LB
1131
.K3

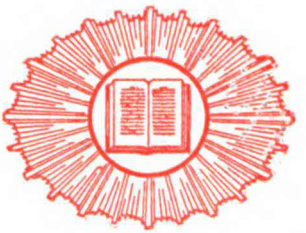
MEASUREMENT AND ADJUSTMENT SERIES

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

Generated on 2020-12-23 00:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

WORLD BOOK COMPANY

THE HOUSE OF APPLIED KNOWLEDGE

Established 1905 by Caspar W. Hodgson

YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Nebuchadnezzar had had a dream, and he remembered only that it was a bad one; yet he demanded that it be told him and interpreted. Daniel saved his own head and the heads of the other wise men of Babylon also, by meeting both requirements. Interpretation, in that case, was a matter of vital importance. What's the use in having Chaldeans and such, thought the old king, unless they can tell us something that we do not already know? What's the use in having tests unless the results of them can be interpreted? In preparing this book on the *Interpretation of Educational Measurements*, the author has established a kind of bureau of standards for the evaluation and interpretation of educational tests and the means of testing. His critical inquiries are designed to eliminate guesswork to the end that tests may tell us more, and that more accurately, about mind in the making

MAS: KIEM-1

Copyright 1927 by World Book Company
Copyright in Great Britain
All rights reserved

PRINTED IN U.S.A.

Generated on 2020-12-23 00:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Sch of Ed.
Idaho
1-26-1928
1395

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

Generated on 2020-12-23 00:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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

Generated on 2020-12-23 00:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

CONTENTS

Generated on 2020-12-23 00:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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-*

Generated on 2020-12-23 00:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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-

Generated on 2020-12-23 00:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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

Generated on 2020-12-23 00:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

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

Generated on 2020-12-23 00:05 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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

Generated on 2020-12-23 00:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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'Annés Psychologique*, Vols. XI-XVII, especially Vol. XI (1905).

Generated on 2020-12-23 00:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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.

Generated on 2020-12-23 00:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

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."

Generated on 2020-12-23 00:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Mental age as originally conceived was as defined by Terman, but as now commonly used it is to be taken as qualified by Pintner. In other words, the Stanford Binet, the Herring Binet, and other Binet scores do not give, for average children above age 14, mental ages which are the same as their chronological ages — a typical or median 16.0-year-old will not secure a Binet mental age of 16.0, but a lesser "mental age." For this reason no simple meaning applicable to young and old, dull and bright children can be attached to the term "mental age." In subsequent chapters, wherever the term is used, it is to be understood that children below the ability of average 14-year-olds are being considered. In this narrower field the definitions of Terman and Pintner hold.[1]

7. Subject and achievement ages. We may at this point define certain other terms. The reading age of a child as determined by a certain reading test is the age of typical or median children who do just as well on this test as the child in question. Arithmetic age, spelling age, etc., all have comparable meanings. Any of these may be designated as a "subject age." If a number of school subjects are incorporated into a single achievement test, the score of the child expressed in terms of the age of average children who do equally well is called an "achievement age." It is obvious that just as the mental age loses its original significance for ages where growth in intelligence is small and becomes meaningless, in the original or defined sense for individuals scoring higher than average adults, so likewise do all subject ages and achievement ages. It is accordingly well to restrict these terms to the abilities of young children.

8. Subject and achievement quotients. When a child's mental age was divided by his chronological age (or 16.0 if

[1] Since this section was written, an important criticism of "The Mental-Age Concept," by L. L. Thurstone (1926), has appeared.

Generated on 2020-12-23 00:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

his chronological age exceeded 16.0), we obtained his intelligence quotient. In a similar manner we may obtain his reading quotient by dividing his reading age by his chronological age; his arithmetic quotient, by dividing his arithmetic age by his chronological age, etc.; and his achievement quotient, by dividing his achievement age by his chronological age. None of these quotients can maintain its original meaning when applied to individuals scoring above average adult, and in practice it will ordinarily be found to have changed its meaning when the individual secures a score above typical 14-year-olds. In this text the use of mental ages, subject ages, achievement ages, and quotients built upon them is restricted to individuals scoring below average 14-year-olds.

9. The accomplishment quotient. In 1920 Franzen devised and popularized the use of the accomplishment quotient. He defined it as the achievement age divided by the mental age, and interpreted a quotient of less than 100 (as is usual, the decimal point has been dropped) as indicating that the child was not achieving up to the level of his ability. This procedure has become rather widespread, and though Dr. Franzen himself now recognizes the dangers of so naïve an interpretation and recommends other interpretative devices, he has as yet been unable to stop the ball he started rolling. Dr. McCall wrote most enthusiastically of the accomplishment quotient in 1922 and said: "The accomplishment quotient is the most exact present-day measure of the efficiency of study, instruction, and supervision; it is the only just basis for reporting to parents and for judging pupils; and it is the best index of what pupils need special attention and spurring, of what pupils need restraining, perhaps, and of what pupils need to be 'let alone.'" ". . . the accomplishment quotient asks the pupil to progress at a rate which is proportional to the mental capacity with which nature endowed him." As the writer differs decidedly with

Generated on 2020-12-23 00:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

this statement of Dr. McCall's, he will endeavor to show later in this text that due to the error of measurement in our intelligence and achievement tests, a trust placed in the accomplishment quotient is largely misplaced, and that an interpretation of a child's accomplishment through other channels is entitled to a greater trust.

In order to distinguish between achievement-age-divided-by-chronological-age, which is sometimes called an accomplishment quotient, and the accomplishment quotient as defined by Franzen, Otis (1925) used the term "accomplishment ratio" in place of accomplishment quotient. This is still rather ambiguous, and since the number of quotients is unlimited, — for we may have a reading age divided by an arithmetic age, a reading age divided by a mental age, etc., — it seems preferable to refer to these quotients by naming both the numerator and the denominator. Thus, reading-age-divided-by-arithmetic-age may be referred to as the "reading-arithmetic" quotient, and similarly for other quotients. This practice will be followed herein, except that reading-age-divided-by-chronological-age (and similarly with other quotients involving chronological age in the denominator) will, in harmony with general practice, simply be referred to as a "reading quotient."

10. Quotients not based upon mental or subject ages. Consider the data of the table below

Age	Reading Test Norms
8.0	60
9.0	68
10.0	75
11.0	82
12.0	88
13.0	93
14.0	98

and the status of a child 10.0 years old who makes a score of 88 on the reading test referred to. The reading age corre-

Generated on 2020-12-23 01:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

sponding is 12.0. Thus, if we divide 12.0 by 10.0, we obtain a reading quotient of 1.20. However, we might have divided the obtained score, 88, by the normal score for 10-year-olds — namely, 75 — and obtained a reading quotient of 117. This is a reading quotient just as truly as is the other. There are no theoretical grounds known to the writer establishing the one quotient as more "true" than the other, the "truth" in this case meaning the actual amount of reading ability possessed by the child as a fraction of the average amount possessed by a fair sampling of children of his chronological age. What constitutes a true quotient in a case like this is a very difficult matter to determine. However, we may say that it is preferable to use the first quotient rather than the second because more people have calculated reading quotients in the first way than in the second. This is admittedly a very inadequate justification, but the writer sees no other, and we may at least have the satisfaction of knowing that if we are in error in this procedure, others are also in error *in the same sense* as we, and thus we shall understand each other, though we shall all be in error in attaching percentage values to our quotient figures. The 10-year-old child scoring 88 and thus obtaining a reading quotient of 120 may in truth have achieved 10 or 50 or some other percentage different from 20 more than average children of his age. We do not know and cannot know until we have established a sound zero point of reading ability. We also do not know whether a 10-year-old securing a reading quotient of 120 is more or less exceptional in his ability than an 8-year-old securing the same quotient. Until sound zero points are established, the same criticism applies to such quotients as the 117 cited, built on other bases. Where possible, it is well to avoid the use of quotients, but if quotients are to be used, the age-basis quotient at present seems preferable for young children because of its explicitness of definition.

Generated on 2020-12-23 00:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

11. The mean. According to Klemm (1914, page 224), Roger Cotes made use of a weighted average in 1722, and Thomas Simpson in 1757 showed that the reliability of the mean increased with the number of observations. Earlier than this, Bernoulli (*Ars Conjectandi*, 1713) had shown that the accuracy of a proportion (frequency in a class as a fraction of the total frequency) increases with the size of the population. Simpson, however, probably got this idea not from Bernoulli, but from a little-known work published by De Moivre in 1733 (see Pearson; hist. 1924). Physical and mental science are indeed deeply indebted to De Moivre for establishing the fact that accuracy varies inversely as the square root of the size of the sample; i.e., as $1/\sqrt{N}$.

12. Individual differences. The importance and presence of individual differences may be considered a corner stone of Plato's philosophy, but a great deal has recently been added to this concept in that the magnitudes of individual differences are now quite commonly stated in quantitative terms. The greatest contributions along this line were made by Galton (1869 and 1889), and a generation later by Thorndike (1904 and 1913) and Cattell in various Columbia studies.

The reality of individual differences and the possibility of measuring them was convincingly and repeatedly presented by Galton. Galton was very modest in making claims for himself, and we may take the following quotation from the introduction to his *Hereditary Genius* (1869) as stating but a part of the truth, as far as his own contribution is concerned. He wrote: "The theory of hereditary genius, though usually scouted, has been advocated by a few writers in past as well as in modern times. But I may claim to be the first to treat the subject in a statistical manner, to arrive at numerical results, and to introduce the law of deviation from an average into discussions on heredity."

Generated on 2020-12-23 00:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

13. The normal distribution. The " law of deviation from an average " as used by Galton is equivalent to the statement that in a homogeneous race the distribution of mentality follows a normal curve. Galton obtained his concept of the normal distribution from Quetelet, but the more remote and primal source is undoubtedly not Quetelet, nor even Gauss or Laplace, but De Moivre (1733).

14. Psychophysical methods and standardized administration. The attempt to score mental reactions in an objective manner was undoubtedly given a great impetus by success in the measurement of sensations and the development of psychophysical methods. The causal connection between these is none too clear, but Cattell has apparently been one important link. Standardized administration has developed as a corollary to objective scoring. Both of these were emphasized by Cattell (1890).

15. Quantitative measurement. The method, appearing almost brutal to the poet, the æsthete, and certain other uncalloused souls, by which different kinds of behavior are given quantitative values upon a single scale, shows, in connection with achievement, a development through Galton, Pearson, and Thorndike, resulting in such products as Thorndike's drawing scale, Hillegas' composition scale, Abbott-Trabue poetry test, Thorndike's æsthetic appreciation test, Upton-Chassell citizenship scales, etc. A second development in connection with intelligence has been through Binet and Simon and their followers. On the whole this phase of the movement has involved, in addition to the difficulties of scaling reactions, added hazards due to the greater uncertainty as to the singleness of function measured.

16. Group measurement. The origin of the group measurement of abilities is lost in the school examinations of the past, and group testing as applied to other than school subjects sprang to life without conscious parenthood from a

Generated on 2020-12-23 00:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

study of individual differences. Galton and Wundt provided the background which is expressed in the group tests used by Bolton (1892). Otis (1920) deserves credit for furthering in 1917 the group testing of intelligence. Thorndike (1918) in 1914 and Norsworthy (1906) antedated him as authors of batteries of group tests of intelligence. However, certain tests, particularly of the opposites and the sentence completion types, — e.g., Ebbinghaus completion test (1895), — used still earlier, can well be called group intelligence tests.

17. Norms. The writer will not attempt to trace to the source the " establishment of norms " procedure. At least four lines of development may be mentioned: the interpretation of scores or records (1) by referring to grade averages, (2) by referring to age standards, (3) by referring to percentile or deviation position in a defined group (usually an age or a grade group), and (4) by position determined by the variability of judgments of " competent judges." These four methods do not serve identical purposes. Galton at various times encouraged the general movement toward the establishment of norms, as did Cattell (1890) a little later. Rice (1897) started a movement based upon grade norms which has extended far. A powerful factor furthering the establishment of such norms has been the school " survey " movement, beginning with the Pittsburgh Survey in 1907, and the New York Survey in 1911–1912, which utilized the Courtis arithmetic tests. The grade norm developed in connection with normal children, and an age norm used by Binet and Simon in 1908, determined by the performances of normal children, were used to interpret the reactions of the abnormal. This early difference in the use of these two types of norms is still very prominent, though the age norm, particularly since the work of Terman (1916) in 1914–1916, is commonly being used in studies of normal children. The utilization of the variability of a defined group for interpreting individual

Generated on 2020-12-23 00:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

scores was well recognized by Galton (1889 and 1907) and is now a very common practice — a few illustrations being the "reduced" measures of Woodworth (1912), the "standard measures" of Kelley (1914 comp. and 1923 stat.), the "T-scores" of McCall (1921), and the "sigma indexes" of Franzen (1924).

18. Standardized judgments. The names of Fechner (1860), Mueller (1878), Fullerton and Cattell (1892), Urban (1909), Thomson (1919), and Thorndike (1910) in his derivation of a handwriting scale should be mentioned in connection with the utilization of judgments in building up standards.

19. Early educational tests. The earliest reported use of objective educational tests is that brought to light by Dr. Isaac L. Kandel and reported by Thorndike (1913). It is therein shown that the Rev. George Fisher, a schoolmaster in England, was the author in 1864 of a scale book wherein questions and samples were given, enabling a numerical grading on an objective scale in "writing, spelling, mathematics, navigation, Scripture, knowledge, grammar, and practical science."

20. Validity and reliability. We may mention two closely related tendencies which are so ill defined that parentage has not been claimed. The older of these two is the "validity" movement, with the attendant problem, "reliability," and the younger is the trait-analysis movement. The question of validity would not be raised so long as one man uses a test or examination of his own devising for his private purposes, but the purposes for which schoolmasters have used tests have been too intimately connected with the weal of their pupils to permit the validity of a test to go unchallenged. The pupil, particularly the modern Dewey self-motivated pupil, is the dynamic force behind the validity movement. The question is thoroughly roused from a slumber of cen-

Generated on 2020-12-23 00:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

turies, probably never to sleep again. Further, now that the same tests are used in widely scattered places and that many very different tests all going by the same name are gently recommended by their respective authors, even the most complacent schoolmen, the most autocratic, and the least in touch with pupils, are beginning to question the real fitness of a test. Could present test devisers but have stenographic reports of the sittings of college entrance examination boards, they would surely find that validity is with them an old issue. If the deliberations of such boards can be supplemented by an adequate statistical technique, the problem of the validity of a test will shortly assume the importance that is its due.

The problem of validity is that of whether a test really measures what it purports to measure, while the question of reliability is that of how accurately a test measures the thing which it does measure. The statistical technique for solving this second question has outrun that for the first. And here two worthy critics, each of the other, provide the strands which commingle so usefully in the reliability coefficient — Pearson in developing the product-moment coefficient of correlation and Spearman (1904 and 1907) in applying it to the correlation between similar tests and in pointing out the significance of this correlation.

21. Analytical measures. The still younger trait-analysis tendency referred to may be called the analytical movement. As a corollary to this is an educational and vocational classification and guidance based upon differential analyses of ability. The value of doing this has been mentioned by many vocational guidance advocates and has been in the minds of certain test devisers, — for example, Rugg and Clark when devising their standardized tests in first-year algebra of 1918, and Minnick in devising his geometry test of 1919, — but the statistical and experimental principles

Generated on 2020-12-23 00:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

underlying an analytical treatment of test scores has been so inadequate that it is proper to describe the movement as in its veriest infancy. This text and earlier contributions (1914 educ., 1919, 1923 princ., 1923 new, and 1923 stat., especially Chapter IX, dealing with estimates of true scores, probable errors of such estimates, and the probable error of a coefficient of correlation corrected for attenuation) constitute the writer's attempts to promote a sound analytical use of test scores.

