearborn III—3	7.20
Myers 1	7.86
Myers 2	8.76
Myers 3	9.06
Myers 4	8.80
Pressey 4	9.00

other tests. This is done by the method of partial correlation. The formula[1] used is:

$$r_{12\cdot3} = \frac{r_{12} - (r_{12})\,(r_{23})}{\sqrt{(1 - r^2{}_{12})\,(1 - r^2{}_{23})}}$$

r_{12} = The individual tests and the final criterion.
r_{13} = The individual tests and the basic tests.
r_{23} = The basic tests and the final criterion.

The results are shown in Table XIII. Test 4 of Pressey Primer Scale has distinctly high partial correlation (.60) with the criterion after the learning and completion elements are partialed out. As to the other tests, the partial correlations vary from −.25 to +.43.

D. SELECTION OF TESTS BY A COMPOSITE METHOD

The rating method and the partial correlation method both indicated the general value of the different tests, but each by itself could not be used as a basis for the selection of the tests. The best way

[1] For a complete discussion on the partial correlation method, see Thorndike, E. L.: *Theory of Mental and Social Measurements*, p. 182; and Kelley, T. L.: "Table to Facilitate the Calculation of Partial Coefficient of Correlation and Regression Equations," *Bulletin of University of Texas*, 1916, No. 27.

was to use a combination of all the available methods together with a consideration of all the other factors. This could be accomplished by first summing up the results obtained from the different methods and then selecting the best tests according to the composite results, which are shown in Table XIV.

TABLE XIII

CORRELATIONS OF THE INDIVIDUAL TESTS WITH THE FINAL CRITERION WITH THE ELEMENTS OF THE BASIC TESTS ELIMINATED ($r_{12.3}$ COLUMN)

r_{23} = .815 (r Final Criterion and Basic Tests)

Tests	Times	r_2	r_3	$r_{12.3}$
Dearborn I—7	2	.45	.26	.42
Dearborn I—10	2	.52	.35	.43
Dearborn I—11	2	.50	.47	.23
Dearborn II—2	5	.58	.47	.38
Dearborn II—4	7	.46	.40	.25
Dearborn II—7	3	.56	.46	.36
Dearborn III—1	5	.58	.44	.43
Dearborn III—3	5	.43	.36	.26
Army 4	2	.44	.58	−.07
Army 5	3	.52	.38	.39
Myers 1	4	.51	.51	.19
Myers 2	4	.47	.39	.28
Myers 3	4	.50	.34	.41
Myers 4	5	.49	.73	−.25
Pressey 4	3	.54	.25	.60

In comparison with the other factors, more weight should be attached to the partial correlations. Consequently, they were multiplied by 50 so as to equalize the values of the "alternative forms," "verbal instruction," and "application to Chinese." The last column is the summing up of the four values. A review of the combined results shows that the following tests have the highest values:

Pressey Scale, test 4	53.70
Army Beta, test 5	43.33

Evidently, Test 4 of the Pressey Scale and Test 5 of the Army Beta Examination are the best among all the individual tests of the five scales to add to the basic tests. These two were consequently definitely selected to be included in the proposed Chinese scale.[1]

With the selection of Test 4 of the Pressey Scale and Test 5 of the Army Beta Examination, to be included in the Chinese scale, it was necessary to consider the character of these two tests in greater detail. A study of their correlations with other elements showed the following results:

TABLE XIV

COMBINED VALUE OF THE INDIVIDUAL TESTS AS DETERMINED BY RATINGS AND PARTIAL *r* METHOD

Tests	Alternative Forms	Instruction	Application	Partial *r* × 50	Combined Value
Dearborn I—7	4.9	1.4	6.0	21.0	33.32
Dearborn I—10	5.3	1.5	7.7	21.5	30.20
Dearborn I—11	2.8	0.7	7.3	11.5	22.30
Dearborn II—2	6.7	2.2	8.5	19.0	36.40
Dearborn II—4	8.5	1.7	8.2	12.5	30.90
Dearborn II—7	5.3	1.7	7.0	18.0	32.00
Dearborn III—1	7.2	3.1	9.8	21.5	41.60
Dearborn III—3	2.9	1.3	7.2	13.0	24.40
Army 4	9.7	7.6	9.6	− 3.5	23.40
Army 5	9.9	5.9	8.03	19.5	43.33
Myers 1	8.3	0.8	7.9	9.5	26.50
Myers 2	8.6	6.8	8.8	14.0	37.70
Myers 3	8.3	3.5	9.1	20.5	41.40
Myers 4	8.0	3.3	8.8	−12.5	7.60
Pressey 4	8.9	5.8	9.0	30.0	53.70

[1] Test 1 of the Dearborn Group Examination III would likewise have been included, had it not closely resembled the basic tests. Other important objections to the Dearborn tests were: first, the value of the test might be due to practice effect; second, the test, comprising three pages of pictures, was too expensive.

CORRELATION BETWEEN:	CORRELATION
Pressey 4 and Criterion	.54
Army 5 and Criterion	.52
Pressey 4 and Basic Tests	.25
Army 5 and Basic Tests	.38
Pressey 4 and Army 5	.49

Thus Pressey 4 and Army Beta 5 both correlate fairly high with the final criterion, and rather low with the basic tests. On the other hand, their correlations with each other were not high. This proved that the two tests were good measures of intelligence, each measuring traits different from those of the basic tests and from each other. Because of these special qualities and characteristics of the Pressey 4 and Army Beta 5, they were chosen, along with the basic tests, to form the proposed Chinese intelligence examination.

E. WEIGHTING BY REGRESSION EQUATION

It has been found that both Army 5 and Pressey 4 should be included in the proposed Chinese examination. The question then arises as to the amount of weight to be attached to the two tests and the basic tests. To solve this problem the regression equation was used. The regression equation follows:

$$X_1 = r_{12\cdot 34}\frac{\sigma_{1\cdot 234}}{\sigma_{2\cdot 134}}X_2 + r_{13\cdot 24}\frac{\sigma_{1\cdot 234}}{\sigma_{3\cdot 124}}X_3 + r_{14\cdot 23}\frac{\sigma_{1\cdot 234}}{\sigma_{4\cdot 123}}X_4$$

$$\sigma_{1\cdot 234} = \sigma_1 \sqrt{1 - r^2_{12}}\ \sqrt{1 - r^2_{13\cdot 2}}\ \sqrt{1 - r^2_{14\cdot 23}}$$

$$\sigma_{2\cdot 134} = \sigma_2 \sqrt{1 - r^2_{24}}\ \sqrt{1 - r^2_{23\cdot 4}}\ \sqrt{1 - r^2_{12\cdot 34}}$$

$$\sigma_{3\cdot 124} = \sigma_3 \sqrt{1 - r^2_{34}}\ \sqrt{1 - r^2_{23\cdot 4}}\ \sqrt{1 - r^2_{13\cdot 24}}$$

$$\sigma_{4\cdot 123} = \sigma_4 \sqrt{1 - r^2_{34}}\ \sqrt{1 - r^2_{24\cdot 3}}\ \sqrt{1 - r^2_{14\cdot 23}}$$

$$r_{12\cdot 34} = r_{12\cdot 4}\,A_{13\cdot 4,\ 23\cdot 4} - B_{13\cdot 4,\ 23\cdot 4}$$

$$r_{13\cdot 24} = r_{13\cdot 4}\,A_{12\cdot 4,\ 23\cdot 4} - B_{12\cdot 4,\ 23\cdot 4}$$

$$r_{14\cdot 23} = r_{14\cdot 3}\,A_{12\cdot 3,\ 24\cdot 3} - B_{12\cdot 3,\ 24\cdot 3}$$

