

MEASURING INTELLIGENCE

A GUIDE TO THE ADMINISTRATION OF THE NEW
REVISED STANFORD-BINET TESTS OF INTELLIGENCE

BY

LEWIS M. TERMAN

*Professor of Psychology
Stanford University*

AND

MAUD A. MERRILL

*Associate Professor of Psychology
Stanford University*



HOUGHTON MIFFLIN COMPANY

BOSTON • NEW YORK • CHICAGO • DALLAS
ATLANTA • SAN FRANCISCO

The Riverside Press Cambridge

EDITOR'S INTRODUCTION

TWENTY-ONE years ago Terman's *The Measurement of Intelligence* was issued as the tenth volume in the Riverside Text-books in Education Series, and in introducing it to the public the editor of the Series made, in part, the following statement, all of which seems pertinent to the present edition:

The present volume appeals to the editor of this series as one of the most significant books, viewed from the standpoint of the future of our educational theory and practice, that has been issued in years. Not only does the volume set forth . . . the large importance for public education of a careful measurement of the intelligence of children, but it also describes the tests which are to be given and the entire procedure of giving them. In a clear and easy style the author sets forth scientific facts of far-reaching importance, facts which it has cost him, his students, and many other scientific workers, years of patient labor to accumulate. . . .

The educational significance of the results to be obtained from careful measurements of the intelligence of children can hardly be overestimated. Questions relating to the choice of studies, vocational guidance, schoolroom procedure, the grading of pupils, promotional schemes, the study of the retardation of children in the schools, juvenile delinquency, and the proper handling of sub-normals on the one hand and of gifted children on the other — all alike acquire new meaning and significance when viewed in the light of the measurement of intelligence as outlined in this volume. As a guide to the interpretation of the results of other forms of investigation relating to the work, progress, and needs of children, intelligence tests form a very valuable aid. More than all other forms of data combined, such tests give the necessary information from which a pupil's possibilities of future mental

growth can be foretold, and upon which his further education can be most profitably directed.

The publication of this revision and extension of the original Binet-Simon scale for measuring intelligence, with the closer adaptation of it to American conditions and needs, should mark a distinct step in advance in our educational procedure. It means the perfection of another and a very important measuring-stick for evaluating educational practices, and in particular for diagnosing individual possibilities and needs. Just now the method is new, and its use somewhat limited, but it is the confident prediction of many students of the subject that, before long, intelligence tests will become as much a matter of necessary routine in schoolroom procedure as a blood-count now is in physical diagnosis. That our schoolroom methods will in turn become much more intelligent, and that all classes of children, but especially the gifted and the slow, will profit by such intellectual diagnosis, there can be but little question....

The original volume and its scales have now been before the public for twenty-one years, and the sales, not only in the United States, but in England and other countries comprising the British Commonwealth of Nations, have been exceedingly gratifying to both author and publishers.

Inspired by a desire not only to perfect and extend the scales, but to make them still more useful by preparing a second scale that should be equivalent in range, difficulty, reliability, and validity, Dr. Terman began, ten years ago, a complete revision and extension of the original Stanford-Binet scales, basing the revision and standardization upon larger and more representative groups. In this undertaking Dr. Merrill joined as collaborator. The work proved of greater magnitude and more laborious than had at first been expected, but now, after ten years of painstaking research, two new and equivalent scales, each more extensive than the original both in range and in number of tests, and each providing for greater objectivity in

EDITOR'S INTRODUCTION

scoring, are at last ready for public use and are here so offered. While the original Stanford-Binet scale remains as valid as before, the new scales, with their wider range and with the advantages each offers for retesting by the other and as a safeguard against coaching, and with their far more accurate standardization, present so much larger possibilities for usefulness that the old scale probably will soon be entirely superseded by the new.

An examination of the manuscript of the present volume leads the editor again to state that the work in its new form, as in the old, represents a distinct contribution to educational procedures, and to predict for the revision a usefulness even greater than that enjoyed by the original edition.

ELLWOOD P. CUBBERLEY

PREFACE

THE Stanford revision of the Binet-Simon intelligence scale has had a wider field of usefulness than anyone could have foreseen at the time of its publication in 1916. It has become the standard clinical method for the evaluation of intellectual status and is used, not only in clinical practice, but also as a tool of research with a wide variety of subjects, including defectives, delinquents, the retarded, the gifted, the normal, and the psychopathic.

