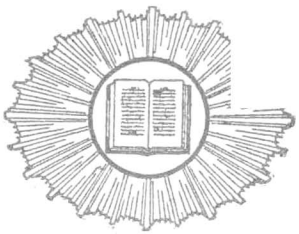
MEASUREMENT AND ADJUSTMENT SERIES

EDITED BY LEWIS M. TERMAN

INTERPRETATION OF EDUCATIONAL MEASUREMENTS

BY TRUMAN LEE KELLEY, PH.D.

Professor of Education and Psychology
Stanford University



WORLD BOOK COMPANY

Yonkers-on-Hudson, New York
and Chicago, Illinois

PREFACE

THE claims put forward for standardized intelligence and educational tests extend from the cradle to the grave. They have been mentioned seriously in connection with the selection of children for adoption and in choosing life partners. They have been charged with undermining democracy and have been hailed as of the greatest aid in solving the complex social problems of present times. It is my thesis that these instruments are potent for good if intelligently used by honest, capable, and socially minded counselors, and it is the purpose of this book to offer certain guides in the interpretation of test scores and to make explicit the errors involved — all with a view to a more sane, a more widespread, and at the same time a more penetrating use of such measures.

The most radical departures from the treatments of earlier texts dealing with mental measurements are, first, a study of achievement and intelligence measures in their mutual relationships and not of either the one or the other separately; second, an emphasis upon measures of reliability and an attempt to determine the trustworthiness of each and every conclusion reached; and third, the publication of the ratings for general excellence for purposes of individual measurement and diagnoses of all the well-known intelligence and educational tests. I am deeply indebted to the judges, Drs. Raymond Franzen, Frank N. Freeman, William A. McCall, Arthur S. Otis, Marion R. Trabue, and Martin J. Van Wagenen, who have so kindly provided me with their opinions. I believe I can speak for a great many and say to these judges that they have rendered a great service to perplexed school men and women by thus making known their individual appraisals of tests. A correspondingly great service has been rendered by authors and others who have so willingly coöperated in supplying measures of reliability of tests. In this

connection I am particularly indebted to Dr. G. M. Ruch for reliability data drawn from his personal files, to Miss M. Alice Cronin for data reported in a master's thesis at Stanford University, and to Dr. G. M. Ruch and Mr. G. D. Stoddard for the extensive data which they have incorporated in their recent work, *Tests and Measurements in High School Instruction.* I am indebted to my colleagues, Dr. Harold Hotelling, for a suggestion followed in Section 5 of Chapter VIII, and Dr. Walter R. Miles, for his counsel in connection with the discussion of Chapter V, dealing with mental types.

That this text presents to the reader more problems than it solves is perhaps merely a sign of the youth and vitality of a movement which I believe is destined to revolutionize the human relationship problems of society.

TRUMAN L. KELLEY

STANFORD UNIVERSITY

EDITOR'S INTRODUCTION

It can no longer be doubted that the recent development and widespread adoption of standard tests for measuring pupil ability and pupil achievement marks the beginning of a new epoch in the history of educational practice. Youthful as the movement is, we have already passed well beyond the stage of question and debate as to the usefulness of mental and achievement tests when they are employed with a due regard for their acknowledged limitations. Unfortunately not all of these limitations are sufficiently well known to the teachers and principals who use tests. Some of them, in fact, are not so well known as they should be even to directors of educational research and to other officers who are charged with the planning and administration of measurement programs in the schools.

The benefits that may come to the individual child from test results correctly interpreted are so real and important, and these benefits are so greatly reduced when the interpretation is incorrect or otherwise faulty, that the established facts regarding the reliability, validity, and practical significance of test scores deserve the most careful study. The editor believes that before many years considerable formal instruction along this line will be regarded as a necessary part of the training of all teachers. Certainly the kind of training here referred to will be materially facilitated by Professor Kelley's admirable textbook, which is really the first of its kind. Earlier books dealing with educational measurements have been for the most part either descriptive and general or else chiefly statistical in nature. There has been great need for a text which would explain and illustrate the application of sound statistical procedure in the interpretation of test scores for purposes of pupil classification and educational guidance. The editor confidently believes that Professor Kelley's *Inter-*

pretation of Educational Measurements will meet this need. Both by his acknowledged leadership in the field of statistics and by his wide experience in the use of tests, the author is ideally fitted for his task. His treatment of the subject throughout is masterly and vigorous.