TABLE XV

DATA FOR CALCULATION OF REGRESSION EQUATION

1 = Criterion. 2 = Basic tests. 3 = Army Beta 5. 4 = Pressey 4.

	1	2	3	4
1				
2	.815			
3	.52	.38		
4	.54	.25	.49	
σ	45.3	28.0	8.1	5.6

In Table XV the figure "1" stands for criterion; "2" for basic tests; "3" for Arma Beta 5; "4" for Pressey 4. These correlations were substituted in the above regression equation and the following result was obtained:

$$X_1 = .81 \times \frac{21.19}{15.11} X_2 + 18 \times \frac{21.19}{6.60} X_3 + 48 \times \frac{21.9}{4.26} X_4 \qquad \text{(or)}$$

$$X_1 = 1.135\, X_2 + 0.578\, X_3 + 2.387\, X_4$$

According to the result of the regression equation, the different tests should be weighted as follows: (*a*) multiplying the basic tests score by 1.14; (*b*) multiplying Army Beta 5 by .58; (*c*) multiplying Pressey 4 by 2.39. In consideration of the general impression of the tests, however, a conservative procedure was adopted. In giving final weights to the tests, the scores of the Basic Tests and of Army 5 were left unchanged, while the score of Pressey 4 was multiplied by 2. The weighted composite scores so obtained (called Composite A) were then correlated with the final criterion, the correlation found being .812. This result was very satisfactory, since it exceeds the goal of r = .80. In order to find out whether the weighting had raised the correlation or not, the raw composite scores of the Basic Tests, Pressey 4, and Army Beta 5 (called Composite B), were also correlated with the final criterion, the correlation found being .789. This showed that the weighting had raised the correlation slightly.

It should be kept in mind that the tests chosen are not based upon an empirical method of a single statistical computation, but upon all the possible available methods, such as correlation, rating by specialists, partial correlation, regression equation. The test elements finally chosen from the five scales for the proposed Chinese Non-verbal Intelligence Examination are:

Test 2 of Pintner Non-language Tests
Test 3 of Pintner Non-language Tests
Test 5 of Army Beta Examination
Test 6 of Army Beta Examination
Test 4 of Pressey Primer Scale

CHAPTER V

RE-TESTING

A. PROCEDURE OF RE-TESTING

The tests to be included in the proposed Chinese intelligence examination having been tentatively selected, the next step was to determine their reliability and practicability. This could be done by giving the above tests to the same children and calculating the correlations of their scores with the final criterion. If the tests are reliable and practicable, they should correlate highly with the old criterion. An effort was made, therefore, to secure the same subjects who the year before had taken all the tests. Some of them had moved out of the district or gone to a higher school and it was impossible to locate all of them, but finally 190 children (from the earlier total of 401) were secured.

The re-testing was done from November 28 to 30, 1921, in the same room where the children were formerly tested. A uniform environment, which was similar to that at the first testing, was maintained throughout the examination. The same principal and the same teacher assisted in timing and policing. As in the first testing, 28 children were tested at a time; the children being sufficiently separated from each other, there was no opportunity for copying. The papers of three children, who continued working after the "stop" signal had been given, were discarded for the computation of the results, leaving papers for 187 children.

The directions for giving and scoring the tests were the same as those of the year before, with the exception of a slight modification in introduction (see Chapter VII for a complete record of the tests). Preceding the testing, four boys and five girls were individually interviewed. Each was questioned whether he could recall anything concerning the tests of the year before. All of them indeed remembered the occasion of the testing—they remembered "the good time they had had with the Chinese teacher," but not one of them could recall any of the tests. In other words, these boys and girls had completely forgotten all about the first test, except for the vague idea of having done it. It is possible that the actual performing of the tests might recall the experience in previous testing, but in young children of this age the likelihood of recalling the tests of the year before seems so slight as to be immaterial. Consequently, the process of re-testing these children cannot be said to be influenced to any noticeable degree by repetition.

In the re-testing, the children appeared to enjoy their work. There was no sign of fatigue; instead, they were very enthusiastic. The writer obtained some interesting information, on the effects of the tests upon the children, by mixing with them during the recess. Joining in their play, he was constantly approached by them with such remarks as, "Mister, play some more games with us." "When will you come back again?", "Oh, I like to see the woman without a nose, and the poor fish without an eye," "There's lots of fun in making zeros and crosses."

The time consumed in testing was from 25 to 30 minutes. It is important to note, in discussing the time necessary for this testing, that none of the groups consumed more than 30 minutes in their testing, nor less than 25 minutes. This, of course, does not include the time taken in the distribution of test material nor for the preliminary remarks by the examiner.

The method of scoring the tests was very simple. Stencils were prepared in order to facilitate the work. With a small amount of practice, test papers could be scored very rapidly, even at the rate of a paper a minute.

B. STATISTICAL STUDY

The first step was to determine the general merits of the selected tests, from now on known as "The proposed Chinese Non-verbal Intelligence Examination." Tables of grade distribution and age distribution were prepared (Tables XVI and XVII), and the medians for the different grades and ages were calculated. The medians found for the different grades were: Grade III, 101.36; Grade IV, 125.26; Grade V, 148.50. The result was encouraging as it showed a fair improvement in central tendencies for the different grades.

The median scores for the different ages were found as follows: Age 8, 85; age 9, 115; age 10, 128; age 11, 142; age 12, 148. The result was also encouraging. The medians for ages 11 and 12 were close to each other, probably because the 12-year-old children in these grades were duller than the average 12-year-old.