In view of the numerous and important services which such a system of tests is called upon to perform, it has seemed to us desirable to make the new revision as thoroughgoing as possible. We have accordingly provided two scales instead of one, have extended them so as to afford a more adequate sampling of abilities at the upper and lower levels, have defined still more meticulously the procedures for administration and scoring, and have based the standardization upon larger and more representative populations. Besides these major improvements, many others of less crucial importance have been incorporated.

Although the task of revision has required ten years for its completion, and has been carried through with constant awareness of the sources of error involved in test construction, nothing approximating perfection can be claimed for the results. The tools of psychology, particularly those dealing with the more complex mental processes, belong to an entirely different order of precision from those employed by the physical scientist. So far as one can now see, they always will. Apart from possible criticisms of the general plan on which the scales have been constructed, there are doubtless minor imperfections which only extended use will disclose. Even so, it is hoped

PREFACE

that this revision of the Binet method will long provide a common standard by which to gauge the intellectual level of human subjects from early childhood to the end of life's span.

In the course of so extensive an undertaking the authors have naturally incurred a heavy load of indebtedness to many persons: to the Social Science Research Council of Stanford University for its generous support of the project; to our highly competent and devoted assistants, Helen Campbell, Helen Green, Lois Kulmann, Barbara Mayer, Margaret Murray, Melita Oden, and Wilhelmina Warkentin, who assisted in gathering the standardization data; and to the school administrators and teachers throughout the country whose unanimous and whole-hearted co-operation greatly facilitated the testing program which provided the fundamental data. Miss Mayer and Mrs. Oden, in addition to their work as field examiners, have carried through the arduous task of scoring all the tests and getting out the major part of the statistical data. Grateful acknowledgment is made of the expert assistance rendered by Dr. Paul Bottenwieser and Dr. Quinn McNemar in the statistical treatment of results, and of the help given by Mrs. Alice McAnulty Horn in formulating the plans for Hollerith coding.

LEWIS M. TERMAN
MAUD A. MERRILL

CONTENTS

PART I — THE NEW REVISION

CHAPTER	PAGE
I. ESSENTIAL FEATURES OF THE REVISION	3
II. DEVELOPMENT AND STANDARDIZATION OF THE SCALES	7
Preliminary Selection of Tests	7
Formation and Use of the Provisional Scales	10
The Selection of Subjects	12
Derivation of the Final Scales from the Standardization Data	21
Units of Measurement	24
Adult Mental Age	29
The Abbreviated Scales	31
III. STATISTICAL ANALYSIS OF SCORES	33
Accuracy of Standardization	33
I.Q. Variability in Relation to Age	38
Standard Scores	42
Practice Effects	43
Reliability of the Scales and Probable Error of Scores	44
Scores in Relation to Socio-Economic Status of Parents	48
Equivalence of Old Stanford-Binet Scores to New Scales	50
IV. GENERAL INSTRUCTIONS	52
Essential Requirements of a Valid Test	52
Adherence to Standard Procedure	52
The Importance of Rapport	56
The Appraisal of Responses	58
Administering the Tests	60
The Testing of Pre-School Children	68

PART II — INSTRUCTIONS FOR ADMINISTERING FORM L

YEAR	PAGE	YEAR	PAGE
II	75	IX	102
II-6	78	X	105
III	80	XI	108
III-6	83	XII	110
IV	86	XIII	113
IV-6	89	XIV	116
V	91	Average Adult	120
VI	94	Superior Adult I	124
VII	97	Superior Adult II	127
VIII	99	Superior Adult III	130

CONTENTS

PART III — INSTRUCTIONS FOR ADMINISTERING FORM M

YEAR	PAGE	YEAR	PAGE
II	135	IX	160
II-6	137	X	163
III	140	XI	166
III-6	142	XII	168
IV	145	XIII	170
IV-6	148	XIV	174
V	151	Average Adult	177
VI	153	Superior Adult I	181
VII	156	Superior Adult II	183
VIII	158	Superior Adult III	186

PART IV — SCORING STANDARDS FOR FORM L

INTRODUCTION	191		
YEAR	PAGE	YEAR	PAGE
II	193	IX	242
II-6	198	X	254
III	200	XI	259
III-6	203	XII	267
IV	209	XIII	272
IV-6	213	XIV	277
V	217	Average Adult	280
VI	222	Superior Adult I	289
VII	226	Superior Adult II	292
VIII	233	Superior Adult III	299
		Vocabulary	302