22. Tested procedures. It may be said that procedures involving the calculation of averages and of measures of variability and the measuring of correlation between tests are well established, accepted devices. In subsequent pages the combining of qualitatively different material is at times resorted to; e.g., obtaining a total score from a number of separate achievement test scores. It is admittedly upon a less sound historical and logical foundation than the other procedures just mentioned. The practical advantages resulting from the use of such composite scores have proved to be very great, so that practice has, in a sense, outrun logical development. Finally, the analytical treatment here elaborated is practically without antecedent history, and it rests primarily for its justification upon the derivations and proofs given by the author in the works just cited.

23. The steps and pitfalls ahead. Although a number of important tendencies and accomplishments of the past have been mentioned, the future beckons alluringly. Most of these fields are still sufficiently untouched to offer a happy hunting ground for the teacher who loves his charges and wishes to guide their footsteps in the path that leads to the fuller life, and also to the searcher after truth for its own majestic harmony and beauty.

It has been said that general intelligence is of significance in many phases of life, but who has as yet defined these

Generated on 2020-12-23 00:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

phases or placed boundaries to this significance? It has been said that the intelligence quotient is constant throughout life, but who is fatalist enough to believe this for himself even though he might for others? And if, as most of us surely believe, it is not strictly constant throughout life, who has shown the reasonable limits of the concept, the situations in which it makes for understanding and can be used for good, or those in which its application leads to an unfruitful resignation, obscurity, and evil?

These issues strike deep in social life and individual philosophy. We think of the "old" methods and the "old" subjects of the curriculum as being hoary with precedent and prejudice, but the ruts of the test movement are already so deep that there are many who do not see beyond them. We assume that there is a trait — for example, reading — varying from child to child. Let us question this assumption, for it may be a dozen traits erroneously called one. We assume that tests as given by different teachers and at different times have called forth equal or approximately equal effort; we assume a sufficient sensory and motor equipment; we assume that the sampling as drawn out by the test questions constitutes a fair and sufficient sampling of ability. If we cannot avoid making these assumptions, we can at least pause long enough to steep our souls in the conviction that they are present and obscure our findings. If the pause is long enough and well spent, we may secure an estimate of the magnitude of the errors introduced. There is a becoming modesty and reserve in the verdict of a tester who has paused this long and to this outcome.

Two plus three has so often totaled five, and two times three so commonly yielded six, that we have assumed test scores may with entire propriety be added, subtracted, multiplied, and divided. They seldom can. Test devisers have apparently been quite successful in obtaining test-score

Generated on 2020-12-23 00:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

units which are substantially equal and can be added and subtracted, but they have failed quite signally in determining reasonable zero points, so that the product or quotient technique rests upon shifting ground. Let us not forget this, and repeatedly ask, "Do I know that the beginning of the scale of measurement is a sound zero point of ability and that I thus may obtain a meaningful quotient?" The very asking of the question has profoundly stirred our mensurative natures, and answering it "No," as we generally must, robs us at once of a very simple method of interpretation, of a very common source of errors in judgment, and of our fellowship with the get-rich-quick variety of mental-test interpreter. It is not to be desired that the quotient technique be completely discarded, but the writer's immediate purpose will have been accomplished if his readers will but think of the height above zero of an average 12-year-old in a dozen mental tests as being comparable to the height above the water of the rail of a rolling ocean liner as measured at twelve different times. This should be — let us hope it is — a concept to make one dizzy, for uncritically to accept any zero point, however derived, as a proper basis for determining quotients is bewildering and mentally loathsome.

The attempt of this chapter has been to give a perspective to the more detailed work of later sections and to encourage a critical approach to the problem of test purposes, selection of measuring instruments, and statistical treatment of results.

Generated on 2020-12-23 00:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER TWO

Purposes Served by Educational Tests

1. Intelligence tests versus achievement tests. One of the most frequent questions arising in connection with the test program is whether to use an intelligence test or an achievement test, or both. The answer cannot be given finally with our present knowledge, for usually the tester does not know whether the intelligence and the achievement tests being considered measure the same or different traits. It has commonly been found that the two tests do not correlate perfectly, but this may of course be due simply to the chance errors involved in each. When allowance for chance errors is made, the correlation between a good battery achievement test and an intelligence test is found to be very high. If a number of children have been together in the same school for a year, it would seem the part of wisdom to judge both of their general ability and their accomplishment and to compare one child with another by means of a good achievement test. If their antecedent histories are quite different, — e.g., if they are transfers from other schools or if they have had private instruction, — it then would be well to judge of general ability and fitness for further work by means of a good general intelligence test.

2. The responsibility of the counselor. If the two tests really measure the same single capacity, there would still be an advantage in using both and averaging the results, for this would give more reliable individual measures. However, the test administrator who should advocate the use of both an intelligence and an achievement test merely and professedly to obtain a more reliable measure of some single ability would probably have to be content with a clear conscience rather than a lucrative employment. The writer, having been an early — probably the first — full-time school con-

Generated on 2020-12-23 00:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

sulting psychologist, can testify to the pressure of administrative authorities, teachers, and parents for an immediate and decisive statement as to the difference between a pupil's native intelligence and his achievement. An honest confession of indecision is not nearly so welcome as a positive assertion of a definite difference, however inadequate the grounds for judgment may be. It is only the well-informed person who knows that positive assertions upon moot and abstruse questions of character analysis are presumptive of bluff and charlatanry.

The future enhanced respect for our profession rests largely upon a greater accuracy and moderation in our prognoses. It behooves us to take a personal sense of responsibility in our utterances and to have a sense of criminal guilt if we mislead by false advice. The uncurbed minister of the Gospel who expresses an unambiguous opinion as to the ultimate destiny of a particular human soul, the palmist who traces a life line through unreal woe to impossible weal, and the psychoanalyst who misinterprets, as referring to one's lover, a nightmare in which appears a black beast with a double face, all find a following in a credulous public because of the flexibility of their imaginations and their lurid substitutes for facts. Let us hope that the school principal and the guidance counselor have entirely other claims for consideration — an imagination that refuses to function at just the time when most imaginations soar the highest; namely, when facts are absent and when the probable error of judgment is large. In other words, their good name rests upon an imagination that neither clouds nor overrides their knowledge and sense of personal responsibility — an imagination that never loses sight of the ubiquitous probable error. A tin can on a dog's tail is a very effective reminder to the dog that he is not a free agent, and a probable error attached to an intellect in such a manner that it proclaims itself when-

Generated on 2020-12-23 00:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

ever the imagination runs rampant would be equally serviceable as a part of a counselor's equipment. It is true that it would be annoying, but it would nevertheless have a very salutary effect.

The layman should not be held responsible for the pressure which he exerts upon the counselor for opinions exceeding his means for making valid judgments, for the layman is not supposed to know what the bounds of valid judgment are. The psychologist alone should set the limits within which he is willing to testify; and very definite limits exist. These are defined by the probable errors of the measurements utilized. Thus, if on a certain intelligence test a child's mental age is 10.0 and on a certain achievement test the achievement age is 9.5, the judgment that " he is not working up to his mental capacity " (or any similar judgment) is sound only in case the probable error of the .5-year difference is small. Unfortunately there have been many test interpreters, so called, who have thought and known little about this probable error.

If a doctor of medicine hastily diagnoses a case as chicken pox and prescribes treatment upon that basis when there are forty chances in one hundred that it is smallpox, we should say either that he was ignorant of the full implications of the symptoms or irresponsible in interpreting them, and we should not forgive him, if wrong, on the ground that he chose the more reasonable diagnosis — the one in which the chances were sixty in one hundred in his favor. Who causes the greater sorrow, the physician who wrongly diagnoses thirty in one hundred ailments or the school principal who wrongly judges intellect and effort and gives unsound advice as to training and vocation to some thirty in one hundred of his graduating class? The onus is great in either case and but little relieved by pointing with pride to the seventy correct diagnoses. Psychologists who do not know the prob-

Generated on 2020-12-23 00:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

able errors of their judgments and qualify them accordingly have no more right to diagnose and prescribe than have equally incompetent physicians.

3. The probable error. Our present uncertainty as to the significance of obtained differences between achievement and intelligence scores is simply an illustration of one problem the solution of which depends upon the knowledge of a probable error. All the problems of the counselor are of this type, as all of his information about mental traits is based upon measures or judgments containing substantial error. The chief contribution of this text is an emphasis of the universality of error in our mental measurements, of the importance of measures of reliability, and an effort to show how to obtain and use them.

4. Community of function. Such experimental evidence as is available points to a high degree of community of function in various tests differently labeled and supposedly measuring different traits. Though this text was not intended to include a technical discussion of this evidence, the matter is so important that the writer has given in Chapter VIII certain evidence bearing upon the community between intelligence and achievement. In the main, however, he must simply state the conclusions (lettered *a*, *b*, *c*, *d*, *e*, *f* in following paragraphs) that he has reached at this time, as they serve as the point of view of the subsequent treatment.

5. Community in achievement tests and general intelligence tests. (*a*) On the average, in the neighborhood of .90 of the capacity measured by an all-round achievement battery score, — reading, arithmetic, science, history, etc., — and of the capacity measured by a general intelligence test is one and the same. If a comprehensive educational achievement test and a general intelligence test each give "fairly reliable" total scores, each would need to be more than ten times as long to yield equally reliable measures of difference between

Generated on 2020-12-23 00:14 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the educational achievement and the intelligence scores. This is true not only because 90 per cent of the tests measure a common function, but also because the chance factors entering into this 90 per cent of each test tend to obscure whatever real difference is being measured by the 10 per cent. This means that a scant one tenth of the tests are involved in the measure of difference and, practically, that judgments of individual differences between intelligence and achievement based upon the commonly available tests are quite unsound, being of an order of accuracy not of the total scores of the tests, but of total scores of tests less than one tenth as long. The possibility of making sound judgments of this sort by utilizing much more refined measures lies before us.

(*b*) The writer is compelled to advise against the common use of an intelligence test and an achievement test for the purpose of drawing conclusions as to the differences found within the individual on the two tests.

6. The accomplishment quotient. This, of course, implies the discarding, as far as individual diagnosis is concerned, of such a concept as the achievement-intelligence quotient. This may seem to be a radical curtailment of a widespread interpretative concept. In one sense it is, for if achievement-intelligence quotients, as determined, below and above 1.00 correspond to real mental structure, important knowledge of the child is available when the quotient is known. More soundly, however, it is no curtailment at all, for it may be shown that with such achievement and intelligence tests as are commonly used, the great majority of such quotients diverge from 1.00 by amounts to be expected as matters of chance. Thus, at present, eliminating the achievement-intelligence-quotient technique is merely eliminating a false guide. That the concept has, in individual cases, been remarkably confirmed by teachers' and parents' judgments should be recalled in connection with its equally great failure in other cases.

Generated on 2020-12-23 01:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

It is, of course, true that outstanding differences in obtained achievement and general intelligence scores are more likely to be significant than medium differences. Thus, if one is accustomed to use the achievement-intelligence quotient and to check it by securing accessory information in regard to those cases yielding exceptional quotients, he will find the quotient to be in one sense verified, and he will thus be prone to attribute a high degree of validity to it. This validity based upon extreme quotients is, however, not a guide for average cases. The artificial data of the accompanying table are chosen to illustrate the typical situation. The scores are subject ages in terms of months; thus A's obtained achievement score of 90 indicates an achievement age of 7 years, 6 months. It is true that the data are hypothetical and were devised to illustrate the present point, — that judgments are likely to be formed from extreme and not from typical cases, — but nevertheless it can be shown that actual data in which are to be found material errors of measurement (and all our achievement and intelligence tests yield such) will as a matter of chance tend to exaggerate extreme differences, just as do these artificial data.

TABLE 1

Pupil	True Achievement Ability	Error in Achievement Measure	X, Obtained Achievement Score	True General Intelligence Ability	Error in General Intelligence Score	Y, Obtained General Intelligence Score	True Accomplishment Quotient	X/Y, Obtained Accomplishment Quotient
A	80	10	90	80	0	80	1.000	1.125
B	70	0	70	60	20	80	1.167	.875
C	60	− 20	40	80	10	90	.750	.444
D	90	− 10	80	90	− 10	80	1.000	1.000
E	70	20	90	70	0	70	1.000	1.286
F	80	0	80	70	− 20	50	1.143	1.600

Generated on 2020-12-23 00:14 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The score the pupil makes in the achievement test is recorded in the " X " column, and as indicated, it is due to his true ability as expressed in the " true achievement ability " column plus an error of measurement as given in the " error in achievement measure " column. The pupil's obtained general intelligence score " Y " is likewise equal to a true ability plus an error of measurement. If we divide X by Y and obtain accomplishment quotients as given in the last column, we shall probably be struck by the record of Pupil C and investigate the case. Since Pupil C is in true accomplishment (true achievement score = 60) below his true general intelligence (true general intelligence score = 80), our investigation of the case will " confirm " the test finding. It is, of course, not a true confirmation, for the true accomplishment quotient is .75 and not .44 as found; but as we are not able to judge of this difference and are able by our accessory investigation to convince ourselves that the achievement ability is less than the general intelligence ability, we consider the quotient to have established itself as correct. This leads us to place confidence in the quotient. If we investigate another case, it will probably be that of Pupil F, and here again we shall find " confirmation " of the quotient. If we now desist in our checking-up process and forthwith trust the remaining quotients, we shall be in substantial error in every instance except that of Pupil D, but we shall not be aware of any of these errors. The writer fears that just such a process as is here described has been unwittingly followed by many who trust the quotient technique. The only sound way to judge of the efficacy of a particular kind of accomplishment quotient is to check up on all cases, average as well as extreme. Such a procedure constitutes a thoroughgoing determination of the probable errors of quotients, and when it is made, the writer predicts from such evidence as is at hand that it will lead to the conclusion that in the majority

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of cases quotients may not be taken as reliable. Even extreme quotients cannot be trusted, for they are systematically overstatements of the amount of divergence between achievement and intelligence.

The observations just made have to do with the reliability of individual achievement-intelligence quotients. Just as the reliability of an average of a number of scores is much greater than that of the single scores separately, so the class average quotient may be trusted when the single quotients of the members are quite unreliable. Though the achievement-quotient technique may be used for group interpretation with fair accuracy, provided the tests employed are of excellent reliability, nevertheless a technique which is inaccurate in a study of individual cases can be discarded generally with little loss.

(*c*) The similarity between a battery achievement score and a general intelligence score is obviously considerably greater than the similarity between the score on a single school subject test and a general intelligence test score. The average community between the separate elementary school subjects, omitting drawing and music, and general intelligence is probably in the neighborhood of 85 per cent.

7. Community in different achievement tests. (*d*) The general intelligence test, tapping as it does a wide range of subjects, is more akin to any one of our common subject-matter tests than are two different subject-matter tests akin to each other. Accordingly, with equally reliable measures, the distinction between a pupil's abilities in two subjects can be made with greater certainty than a distinction between his general ability and his ability in either one of these subjects. We may tentatively think of the community of function between any two of the common school subjects (not including drawing or music) as being in the neighborhood of 80 per cent.

Generated on 2020-12-23 00:14 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

8. The prognostic value of achievement and intelligence scores. (*e*) If a specific subject-matter test and a general intelligence test are equally reliable, the former will give slightly better evidence of later performance in the specific subject than the latter. Therefore, for school purposes, where equally reliable tests are available, achievement tests are generally preferable to general intelligence tests.

9. Primary and university tests. (*f*) Subject-matter tests are at present not so reliable as available intelligence tests for the primary and upper high school and university grades, so that intelligence tests for the kindergarten, first grade, and possibly second grade, and for the last two years of the high school and for the university are commonly preferable to achievement tests, if the purpose is a classification of pupils according to school promise.