The last step was to find out how closely the test scores of the selected tests corresponded with the old final criterion, which was used as a standard for the measure of general intelligence. Consequently, a scattergram was made and the correlation found, by the product-moment method, to be .8768. The result was very satisfactory. Theoretically the correlation between the selected tests and the old criterion should be higher than the correlation between

TABLE XVI

Distribution of Re-testing Scores by Grades

Re-testing Scores	Grade III	Grade IV	Grade V
170–180	..	1	2
160–170	..	2	12
150–160	2	4	13
140–150	2	11	10
130–140	3	15	11
120–130	4	18	9
110–120	3	11	..
100–110	11	8	..
90–100	5	6	..
80– 90	3	2	..
70– 80	6	1	..
60– 70	2	2	..
50– 60	2	..	..
40– 50	1	1	..
30– 40	1	1	..
20– 30	2	..	..
10– 20	..	..	..
0– 10	..	..	..
Number of Cases	47	83	57
Median	101.36	125.28	148.5

TABLE XVII

Distribution of Re-testing Scores by Ages

Re-testing Scores	Age 7	Age 8	Age 9	Age 10	Age 11	Age 12	Age 13
170–180 . .	. .	. .	1	2	. .	. .	. .
160–170 . .	. .	. .	4	6	3	1	. .
150–160 . .	. .	3	8	7	1	. .	. .
140–150 . .	. .	6	5	9	2	. .	1
130–140 . .	. .	7	14	6	. .	1	. .
120–130 . .	. .	10	9	8	1	1	1
110–120 . .	1	5	6	. .	1	. .	. .
100–110 . .	. .	9	9	2	. .	1	. .
90–100 . .	. .	7	2	2	. .	. .	. .
80– 90 . .	. .	1	4	. .	. .	. .	. .
70– 80 . .	. .	2	3	1	1	. .	. .
60– 70 . .	. .	2	1	1	. .	. .	. .
50– 60 . .	1	. .	1	. .	. .	. .	. .
40– 50 . .	. .	1	1	. .	. .	. .	. .
30– 40 . .	. .	2	. .	. .	. .	. .	. .
20– 30 . .	. .	2	. .	. .	. .	. .	. .
10– 20 . .	. .	. .	. .	. .	. .	. .	. .
0– 10 . .	. .	. .	. .	. .	. .	. .	. .
No. of Cases	2	57	68	44	9	4	2
Median . .	85	115	127.7	142.2	147.5	. .	. .

any of the five scales with the old final criterion. This was proven true, as shown in the following:[1]

Correlations Between the Final Criterion and the Different Scales

Scales	Correlation
The Selected Tests	.88
Dearborn Group Tests	.80
Pintner Tests	78
Army Beta Examination	.75
Myers Mental Measure	.65
Pressey Primer Scale	.58

[1] The first correlation is not strictly comparable to the others since it was obtained from the 187 cases of re-testing while the others were from the more than 250 cases in the first testing.

Judging by the results of the correlation of the selected tests with the old final criterion, by the comparatively short time to give the tests, and by the deep interest displayed by the children indoing the tests, together with their other merits, it seems fair to conclude that the selected five tests which are included in the proposed Chinese Non-verbal Intelligence Examination give better results than any of the five scales used in this experiment.

CHAPTER VI

ALTERNATIVE FORMS AND STANDARDIZATION

A. ALTERNATIVE FORMS

Although the selected five tests are to be considered the best among the five scales used in the experiment, they cannot be applied to Chinese as satisfactorily as to American children. For instance in tests 2, 3 and 5, Arabic figures are used in substitution and number-checking. Arabic figures are taught in all of the modern Chinese schools, but the children who have not attended a modern school or learned the Western arithmetic are wholly ignorant of the meaning of them. Chinese children, not of better class families in some modern city such as Shanghai, also will be greatly handicapped in performing test 1. They can hardly be expected to draw the filament of an electric bulb. They cannot place a postage stamp in its proper American position on the envelope, nor complete the drawing of a pistol, a bowling game, a phonograph or a tennis net, for these objects are rare in China. The same may be said of Test 4, the telephone, the gloves, the *ABC*, the American flag, the music scale, and so on, are most likely unknown to 99 per cent of Chinese children. Consequently, these tests cannot be applied unless alternative forms are devised. As explained in previous chapters, alternative forms have distinct advantages, besides their application to Chinese, such as the prevention of coaching and the provision of material for retesting.

In preparing the alternative forms, the criterions first adopted were strictly observed. One point was especially emphasized; namely, that the test material should be drawn from a social environment common to all people and the test should measure only those mental traits which every child has an equal opportunity to develop. This means that the test material selected should not be dependent upon any social or educational advantages. An attempt also was made to bring all of the alternative forms to yield the same result. The writer, however, cannot claim credit for such an achievement as yet, because the tests have not been tried out in China.

The first step in preparing the alternative forms was to devise a large number of test items. These were then submitted to ten graduate students originally from different parts of China. They were asked, "Is this common in your locality?" All the test items which were marked "Not common" in any of the localities were discarded. The selected test-elements were submitted to 2 Japanese, 2 Filipinos, 2 Indians, 2 Britons, and 2 Americans; and they were asked the same question. "Is this common in your country?" All those which were marked "Not common" were again discarded. These remaining from this double sifting were finally gathered and sorted into forms.

Different methods are required for placing the test-items in the different individual tests. For tests 2, 3 and 5, the selection of the symbols was made by the chance method of tossing coins. For tests 1 and 4, the pictures were arranged, by the combined judgments of three experts, according to their degree of difficulty, beginning with the easiest, and ending with the most difficult ones. The best method for arranging the tests in the order of their difficulty would be one in which the tests are given to several hundred children, with the answers scored either right or wrong, and the per cent of correct answers obtained.

In Tests 1 and 5 of the Chinese non-verbal forms, the preliminary demonstration is modified. To be uniform with the other tests, the marks and pictures to be used for the preliminary demonstration are printed at the top of each sheet. This is an improvement also because the use of a blackboard may be inconvenient or unfair.

The alternative forms thus devised cannot be claimed as the final forms. They must yet be tried out upon a large number of children, the norms for ages and grades must be computed, and the tests scaled; but judging by the results of the experiment, there is every reason to believe that the tests will prove reliable and useful.

B. STANDARDIZATION

The last step of scale construction is standardization—the obtaining of norms and scaling of the tests. In order to do this for the proposed Chinese Non-verbal Intelligence Scale, it is necessary to give it to a large number of Chinese subjects, perhaps 5000. The selected tests were only applied to about 200 pupils, very few of whom were Chinese. The devised alternative forms, furthermore, cannot be tried out in America. It was thus impossible to secure any age or grade norms to be reported here or to scale the tests. The final standardization must be done in China. However, the technique may be briefly discussed here.

1. *Norms*

The purpose of mental measurement is to reveal individual and group differences of intelligence. To perform such a function, norms or standards of achievement for different ages and possibly grades are required. We cannot, however, test all the Chinese people between certain ages and compute the average achievement of each age. This is unnecessary as well as impracticable. The obtaining of reliable norms does not require the test of every child in the country, but it is essential that the subjects selected should be in random sampling, representing the whole range of intelligence from a low degree of moron to a high degree of genius. It is also essential that the subjects should be representative of all types of social environment in different parts of the country.