It can hardly be expected that either the novice or the so-called expert in educational measurements will always find himself in complete agreement with the author, a fact which perhaps enhances rather than limits the value of the work for textbook purposes. It is thought-provoking and challenging. At the same time the author's objectivity and freedom from bias will be evident to all. It would matter little if some should feel that Professor Kelley has underrated the usefulness of intelligence tests or the practical value of the achievement quotient technique. One who disagrees with the author on these questions, or any other, feels challenged to justify his dissent by careful reëxamination of the facts and arguments. Whether one ends by agreeing with the author or not, the main purpose of the book has been served — one's sensitivity to the existence of the ubiquitous probable error has been heightened.

Although the keynote of this book is the universality of error in our educational measurements, its tone is never one of discouragement with reference to the practical value of the test movement. Quite the reverse. When we become as conscious of the probable error as Professor Kelley would have us, our tests are certain to undergo rapid and marked improvements. The first step in progress will be to admit that for purposes of individual diagnosis, the majority of our tests are of questionable value. Chapter IV, on "The Measurement of Individual Achievement," and Chapter V, on "The Determination of Individual Idiosyncrasy," are of outstanding value. Indeed, in the judgment of the editor, these chapters are classics hardly to be matched in the litera-

ture of educational measurements. For reference purposes Chapters IX and X are well-nigh invaluable, for there is no other source giving similar information. The temerity of the author in herein presenting ratings of tests for general merit as instruments of individual measurement is surely justified by the names of the judges contributing them. The ratings are unquestionably based upon a wide knowledge of the technique of mental measurement and of the needs of school men and counselors.

This book will doubtless find a wide field of usefulness as a text in teachers' colleges and universities and as a *vade mecum* for school principals, school counselors, and research directors in the daily interpretation and use of test results.

LEWIS M. TERMAN

CONTENTS

INTERPRETATION OF EDUCATIONAL MEASUREMENTS

CHAPTER ONE

Historical Survey of Mental Measurement

1. Sources. The origins of the test movement as applied to mental capacity are lost in the distant past. We can find in the initiation ceremonies of primitive and savage peoples tasks involving mental as well as physical prowess, and we have in early Greek history mention of a very momentous mental test. In the year 413 B.C. some seven thousand survivors of the ill-fated Athenian army in Sicily were thrown into the quarries near Syracuse, and it is recorded that in many cases their very lives and their release from the agonies of their imprisonment depended upon their ability to repeat verses of Euripides. Let the candidate trembling before a college entrance examination of today contemplate the nerve strain of this Sicilian mental test and be happy that in the present generation the results, fail or pass, of mental testing are beneficent and directed to his individual good.

2. Written examinations. Even the formal setting of written examinations dates back centuries — certainly for more than thirteen centuries in China. Probably, of the cultures still thriving, the Chinese has the first claim to being considered the mother of the achievement test. The eagerness with which China welcomes modern improvements in test procedure and the facility and rapidity with which she adjusts them to her own tongue and requirements shows that hers is still a very fertile and congenial soil.

3. Diverse and mingled origins. The writer will not attempt a historical account covering the early origins of the modern movement, nor even its more recent developments. Any claim to having done this in a brief account

would be more misleading than otherwise, because almost innumerable strands have been woven together in the creation of our present test products. Klemm (1914, page 218), writing in 1910, states: "It is certain that there is not one of the methods of psychical measurement that did not exist in its broad outlines before the time of Fechner. Yet it was only through him that these methods became a recognized part of experimental psychology. Even the concept of the psychical measure is much older than Fechner." There is even greater difficulty at the present time in tracing movements because there are now so many contributors in the field of mental measurement that it is generally hazardous to say that it is only through a certain one that a specific procedure has been handed on. The writer will, then, at most attempt to gather up only a few strands and mention a few names and movements that would be found in any adequate historical study of test development.

If in our strenuous and frequently uncritical attempts to improve upon the past we pause long enough to ask what are the concepts that seem to be the most dependable, that have most firmly stood the test of time, and that offer the greatest promise in the synthesis, analysis, and general understanding of human character, we shall probably be struck by the number of things that we use quite unconsciously, but which have been acquired by the arduous labors of those who have preceded us. To give a simple illustration:

"John's intelligence quotient is 110." We take this as a starting point for further reasoning, but let us for a moment deliberate upon it. At least the following things are implicit in the statement:

1. There is such a thing as general intelligence.
2. On the average it increases with age; so we reach the concept "mental age."
3. General intelligence is in fact quantitative, even though

it may manifest itself at different ages in acts which at first sight seem to be qualitatively different. Thus numerical measures may, with correctness, be assigned to measures of intelligence and of mental age, and these may be manipulated in an algebraic and arithmetical manner.