10. Endowment, training, and the problems of measurement. There has been in some quarters a very naïve use of intelligence tests for the purpose of measuring a child's endowment independent of training. In fact, such use of Binet tests and group intelligence tests has been very common and has not led to obviously absurd interpretations. The reason for this is probably due to the fact that training of different children is fairly constant — that is, growing up as a child in a country speaking a single tongue, the several sections of which respond in the main to the same impulses of right and wrong, to the same Sunday supplements, to the same attitudes of leadership and submission, to the same vocabulary, coinage, and methods for measuring time, constitutes a training in which the constant elements far outweigh the variable. We observe that this child is a Catholic and that child a Protestant, and forget the more universal and intimate common elements of life: that each is sometimes punished and sometimes not, when he prevaricates, and that the general likelihood of social punishment bears a

Generated on 2020-12-23 00:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

fairly constant inverse ratio to the excellence of the prevarication; that each looks with the same envy upon a red apple, whether on a tree over the fence or on an Italian's pushcart; that each comes into close contact with those who bully and those who can be bullied; and finally and probably most important, that each from a very early age, and guided by his individual urge, chooses his friends out of many available and develops his interests out of the great richness of life's offerings.

The writer finds not so much occasion to criticize the practical conclusions of those who consider that intelligence tests in the main measure innate differences as he does the conclusions of those who consider achievement tests to measure in the main acquired differences. The theoretical issues here involved are important, but the practical issues confronting the guidance counselor are generally such that answers do not depend upon whether the traits measured are innate or acquired. One is prone to feel that an individual difference definitely determined to be innate is a more important difference than one known to be acquired. From the standpoint of heredity and eugenics this may be so, but from that of vocation and success, it matters not a whit how one has reached his present stature. If an ability is actually present, what employer cares whence derived? If A can build a good bridge and so also can B, no horse or automobile will break through A's bridge and not through B's because A was less gifted innately than was B. We have no evidence whatever that in the case of children who have had roughly similar educational opportunity, tendencies considered innate (that is, measured by intelligence tests) are better measures of future success than tendencies quite commonly considered acquired (that is, measured by achievement tests). In fact, there is certain evidence that might be considered by some to point the other way, in that achievement records at an early

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

date correlate more highly with achievement records of the same sort at a later date than do early general intelligence records. For our purposes, then, we shall make no effort to distinguish between innate and acquired individual differences. The question is raised here, only to draw the conclusion that, could we devise them, there would be no special merit for prognosis purposes in tests of innate capacities as opposed to tests of acquired capacities, and this for the simple reason that, so far as we know, acquired capacities, after being once acquired, are just as likely to persist into the later life of the individual as are innate capacities. Otherwise stated, habits once acquired are from thence on indistinguishable from instincts. Even the possessor himself, if he can but forget the origin, is unable to distinguish between his habits and his instincts.

11. The adequacy of the achievement test. We have reached the conclusion that achievement tests, if of satisfactory reliability, do not commonly need supplementing by intelligence tests in the classification of pupils for school purposes and for prognoses as to school success. We shall not need to alter this conclusion after we examine more closely into the specific purposes of school examination programs. We may state these purposes as six in number, and in each the test is both a measure of the past and evidence of the promise of the future. These six fall into two groups of three each, depending upon whether group or individual diagnoses are involved.

12. Six purposes. For the group we have:

(1) The measurement of the general group (grade or school) accomplishment and an estimate of the probable future general group success in school work.

(2) The measurement of a school group in some specific subject and an estimate of the future group promise in the same or a closely related subject.

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(3) The measurement of the relative differences in achievement of the group in two or more scholastic lines and an estimate of the significance of such differences.

The same three purposes as pertaining to the individual give:

(4) The measurement of the past general scholastic success and the future promise of an individual.

(5) The measurement of the success of an individual in a specific school subject and an estimate of his future promise in the same or a closely related subject.

(6) The measurement of differences in the individual of abilities and accomplishments in several scholastic lines and an estimate of the probability of persistence of differences, of the sort revealed, in future school work or vocation.

13. Reliabilities requisite to each purpose. These six purposes are listed in the order of the excellence of the tests demanded in their solution. As dealt with more extensively in Chapter VIII, the minimal satisfactory reliabilities as measured by a reliability coefficient determined from the pupils in a single school grade, of tests serving these six purposes, are as follows: .50, .50, .90, .94, .94, and .98, respectively. Under certain conditions, with various procedures and with certain school subjects, these figures will need a slight modification, but on the average we are quite safe in taking them as minimal reliability requirements.

14. Validity. Certain other test desiderata than that of reliability also change with the purpose, and still others scarcely change at all. Thus the validity of a test is of high importance in all five of these purposes, and hardly more important in one than in another. The evidence that a test measures a worth-while function, which statement includes in itself by implication the idea that it does not, in material part, measure minor or inconsequential functions, and that it does not give improper relative importance to the various

Generated on 2020-12-23 00:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

phases of the subject named, rests in the first instance upon the opinion of those competent to judge of what the functions are that are measured by the various questions, as affected by the conditions of administration of the test — directions, time limits, etc. In the second and final instance the validity of a test is measured by the extent to which it accomplishes the purpose claimed for it. The correlation between a test proposed as one having prognostic value (and what test is not so proposed?) and later demonstrated degrees of success or failure constitutes the final measure of whether the test is actually valid for the purpose claimed. Many tests, and some of these have stood the trial of time fairly well, have been put out supported by evidence of validity of the first sort only, while others have had much data of the second sort to support them before they first were offered to the public. Though many would be inclined to accept the judgment of some eminent psychologist that the test was valid in preference to the figures of an uncertain tabulator and interpreter of correlation data, nevertheless it is not now too much to demand that the validity of forthcoming tests be adequately supported by indubitable correlation results as well as by, or over and above, the opinions of their authors. In the case of certain recent school achievement tests, not only the tests as a whole, but every item separately in them has been selected because of its correlation with school records of achievement. This is a definite advance in method and an added insurance as to validity. Even though such care has been taken, correlations with criteria have not been perfect, even when chance errors have been allowed for, so that with the best of the educational tests there is still lacking the guarantee that extraneous elements are not, to an extent, included in the measure.

The establishment of the fact that a given test is valid for a specifically named purpose is at present one of the most, if

Generated on 2020-12-23 00:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

not in fact the most, difficult of the problems confronting the test deviser. It is proper for the test user to exercise his individual judgment in this matter, though he should hardly accept it as being on a par with, or as worthy of credence as, experimentally established facts showing validity.

Important, and as yet but partially settled, issues concerned with the nature and significance of the function measured are tied up with questions of "speed" and "power." The ideal speed test, also called the time-limit test, is one composed of homogeneous material; that is, many exercises, all measuring the same capacity and of equal difficulty, given with so short a time limit that none or few of the subjects finish. With such a test the number of exercises done (or correctly done, — there is usually little difference in such a test between the number worked and the number worked correctly) constitutes the score. Obviously, "speed" is an essential phase of whatever is measured. A good illustration of this type of test is the Courtis Standard Research Tests in Arithmetic, Series B. The ideal power test, also called the work-limit test, is one composed of items increasing in difficulty by regular steps, given either with no time limit or with so long a time limit that speed of performance is not a material factor. In some power tests the number of exercises correctly done constitutes the score, while in others the difficulty level reached is the score. Clearly, intellectual mastery, or power to do more and more difficult tasks, is the thing measured. Practically all of the spelling tests are good illustrations of power tests. Our knowledge as to the educational and social situations in which speed is of prime importance and those in which power is especially demanded is quite limited. This question is not to be settled by speculation, and relatively few experimental correlation studies comparing the merits of these two functions have been made. Such data as are available incline

Generated on 2020-12-23 00:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the writer to the view that power is generally the more important.

15. Other desiderata. Certain characteristics of a test are unlike validity in that they have quite different degrees of importance, depending upon the purpose in hand. The degree to which the conditions for giving a test are standardized and can readily and uniformly be followed by different test administrators, the time required to give it, its cost, its ease of scoring, the existence of extensive norms, and to a lesser degree the objectivity of scoring, all assume different values, depending upon the purpose. The existence of reliability coefficients and probable errors of scores is an important consideration wherever refinement and accuracy of interpretation are sought.

As tests become more widely used in determining promotions, there will be certain shortsighted individuals who will coach up upon the specific test to be employed. Of course, if homogeneous classification results from test programs, a child's educational position is injured if by chance or unfair methods he secures a higher score on a test than his talent rightfully entitles him to secure. Some tests are more easily coached for than others, and when the danger of such practice is imminent, the degree to which a test is noncoachable is an important item in determining its worth. Of the extant tests, the Thorndike College Entrance Examination, in which new forms appear every year, is as nearly uncoachable as any of our tests, but even in this case, with bright subjects there is probably an average gain of from three to five points in the score made upon a second form, from the mere practice of having taken the first form.

A test program dominated by the desire to appraise group accomplishment may well, because of the numbers involved, the cost, and the time demanded, be served by means of a short test costing but a trifle, having "fool-

Generated on 2020-12-23 01:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

proof" scoring devices, and having extensive norms for interpretative purposes. All of these conditions can commonly be met by means of some of our better low-reliability tests. For individual guidance, the necessity for higher reliability requires a long test, costing more, having as objective scoring devices as possible, and just sufficient norms to enable a comparison of the child with his peers. The mean and variability of the class in which the child is located need to be known, and the means and variabilities of the classes immediately above and below are desirable in order to locate the child with reference to his immediate environment, but there is no need for extensive norms from other communities. A reliable test rather than one having extensive published norms is the serviceable instrument for individual diagnosis.

16. Requisite reliability for group measurement. If group measurement is all that is undertaken, scores which are individually reliable are not demanded, for the reliability of an average score is much greater than that of the single score. Specifically, if the probable deviation of an individual's obtained score from his true ability score is a certain number of units then the probable deviation of the group average from the true average for this particular group is only $1/\sqrt{N}$ times as large, — N being the population or number of individuals in the group. For example, if a certain individual reading-test score has a probable error of 12, and if there are 36 children in the class, then the probable error of the average score of the class is $12/\sqrt{36}$, or 2. Thus, it is seen that a test so unreliable that it will not be serviceable in making individual diagnoses may be very serviceable for group diagnosis. This lessened need for high reliability in group investigation makes possible the use of tests which require but a short time to give and to score and which cost little. A two- or three-hour examination is needed to determine, approximately, individual fitness for college work,

Generated on 2020-12-23 00:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

but a carefully devised five-minute examination given to all the entering students of two universities would easily enable one to tell which of the universities drew the more capable students.

17. Age and grade norms. It is very commonly thought that the existence of extensive norms is essential, but this is true only for certain purposes. Let us consider the uses of grade and age norms. If the superintendent of Westernville desires to compare his city with Easternburg, he must have available grade norms from Easternburg. Suppose he makes this comparison. What next? Why, nothing next, except that he publishes the results in the *Widely Read School Journal.* This comparison does not change instruction, does not improve his teachers, nor classify his pupils into such groups that they will profit more by such instruction as is given. It is usually merely an entertaining, useless bit of information. It is not always useless, for a university receiving students from many high schools could make much use of average scores made by pupils from different communities, and if a state-wide or nation-wide average or norm is available, the university can interpret individual scores with reference to it. However, the real value of the test score, as far as Westernville is concerned, lies not in comparison with city, state, or national norms, but in knowledge of differences in accomplishment found within the school system of Westernville itself. A fifth grade in Westernville is to be judged in comparison with other fifth grades of the city, and a pupil in some one fifth grade is to be judged by comparison with his peers in the same fifth grade. Ordinarily, extensive grade norms are of no importance in an educational test program, and the lack of published norms, if the test is otherwise suitable, is no hindrance to its complete serviceability in meeting the six major purposes listed. This is particularly true with reference to the very important individual purposes, 4, 5, and 6.

Generated on 2020-12-23 00:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

18. The substitution of national for local norms. A modification of the point of view is necessary in case a single pupil is measured by some test. In this case the child has no similarly environed peers and must be compared with children in general of his age or grade. Accordingly we should here need age or grade norms. It is common practice to compare successively the children of a given grade with published national age or grade norms, quite neglecting the average local record. Such a procedure ignores the common environmental factor present for all those of a given grade and is therefore not the preferred treatment. It is, however, a serviceable method, generally leading to the same practical conclusions as would one utilizing the records of peers. National age norms are as inferior to local records in solving local problems as are national grade norms to local grade norms. For the reasons stated we shall consider the existence of norms derived from a huge number of cases a very slight asset, and the absence of them altogether scarcely a debit in comparing the general merit of tests. The illustrative treatment of this text does utilize national norms, for some point of reference is demanded, but it is hoped that the reader will readily see how he can develop a treatment based upon local data and specifically meeting all local needs.

19. The objectivity or reliability of scoring. The objectivity with which a test can be scored plays a very important part in the test program. If the score on a test as determined by one scorer agrees with the score as determined by another scorer working entirely independently, it is said that the scoring is objective. If a scorer is guilty of some systematic error in procedure, a rescoring of the same set of papers by the same scorer will not reveal this fault, so that the proper way to determine the reliability of scoring is to have two persons, entirely without agreement or discussion between themselves

Generated on 2020-12-23 00:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and merely guided by reading the Manual of Directions, score the same set of papers in such a manner that the scores given by the first scorer are not visible to the second, and then to check item by item and determine the discrepancies. If they are many, the scoring is not objective. It is desirable for all purposes that scoring be objective, but more essential for certain purposes than for others. If a single teacher does all the scoring and uses the results for his single class, it is less essential that strict objectivity be present than if many different classes are scored by different scorers and the results thrown together into a single comparison. Some traits are very difficult to score objectively. No one has as yet devised a really objective method of scoring English composition. If the method of scoring were to add the number of long words — those having more than seven letters — and subtract the number of short words, a very objective scheme would be built up, but objectivity would be obtained at the expense of validity, for the trait measured would no longer be English composition as ordinarily understood.

There is danger that very objective grading of composition, reading, and certain other subjects may be obtained at the expense of validity. Some of the most important functions — for example, that measured by the Thorndike-McCall Reading Test, or that involved in the Trabue Completion Exercises — do not permit of strictly objective scoring. In such instances a balance must be struck between these two values, validity and objectivity, and an endeavor made to devise a test which does not get away from the important function, but is at the same time amenable to fairly objective marking. A general rule can scarcely be laid down, but it is wise to be wary of a test claiming entirely objective scoring if the function involves freedom of association on the part of the subject. If the child must choose from a certain number of options, the scoring can be made objective, but if he is at

Generated on 2020-12-23 00:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

liberty to exercise imagination and freedom of association, then the scoring can scarcely be entirely so. Since many traits — initiative, independence, constructive imagination, etc. — involve complete or nearly complete freedom of association, and since these are traits of high importance, we should keep them as things to be tested and not expect the objectivity in scoring that can be easily attained in spelling, arithmetic, history information, etc.

The measure of objectivity of scoring is the correlation between the scores given to the same set of papers by two equally competent independent scorers. For single school grades this correlation runs from about .60 for composition, which is very low for a reliability of scoring coefficient, to .99 or higher for algebra and arithmetic tests. Tests of the other school subjects generally lie between these values. One would expect spelling to yield a high reliability of scoring coefficient, but it quite commonly does not do so, due to carelessness of scorers in noting details.

20. The reliability of a test score. The unreliability of a test score is of course influenced by the unreliability of scoring, but this is only one of the causes. Generally a more potent cause is the unreliability of the sampling of the child's capacity. In a "free" test, such as the assignment that the child write a composition, we may call this the variability-of-performance factor. The child "just happens" to get started well and write a better than usual composition on "What I Should Like to Do Next Saturday," or again, he "just happens" to find little to say upon "The Most Exciting Ride I Ever Had," and so it goes. Under these conditions there is wide variability in the merit of performances at different sittings, and accordingly any single composition, even if it could be very accurately scored, would be an inaccurate index of the child's average ability. In a "controlled" association test the number of items or ques-

Generated on 2020-12-23 00:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

tions is of course limited, and again, the score made will likely vary considerably from average or true ability. If it is known as a result of a careful investigation that one hundred geometry exercises are of equal difficulty for high school students of geometry, nevertheless the score a child makes on any ten of them will likely vary from his score on another ten. This is because sampling a child's ability through the medium of ten questions is not a sufficiently extensive sampling to yield a satisfactory score. Thus, finally, we shall conclude that the unreliability of a test score is caused (*a*) by too limited a sampling of individual ability, (*b*) by variability in individual performance, and (*c*) by unreliability of scoring. All three of these influences must be small before we have a highly reliable scoring. The unreliability of an arithmetic test is almost entirely due to the limitation of sampling, while that of a composition score is decidedly affected by all three, though more by the variability of performance than by the other two. It is desirable in attempting to locate errors to think of these three causes, but for most purposes the measure of unreliability needed is one that combines all three. This we have in the reliability coefficient.