Norms are more valuable when they are stable. When a norm is stable, it indicates that the subjects are selected from random sampling and the number of cases is sufficient. As a rule, the greater the number of cases taken, the more stable are the norms; certainly a norm can be claimed to be stable only when it reaches the point where the addition of new cases does not materially alter the previous determination. The safest way to tell whether the norms are stable or not is to average the scores of a varying number of cases and watch the resulting fluctuations in the average. McCall states that "when the addition of, say, 100 cases does not materially alter the previously determined norm, the norm has stabilized." [1]

Norms for both age and grade should be worked out. However, in China the age norms will be more important than the grade norms, as the grades are not uniform in the schools. Care must be taken, however, in obtaining ages to record the actual date of birth accord-

[1] McCall, W. A.: *How to Measure in Education*, p. 315.

ing to both old and new calendar, as many subjects, undoubtedly, will follow the custom of reporting ages by years although they may be born in the end of the year.[1]

2. *Scaling*

After the tests have been applied to a large number of subjects and the norms are obtained, scaling is comparatively an easy task. There are numerous methods of scaling tests. For the Chinese scale, the writer plans to adopt one or both of the two most commonly used methods—an age scale and a percentile scale.

(*a*) Age scale: The construction of an age scale merely requires the determination of stable norms. Given a norm for each age, any pupil's test-score may be transmuted into a mental age and intelligence quotient. Mental age is obtained from a comparison of the subject's performances with the standard for normal children of the same age. Let us suppose the subject tested is 10 years of age. If he can do as much as normal 10-year-old children do, the child has a mental age of 10, which in this case is normal. If he goes as far as normal 8-year-old children go, his mental age is 8. In this case, he is subnormal. In like manner a mental defective 10 years old may have only a mental age of five, and a genius of the same age may have a mental age of 13 or 14.

The intelligence quotient, often designated as I Q, is the ratio of mental age to chronological age. It is a valid expression of intelligence. On this basis of the Stanford Revision of the Binet Scale, Terman [2] suggests this classification of intelligence quotients:

I Q	Classification
Above 140	"Near" genius or genius
120–140	Very superior in intelligence
110–120	Superior intelligence
90–110	Normal, or average intelligence
80– 90	Dullness
70– 80	Border-line deficiency
Below	Feeble-mindedness

[1] According to the old custom in China, which still prevails in many portions of the country, age is reckoned in years, according to the calendar. For example, a man whose 25th birthday comes in December would be considered as already 25 years of age in the preceding January. This may be explained as resulting from the literal translation of Chinese into English. In the Chinese language, age or "sui" is expressed in the phrase "in the 25th year," whereas in America this would be translated as "25 years old."

[2] Terman, L. M.: *The Measurement of Intelligence*, p. 79.

(*b*) Percentile scale: The technique of percentile scale construction is described in detail by Pintner.[1] After the test papers have been scored, a distribution table for each test is made. The percentiles are then calculated for each test counting usually from the lower end of the table. The 25-percentile or Q_1 is that score which is found by counting one-fourth of the score. The 75-percentile is found by counting three-fourths of the scores. Similarly, the 10-percentile is found by counting one-tenth of the score, the 20-percentile by counting one-fifth of the scores and similarly for any other percentiles. After the percentiles are calculated, the percentile table for each test should be prepared. To get the mental index of any individual his percentile placement for each test is found by comparing his score with those found in the table, and then the median of these various placements is found. Similarly the mental index for the class, for the grade, and for the entire school can be found. For purposes of rough classification, Pintner has adopted the following scheme:

PERCENTILE	CLASSIFICATION
84—100	Very bright
72— 83	Bright
39— 71	Average
22— 38	Backward
0— 21	Dull

[1] Pintner, R.: *The Mental Survey*, p. 28, ff.

CHAPTER VII

THE CHINESE NON-VERBAL TESTS[1]

A. THE NATURE OF THE TESTS

The measurement of intelligence has recently become widespread in America. It has been proved very helpful in solving many administrative problems. With the hope of facilitating Chinese educational work, these tests are therefore introduced.

The tests were scientifically constructed for the measurement of mental ability. They are applicable to a large number of children at a time, who are in the Citizens' Schools or Higher Primary Schools. There are four forms, all of equal value. It is advisable to use different forms in various grades, so as to prevent coaching. The period of testing does not exceed thirty minutes. It will enable the teacher or school administrator to measure the mental ability of pupils in groups for the following purposes:

1. *Classification*. The object of classification is to divide into homogeneous groups the pupils whose needs are similar, in order that work can be more exactly adapted to them. With the application of these tests, a teacher can scientifically determine the mental ability of his pupils in a rapid and accurate manner.

2. *Promotion*. The variability in the ability to learn among children of any grade is great, and their progress is not at an equal rate. It is obviously unwise to attempt to force all of them to keep the same pace in their class work at one time. The bright pupils therefore should be promoted as fast as their ability permits them to absorb their work, or their courses of study should be enriched; while slow ones may be given more time or requirements upon them may be reduced to the minimum essentials.

3. *Provision for the Backward*. These tests may give a valuable indication of the probable causes of difficulty with troublesome backward children. Their restlessness, incorrigibility, and lack of school progress may be due to a mentality unequal to the strain of ordinary

[1] The material in this chapter is translated from the Manual of Directions.

school work. The tests may therefore indicate those who should be segregated from the normal class and given special courses of study.

4. *Vocational Guidance.* These tests will not give prognosis of fitness for specific trades or professions except along broad lines; they are selective. The test scores will show whether a child should be encouraged to take a profession or do unskilled work. For instance, it would be absurd to encourage a child whose test indicates feeble-mindedness to study medicine or one with a genius to be a riksha coolie.

Although these tests are primarily devised for the use of school children, they will be of aid to the employer in making a hasty classification of his employees, especially the unskilled laborers; and will aid the employee to find early the place for which he is best fitted.

B. INSTRUCTIONS TO EXAMINER

1. Any intelligent person who has a pleasing personality can conduct a group examination with these non-verbal tests in a reasonably satisfactory manner and obtain fairly reliable results.

2. The examiner cannot give the examination satisfactorily until he has thoroughly mastered the technique. He should try the tests out on a smaller group of children than the one to be tested and then memorize the procedure. However, he should always read the directions from the manual.

3. The room for testing should be provided with chairs and desks. It should be free from distracting noises within or without. No visitor, school authority, or pupil should be permitted to enter or leave the room during an examination unless the reason for so doing is imperative. The school administration should so arrange the place and time of testing that no one be permitted to weaken the value of tests by distracting the attention of the children in any manner.