PART V — SCORING STANDARDS FOR FORM M

YEAR	PAGE	YEAR	PAGE
II	327	IX	362
II-6	330	X	376
III	332	XI	383
III-6	334	XII	386
IV	340	XIII	389
IV-6	342	XIV	392
V	346	Average Adult	397
VI	348	Superior Adult I	403
VII	351	Superior Adult II	405
VIII	355	Superior Adult III	408
APPENDIX: I.Q. TABLES			413
INDEXES			451

LIST OF FIGURES

FIGURE	PAGE
Fig. 1. Distributions of Composite L-M I.Q.'s of Standardization Group	37
Fig. 2. Distributions of Composite L-M I.Q.'s at Three Age Levels	41
Fig. 3. Scatter Plot of Correlation Between L and M I.Q.'s at C.A. 7	45
Fig. 4. Distributions of I.Q.'s of Rural and Urban Groups . . .	49
Copying a Circle (L III, 5)	202
Drawing a Cross (L III-6, Alt.)	208
Picture Completion: Man (L IV, 3)	211
Copying a Square (L V, 4)	220
Copying a Diamond (L VII, 3)	231
Paper Cutting I (L IX, 1)	243, 245
Memory for Designs (L IX, 3)	250-253
Plan of Search (L XIII, 1)	273
Paper Cutting II (L S.A. III, 4)	301
Picture Completion: Bird (M IV-6, 4)	344
Memory for Designs I (M IX, 1)	364-367
Memory for Designs II (M XII, 1)	387
Plan of Search (M XIII, 1)	390

LIST OF TABLES

TABLE	PAGE
1. Occupational Classification of Fathers of Revision Cases . . .	14
2. Age-Grade Distribution of Subjects in School	17
3. Birthplace of Parents	21
4. Correction Table for Higher Chronological Ages	31
5. I.Q. Means by Age for the Standardization Group . . .	35
6. Composite L-M I.Q. Means Adjusted for 1930 Census Frequen- cies of Occupational Groupings	36
7. I.Q. Variability in Relation to Age	40
8. I.Q. Equivalents of Standard Scores	42
9. Practice Effects in Relation to Age	43
10. Average Difference Between L and M I.Q.'s	46
11. Errors of Measurement and Derived Reliabilities for Various I.Q. Levels	46
12. Mean I.Q.'s (L-M Composite) According to Father's Occupation	48
13. Correspondence Between Stanford-Binet I.Q.'s and L-M Com- posite I.Q.'s	50

PART I
THE NEW REVISION

CHAPTER I

ESSENTIAL FEATURES OF THE REVISION

THE major faults of the original Stanford-Binet scale have long been recognized. Although affording a satisfactorily valid and reliable measure over a fairly wide intermediate range, it was especially defective at both extremes. Abilities below the mental level of four years or above that of the average adult were very inadequately sampled. In the range from five to ten years the standardization was surprisingly correct, considering the rather small number of subjects on which it was based, but above ten it yielded scores that were progressively too low. A number of tests in the scale were unsatisfactory because of low validity, difficulty of scoring, susceptibility to coaching, etc. The instructions both for administration and scoring in numerous instances lacked the precision which is necessary to insure objectivity and comparability of results. Finally, one of the severest limitations to the usefulness of the scale was the fact that no alternative form was available for use in retesting or as a safeguard against coaching.

In the revision here offered we have provided two scales which differ almost completely in content, but are mutually equivalent with respect to difficulty, range, reliability, and validity. The scales are designated as Form L and Form M. In content Form L bears greater resemblance to the original Stanford-Binet, but neither form can be recommended above the other. Both, we believe, are relatively free from the grosser faults of the old scale. They cover a far wider range, they are more accurately standardized throughout, the tests provide a richer sampling of abilities, and the procedures have

been more rigidly defined. On the whole they are somewhat less verbal than the old scale, especially in the lower years.

The revision utilizes the assumptions, methods, and principles of the age scale as conceived by Binet. There are of course other systems of tests which are meritorious, but for the all-round clinical appraisal of a subject's intellectual level the Binet type of scale has no serious rival. It is not merely an intelligence test; it is a method of standardized interview which is highly interesting to the subject and calls forth his natural responses to an extraordinary variety of situations. The arrangement of the tests in year groups makes the examination more interesting to the examiner by enabling him to grasp the evidence as it comes in. There is a fascination in the use of an age scale that does not fade out with experience. Each examination is a new adventure in which every step is interesting and meaningful. The variety provided by the ever-changing tasks insures the zestful cooperation of subjects and is at the same time based upon what we believe to be sound psychological theory. It is a method which, to paraphrase an oft-quoted statement by Galton,¹ attempts to obtain a general knowledge of the capacities of a subject by the sinking of shafts at critical points. In our revision we have greatly increased the number of shafts and have sunk them at points which wider experience with tests has shown to be critical.