21. The reliability coefficient. The reliability coefficient is the correlation of the scores of the same individuals upon two successive similar tests. To illustrate: Let us have available two reading tests, which are equally difficult and basically measure the same function. That these conditions are met is a matter that has concerned their author, and we shall not investigate this here. One of these, Form 1, is given to a class under the conditions as laid down in the Manual of Directions, and scored. A day or a week later (but not so much later that decided growth in the function has taken place[1]) the second test, Form 2, is given and

[1] A recent study by Dr. Ella Woodyard (1926) indicates that an elapsed time of a year between tests is not too great.

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

scored, by an equally competent but different scorer. We then have two scores for each pupil. The conditions of giving, the average condition of the pupils, the difficulty of the questions, and the conditions of scoring have been equally excellent throughout. The correlation between these two sets of scores is the reliability coefficient, because it is the correlation coefficient between two sets of similar measures.

22. Similar forms. It is frequently difficult to insure strict similarity in our measures. The first taking of the test changes the nature of the second test. It affects it in two ways, — added familiarity and practice lead to an improved score, and lack of novelty leads generally to lessened effort. If these two influences affect every child alike, it will make the average score on the second test somewhat different from that on the first, but it will not change the correlation between the two, so that we should still have a correct reliability coefficient. The several children in a homogeneous grade group are probably influenced at the time of the second test in much the same manner because of having taken the first, so that the correlation coefficient obtained is a quite reasonable measure of the reliability; but we cannot establish this point beyond a doubt.

23. The retesting coefficient. At times it has been attempted to obtain the reliability coefficient, when but a single form of a test was available, by giving it twice, but the correlation coefficient hereby found is very misleading and in general higher numerically than the correct reliability coefficient. This is because there is a correlation between errors. If a child is confronted with a question on Monday and reaches an answer by a certain mental process, there is a strong mental tendency for him to repeat the process when given the same question on Tuesday. Thus, whether it is right or wrong, it is merely a repeated process. The mental operation on Tuesday is not at all of the same sort as the opera-

Generated on 2020-12-23 00:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

tion on Monday. On Tuesday the main feature is memory, or a reinstating of what took place Monday, while on Monday the question was a typical problem situation not involving these memory elements at all. We shall call the correlation between repeated tests a retesting coefficient and attach little importance to it. If known, we shall consider it as a value which is greater than the correct reliability coefficient.

The objection here made to retesting coefficients does not hold in certain situations. Thus, if the test on Monday is to make as many dots as possible in 30 seconds, and the test on Tuesday is the same, then the correlation between these two results may be considered a true reliability coefficient, for it is absurd to think that there could be any memory transfer specifically influencing the second result. This sort of test is, however, not the typical school subject-matter test, and we may therefore in general object to the use of retesting coefficients as reliability coefficients.

24. The split-test method. A much better procedure, if but a single form of a test is available, is to split it into two comparable halves, determine the score on each half, correlate these, and then by the Spearman-Brown formula given below estimate what the correlation would be if the entire first form had been correlated with a second similar form, had it been available. Let us consider the possibility of splitting a test into two comparable halves. Many tests are built up of elements or questions 1, 2, 3, 4, 5 . . . of increasing difficulty, each of which is thought to measure the single capacity represented by the name of the test. A very good illustration of this type is a 20-word spelling test, the words having been chosen so as to increase regularly in difficulty. Such a test may easily be split into comparable halves by taking the odd-numbered words as one half and the even-numbered words as the other, or by taking words 1, 4, 5, 8, 9, 12, 13, 16, 17, 20 as one half and 2, 3, 6, 7, 10, 11, 14, 15, 18, 19 as the

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

other half. The entire test may be given as usual, but the score on each half is to be determined separately and recorded, giving two scores — or let us call them half scores — for each pupil. The correlation between these half scores may be represented by $r_{\frac{1}{2}\mathrm{I}\,\frac{1}{2}\mathrm{II}}$. Having this correlation, we may easily obtain an excellent estimate of what the correlation between the total score on the 20-words and a second total score on 20 other equally difficult words would be had we given and scored the second list. This correlation we shall call the reliability coefficient of the 20-word spelling test and designate it by $r_{1\mathrm{I}}$. It is given by the Spearman-Brown formula:

$$r_{1\mathrm{I}} = \frac{2r_{\frac{1}{2}\mathrm{I}\,\frac{1}{2}\mathrm{II}}}{1 + r_{\frac{1}{2}\mathrm{I}\,\frac{1}{2}\mathrm{II}}}$$

(For estimating the reliability of an entire test, knowing the reliability of the half test) . . . . . [1]

It not infrequently happens that a test cannot be split into comparable halves. If but one form of a test which is largely a speed test, such as Courtis's test in fundamentals in arithmetic, is given, there is no way of dividing it into comparable halves. In the Courtis test the number of problems correctly added in 8 minutes is the total addition score. Suppose this number is 9, 5 odd-numbered and 4 even-numbered problems. The 5 and 4 are not independent measures, as both have been affected by the same time limit and the same idiosyncrasies pertaining to the particular performance. Thus, if the child got confused on Problem No. 3 and took a long time for it, he has lowered his score not only on the odd problems, but also on the even problems. Clearly, this test cannot be split into independent halves. In general, speed tests cannot be so divided, and therefore the only sound way to obtain a reliability coefficient is to give at a later time a second similar form of the test, correlate the score on the two forms, and thus directly obtain $r_{1\mathrm{I}}$.

Generated on 2020-12-23 00:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The reliability of the measures used enables us to determine the reliability of judgments based upon test scores, and is therefore an important feature to know. Before considering a typical test problem, the mechanical steps involved in calculating a correlation coefficient and the probable error of a score should be at hand. If the reader is not already familiar with the simple statistical techniques involved, particularly those concerned with probable errors, he should at this point read Chapter VII on "Elementary Statistical Procedures." No attempt is made in Chapter VII to prove the formulas involved. For this the reader is referred to texts on statistics (Chaddock, 1925; Chambers, 1925; Garrett, 1926; Jones, 1921; Kelley, 1923 stat.; Odell, 1925; Pearl, 1923; Rietz et al., 1924; Rugg, 1917; Thorndike, 1904 and 1913 ment.; Thurstone, 1925).

Generated on 2020-12-23 00:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER THREE

The Measurement of Group Achievement

1. Two types of survey tests. If it is desired to compare grade with grade, we need a test which can be given to several successive grades. Either the same test must be given to the pupils in the different grades or a very careful preliminary study must have been made enabling a comparison, let us say, of the third-grade scores on the third-grade test with the fourth-grade scores on the fourth-grade test, etc. Such comparative studies have been made by certain authors of tests, and there are decided advantages in this procedure in that the test given to the third grade need not be encumbered with second-, fourth-, etc., grade material; the test given to the fourth grade may be specifically adjusted to their needs, and so forth. An excellent sample of this type of test is the Monroe Standardized Reasoning Test in Arithmetic.

However, the mechanical difficulties arising from the fact that different test blanks are required for different grades, and the statistical difficulties of comparing second-, third-, fourth-, etc., grade scores made on different tests, have resulted in this type of test being much less common than the achievement test which begins with easy material and continues on into much more difficult subject matter, permitting the same test to be given to a wide range of school grades and making possible a direct comparison of gross scores. Though the first type has very real advantages (and we may expect to see still better ones of this sort devised and widely used), we shall here consider making a school survey based upon a test of the second kind. A good illustration of this second type of test is the Woody-McCall Mixed Fundamentals Arithmetic Test, devised to be applicable in the third to eighth grades inclusive.

Generated on 2020-12-23 00:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

2. The relation between test used and purpose. In considering a survey, the first and most important question to ask is, "What is its purpose?" If the answer is (*a*) "to secure an idea of the difference in general scholastic success of different grades and classes throughout the system," we should have to conclude that the Woody-McCall test was not adequate for this purpose, as it measures but certain phases of a single subject. If we desire (*b*) "to measure the differences of classes in arithmetic ability," we must examine the test closely, for it measures only certain phases of arithmetic. Reference to the information about the test given in Chapter X, page 323, shows that one of the authors specifically states that the test does not measure "(1) arithmetic of the problem variety; (2) arithmetic beyond fundamentals in integers, fractions, and decimals; and (3) exact measures of rate." An examination of the specific questions of the test and of the time limits would seem to confirm this view, and we shall conclude that the test is not adequate for purpose (*b*) unless abilities (1), (2), and (3) are so similar or highly correlated with ability in fundamentals in arithmetic as measured by the test that they do not need to be measured separately. Few would be inclined to conclude that speed in computation, problem solving, and fundamentals in computation are each adequately measured by a mixed-fundamentals test; so we shall rule the test out as an instrument of measurement for purpose (*b*). If the purpose is (*c*) to measure the differences in computation ability between classes, irrespective of speed in the fundamental arithmetic operations and of ability with written problems in arithmetic, we may safely conclude that the test is appropriate, provided only that it is sufficiently reliable. The reliability coefficient of the test for a single grade range is in the neighborhood of .60 (as given in Chapter X), and as we require a reliability of only .50, it is entirely satisfactory for

Generated on 2020-12-23 01:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

this purpose. We shall thus assert that our purpose is purpose (*c*) and proceed with the testing program.

3. Giving the test. The directions accompanying the test are explicit, and with such instruction as might be given by the superintendent or principal verbally at a teacher's meeting or in a circular letter, it might be expected that the class teachers would be competent to give it properly. This, however, assumes a knowledge of experimental technique that teachers are not commonly equipped with, not to mention a very rigorous sense of honesty. Miss Blank, teacher of the fourth-grade class, finds in the test certain very simple things that she has not taught her pupils, and without any real thought of dishonesty, informs them that a certain thing means "take away," which way of expressing it, of course, they all understand, and furthermore the time limit is not quite fair, — things do not get started well, "they asked questions and were nervous," so she adds ten seconds to the stipulated time, again not feeling that it is any more than fair to her pupils. These things, of course, are not to be tolerated in a standardized test, and the only way to insure against them is to have the testing done by some one other than the teacher. It is an improvement to interchange teachers, but still better to employ a small group, specially trained by the superintendent or principal, to do all the testing in all the classes. It seems that only by so doing can uniform procedure and entirely comparable results be secured.

The reader must not assume that no words with pupils other than those printed on the directions sheet are admissible. Such statements as the following in response to questions are generally in order: "Work on the margin. You do not need scratch paper"; "You may use either pen or pencil"; "If you do not know what that word (symbol, question, sign) means, go on to the next question, because I must not tell you. You will probably know the next one";

Generated on 2020-12-23 00:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

" Yes, when you have finished the first column, go on to the next one "; etc. All comments such as these which are outside of the printed directions should be made in a low voice to individual pupils as need arises. It is a good practice to forbid the asking of questions while the test is in progress except upon raising the hand and after the tester has reached the desk of the pupil so that the question can be put in a whisper. To refuse to attend to a child who is frantically perplexed because he does not know whether to write his answers under the questions or in the margin is not good standardized procedure. The examiner should be free to say or do anything that does not disturb or delay pupils at work, that does not help the individual child in the thing in which he is being tested, and that does set him to work again after some foolish or trivial issue has troubled him. Teachers have been known to translate sentences into juvenile or baby talk that they may be understood; to say in effect, " This is like what we did yesterday "; or sympathetically to encourage a pupil by saying, " That is all wrong. You ought to know better than that "; yes, even to say, " Now, Johnny Jones, don't you dare cheat today "; and to say all these things in a loud, penetrating voice, oblivious of the fact that they are thereby the worst of violators of standardized procedure requirements. No set of rules laid down here can meet the odd and ridiculous situations that arise in class. To the competent, level-headed examiner these situations are not even annoying — in fact, the humor in them is generally one of the enjoyable features of the work.

The test should be given to all the classes upon the same day, or at least within a few days. The period of the day in which the testing is done is immaterial, provided only that there are no interruptions such as an assembly cutting in, a boisterous mob on the playground outside, etc. In the high school such things as dances late into the preceding night

Generated on 2020-12-23 01:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

should be considered. It is impossible to anticipate discommoding circumstances, and it is well to change a testing program even at the last minute if unforeseen situations arise.

4. Scoring the papers. The test given, the papers are to be scored. Here again it is not advisable to ask the teachers to score the papers of their own pupils. An interchange of papers between teachers is a slight improvement, but much greater accuracy is secured and less total effort on scoring spent if all scoring is done in a central office by a specially selected group of teachers or clerks. Generally speaking, if a careful scoring plan is followed, the speed of scoring may be at least doubled over the speed of a single teacher scoring a single class, and the accuracy will be very much more than doubled. It is, accordingly, generally not an expensive task to have all scoring and tabulating done by clerks whose work, while learning, is carefully checked by a competent supervisor.

5. Tabulations and computations. We may perform the requisite tabulations and computations for a single class as a sample of what is to be done for each class. (See Table 2, on following page.) The raw data given in the table are the actual records of a high eighth-grade class. The mean records given later for other grades and classes are hypothetical, but we shall consider them to be the actual records for an entire school system, in order to bring out the appropriate steps of interpretation.

Should the reader compare the scores in the table with the published norms for this test, he may be surprised at the wide spread found. It should be said, however, that these are the scores made by an actual eighth-grade group, which, as far as the writer can otherwise judge, is a typical grade.

The sum of these scores, divided by the number of them, gives the mean. This is not a long process, but as the " method of moments " is a numerically simpler process and is very serviceable in further work, it will be followed here.

Generated on 2020-12-23 01:12 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 2

SCORES MADE BY THE HIGH EIGHTH-GRADE CLASS OF THE DEWEY JUNIOR HIGH SCHOOL, MAY 28, 1926, ON THE WOODY-MCCALL MIXED FUNDAMENTALS ARITHMETIC TEST: FORM 2

NAME OF PUPIL	SCORE: NUMBER RIGHT
Ida A.	30
Robert B.	24
Albert B.	33
Frank B.	24
Letha C.	33
George C.	23
Lucy C.	31
Grace C.	33
Gladys C.	31
Doris C.	32
Wayne D.	26
Alice E.	31
Jonathan F.	28
Horace H.	32
Franklin H.	29
Jack K.	28
Clark L.	32
Jeannette L.	30
Carmine L.	28
James M.	34
Helen N.	30
Sarah P.	31
Florence P.	29
Alton P.	23
Anna R.	34
Marion R.	30
May S.	32
Emily S.	31
Ethel S.	31
Earl S.	27
Fred S.	30
Alice S.	32
Elbert T.	33
Ethel T.	28
Marion V.	29
George Z.	27

Generated on 2020-12-23 01:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

From the preceding scores we obtain the following:

TABLE 8

Tally Sheet			Computation of Mean			
	Scores	Tally	f	ξ	$f\xi$	
	23	||	2	− 7	− 14	
	24	||	2	− 6	− 12	
	25		0			
	26	|	1	− 4	− 4	
	27	||	2	− 3	− 6	
	28	||||	4	− 2	− 8	
	29	|||	3	− 1	− 3	
Arbitrary Origin—						− 47
	30	𝍸	5	0		
	31	𝍸 |	6	1	6	
	32	𝍸	5	2	10	
	33	||||	4	3	12	
	34	||	2	4	8	
			36			36
						− 11

$$M = \text{Arb. Orig.} + i\,\frac{\Sigma\xi}{N} \qquad \text{(The mean computed from an arbitrary origin)} \quad . \quad . \quad . \quad . \quad [2]$$

The computation of the mean given herewith follows exactly the same lines as in Chapter VII, Section 2, where it is explained in much greater detail.