4. Children should used pencils rather than pens. Each child should be provided with two pencils (with eraser) and the examiner should always have on hand a supply of sharpened pencils to be used if needed. If a child breaks his pencil, the examiner should supply another with entire quietness and as little loss of time to the child as is possible.

5. It is better for the examiner to remain at the front of the room

during the entire testing. He should ask the assistant, or appoint several pupils, to distribute the test papers.

6. Before the examination begins, those to be tested should be made to feel comfortable, and in an easy, contented but responsive form of mind. Every effort should be made to make the testing as informal and as much like a game as possible, yet precision and exactness in obeying all the rules that have been worked out for administering the tests is essential. Otherwise the results obtained in different schools will be untrustworthy and not comparable.

7. In a given school, children should be tested in order from the lower grade upward. So far as possible, the same examiner should give all the examinations within the school.

8. The examiner should give the directions in a clear, energetic voice. He should speak distinctly, at moderate speed and loud enough to make his voice clearly audible to all the pupils in the room. He must make sure that each step is understood by all, that they turn to the proper page when the new test is to be begun, and that they give instant obedience to his directions.

9. The directions for giving the tests should be followed literally. Avoid all impromptu directions since such variations may modify the results. Even though the directions are memorized, they should always be read from the manual in giving a test.

10. All should start and stop together. If a child comes in late or leaves the room early, or his work otherwise is interfered with, a note of the fact should be put on his paper at the time.

11. Accurate timing of the results is of great importance. Use a watch with a second hand. Have an assistant to act as timer if convenient.

12. The children must be constantly watched for copying. Every precaution against cheating should be taken, yet the manner of the examiner should not be accusing or offensive to the self-respect of the pupils.

C. DIRECTIONS FOR GIVING THE TESTS

Read: "Would you like to play a game?"

(*For pupils who can read and write*), "Before we begin I must ask you a few questions. First, I want to know your name. Please write your name at the upper right corner." (Hold up test blank and point.) (Pause.) "Have you all done that?" (Pause.) "That ls fine! Now answer all the questions on the whole page." (Pause.) "Who has not as yet finished?" (Pause.)

(*If subjects cannot read or write at all, begin here.*)

After all the test blanks are filled out, the examiner should say: "Now I want to tell you something about the game. I am going to ask you to do things for me. Some of them will be very very easy and some will be hard. You will not be expected to do all of them, but do the very best you can. You must listen carefully to what I say or you will not know what to do. After I say, 'Go,' don't ask any questions and don't look at anybody's paper but your own. When I say, 'Stop,' you are to quit work at once, even if you have not finished. If you finish before I say, 'Stop,' put your hands back of your head."

TEST I. PICTURE COMPLETION

"Now turn the first page like this." (Hold up the test blank and point.) "Here," (pointing) "is the first game. Have you all found it? That is fine." (Pointing to it.) "Now look at the pictures at the top of the page" (pointing). "There is something gone or missing from each of these pictures. What is the matter with the hand? What is left out?" (Pause.) "Yes, one finger is gone. Take your pencil and put in the finger." (Pause.) "What is the matter with the fish? What is left out?" (Pause.) "Yes, that is right. The eye is gone. Put in the eye." (Pause.) "What is the matter with the table?" (Pause.) "That is right. One leg is gone. Draw the leg on." (Pause.) "Now listen! There are other pictures on this page. None of them are finished. Everyone has something gone or left out. I want you to find out what is gone in every picture and then put it in. Ready. . . . Go!" (Time limit is three minutes.) "Stop! Hands back of your heads!"

TEST 2. EASY LEARNING

"Turn over the page and fold your book like this." (Show how to do it.) "Here is the second game" (pointing). "Have you all found it? That is fine. Look at the three boxes at the top of the page. Now watch me." (Hold up the test blank and point.) "The two marks in this box must always go together" (pointing to the first box at the top of the page). "The two marks in this box must always go together" (pointing to the second box at the top of the page), "and the two marks in this box" (pointing to the third box at the top of the page) "must always go together. Now you must put in all these boxes that have only one mark" (pointing), "the other mark that belongs with that one. Do you understand?" (In case the children do not understand, reread the directions beginning at the "Some of the boxes in the first row have already been filled in the way they should be." "When I say Go," I want you to put in each box the mark that belongs with the mark that is there. Ready. . . . Go!" (Time limit is three minutes.) "Stop! Hands behind your heads!"

Test 3. Hard Learning

"Turn over your book and fold it like this." (Show how to do it.) "Here is the third game. It is like the second one, only there are more boxes with more kinds of marks to make. Have you all seen the boxes at the top of the page?" (pointing and waiting a moment). "Now watch me. The two marks in this box" (hold up the test blank and point to the first box at the top of the page), "must always go together, the two marks in this box" (point to the second box at the top of the page), "must always go together, the two marks in this box" (point to the third box at the top of the page) "must always go together, and so on" (point to all the rest of the boxes at the top of the page). "Do you understand? That is fine. Some of the boxes in the first row have already been filled in the way they should be. When I say 'Go,' I want you to put in each box the mark that belongs with the mark which is there. Ready. . . . Go!" (Time limit is three minutes.) "Stop! Hands back of your heads!"

Test 4. Absurdities

"Now turn over the page and fold your book like this." (Show how to do it.) "Here are pictures. Every one of them has something wrong. I want you to find out what is wrong and cross it out with your pencil. Look at the first picture. What is wrong with the boy's face?" (Pause.) "Yes, the eye. Cross out the eye, because it is wrong." (Make a gesture to show how to make a cross.) "What is wrong with the bird in the next picture?" (Pause.) "Yes, the bird has two heads. Which one is wrong?" (Pause.) "That is right. Cross out the upper head." (Pause.) "What is wrong with the third picture?" (Pause.) "Yes, his foot is turned the wrong way. Cross out the foot." (Pause.) "Now listen! Mark the other pictures on the whole page in the same way. In each picture, cross out the one part that is wrong. Ready. . . . Go!" (Time limit is three minutes.) "Stop! Hands back of your heads!"

Test 5. Mark Checking

"Turn over your book like this." (Show.) "Have you all seen the rows of marks and the little boxes?" (Point.) "That is fine! I want you to find out whether the marks in each row are the same. If they are the same you are to put 'x' (make a gesture to show how the 'x' is made), in the little box at the side.. If they are different you are to put in 'o' (make a gesture to show how the 'o' is made), "in the little box at the side. Look at the ones at the top of the page. Are these first two the same? (pointing and pause). Yes, so the mark in the box is 'x.' Look at the next ones (pointing), are they the same?" (pause). "No, so the mark in the box is 'o.' The next ones are not the same, so the mark is 'o.' In the next one which should we put in, 'x' or 'o'?" (when some child gives 'x' say) "Yes, all of you put 'x' in the little box. In the next one which should we put in?" (Pause.) "Yes, 'o' is right. All of you put 'o' in the little box. Now when I say 'Go' you are to put the right mark in all the little boxes all the way down the page, on one side, and then all the way down on the other side." (Pointing) "Ready. . . . Go!" (Time limit is three minutes) "Stop! Hands back of your heads!"