The scale devised by Binet contained 54 tests, and the first Stanford revision increased the number to 90. Each form of the new revision contains 129 tests. Below the five-year level tests are now located at half-year intervals, the gaps which existed at years eleven and thirteen have been filled, and the scale has been given more top by the addition of two supplementary superior adult levels. In the selection of tests we

¹ In a footnote to an article by James McKeen Cattell, "Mental Tests and Measurement," *Mind*, 1890, 15, 373.

have tried to correct such inadequacies of the old scale as its too verbal character at the lower levels and its too great dependence upon rote memory at the upper. For the younger subjects the scale has been made incomparably more interesting and also more valid by the liberal use of diminutive objects, brightly colored cubes, wooden beads, and other attractive materials. In general, however, the content of the new scales resembles that of the old and includes such well-known tests as comprehension, absurdities, word-naming, drawing designs, memory for digits, giving differences and similarities, defining abstract terms, etc.

Our efforts to increase the number of non-verbal tests were successful chiefly at the lower levels. Like other investigators we have found that it is extremely difficult to devise non-verbal tests for the upper levels which satisfy the requirements of validity, reliability, and time economy. At these levels the major intellectual differences between subjects reduce largely to differences in the ability to do conceptual thinking, and facility in dealing with concepts is most readily sampled by the use of verbal tests. Language, essentially, is the shorthand of the higher thought processes, and the level at which this shorthand functions is one of the most important determinants of the level of the processes themselves.

One of the important aims of the revision was to secure greater objectivity of scoring. Where judgment is involved in evaluating responses to an item, definite principles and classified illustrations have been given to guide the examiner. Ease and objectivity of scoring have in fact often played a crucial rôle in the selection and rejection of test items. The part played by subjective judgment cannot be wholly eliminated from a test of the Binet type, but we have tried to bring it as near as possible to the irreducible minimum.

Hardly less important than the selection of suitable tests has been the selection of subjects for use in the standardiza-

tion of the scales. We have devoted more than ordinary effort to secure a representative sampling of the white child population in the United States between the ages of two and eighteen years. Besides increasing the number of subjects tested to 100 at each half-year level below six, to 200 at each age between six and fourteen, and to 100 at each age from fifteen to eighteen, we have made a stubborn attempt to avoid sampling errors inherent in age, grade location, nationality, and geographical distribution. We do not flatter ourselves that we have been entirely successful, but our data represent a much closer approximation to an unbiased sampling than has heretofore been attained in the standardization of any scale for individual examining. The fact that the same subjects were used in the standardization of Form L and Form M has made it possible to guarantee almost perfect equivalence of the scores yielded by the two scales.

For reasons elsewhere set forth (page 24 ff.) we have retained both M.A. and I.Q. scores. We have provided, however, a table whereby the latter (and, indirectly, the former) may be readily converted into standard scores.

CHAPTER II

DEVELOPMENT AND STANDARDIZATION OF THE SCALES

PRELIMINARY SELECTION OF TESTS

WORK on the revision was begun with a survey of the literature on the old Stanford-Binet and a study of every kind of intelligence test item that had been used or suggested. The search for suitable material yielded thousands of test items, some of them of unknown value and most of them of unknown difficulty. The first principle of sifting was to give preference, other things equal, to types of test items that experience had shown to yield high correlations with acceptable criteria of intelligence. Such items were assembled in as great variety as possible and with special attention to promising types of non-verbal tests. Practical considerations which had to be taken into account included ease of scoring, appeal to the subject, time requirement, and convenience of administration. For one or another of these reasons a large number of otherwise excellent tests had to be rejected. Prominent among tests which have universally proved their worth are analogies, opposites, comprehension, vocabulary, similarities and differences, verbal and pictorial completion, absurdities, drawing designs from copy and from memory, memory for meaningful material and for digits, etc.

General preliminary researches on particular types of tests and special problems of method were undertaken by qualified graduate students under our direction in the laboratory. Two of these ¹ which dealt with scoring methods for the vocabulary test will be described in connection with a discussion of that

¹ Cf. Part II, p. 303.