In this formula "Arb. Orig." is the value of the gross score from which deviations are taken (any convenient gross score may be chosen); ξ stands for a score as a deviation from this arbitrary origin; i is the size of the ξ interval (that is, the number of X, or original test score, units corresponding to one ξ unit); $\Sigma\xi$ is the sum of all the ξ deviations, taking each deviation as many times as there are individuals having this deviation. This summation, $\Sigma\xi$, is sometimes written $\Sigma f\xi$. The two things are identical in meaning. In the present

Generated on 2020-12-23 00:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

problem Arb. Orig. = 30, i = 1.00, N = 36, and $\Sigma\xi = -11$. Thus the required value of the mean is:

$$M = 30 + 1\frac{(-11)}{36} = 29.7$$

This illustrates the computation for a single grade. The results for the several schools and classes may be brought together as in Table 4.

TABLE 4

THE MEAN SCORES MADE BY EACH CLASS IN THE WESTERNVILLE ELEMENTARY SCHOOLS, MAY 25–29, 1926, ON THE WOODY-MCCALL MIXED FUNDAMENTALS TEST

(Number of pupils given in parentheses)

SCHOOL GRADE	CUBBERLEY SCHOOL	DEWEY SCHOOL	THORNDIKE SCHOOL
Low third	13.4 (30)	10.5 (42)	16.0 (43)
High third	16.7 (34)	15.0 (60)	18.5 (40)
Low fourth	18.2 (29)	17.9 (41)	20.8 (33)
High fourth	21.0 (36)	21.1 (65)	21.9 (36)
Low fifth	21.9 (27)	20.7 (35)	24.2 (32)
High fifth	23.7 (30)	24.3 (56)	26.7 (29)
Low sixth	26.7 (22)	27.1 (28)	28.6 (31)
High sixth	27.4 (28)	28.8 (44)	29.3 (38)
Low seventh	28.8 (23)	29.4 (34)	29.2 (44)
High seventh	29.8 (32)	29.0 (40)	31.4 (32)
Low eighth	31.1 (26)	30.1 (30)	31.0 (43)
High eighth	33.0 (29)	29.7 (36)	32.8 (33)

6. Use of local norms. The first comparison which is ordinarily of value is that of each grade with the average of such grades for the city. The Cubberley School low third makes an average score of 13.4; the Dewey low third scores 10.5; and the Thorndike low third, 16.0. If we add these three scores and divide by 3, we shall obtain a city average giving just as much importance, or weight, to the Cubberley School low third-grade record as to the Dewey and Thorn-

Generated on 2020-12-23 00:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

dike low third-grade means. This method of averaging is frequently followed, but in general it is more equitable to calculate the city average after weighting the separate school-grade records according to the number of pupils in each grade. We shall thus calculate the low-third city average, as in Table 5:

TABLE 5

CALCULATION OF A CITY GRADE MEAN

GRADE	f NO. OF PUPILS	X GRADE AVERAGE	fX
Cubberley low third	30	13.4	402.0
Dewey low third	42	10.5	441.0
Thorndike low third	43	16.0	688.0
	$115 = N$		$1531.0 = \Sigma X$

$$\text{Grade mean for entire city} = \frac{\Sigma X}{N} = \frac{1531.0}{115} = 13.3$$

We thus see that the Cubberley School is very close to the city average, the Dewey School about three units below, and the Thorndike School about an equal amount above. Proceeding in the same manner for all the other grades, we obtain city average grade scores, as in Table 6 (page 53).

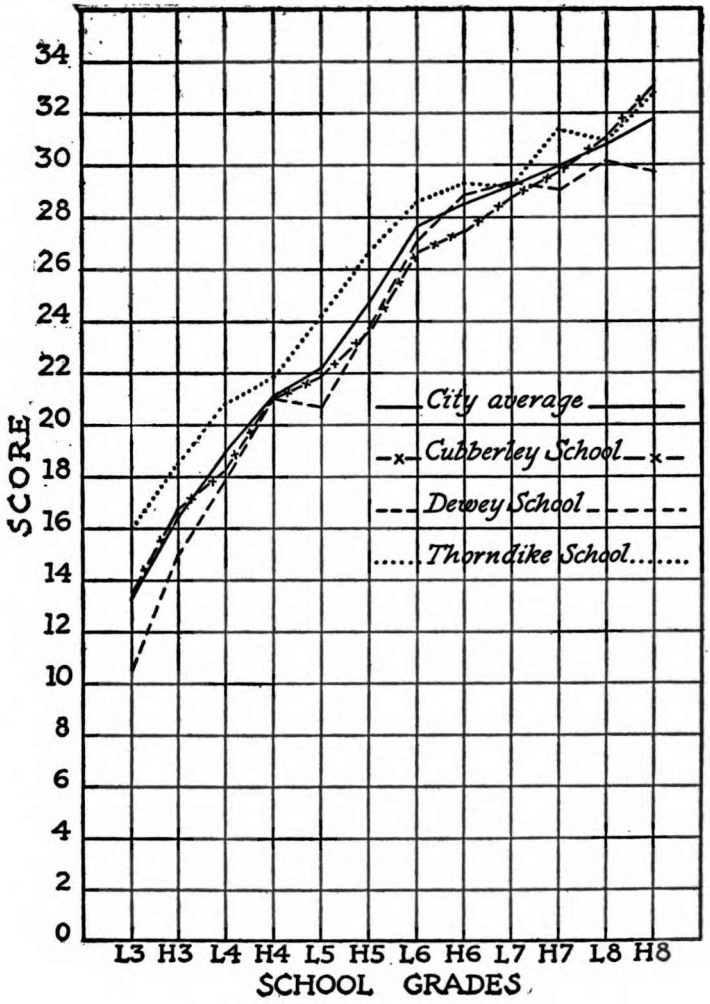
A comparison of each school grade with the city standards is readily made by means of a graph, as shown in Chart 1, on the next page. Before attempting to interpret the differences between schools revealed by this chart, we should first secure some idea as to the probable error of our mean grade scores.

7. The probable error of class means. Let us calculate the probable error of the Dewey eighth-grade mean, for

Generated on 2020-12-23 00:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHART I

AVERAGE SCORES BY GRADE, WOODY-MCCALL MIXED FUNDAMENTALS, WESTERNVILLE, MAY 25–29, 1926



Generated on 2020-12-23 00:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 6

CITY GRADE MEANS

GRADE	CITY NORM
Low third	13.3
High third	16.5
Low fourth	18.9
High fourth	21.3
Low fifth	22.2
High fifth	24.7
Low sixth	27.6
High sixth	28.6
Low seventh	29.2
High seventh	30.0
Low eighth	30.8
High eighth	31.7

which we have the necessary data conveniently recorded in Table 3. The probable error of the mean is given by the formula,

$$\text{P. E.}_M = .6745\frac{\sigma}{\sqrt{N}} \quad . \quad . \quad . \quad . \quad . \quad [3]$$

in which N is the population, 36, and σ is the standard deviation of the scores of the members of the class. In calculating this standard deviation, we may utilize the steps already performed in the calculation of the mean (Table 3).

The computation of the standard deviation shown in Table 7, on the following page, parallels that of Chapter VII, Section 3, where it is described in much greater detail. The formula for the standard deviation is:

$$\sigma = i\sqrt{\frac{\Sigma\xi^2}{N} - \left(\frac{\Sigma\xi}{N}\right)^2} \quad . \quad . \quad . \quad . \quad . \quad [4]$$

in which N, i, and $\Sigma\xi$ have already been defined. The quantity $\Sigma\xi^2$ is to be calculated as shown in the last col-

Generated on 2020-12-23 00:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 7

COMPUTATION OF THE STANDARD DEVIATION

x	f	ξ	$f\xi$	$f\xi^2$
23	2	− 7	− 14	98
24	2	− 6	− 12	72
25	0			
26	1	− 4	− 4	16
27	2	− 3	− 6	18
28	4	− 2	− 8	16
29	3	− 1	− 3	3
30	5	0	− 47	
31	6	1	6	6
32	5	2	10	20
33	4	3	12	36
34	2	4	8	32
	$36 = N$		$-11 = \Sigma\xi$	$317 = \Sigma\xi^2$

umn of Table 7, and in our problem equals 317. Thus we have:

$$\sigma = 1.00\sqrt{\frac{317}{36} - \left(\frac{-11}{36}\right)^2} = 2.95$$

We may now use Formula 3 for the probable error of the mean, and for this eighth-grade class of 36 we obtain:

$$\text{P.E.}_M = \frac{.6745(2.95)}{\sqrt{36}} = .33$$

The populations of the other classes do not differ greatly from 36, and probably the standard deviations of the scores of the other classes will be in the general neighborhood of the standard deviation for this eighth grade, — namely, 2.95, — so we shall not be far astray if we take .33 as the probable error of each of the grade means.

8. The interpretation of differences in class means. If we now look again at Chart 1 and keep in mind that one third

Generated on 2020-12-23 00:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of a unit is the approximate probable error of means, we see that the Cubberley School, except in the eighth grade, scarcely differs enough from the city standards for the amount to be significant, though there is some evidence that the Cubberley eighth grade is superior to the city average. The Thorndike School is superior, with a few exceptions, all the way from the third grade to the eighth grade, while the Dewey School is inferior in the low and high third, low fourth, low fifth, high seventh, and high eighth grades.

Conclusions as stated complete the statistical survey. When we go beyond these conclusions and assign causes to the differences found, we pass beyond the field of statistical evidence. If we conclude that the Dewey eighth-grade teacher is a poor teacher of computation, we may be right and again we may be wrong, for there are other possible explanations of the poor showing of the Dewey eighth-grade class — the children may be natively less well endowed than the children of the other eighth grades; they may have an enriched curriculum that cuts short the time that they can devote to computation; etc. We must know the facts of class achievement before further reasoning is possible; but let us clearly distinguish between the facts provided by statistics and the further deductions, and not be guilty, as superintendents have been known to be, and report to the teacher of the Dewey eighth-grade class that "statistics prove you are a poor teacher." Without making or implying any such judgment, the superintendent may very reasonably say to this teacher: "Your children are not doing as well in computation as we expect in this city. Do you have an explanation for this, and can you improve the situation?"

The survey as outlined is complete, in the sense that the achievement data are available for all intra-city grade and school problems dealing with computation.

Generated on 2020-12-23 00:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

If in addition to this the superintendent desires to compare his schools with those of other cities, he will need outside norms. Such a comparison is of doubtful value or even noxious in its effects, unless very careful steps are taken to insure that age and race are not the chief causal factors determining the showing made by a city instead of, as is usually assumed to be the case, the excellence of the instruction. A superintendent may, by bringing about excessive retardation of his pupils, raise almost indefinitely the levels of attainment of his various school grades. A good showing created in this manner is most unwholesome, as it ultimately leads to elimination of pupils from school long before they have completed the high school and before they have had the advantages of a differentiated and a partly elective curriculum. These advantages are of special value to the child who does not continue into higher education.

In Chapter II, six purposes served by educational tests were listed, three of them group purposes, as follows:

1. Group survey and prognosis, with reference to general group success in school work.

2. Group survey and prognosis, with reference to a single subject.

3. Group survey and prognosis, with reference to group differences in ability and achievement in two or more specific subjects.

The illustrative computation just completed is typical of a study of the second sort. A study meeting the first purpose would involve the same steps as are here illustrated, the difference being in that the test employed would be a general, all-round, school achievement test instead of a specific subject test. Several such achievement batteries are listed in Chapters IX and X. A survey of the third sort involves two or more achievement tests, each of considerable reliability and chosen so as to reveal differences in achievement along the

Generated on 2020-12-23 00:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

line, or lines, of interest to the experimenter. It involves for each test the same steps of calculation as are here performed for the computation survey, plus certain additional features as outlined herewith. Let us suppose an investigation is being made of the difference in achievement in reading and in computation of the Dewey high eighth grade. Suppose the reading ability is measured by the ABC Reading Test and the computation ability by the Woody-McCall test. We shall let symbols with the subscript 1 stand for reading, and those with the subscript 2, for computation. After giving the test, scoring the papers, and making calculations as already outlined, we shall have:

N = Number of pupils tested with both tests (omit from all calculations those pupils who took one test only)
M_1 = Mean score of class in reading
σ_1 = Standard deviation of scores of class in reading
M_2 = Mean score of class in computation
σ_2 = Standard deviation of scores of class in computation

In addition to these constants we need r_{12}, the correlation between the scores in reading and those in computation. This is to be calculated as illustrated in Chapter VII. Having determined these things, we shall have numerical values, let us say, as follows:

N = 36 (Population)
M_1 = 84.0 (Mean score in reading test)
σ_1 = 8.00 (Standard deviation of scores in reading test)
M_2 = 29.7 (Mean score in computation test)
σ_2 = 2.95 (Standard deviation of scores in computation test)
r_{12} = .60 (The correlation between reading and computation test scores)

Generated on 2020-12-23 00:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

It is very clear that without more information than is represented by these six constants, it is impossible to say whether or not the class stands relatively higher in reading than in computation. The mean reading and computation scores must each be compared with some sort of standards. Here again city norms will be the most meaningful. Let us say that such have been determined and are as follows:

TABLE 8

WESTERNVILLE SCHOOLS: MEAN GRADE SCORES IN THE ABC READING TEST AND THE WOODY-MCCALL COMPUTATION TEST, MAY 28, 1926

GRADE	READING	COMPUTATION
Low sixth	73.5	27.6
High sixth	75.7	28.6
Low seventh . . .	78.1	29.2
High seventh . . .	80.4	30.0
Low eighth	82.0	30.8
High eighth	84.2	31.7

If we now compare the scores of the Dewey high eighth-grade class with the grade means, we see that the reading score is below the high eighth local norm by .2 reading-test units and that the computation score is below the high eighth norm by 2.0 computation-test units. We can now say that both scores are below the city average. Further, the Dewey high eighth grade is, in reading, much closer to the high eighth norm than to the low eighth, while in computation it is between the low seventh- and the high seventh-grade norms. We may therefore say that the class is considerably lower in computation than it is in reading.

A somewhat different procedure, based upon high eighth-grade norms only, will lead to a conclusion of the same general import. If we have the high eighth-grade norms only, it is not immediately obvious which mean is the poorer,

Generated on 2020-12-23 00:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

because 1.0 reading-test unit is not comparable to 1.0 computation-test unit. Thus, it is not clear which has the greater significance, .2 reading-test unit or 2.0 computation-test units. We may secure comparable units by dividing the reading-test difference by the standard deviation of the high eighth-grade reading-test scores, and also by dividing the computation-test difference by the standard deviation of the computation-test scores. Thus:

$\frac{-.2}{8.0} = -.025$ (Deviation of average reading-test score from the norm measured in reading-test standard deviations)

$\frac{-2.0}{2.95} = -.678$ (Deviation of average computation-test score from the norm measured in computation-test standard deviations)

.653 (The number of standard deviations that reading score is superior to computation score)

Let us express these steps in symbols. The symbol ${}_{8}M_1$ will stand for the high eighth-grade city norm in the reading test, and ${}_{8}M_2$ for the high eighth-grade city norm in the computation test. With this notation, the $-.2$ reading-test difference is represented by $(M_1 - {}_{8}M_1)$, and the -2.0 computation-test difference is represented by $(M_2 - {}_{8}M_2)$. The standard deviations 8.0 and 2.95 are represented by σ_1 and σ_2, respectively. Thus, the difference .653, which we will represent by the letter d, is given by:

$$d = \frac{(M_1 - {}_{8}M_1)}{\sigma_1} - \frac{(M_2 - {}_{8}M_2)}{\sigma_2} \quad . \quad . \quad . \quad [5]$$

To know whether the difference is significant or not, one must have the probable error of d. If ${}_{8}M_1$ and ${}_{8}M_2$ are based on a rather small number of classes, then the probable error of d is rather difficult to calculate. A formula for the probable error under these conditions will not be given here.