(Collect the test booklets at once, not permitting any time for further work.)

D. DIRECTIONS FOR SCORING THE TESTS

Keep the test papers for each group together and score test by test in one whole set, rather than running through all the tests in each paper.

Use keys and stencil for scoring.

Where accuracy is desired, all scoring should be checked by a second scorer.

In scoring, mark the correct items by a check (√) and indicate an error by o.

When an item evidently has been corrected by the pupil, the correction is the answer to be scored.

The score for each test should be entered in the upper right-hand corner of the test paper, and encircled. When the scoring has been checked, a check mark may be made beside the circle.

Test 1

1. Score is number right.

2. Allow much awkwardness in drawing. Writing in name of missing part, or other way of indicating it, receives credit, if idea is clear.

3. Additional parts do not make items wrong, if the proper missing part is also inserted.

Tests 2 and 3

1. Score is number correct. Maximum score is 50.

2. Lay the transparent stencil on the paper. The correct symbols will appear just below the child's symbols.

Test 4

1. Score is number right.

2. Any mark that clearly indicates correctly the absurd object should be scored as correct.

Test 5

1. Score is right minus wrong number (number of items checked that are correctly checked minus number of items checked that are wrong). Pay no attention to omissions.

2. If other clear indications are used, instead of crosses and zeros, give credit.

3. If pupils give nothing but crosses or zeros the score of the test is zero.

E. TREATMENT OF RESULTS

1. When the scoring is finished the test papers should be arranged or grouped according to the age or grade of the children.

2. Then the mean for each group is calculated by adding together all the scores of the individuals in that group. The mean so obtained may be used to represent the attainment of the age or of the class.

F. CAUTION

A caution should be urged against relying too exclusively on the bald test scores as a basis for administrative action. These tests when properly administered are fairly reliable as a measure of intelligence. There is always the possibility, however, that the child for some reason may have failed to do himself full justice in the test. He may have been sick, or he may not have taken the testing seriously. There is also the possibility that the examiner, or scorer, has made a statistical error. So the results of the tests should be interpreted in the light of all such supplementary information as may be available. In the small number of cases where there is a clear disagreement between the results of the tests and other data, such as school marks, teachers' estimates, and so on, an alternative form of the test may be repeated and the scores compared. It should be especially pointed out that the tests are not a substitute for common sense on the part of teacher or principal.

CHAPTER VIII

SUMMARY AND CONCLUSIONS

1. Progressive Chinese educators who are planning to introduce the measurement movement into China are confronted with the problem of procuring suitable test material. China, with her distinctive civilization, with her numerous dialects and her lack of universal education, encounters great difficulty in the application of language tests. This study is an attempt to develop a non-verbal scale which, because of the elimination of language, environmental and educational factors, may be used either as an independent measure of general intelligence or as a supplement to a language test.

2. Instead of forming a purely Chinese test, it was decided to select the most useful elements from the best known American non-verbal tests which have already been standardized. The following tests were chosen for experimentation: The Myers Mental Measure, The Pressey Primer Scale, The Pintner Non-language Tests, The Army Beta Examination, and The Dearborn Group Tests of Intelligence (Examinations I, II, and III).

3. The selected tests were given in Public School No. 108 of the Chinese section of New York City to 401 children of Chinese, Italian, and Hebrew descent. Most of the children were Italians. Since the purpose of the study was to select the best test elements and since it was not intended to derive norms, it made no difference whether the subjects were Chinese or of any other nationality.

4. The criterion, after many trials, was decided to be a weighted composite of age, teachers' estimates, school marks, school progress and test scores. Each of these measures general intelligence, to some degree, in a different way, so their combination should be reliable. The criterion is the standard extensively used for the selection of the best test elements from the five scales.

5. By the methods of correlation, rating, partial correlation, and regression equation, the test elements of the five scales were checked against the criterion to determine their validity. The foremost

valid tests were selected to form the Chinese Non-verbal Intelligence Examination. Those so chosen are Tests 2 and 3 of the Pintner Non-language Tests, Tests 5 and 6 of Army Beta Examination, and Test 4 of the Pressey Primer Scale.

6. The tests thus selected were given to the children who the year before had taken all of the five tests. The correlation between these test scores and the final criterion was .8768, which was higher than the correlation of any one of the original examinations with the criterion.

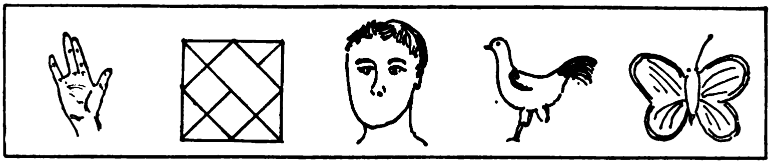
7. Although the selected tests are the best among the five scales, they cannot be applied to Chinese children as successfully as to American children, because of the unfamiliarity of the Chinese children with the objects shown or situations represented. Consequently alternative forms were prepared which are international in nature and are not influenced by schooling or civilization. The alternative forms will also prevent coaching.

8. The Chinese Non-verbal Tests so constructed are applicable to a number of children at one time in the elementary school. The period of examination does not exceed thirty minutes, but this short time will enable a teacher or school administrator to measure the native ability of the pupils as an aid in classification, promotion, provision for the backward, improvement in methods of teaching, and vocational guidance.

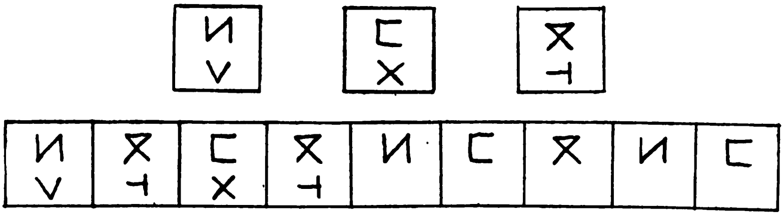
9. The norms are to be established in China and the final standardization will take place there.