4. General intelligence is not merely a function of chronological age.

5. There is a valuable concept corresponding to the quotient of mental age and chronological age.

If we examine more closely, we shall find still other things tacitly agreed to:

6. The average is a particularly valuable point of reference, and it has exceptional stability.

7. People differ greatly in mental ability.

Some of these are deeply rooted concepts, but not one of them is a part of our original nature. Each has been acquired. Each has a social history which it is profitable to study, for, as is very common, the originator and early user of a concept is commonly more keenly aware of its limitations than later followers.

4. General intelligence. The writer does not know to whom the concept "general intelligence" first presented itself. It was undoubtedly a very common concept long before any one thought of measuring intelligence in a numerical manner. The numerical treatment of different evidences of intelligence seems to have been a consequence of Binet's [1] experimental and analytical approach, and not even in his own mind to have preceded it. We thus find Binet and Simon verbally proclaiming many discrete functions, "judgment," "memory," "sensorial intelligence," etc., but actually throwing all of these together in their "mental age" measure. Terman, in the Stanford Binet, does the same,

[1] Binet and Simon (1908), and also several other articles by the same authors in *L'Année Psychologique*, Vols. XI-XVII, especially Vol. XI (1905).

though, as he seems to lean logically toward Spearman's single-general-mental-function view, this does not carry with it the inconsistency found in Binet and Simon. In other words, the differences which Binet noted as being concomitant with age differences appeared to him as qualitative differences. The composite mental-age concept which is commonly thought of as Binet's most important contribution seems, as pointed out by Spearman (1923), to be one whose logical implications Binet himself did not appreciate. Goddard (1911) in this country early made a thoroughgoing and systematic use of " mental age."

That general intelligence is in fact quantitative, even though the characteristics manifested in varying situations are seemingly different, is a concept that Spearman has ably presented and has defended for the last two decades. In fact, he and others who agree with his philosophy are the only persons who logically defend the use of widely varying measures as being measures of a single intellectual function.

That intelligence is in part a function of other things than age is not recognized in the practice of the Church, dealing with communion, or in the laws of the land concerning franchise, the age of consent, compulsory or part-time education, etc. It may be that the reason for this lies not so much in a common failure to recognize individual differences in intelligence which are independent of age as in the popular belief that such differences cannot be measured. As the laws of the country today reflect the genius of an earlier generation, so when the leaders of the present day have become revered memories whose crude methods and mistakes cause not ire but amusement, and when Army Alpha has taken its place with Magna Charta, then regulation based upon individual mental differences not correlated with age will be a commonplace in law and custom. But to return to the past.

5. The intelligence quotient. Stern (1914) in 1912 was the first to use in print the term " mental quotient," meaning thereby the mental age divided by the chronological age. Bobertag (1912) also suggested such use in 1912. Kuhlmann (1913) independently, in the spring of 1912, hit upon the same device, and published a little later. The concept here discussed is the now familiar IQ (intelligence quotient). Terman (1916) and others have adopted the term and investigated the concept. As a result of these studies it appears that one's intelligence quotient is, at least to quite a marked degree, constant throughout life. This relative constancy appears when mental age 16 is taken as the average adult mental age, thus giving all chronological ages above 16 the value 16. More searching investigation of the IQ is required, but it seems at the present time that the term is with us to stay.

6. Mental age. The description of the intelligence quotient of the last paragraph used the term " mental age." This concept was first extensively used by Binet in 1908. It was originally developed in connection with young children (those under 14), and in connection with them the definition given by Terman (1919, page 7) holds: " By a given mental age we mean that degree of general mental ability which is possessed by the average child of corresponding chronological age." Pintner, however, qualifies this statement when dealing with the Stanford Binet and with older children. He writes (1923, page 74): " . . . there is a possibility that the higher ages (12, 14, 16) are too hard for the average child of those ages; nevertheless, constant use of the scale gives us a familiarity with its meaning, and something like conventional significance is attached to the different mental ages on the Stanford Revision. They are beginning to stand for specific degrees of intelligence even though they may not in every case actually measure the average ability of the age in question."