Generated on 2020-12-23 00:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

If, however, the grade norms ${}_{8}M_1$ and ${}_{8}M_2$ are determined from very much more extensive data than are M_1 and M_2 (for example, if ${}_{8}M_1$ is based on 25 high eighth-grade classes and M_1 based on one such), then the formula for the probable error of d is readily obtained and easy to use. It is given here without proof:

$$\text{P. E.}_d = .6745\sqrt{\frac{2 - 2r_{12}}{N}}$$

(The probable error of the difference between two mean scores, when each mean is expressed in standard deviation units) . . . . . [6]

If $N = 36$ and $r_{12} = .60$, we immediately obtain for our present problem:

$$\text{P. E.}_d = .6745\sqrt{\frac{.80}{36}} = .10$$

We may thus write:

$$d = -.65 \pm .10$$

The superiority of the Dewey eighth grade in reading to computation is clearly established, for the difference here found, $-.65$, is six and one half times its probable error. This finding has required the use of a standard as to what constitutes equal achievement along these two lines. The mean high eighth-grade reading score for the entire city, ${}_{8}M_1$, is taken as representing a level in reading equal to that in computation given by the mean high eighth-grade computation score, ${}_{8}M_2$. If one desires some other, say a national, standard, it may of course be used. The statistical treatment and argument would be of the same type throughout. Since $d = -.65 \pm .10$, the statistical conclusion is that, accepting the standards in the two subjects as being equal in a developmental sense, it is then established that there is unequal achievement upon the part of this particular high eighth grade in these two school subjects. The cause of this variance is not revealed, and one should be slow in attributing

Generated on 2020-12-23 00:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

it to the pupils, the teacher, the climate, or any other specific thing. The writer is familiar with one school system in which repeated testings of different classes and for different years have quite uniformly yielded higher computation scores than reading scores as judged by a comparison with national norms. He has also been informed by several teachers of this system that the superintendent exerts a more uniform pressure for good work in mathematics than for good work in any other school subject. This suggests that the cause of the difference found is, in this case, the superintendent. Thus, it is further suggested that relative group differences of this sort are quite definitely amenable to environmental influences. Other data at hand suggest a strong hereditary influence affecting differences in achievement within the individual, in reading and computation. An accurate evaluation of the various causes of individual and group differences is still to be made. We certainly must not take the position that the child's inheritance determines his general level of ability and his environment determines the differences found between his abilities, nor should we believe that all of his special abilities are given by a differentiated inheritance. Undoubtedly, a middle view between these two, located just where we do not as yet know, pictures reality — the alluring, pulsating battlefield and playground of the developing child.

A real understanding of group achievement must ultimately be based upon a grasp of individual achievement and differences in achievement, which topics are investigated in Chapters IV, V, and VI.

Generated on 2020-12-23 01:13 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER FOUR

The Measurement of Individual Achievement

1. The problems of individual measurement. It has already been mentioned that there are three kinds of problems which are commonly of interest in connection with individual measurement: the measurement (*a*) of general all-round achievement; (*b*) of achievement along some one special line; and (*c*) of oddity or singularity in achievement. Problems (*a*) and (*b*) are dealt with in the next three sections, and problem (*c*) in Chapters V and VI.

The expression "general all-round achievement" may be somewhat too broad, for as used in this chapter it refers essentially to intellectual achievement. Man's all-round usefulness to society includes his physical fitness for labor and military service, his mental assets and talents, his willingness to devote his physical and mental talents to social ends, and his eugenic fitness for parenthood. We shall here assume a physical fitness and moral willingness to serve society, omit the most important question of all, — that of eugenic fitness, — and confine our attention to the measurement of all-round intellectual achievement and promise.

2. The measurement of achievement and of intelligence; "jingle" and "jangle" fallacies. Though the mutual resemblance of achievement and intelligence test measures has been broached several times in earlier chapters, we must now attempt to secure a more accurate idea of this similarity. We must have at least tentative answers to the highly important questions, "How much of achievement is intelligence?" and "How much of intelligence is achievement?" before we can intelligently interpret scores called by these two different names. The detailed answer to this problem must ultimately be made in terms of specific tests. Thus x per cent of the

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

ability measured by reading test A is the same thing as the ability measured by intelligence test B; y per cent of the ability measured by arithmetic test C is the same ability as is measured by intelligence test D; etc. We shall here forgo all these refinements of statement, important though they are, hoping that many of them will be answered in the not distant future, and confine our attention to the one issue, "How much of all-round scholastic achievement (the thing measured by battery school and subject matter tests) is the same as all-round general intelligence (the thing measured by tests now carrying the 'intelligence' label)?" An approximate answer to this question is reached in Sections 1, 2, and 3 of Chapter VIII, and we shall here concern ourselves with the result which, as concerning general scholastic achievement and general intelligence, is that no less than 90 per cent of the one is the same in its nature as the other. When we speak of a school child's "intelligence," meaning thereby the thing measured by intelligence tests, we are, whether we know it or not, in the same breath, to the extent of 90 per cent of the meaning conveyed, discussing his general scholastic achievement; and when we speak of a school child's "achievement," we are actually concerning ourselves in the main with his "general intelligence." The community between these two functions is nine times as great as the disparity between them, and any judgment of difference between achievement and intelligence must be based upon the 10 per cent of each not represented in the other, or it is a spurious judgment.

The glibness with which we differentiate between achievement and intelligence is explained in part by the fact that our language is at fault. To use an illustration given by Thorndike (1904, page 14), the expression "college student," found so frequently in general discussions, covers a multitude of classes: male and female; part time, full time; extension students and those in residence; native, foreign; lower class-

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

men, upper classmen, graduates; etc. In each connection the expression " college student " sounds the same, and thus we come to treat it as a single concept. Dr. Thorndike quotes Professor Aikins as describing this as the " jingle " fallacy because there is merely a verbal resemblance and no sufficient underlying factual similarity between the classes.

Equally contaminating to clear thinking is the use of two separate words or expressions covering in fact the same basic situation, but sounding different, as though they were in truth different. The doing of this latter the writer will call the " jangle " fallacy. " Achievement " and " intelligence " sound as though they were different; they have different " jangles," and thus we treat them as though they were different in truth. There is a modicum of difference between them, and in so far as this only is the issue, it is proper to distinguish between them, just as we may use two nearly related words to draw a fine distinction; thus, " He is upright but not honorable " or " He is fearful but not cowardly," etc. Literary ingenuity creates for our entertainment the man who is fearful but not a coward. It may be that such men exist in blood and bone, but certainly by no known means can the rank and file be classified separately upon these two traits. Nor can they upon the bases of achievement and intelligence. We can mentally conceive of individuals differing in these two traits, and we can occasionally actually find such by using the best of our instruments of mental measurement, but to classify all the members of a single school grade upon the basis of their difference in these two traits is a sheer absurdity. The deviation of achievement-age-minus-mental-age from zero, or of achievement-age-divided-by-mental-age from 1.00, are such measures of difference, and neither is ordinarily to be trusted.[1]

[1] Utilizing Symonds' data (1924), I find, as explained in Section 3 of Chapter VIII, strong support for the point here made, which, however, is just the opposite of the conclusion reached by Dr. Symonds.

Generated on 2020-12-23 00:23 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Though the accomplishment quotient, achievement-age-divided-by-mental-age, is not recommended or used in this text, for the reasons just given, we should consider under what extraordinary conditions its use is warranted. First, as Franzen has pointed out (1924), if the difference between a person's true achievement and his true intelligence is very great, then the evidence that there is a difference between the two is more readily demonstrable by means of achievement and intelligence tests. If we confine our attention to individuals who show by means of their scores on the best of our available tests wide differences between achievement and intelligence, and if we heavily discount such differences as found, — i.e., if we take an obtained accomplishment quotient of 140 as probably representing a true accomplishment quotient somewhere between 110 and 120, — we may then expect our judgment to be right considerably more often than wrong and proceed accordingly. Secondly, if the achievement capacity in which we are interested is not general but special, — e.g., music, computation, spelling, handwriting, etc., — then a quotient such as music-age-divided-by-mental-age has considerable likelihood of being significant, though we should note in passing that music-age-divided-by-general-achievement-age is in this case also likely to be truly significant. A consideration of differences of this second sort will be found in Chapters V and VI.

The preceding discussion has contributed only negatively to the progress of this chapter. Due to the nature of widespread practice, it has seemed necessary to give the reasons for abstaining from a type study involving the comparison of scholastic achievement with general intelligence. Having given them, we shall now proceed to a sample study of measurement of general all-round scholastic achievement by procedures to which the writer believes even those having a fondness for accomplishment quotients will not take exception.

Generated on 2020-12-23 00:24 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

3. The interpretation of individual scores made upon a battery of achievement tests. If the battery of educational tests that are used is published as a unit, — as, for example, is the case with the Stanford Achievement Tests, — then the means of obtaining a " total " score is to be looked for in the Manual of Directions accompanying the tests. Any one can, however, select a battery of educational tests as he sees fit and combine them in a reasonable manner to obtain a total achievement score. Let us do this, using the battery recommended by one of the judges, as indicated in the footnote, page 230. The tests recommended are the Thorndike-McCall Reading, the Woody-McCall Arithmetic, and the Morrison-McCall Spelling tests. Let us be given scores as indicated in Table 9 on these three achievement tests; let us build up a scheme for combining the separate scores into a total score; and let us determine total achievement scores for each pupil.

The means and standard deviations of Table 9 have been calculated by the usual methods.

There are a number of things which should be attended to in combining the scores of the three tests into a grand total achievement score. In order that the particular units of measurement may not be a determining factor, we must " weight," or give an importance to, each test inversely as its standard deviation. We should also weight each test approximately as its importance for the composite desired. Thus, if handwriting is considered less important than reading when measuring general all-round achievement, we should weight handwriting much less than reading.

Further, we should weight each test greater the higher its reliability, and finally, we should weight each in accordance with its independence of the others. Thus, if we are combining three tests, but two of them are almost identical in what they measure, we should weight each of these two less in com-

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 9

SCORES[1] OF PUPILS IN THE HIGH EIGHTH-GRADE CLASS, DEWEY JUNIOR HIGH SCHOOL, WESTERNVILLE, TESTED MAY 25–29, 1926

NAME OF PUPIL	THORNDIKE-MCCALL READING TEST	WOODY-MCCALL ARITHMETIC TEST	MORRISON-MCCALL SPELLING TEST
Ida A.	27	30	34
Robert B.	22	24	25
Albert B.	30	33	37
Frank B.	29	24	40
Letha C.	30	33	36
George C.	24	23	38
Lucy C.	28	31	36
Grace C.	33	33	49
Gladys C.	28	31	50
Doris C.	27	32	38
Wayne D.	20	26	29
Alice E.	32	31	35
Jonathan F.	30	28	43
Horace H.	26	32	31
Franklin H.	21	29	22
Jack K.	28	28	27
Clark L.	30	32	48
Jeannette L.	26	30	42
Carmine L.	25	28	20
James M.	31	34	44
Helen N.	24	30	41
Sarah P.	22	31	29
Florence P.	33	29	50
Alton P.	27	23	32
Anna R.	29	34	47
Marion R.	29	30	50
May S.	30	32	30
Emily S.	29	31	42
Ethel S.	29	31	40
Earl S.	32	27	32
Fred S.	27	30	46
Alice S.	35	32	50
Elbert T.	29	33	43
Ethel T.	28	28	43
Marion V.	32	29	46
George Z.	31	27	38
Means { Raw scores	28.1	29.7	38.4
Means { McCall *T* scores	61		
Standard deviations: Raw scores	3.45	2.95	8.18

[1] Except for the Woody-McCall arithmetic scores, the data are hypothetical.

Generated on 2020-12-23 00:24 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

parison with the third than we should if they were quite independent of each other. The combination of these four weighting factors into a single weight for each test is accomplished most neatly and accurately by means of a multiple regression equation, connecting the three measures with an independently determined criterion measure of general achievement. If, however, we do not have this criterion measure, we must use our best judgment in the matter in lieu of the appropriate regression equation. A table for recording judgments on the items mentioned may be conveniently drawn up in connection with one giving standard deviations and reliability coefficients. Table 10 provides a convenient layout for the work.

TABLE 10

A	B	C	D	E	F	G	H	I
Test	Standard Deviations, σ	Reliability Coefficients for One-half Grade Range, r_{11}	$\frac{\sqrt{r_{11}}}{1-r_{11}}$	Judgments of Person Combining the Three Tests, with Reference to:		D(E+F)	$\frac{D(E+F)}{\sigma}$	Final or Nominal Weights
				Importance of Function Measured	Independence of Each Measure from the Other Two			
Thornike-McCall Reading Test	3.45	.65	2.30	5.0	3.3	19.09	5.53	1
Woody-McCall Arithmetic Test	2.95	.62	2.07	3.3	4.2	15.52	5.26	1
Morrison-McCall Spelling Test	8.18	.70	2.79	$\frac{1.7}{10.0}$	$\frac{2.5}{10.0}$	11.72	1.43	$\frac{1}{2}$

The standard deviations recorded in column B are those for the Dewey School, Westernville, high eighth-grade class,

Generated on 2020-12-23 00:24 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and the reliability coefficients of Column C should theoretically be determined from the same class. Not having calculated them for this class and having the data on reliability for these tests as given in Chapter X, we find it possible to obtain a fairly close estimate of the reliabilities for this class. We obtain the following information from Chapter X[1]:

The reliability for a population of 500 12-year-olds was found by McCall to be .80, and the standard deviation was 10 *T*-score units, or approximately, as indicated by reference to table of equivalents on directions sheet of test, 4 raw test units.

The reliability for unselected age groups is reported by Thorndike to be about .70.

The reliability for a population of 75 high seventh, low eighth, and high eighth pupils having a standard deviation of 9.1 *T*-score units is reported by Cronin to be .57.

From these three items we estimate that the reliability for a group of high eighth-grade pupils whose standard deviation is 3.45 (or approximately 9 *T*-score units) is about .65. Judging by McCall's data alone, we should have estimated a considerably larger value, and judging by Cronin's data alone, a considerably smaller value, whereas had Thorndike's report been the only evidence available, we should have estimated a slightly smaller value, since an unselected age group is commonly much more variable than a grade group. Consequently, if Thorndike found .70 as the reliability for an age group, we should expect between .60 and .65 as the value for a grade group. The value .65 recorded in column C is admittedly an estimate based upon the three available sources of information. Very commonly such estimates need to be made, for ordinarily it is not feasible to determine the reliability coefficient for each class tested. If, however, as complete data as are here published for the Thorndike-McCall

[1] When these calculations were made, the reliability coefficients reported by G.M. Ruch, now given in Chapter X, were not available. They would only slightly alter the result.

Generated on 2020-12-23 00:24 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Reading Test are available, one may make such estimates, with great assurance that he is thereby increasing the accuracy of his general procedure.

Estimates based on such data as are available, and given in Chapter X, for the other two tests used, yield the other figures, .62 and .70, of column C. The entries in column D are readily obtained from the reliability coefficients. The proof that this factor, $\sqrt{r_{11}}/(1 - r_{11})$, is the appropriate multiplier to allow for differences in reliability is given in Section 5 of Chapter VIII.

Columns E and F are personal estimates made successively and, as nearly as possible, independently of each other. Thus the writer judged that if 10 points are to be distributed on the basis of importance among reading, arithmetic, and spelling, in securing a total all-round achievement score, half of them should be assigned to reading, one third to arithmetic, and the balance, 1.7, to spelling. Further, if 10 points are to be distributed among these three tests upon the basis of their independence of each other, the writer judges the arithmetic test to be more dissimilar to reading and spelling than either of these is to the other two. He has thus assigned 4.2 of the 10 points to arithmetic; he has divided the remaining points between reading and spelling in the ratio of 3.3 to 2.5, because spelling, due to the memory factor, seemed to him more dependent upon arithmetic than is reading.