APPENDICES

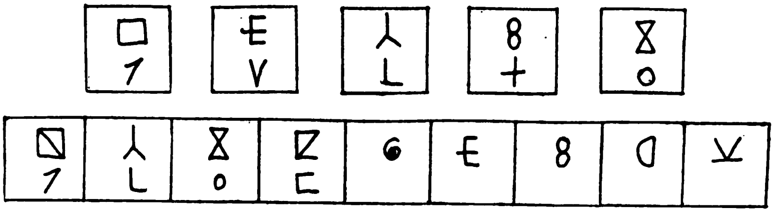
A. Sample of Form A of the Chinese Non-Verbal Intelligence Examination



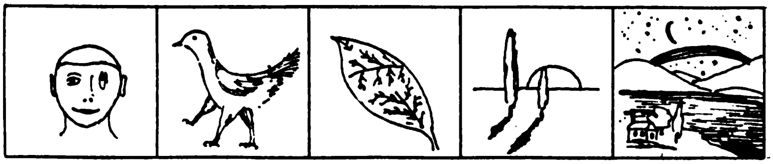
Section of Test 1. Picture Completion.



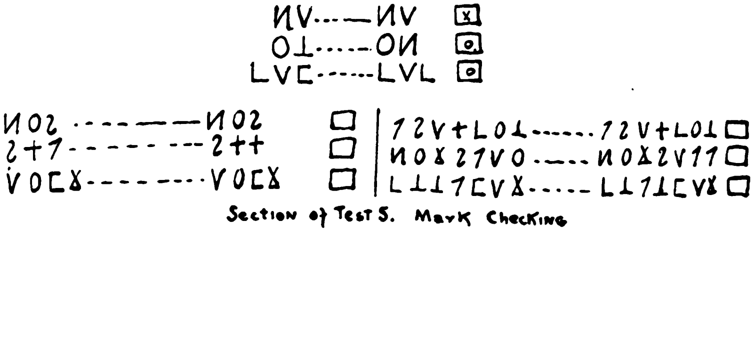
Section of Test 2. Easy Learning



Section of Test 3. Hard Learning



Section of Test 4. Absurdities



Section of Test 5. Mark Checking

B. Sample of Records Kept

The following is a sample from the original record book which is now kept in the library of Teachers College, Columbia University. All those who are interested in the full record may have access to it by communication with the proper authorities.

1.

Boys' Names	No.	Age Yr. Mo.	Nation-ality	Health	Promo-tion	School Marks	Teachers' Estimates	Grade
		1	2	3	4	5	6	7
Jo. G.	118	9 —	Italian	Teeth	Yes	B	C C B	III A
Wi. L.	119	9 —	Chinese	Teeth	Yes	B	B B B	III A
Jo. M.	120	9 —	Italian	Teeth	Yes	B	B B B	III A
De. M.	121	9 —	Italian	Tonsils	Yes	B+	A B A	III A
Ai. N.	122	9 —	Italian	Teeth	Yes	B	D D D	III A

2.

Thorndike-McCall Reading Scale			Credit Assigned to			
T Score	Reading Age	R. Q.	Age	School Marks	Teachers' Estimates	School Progress
8	9	10	11	12	13	14
38.5	88.5	89	7	7	2 2 3	5
35.0	101.0	104	7	7	3 3 3	5
37.0	107.0	106	7	7	3 3 3	5
....	....	...	7	8	4 3 4	5
29.0	90.0	89	7	7	1 1 1	5

3.

School Criterion Total	Pressey+Pintner+ 2×Myers+2×Beta+ Dearborn	Tests Total	Final Criterion (3 × Sch. Crit. + 2 × Test Total)
15	16	17	18
26	484	48	174
28	582	58	200
28	513	51	186
31	389	39	171
22	337	34	134

PRESSEY PRIMER SCALE

4.

19	20	21	22	23	24
Total Score	Test I	Test II	Test III	Test IV	Test II–IV
38	21	0	17	0	0
80	21	21	18	20	41
81	24	19	21	17	36
44	8	11	14	11	22
74	22	18	18	16	34

PINTNER NON-LANGUAGE TESTS

5.

25	26	27	28	29	30	31	32
Test Total	Test I	Test II	Test III	Test IV	Test V	Test VI	II-III
93	2	35	38	14	4	0	73
101	1	38	33	18	6	5	71
80	4	27	27	14	4	4	54
58	0	22	22	14	0	0	44
13	0	0	0	10	1	2	0

Myers Mental Measure

6.

33	34	35	36	37
Test Total	Test I	Test II	Test III	Test IV
13	1	3	7	2
29	3	13	8	5
14	2	3	4	5
14	3	9	2	0
12	3	3	3	3

Army Beta Examination

7.

38	39	40	41	42	43	44	45	46	47
Test Total	I	II	III	IV	V	VI	VII	III-IV V-VI	IV–VI
55⅔	0	7	6	15⅔	13	12	0	47	28
77⅔	9	0	6	21⅔	27	14	0	69	36
82⅔	5	0	8	25⅔	28	13	3	75	39
70⅔	8	9	8	17⅔	14	8	6	48	26
55⅔	6	0	4	16⅔	17	12	0	50	29

Dearborn Group Tests, Series I

8.

48	49	50	51
Grand Total	Exam. I	Exam. II	Exam. III
171	60	86	25
179	78	59	42
182	58	89	34
147	53	81	13
150	51	74	25

Dearborn Examination 1

9.

52	53	54	55	56	57	58	59	60	61	62	63	64	65	66	67	68	69
Total	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	XII	XIII	XIV	XV	XVII	XVIII
60	3	3	3	1	2	2	4	3	1	4	3	4	4	3	3	8	9
79	3	3	3	1	2	2	4	3	1	4	4	4	6	0	3	16	20
58	3	3	3	1	2	2	0	3	2	4	0	4	6	3	0	14	8
51	3	3	3	1	0	2	3	3	0	4	0	4	4	0	3	8	10
52	3	2	3	1	2	2	2	3	1	1	0	1	4	3	0	14	10

Dearborn Examination 2

10.

70	71	72	73	74	75	76	77
Total	I	II	III	IV	V	VI	VII
86	9½	15	21	14	11	12	4
59½	8½	15	15	2	9	9	1
89½	9½	15	24	11	9	13	0
81½	9½	12	24	11	14	11	0
74	10	15	15	10	8	11	5

Dearborn Examination 3

11.

78	79	80	81	82
Total	I	II	III	IV
25	14	0	11	0
42	15	0	4	23
34	14	0	9	11
13	3	0	2	8
25	9	0	0	16

TESTS COMBINATIONS

12.

83	84	85	86
Composite A	Composite B	Beta III-IV-V VI Pressey II-IV	Beta IV-VI Pressey II-IV
104	104	69	36
132	152	116	80
93	110	84	62
94	105	72	51
36	52	67	53

TESTS COMBINATIONS

13.

87	88	89
Pintner II-III Beta VI	Dearborn I I-VI	Dearborn I VII-XV
87	14	25
84	14	29
62	14	29
66	12	22
8	14	21

RE-TESTING

14.

90	91	92	93	94	95
Total	I	II	III	IV	V
166	50	48	33	14	21
170	44	50	35	18	23
143	50	49	15	8	21
116	23	49	11	13	20
116	49	38	6	8	15

BIBLIOGRAPHY

Ayers, L. P. "The Binet-Simon Measuring Scale for Intelligence: Some Criticisms and Suggestions." *Psychological Clinic*, Vol. V (1911), pp. 187–96.