Having the values of columns D, E, and F, column G, giving best estimated effective weights, is immediately obtained by adding E and F and multiplying by D. The values of this column indicate the actual importance attributed to each of the tests by the judge. As yet, no account has been made of the units of measurement, and just as it is not sound to compare height measured in inches with height measured in centimeters, so here an allowance must be made for the particular test units employed. Proper allowance is made if

Generated on 2020-12-23 00:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the effective weight values of column G are divided by the standard deviations of the class scores in the respective tests. This has been done and the answers recorded in column H. However, the values of column H are cumbersome to work with. We shall therefore choose other numbers roughly proportionate to the values of column H for the final, usable weights. Let us note that 1.00 is to 1.00 is to .25 approximately as 5.53 is to 5.26 is to 1.43. Though the approximation is not very close, it is sufficient for practical purposes, and as it gives simple multipliers to work with, the actual combination of the three scores into a total score, as shown in Table 11 (on page 72), is accomplished very rapidly and with relatively small chance of numerical error.

The scores of the composite, X_0, recorded in Table 11 are, as has just been explained, calculated by the formula:

$$X_0 = X_1 + X_2 + .25\ X_3$$

in which X_0 is the composite score, X_1 the score on the reading test, X_2 that on the arithmetic test, and X_3 that on the spelling test. The multipliers of X_1, X_2, and X_3 — namely, 1, 1, and $\frac{1}{4}$ — are the nominal or used weights, but the actual importance that has been given to these three separate tests when thus combined is represented by the product of these nominal weights and the standard deviations of the tests. These products may be called the effective weights, as they represent the actual importance given to the several tests.

Importance given to the reading test
$= 3.45 \times 1.00 = 3.45$

Importance given to the arithmetic test
$= 2.95 \times 1.00 = 2.95$

Importance given to the spelling test
$= 8.18 \times .25 = 2.04$

Generated on 2020-12-23 00:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 11

COMPOSITE SCORE OBTAINED ON THORNDIKE-MCCALL READING, WOODY-MCCALL ARITHMETIC, AND MORRISON-MCCALL SPELLING TESTS; TESTS WEIGHTED 1, 1, ½, RESPECTIVELY

NAME OF PUPIL	TOTAL ACHIEVEMENT SCORE
Ida A.	65
Robert B.	52
Albert B.	72
Frank B.	63
Letha C.	72
George C.	57
Lucy C.	68
Grace C.	78
Gladys C.	71
Doris C.	69
Wayne D.	53
Alice E.	72
Jonathan F.	69
Horace H.	66
Franklin H.	56
Jack K.	63
Clark L.	74
Jeannette L.	66
Carmine L.	58
James M.	76
Helen N.	64
Sarah P.	60
Florence P.	74
Alton P.	58
Anna R.	75
Marion R.	71
May S.	70
Emily S.	70
Ethel S.	70
Earl S.	67
Fred S.	69
Alice S.	79
Elbert T.	73
Ethel T.	67
Marion V.	73
George Z.	68
Mean [1]	67.3
Standard deviation [1]	6.73

[1] Mean and standard deviation calculated after first grouping total scores into intervals of three, 50–52 being the first interval; 53–55, the second; etc.

Generated on 2020-12-23 00:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

A composite score such as that just obtained is, of course, considerably more reliable than the scores on the separate parts. The statistical methods are available for exactly determining the reliability of a composite, but we shall here resort to an approximate method. Let r_{gG} be the desired reliability coefficient of the grand total score; let r_{1I} be the reliability coefficient of the first test; r_{2II}, of the second; etc. Then, if a different tests have been combined to yield the composite and if r is the average reliability of the a tests, then the reliability coefficient of the grand total is approximately given by the following formula:

$$r_{gG} = \frac{ar}{1 + (a - 1)r} \quad . \quad . \quad . \quad . \quad [7]$$

If we apply this formula to our present grand total score, we have:

$a = 3$, because we have combined three different tests

$$r = \frac{.65 + .62 + .70}{3} = .657$$

$$r_{gG} = \frac{3(.657)}{1 + 2(.657)} = .85$$

We thus see that the reliability of our composite measure is decidedly greater than that of the parts, but even so, the composite score is scarcely as reliable as we should demand if we are to make individual diagnoses. In spite of the fact that the reliability does not reach .90, we shall, for illustrative purposes, use these total composite scores for the purpose of classifying the pupils into sections. The classification will be upon the basis of general scholastic achievement.

If great flexibility of classification were administratively possible, it would be desirable to have separate classifications for reading, arithmetic, spelling, history, science, etc. Ordinarily there are practical difficulties in the way of placing an eighth-grade pupil in one class in reading, in a second in

Generated on 2020-12-23 00:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

arithmetic, etc. Ingenious administrators can overcome these difficulties, but there is another reason why we should not attempt to classify these particular eighth-grade pupils by separate subjects. The reason is that we should have a reliability for each test higher than we have found to be the reliability of the composite before such differential classification would be reasonably trustworthy. If each of the tests had a reliability of .95, we could then proceed with fair assurance to a separate classification in each of these three subjects, because the subjects are known to be fairly disparate as far as the basic capacities demanded are concerned. With a reliability as high as .95, we should not attempt separate classifications if the subjects are quite similar — as, for example, are paragraph meaning and word meaning.

We shall therefore proceed to a classification upon the basis of total score only, and shall recommend that Frank B., whose total score is 63, and Jack K., having the same total score, be classified together, though Frank's score is made up of reading, 29; arithmetic, 24; and spelling, 40; while Jack's comes from: reading, 28; arithmetic, 28; and spelling, 27. If we could place implicit trust in these scores, we should place Frank higher than Jack in spelling and lower in arithmetic. On using tests of the reliabilities of these, such a judgment would have so great a chance of being wrong that it is better not to make it but simply to classify upon the basis of total score.

We need to know the norms for higher and lower grades upon this same battery of tests, and if the tests have been drawn from several sources, it is probable that there are no such norms published. The best procedure is to determine norms on this particular battery for the school system concerned. If the testing program has not extended to lower and higher school grades, it is necessary to make certain estimates of the norms for these grades, knowing the local

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

norms for the one grade tested and knowing the national norms which ordinarily are to be found in the manuals of directions accompanying the tests. Having the national norms on the tests as put out by the publishers of the tests, we can build up national norms on the composite, as is illustrated in Table 12.

TABLE 12

NORMS FOR THE BATTERY COMPOSITE SCORE — DERIVED FROM THE NORMS AS PUBLISHED FOR THE SEPARATE TESTS

GRADE	PUBLISHED THORNDIKE-MCCALL READING-TEST NORMS		PUBLISHED WOODY-MCCALL ARITHMETIC-TEST NORMS, X_2	PUBLISHED MORRISON-MCCALL SPELLING-TEST NORMS, X_3	NATIONAL COMPOSITE NORMS = $X_1 + X_2 + \frac{1}{2} X_3 = X_c$
	T-Scores	Equivalent Raw Score, X_1			
3 Mid year				18	
3 End of year	30.0	7	14.25	(21)	26
4 Mid year				24	
4 End of year	37.3	15	19.50	(27)	41
5 Mid year				30	
5 End of year	48.0	22	25.00	(32.5)	55
6 Mid year				35	
6 End of year	53.7	24	29.50	(37)	63
7 Mid year				39	
7 End of year	58.3	27	32.2	(40.5)	69
8 Mid year				42	
8 End of year	60.9	28	33.5	(43.5)	72

Numbers recorded in parentheses for the Morrison-McCall Spelling Test are interpolated values secured from the neighboring values copied from the published norms. We see that the national norm is 72 for the end of the eighth grade and 69 for the end of the seventh grade, whereas the eighth-grade class of the Dewey School, Westernville, made an average score in May — i.e., near the close of the school year — of

Generated on 2020-12-23 00:26 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

but 67.3. Thus the Dewey School is over a year below the national norm. Since, however, we are called upon to classify the pupils for a location in the Dewey School and not with reference to schools elsewhere or schools in general, we shall assume norms for the Dewey School as in the last line of Table 13.

TABLE 13

NORMS FOR END OF SCHOOL YEAR

Grade	3	4	5	6	7	8	9
National norms	26	41	55	63	69	72	—
Dewey School norms (Experimentally determined for Grade 8 and estimated for Grades 7 and 9)	—	—	—	—	64.3	67.3	70.3

The estimated norm for the Dewey seventh grade is 64.3, which is 3 units below 67.3, the actual eighth-grade norm, because the distance apart of the seventh- and eighth-grade national norms is 3 units. It is also reasonable to estimate the ninth-grade norm as 3 units above the eighth-grade norm, as has been done. We finally obtain norms for the Dewey School which, though decidedly below the national norms, are much more reasonable for purposes of classifying Dewey eighth-grade pupils than would be national norms. The argument employed is simply that, the Dewey eighth grade having been found to stand below the national record, it is reasonable to expect the Dewey seventh and ninth grades to be below likewise. As our immediate purpose is not to improve grade records but to classify pupils into homogeneous groups, we disregard national norms entirely in favor of local norms. With these local figures, we may list critical June scores, as follows:

Generated on 2020-12-23 00:26 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

64.3 — high seventh-grade norm
65.05 — point midway between high seventh- and low eighth-grade norms
65.8 — point midway between high seventh- and high eighth-grade norms; i.e., low eighth-grade norm
66.55 — point midway between low eighth- and high eighth-grade norms
67.3 — high eighth-grade norm
68.05 — point midway between high eighth- and low ninth-grade norms
68.8 — point midway between high eighth- and high ninth-grade norms; i.e., low ninth-grade norm
69.55 — point midway between low ninth- and high ninth-grade norms
70.3 — high ninth-grade norm

From the preceding we see that if a child receives a score below 66.55, he scores closer to the low eighth- than to the high eighth-grade norm and should therefore not be promoted regularly with the class.[1] If he receives a score between 66.55 and 68.05, he should be given a single regular promotion; while if he receives a score above 68.05, he should receive a double promotion — i.e., skip one half of a school year in order to be placed with the pupils with whom he is most closely allied in general capacity.

We thus have the following relationship between battery-test score and grade in which located at the time of testing:

[1] A slight error in this statement, due to the unreliability of the test employed, will be apparent to those familiar with the principle of regression. This error is slight if the reliability is .9 or greater, and even with the battery here employed, having a reliability of about .85, the error is not serious.

Generated on 2020-12-23 00:26 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Scores below 63.55	Scores from 63.55 to 65.05	Scores from 65.05 to 66.55	Scores from 66.55 to 68.05	Scores from 68.05 to 69.55	Scores from 69.55 to 71.05	Scores above 71.05
Corresponding school grade } below h 7	h 7	l 8	h 8	l 9	h 9	above h 9

Since an average high eighth-grade pupil in June should enter the low ninth grade the following September, the classification indicated by the test scores is one half of a grade higher than given in the preceding table.

The reader has probably asked himself, "Should one actually classify a pupil in so rigid a manner as described?" The writer would advocate a pretty strict rule-of-thumb procedure where the essential purpose is to secure homogeneous groups and if the test battery used is well chosen. The misplacements consequent to such a procedure would be fewer than is commonly the case where facts of doubtful pertinence — "maturity," "health," "size," "conduct," etc. — and such tenuous considerations as "general worthiness to promotion," "spirit," "attitude," etc., play a large part. Other influences than sheer scholastic achievement are commonly considered by teachers and principals in making promotions. They well should be in the junior and senior high schools, though it is doubtful if they are entitled to an important place in the elementary school. We shall provide for them in the general scheme as shown in Table 14, but we may feel confident that if we have a comprehensive general achievement test and one having a high reliability, we shall secure a very serviceable and workable classification in Grades 1 to 6 if these extra considerations play no part whatever. Given the judgment to include these extra considerations at their proper valuations, one would of course then always improve his classification by using them.

Generated on 2020-12-23 00:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

It will prove convenient to list the pupils in the order of their total scores, as in Table 14 (pages 80–83).

A glance at the thirty-six test scores given in Table 14 shows a range so wide that if lower and higher grade norms were available, it would probably extend from the fifth grade to the eleventh. A situation much like this, though probably not quite so extreme, is very commonly found not only when single school subject tests, but even when comprehensive and highly reliable achievement test batteries, are given. A part of the deviations of the individuals from the average is always to be assigned to unknown or chance factors, but most of it is to be attributed to real differences of abilities of the pupils. Table 14 shows but four children (Lucy C., George Z., Earl S., and Ethel T.) properly classified as judged by the test scores. Many a teacher and principal confronted with this situation would be inclined to consider the test scores all wrong. In this they are hasty. The one truly expert in test interpretation will not take them just as they stand, and this for two reasons: (1) The test has a reliability of .85, and there is therefore a substantial error of estimate when the obtained score is taken as a pupil's true ability score; the standard error of estimate, as given by Formula 16 of Chapter VII, Section 8, is:

$$\sigma_1\sqrt{1 - r_{1_I}} = 6.73\sqrt{1 - .85} = 3.5$$

and the probable error of estimate is:

$$.6745\,\sigma_1\sqrt{1 - r_{1_I}} = 2.4$$

Thus, the chances are fifty in one hundred that the pupil's obtained score differs from his true ability score by an amount greater than 2.4 units. This makes considerable difference in interpretation, so that we may be assured that if we use the obtained scores for classification purposes, we shall be in error by one school year in about half of the instances. (2) The second reason why the test expert will not place im-

Generated on 2020-12-23 00:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 14. CLASS RECORD AND

DEWEY HIGH EIGHTH GRADE, WESTERNVILLE, TESTED MAY 25–29, 1926, READING (NOMINAL WEIGHT, 1), WOODY-MCCALL ARITHMETIC WEIGHT, ½) TESTS

AGE OF PUPILS JUNE 1, 1926	NAMES ARRANGED IN ORDER OF TEST SCORES	TEST SCORES	GRADE INDICATED BY TEST SCORE	TERM MARK GIVEN BY TEACHER
13–4	Alice S.	79	above low 10	A
14–0	Grace C.	78	"	B
14–6	James M.	76	"	A +
15–2	Anna R.	75	"	B
12–4	Clark L.	74	"	D
13–11	Florence P.	74	"	C
14–2	Elbert T.	73	"	D −
14–9	Marion V.	73	"	D
14–11	Albert B.	72	"	B
15–11	Letha C.	72	"	A
12–9	Alice E.	72	"	B
15–5	Gladys C.	71	low 10	D
16–8	Marion R.	71	"	B
15–6	May S.	70	"	C
16–0	Emily S.	70	"	B
14–7	Ethel S.	70	"	B
15–0	Doris C.	69	high 9	C
13–5	Jonathan F.	69	"	C
13–10	Fred S.	69	"	A
14–10	Lucy C.	68	low 9	C
15–8	George Z.	68	"	B

Generated on 2020-12-23 01:14 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

JUNE, 1926, PROMOTION SHEET

WITH ACHIEVEMENT BATTERY CONSISTING OF THORNDIKE-McCALL (NOMINAL WEIGHT, 1), AND MORRISON-McCALL SPELLING (NOMINAL

TEACHER: MISS ROSALIE SWEET COMMENTS OF TEACHER IN REGARD TO PROMOTION	TENTATIVE PROMOTION OF CHILD MADE BY TEACHER	COMMENTS OF PRINCIPAL FOR INFORMATION OF TEACHER OF GRADE TO WHICH PUPIL IS ASSIGNED
Mother insists that Alice go ahead. Alice is bright but really too young for high school	h 9	
Shy; would feel ill at ease with older children	l 9	Send in report on Grace C. in 4 weeks
Best student in the class	h 9 l 9	
Bright, but too immature for 9th grade	h 8	Send in report on Clark L. in 4 weeks
Hates arithmetic	l 9	Florence is bright enough for college if she will do better in mathematics. Endeavor to arouse her interest in high school algebra
Troublesome and doesn't work	h 8	Send in report in 4 weeks
Absent a great deal on account of sickness	h 8	Send in report in 4 weeks
	l 9	Send in report in 4 weeks
	l 9	Send in report in 4 weeks
	l 9	Send in report in 4 weeks
	l 9	
	l 9	
	l 9	
	l 9	
	l 9	
	l 9	
Too immature for 9th grade	h 8	Send in report in 4 weeks
	l 9	
	l 9	
	l 9	