Binet, A. and Simon, T. "Le développement de l'intelligence chez les enfants." *L'Annèe psychologique*, 14 (1908), pp. 1–94.

Binet, A and Simon, T. "L'intelligence des imbéciles." *L'Annèe psychologique*. (1909), pp. 1–47.

Chen, H. C. "Educational Research in China." *Journal of Educational Research* (May, 1921), Vol. III, No. 5, p. 394.

Chen, H. C. and Liao, C. S. *Mental Tests*. Commercial Press, Shanghai, China (1922).

Dearborn, W. F. *The Dearborn Group Tests of Intelligence*. J. B. Lippincott Co. (1920).

Fretwell, E. K. *A Study in Educational Prognosis*. Teachers College Contributions to Education, No. 99 (1919).

Goddard, H. H. "The Binet-Simon Measuring Scale of Intelligence Revised." *Training School Bulletin*, Vol. VIII (1911), pp. 56–62.

Healy, W. and Fernald, G. M. "Tests for Practical Mental Classifications." *Psychological Monographs*, Vol. No. 2, 54 (1911), pp. 4–5.

Herring, John P. "Significance of Certain Elements in Intelligence Examination." Unpublished Ph.D. dissertation, Columbia University (1921).

Kelley, T. L. *Educational Guidance*. Teachers College Contributions to Education, No. 71 (1914).

Kelley, T. L. "Table to Facilitate the Calculation of Partial Coefficient of Correlation and Regression Equation." *Bulletin of the University of Texas* (1916), No. 27.

Knox, H. A. "A Scale Based on the Work at Ellis Island for Establishing Mental Defects." *Journal of the American Medical Association*, Vol. LXII (March 7, 1914), pp. 741–747.

Kuhlman, F. "A Revision of the Binet-Simon System for Measuring the Intelligence of Children." *Journal of Psycho-Asthenics*, Monograph Supplement, No. 1 (1912), p. 14.

McCall, W. A. *How to Measure in Education*. Macmillan Co. (1922).

Myers, Caroline E. and Garry C. "A Group Intelligence Test." *School and Society* (1919), Vol. 10, pp. 355–360.

Peking Teachers College Weekly, No. 132 (Sept. 11, 1921, p. 3.)

Pintner, R. "A Non-language Group Intelligence Tests." *Journal of Applied Psychology*, Vol. III (Sept., 1919).

Pintner, R. *The Mental Survey*. D. Appleton Co. (1918).

Pintner, R. and Patterson, D. G. "The Binet Scale and the Deaf Child." *Journal of Educational Psychology*, Vol. VI (1915), pp. 202 ff.

PRESSEY, S. L. and PRESSEY, L. W. "Cross-out Tests." *Journal of Applied Psychology*, Vol. 3 (1919), pp. 143–150.

PYLE, W. H. "A Study of the Mental and Physical Characteristics of the Chinese." *School and Society*, Vol. VIII, No. 132 (August 31, 1918), pp. 264-269.

STERN, W. *The Psychological Methods of Testing Intelligence.* Translated by G. M. Whipple. Warwick and York, Baltimore (1914).

TERMAN, LEWIS M. *The Measurement of Intelligence.* Riverside Textbooks in Education. Houghton Miffin Co. (1918).

THORNDIKE, E. L. "A Standard Group Examination of Intelligence Independent of Language." *Journal of Applied Psychology*, Vol. III, No. 1 (March, 1919), pp. 13–32.

THORNDIKE, E. L. *Mental and Social Measurements.* Teachers College, Columbia University (1919).

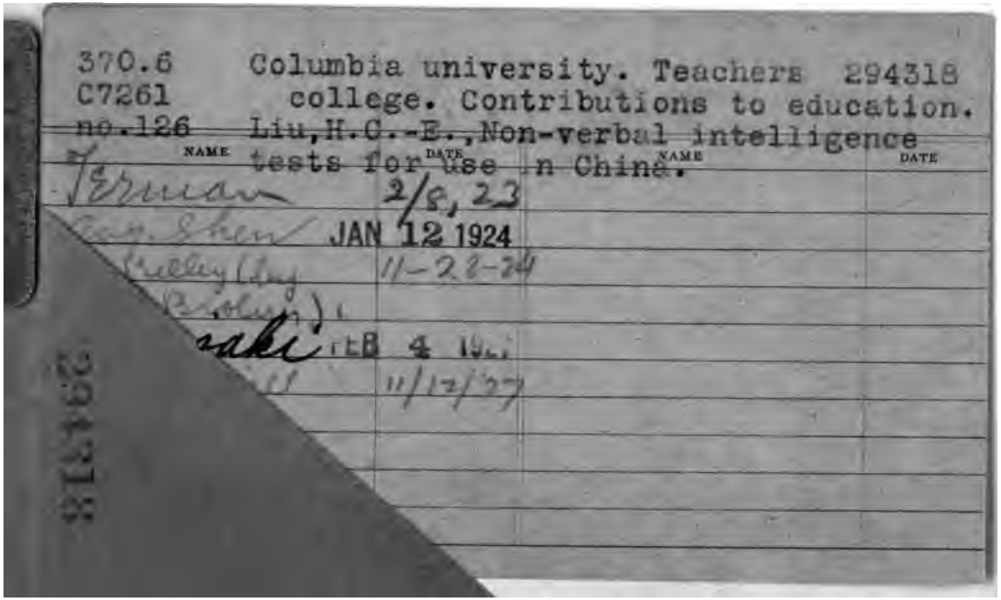
WALCOTT, G. D. "The Intelligence of Chinese Students." *School and Society*, Vol. 11 (1920), pp. 474–480.

WALLIN, J. E. *Experimental Studies of Mental Defectives: a Critique of the Binet-Simon Tests.* Warwick and York, Baltimore (1912).

YERKES, R. M. "Psychological Examining in the United States Army." *Memoirs of the National Academy of Science*, Vol. XV (1921).

YOAKUM, C. S. and YERKES, R. M. *Army Mental Tests.* Henry Holt Co. (1920).

YERKES, R. M., BRIDGES, J. W. and HARDWICK, P. S. *A Point Scale for Measuring Mental Ability.* Warwick and York, Baltimore (1912).



370.6 Columbia university. Teachers 294318
C7261 college. Contributions to education.
no.126 Liu, H. C.-E., Non-verbal intelligence
tests for use n China.

NAME	DATE	NAME	DATE
Terman	2/8, 23		
[illegible] Shen	JAN 12 1924		
Kelley (Sug	11-28-24		
(Brolish)).			
saki	FEB 4 192		
"	11/12/27		