Generated on 2020-12-23 01:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 14

Age of Pupils June 1, 1926	Names Arranged in Order of Test Scores	Test Scores	Grade Indicated by Test Score	Term Mark Given by Teacher
13–10	Earl S.	67	low 9	D
15–3	Ethel T.	67	"	A
14–4	Horace H.	66	high 8	D
15–0	Jeannette L.	66	"	B
16–4	Ida A.	65	low 8	B
14–9	Helen N.	64	"	C
14–5	Frank B.	63	below low 8	D
16–0	Jack K.	63	"	B
17–6	Sarah P.	60	"	B
15–10	Carmine L.	58	"	C
14–5	Alton P.	58	"	D −
16–1	George C.	57	"	D −
14–8	Franklin H.	56	"	D
16–6	Wayne D.	53	"	D −
17–1	Robert B.	52	"	D −
14–10½ = median age		69 = median score		C + = median mark

plicit trust in the total achievement score is because it is a measure of but a part of the subject matter of the grade tested. In other words, even were it perfectly reliable, it would not be a completely valid measure of the traits and capacities which should be considered in determining promo-

Generated on 2020-12-23 00:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(*Continued*)

Teacher: Miss Rosalie Sweet Comments of Teacher in Regard to Promotion	Tentative Promotion of Child Made by Teacher	Comments of Principal for Information of Teacher of Grade to Which Pupil Is Assigned
	h 8 l 9	Send in report in 4 weeks
Too much interested in baseball	h 8 l 9	
Very industrious	l 9	Vocational guidance in 9th grade
Such a little girl, she ought to repeat	h 8	
Has trouble with arithmetic	h 8	
Says he will study spelling this summer	l 9	Test spelling in September and report to office
Pretty good in arithmetic	l 9	Vocational guidance in 9th grade especially needed
Has to work at home	l 9	Vocational guidance in 9th grade especially needed
Failed	h 8	
Failed	h 8	Vocational guidance in 8th grade especially needed
Doesn't like history	h 8	Vocational guidance in 8th grade especially needed
Can't do the work	l 8	Vocational guidance in 8th grade especially needed
Doesn't try	l 8	Vocational guidance in 8th grade especially needed
Number sent to h 9: 2 Number sent to l 9: 21 Number sent to h 8: 11 Number sent to l 8: 2		

tions. For these two reasons, then, the test results should be taken with a good deal of circumspection, and classifications based upon them should be tentative. In the face of this poor showing, why should we use the test results at all? The answer is somewhat disheartening, for it is that, unreli-

Generated on 2020-12-23 00:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

able as are the test results, the teachers' judgments are worse. School marks given by a single average elementary school teacher have a reliability in the neighborhood of .4 or .5, and in addition to this they are almost universally possessed of a constant bias tending to keep together groups which happen to be together. Thus, if an average 11-year-old in the third grade should transfer to a new school, and if the " 3 " on his record card should look like a " 5," so that he was by mistake placed in a fifth grade, the chances are that it would never be discovered unless the child himself made the fact known. This constant bias is just another name for conservatism and narrowness of experience. Commonly a fifth-grade teacher does not know the standards of upper and lower grades well enough to realize that her bright and dull pupils are reacting in the manner of pupils of these grades. To promote children regularly with the class is the easy thing to do, generally satisfying pupil, parent, and principal, though in reality it is very unjust to the backward or precocious child, and also to the average child who for any reason — such as late or early entrance to school, loss of time due to moving, etc. — happens to be poorly placed. Because of the unreliability of test marks and of teachers' marks, every classification, whether based on the one or the other, should be looked upon as tentative and a thing to be reviewed soon and probably revised. Could a superintendent in the middle of some term require by edict that 33 per cent of the pupils in each grade be moved to a lower or higher grade, the salutary effect upon the average school system would be great. Should a superintendent do this for one year, there would be a lessening need of repeating it the next year. However, since in dealing with the average school system a very great amount of shifting needs to be done before even approximate homogeneity of talent in the separate grades is brought about, there is little likelihood of an immediate overdoing of

Generated on 2020-12-23 00:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the matter. Let us review the data and promotions as decided upon by the teacher of the eighth-grade class in the Dewey Junior High School.

Judged by capacity, there is entirely negligible likelihood that Alice S. is (in 1926) or was (in 1925) in the class best fitted to her talents. An observing mother seems to sense the situation because she " insists " that Alice be permitted to skip half of a school year. The teacher reluctantly agrees to this. The high ninth grade is a good place for Alice for about six months, and then she should probably be permitted to skip another half year, for it is to be expected that she is fully as capable as and, in terms of facts, knows as much as pupils two or more grades ahead of her.

Grace C., as well as others, should be given the same opportunity, but she is not given it because the teacher thinks she would " feel ill at ease with older children." The chances are that mentally more developed comrades are just what is needed to set Grace at ease, for she is probably now leading a double life, a happy one in make-believe and story-books, and a troubled one with raucous youngsters of her own age but not of her own mental maturity. It would be good for her to have school work more worthy of her serious effort and attention.

The teacher, Miss Sweet, has probably considered herself very progressive in recommending that James M. be allowed to skip a half grade. So she is, for the idea scarcely occurs to most teachers, but she has made the recommendation with reference to one pupil, whereas she should have made it with reference to eight or ten.

Clark L., probably one of the two brightest in the class when judged by age (Alice S. being the other), is " immature " and is therefore required to repeat the work of the high eighth grade. It is probably true that he has not a bristle on his upper lip, that he plays tag with the girls, recites poetry in

Generated on 2020-12-23 01:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Sunday school, and is mamma's darling and papa's boy in family circles. Why not let him do and be all of these things just as long as he can — in fact, coöperate with him by letting him get a little joy out of school life instead of blighting it with another dull dose of the eighth grade? It is quite possible that you, reader, know no more of the content peculiar to the eighth grade than does Clark, so ask yourself if you would find the eighth-grade pabulum peculiarly thrilling. A second subjection to it may be even more galling to Clark than it would be to you, for he is still possessed of the enthusiasm of youth in its quest for new knowledge. With the torch that within him burns he may now master two years' work in one and throughout life have the confidence, self-respect, and ideals and gratifications of mental work that go with this accomplishment. These are the things that preserve youth and make it meaningful to the bright and studious child. Don't rob a youngster of this opportunity because he is small and buoyant and you think him "immature." You, Miss Sweet, are probably wrong, first in calling him immature, and secondly, in thinking it makes any difference as far as his life of mental values is concerned whether he is physiologically immature or not.[1] The reliability of your judgment upon a matter of which you are fully cognizant — namely, scholastic achievement — is probably about .5, and in the judgment of maturity and knowledge of its significance you are probably about as accurate as a country doctor gazing at a milk tooth. Give Clark L. a chance at the high ninth grade. If he doesn't immediately become "mature" it will not matter, for time will take care of that. If he does not master the scholastic assignments, demote him, which outcome, however, is very improbable. One further fact:

[1] The writer is aware of the studies of Baldwin and others dealing with this matter, but has not been convinced by them that physiological maturity, independent of mental maturity, is an important factor in school work (except for physical training and possibly manual training).

Generated on 2020-12-23 01:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Clark is only of age 12–4, while the average for the class is 14–10½. The rate of growth in mental functions is much greater at this younger age than at the older, so that if Clark is now 2½ years above the average, we may expect him, because of his more rapid growth, to be even farther above in a year's time, provided he is not stunted meanwhile by a very poor educational stimulus. Let us consider an extreme case: three children of ages 8, 12, and 16, equal in general achievement (not in IQ) and of ability represented by a score of 75 (where 100 is the average adult score). If given fair environment, these three children will develop into adults with abilities represented approximately by the scores of 150, 100, and 80. The three are together in ability at this moment, but in one year's time their abilities will be approximately 86, 81, and 77. The younger has outstripped the older by more than a full school grade; thus if there is any question based upon considerations of ability as to which shall be promoted, the younger child and not the older is the one more entitled to advancement. The tool subjects of the elementary school — reading, arithmetic, spelling, history, language, etc. — are subjects that Clark can and will pick up more or less incidentally for himself, and there is no need to waste his time upon them. The writer, when a teacher of college mathematics, was convinced that his subject could be picked up incidentally by competent students. The same is surely true of the content of earlier grades, so that the elementary or high school teacher need not feel that there is any disparagement of his instruction when it is remarked that a bright pupil will suffer no permanent handicap by skipping his work. It comes hard to certain college teachers to advise inquiring students not to take their own offerings, but if they will but think of the number of young persons who have turned out at least fairly well without them, they may bring themselves to do so. Let elementary and high school

Generated on 2020-12-23 00:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

teachers do likewise. To the bright pupil successive grades in the elementary school do not constitute successive links in a chain, but rather routes through a forest in which are many paths and crossings of paths. Many alternative paths are pleasing and useful to him who goes not far into the unknown. But this variety of route, though known and enjoyed in moments of relaxation, is not essential to him who goes on and on in search of the deeper mysteries of the wood. We must leave Clark feeling that an unmeasured, possibly unmeasurable, amount of intellectual inspiration has been taken from him by the classification imposed upon him.

Florence P. is promoted in spite of a deplorable shortcoming — she " hates arithmetic." If she really whole-heartedly dislikes it, it probably means that she likes something else, perhaps literature, with a comparable intensity. If this something else is worthy of encouragement and if it is, in fact, stoutly tied to arithmetic (as, for example, is mathematics tied to literature by the college entrance requirement that every matriculant must have taken and passed both in the high school), then an elucidation of the facts would probably stimulate Florence's interest in mathematics. If there is no such bond between the subjects created by dictum of higher educational authorities or by the natural and inescapable relationships between them, do not conjure up a linkage, but rather, since Florence is possessed of enough mathematics for everyday needs, let the weakness alone and cater to her fortes. Let her day be full of interest and of tasks that tax her and fit her for a serviceable place in adult society.

Elbert T. is a problem such as every teacher has, and the problem is commonly " solved " as Miss Sweet solved it, by not promoting Elbert, though he knows enough to understand the instruction of the low ninth or even of a higher grade. For misconduct to secure the deserts of dullness is a vicarious punishment. No teacher would think of saying to

Generated on 2020-12-23 00:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

a child, "If you fight on the playground, I will mark your arithmetic example wrong"; yet this is, in fact, what is done when promotion, presumably dependent upon scholastic achievement, is made to depend upon deportment. Certainly the problems of instruction which are so greatly lessened by homogeneous mental classification are greatly aggravated by the use of deportment as a basis for advancement, and it is very doubtful even if the problems of discipline themselves are lightened in any true sense by tying them up with scholastic achievement.

Measurement studies show that very commonly laziness, lack of interest in studies, interest in other matters, and teacher-baiting proclivities lead to a scholastic classification which is lower than that warranted by achievement records. It is an injustice to reward docility or punish misbehavior by a mark supposedly indicative of scholastic achievement, and such a procedure merely aggravates problem cases. If a child of average sixth-grade ability (not achievement) is lazy and not interested in school work, he has already penalized himself most markedly, with the result that his achievement record is much below his capacity to achieve. If because of his working against his own interest he makes but a fifth-grade scholastic record, then, if the teacher because of his unsatisfactory attitude places him in the fourth grade, there results a double displacement. The child has placed himself one grade lower than his talents warrant, and then the teacher puts him down another grade, so that he finally ends up in a scholastic position out of all harmony with his mental capacity. The child now, instead of being an educational problem, is merely a disciplinary one.

The writer advocates placement in the elementary school according to achievement irrespective of disciplinary issues, but he would like to see tried out as an experiment, and with appropriate checks, the placing of all lack of interest and con-

Generated on 2020-12-23 00:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

duct cases in classes one half year *above* their scholastic achievements. This is defensible from the standpoint of intellectual homogeneity, but whether the reward for laziness, indifference, and misbehavior seemingly present is so in fact and will have a bad effect can better be told after a careful experiment with the plan. Though such a reward for a lack of interest is not seriously proposed, except as an experiment, the writer considers it definitely unfortunate, unfair, and provocative of disciplinary difficulties to penalize scholastic standing because of character shortcomings. Equations of the sort:

Average knowledge + good conduct = pass
Average knowledge + poor conduct = failure

are indefensible.

Apparently Marion V. has been caught by another type of faulty reasoning:

Average knowledge + promptness and attendance = pass
Average knowledge + sickness and absence = failure

Unfortunate Marion has probably aggravated her sickness by fear that she would not pass, and she has either studied at home or is natively of more than average ability, for she is quite certainly above the average of the class in attainment. Her teacher may have reasoned, "Poor Marion is not very well, and it is not right that she be made to work as hard as a sick child would have to in the ninth grade; so I will just keep her with me in the eighth grade and make it easy for her." This is false kindness and poor reasoning. The basic assumption seems to be that mental activity sufficiently involved to be interesting is unhealthful and that a sick person should not engage in it. We may all subscribe to the doctrine that if a child's health is poor, the improvement of it is the thing of first importance. School attendance and lesson

Generated on 2020-12-23 00:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

assignments must wait, for proper food, sleep, sunshine, and exercise come first. But it is very important that going with these should be feelings of hope and usefulness, and not of failure. Study which is not tiring but is sufficient to show the child that mental growth is taking place even though the body is frail becomes a source of satisfaction in a life where satisfactions are few and the general outlook is discouraging. Marion V. has scholastically earned a promotion. The air and sunshine in the ninth grade are just as invigorating as in the eighth, and the mental joys to Marion would be greater there. She is entitled to them.

The next three children, Albert B., Letha C., and Alice E., would probably succeed in the high ninth grade — it would be well to give them the opportunity.

Jonathan F. seems to be another youngster suffering from "immaturity." He should probably be placed in the high ninth grade or certainly not lower than the low ninth grade.

Ethel T. has certainly impressed the teacher (mark A) much more than she did the achievement test blanks (score 67). In view of her age, it is doubtful if she will ever graduate from high school, and she should be placed in that grade offering the greatest opportunity for immediate vocational equipment. The low ninth grade is probably the proper place for her.

Horace H. is sixth months younger than the average of the class and only a trifle below the average in test score. It is probable that with his slightly above average IQ he would catch his classmates within a year if given the chance. He should probably be promoted to the low ninth grade.

The case of Jeannette L. is much like that of Ethel T.

Ida A., irrespective of her school mark, should, on account of her age, be promoted and given work of value for vocational equipment. Her arithmetic record is the best; so she might become a successful restaurant cashier or bank clerk.

Generated on 2020-12-23 00:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

It seems entirely reasonable that Helen N. should be placed in the high eighth grade simply because her scholastic record is poor and her age average.

Frank B. requires the same treatment, unless it is administratively possible to place him in the low ninth in all subjects except mathematics, and in the high eighth in that. This scheme would probably secure maximum effort upon his part.

Jack K., Sarah P., Carmine L., George C., Wayne D., and Robert B. all need vocational work, probably best found in the ninth grade. If there is any offering anywhere in the school that will help Robert B. to become a self-supporting citizen, he should be given it, irrespective of the grade classification. At the present moment he is probably not equipped for any vocation, and if the school turns him away right now, it must be charged with one failure — if Robert becomes a delinquent, the school is an accomplice in fact. His case has probably been increasingly critical for the last three years, and it is pretty late to remedy the situation now, but now is almost certainly the last chance. Robert is in the junior high school and possibly can be kept in school until the end of the ninth grade, but certainly not beyond that. Do not argue that the world needs hewers of wood and drawers of water; the world needs fewer of these than ever before, and it needs none who are instilled with a sense of failure, as apparently is Robert B. The writer has seen self-confident morons happy as the day is long and useful (hoeing corn) as their mechanisms working at about 100 per cent efficiency permitted, and surely the educative process that has this outcome is the ideal to strive for. It cannot be attained if all children are put through the same educational mill. A skillful administrator can accomplish much for the exceptional child in a junior high school, even though its main purpose is the education of average children.

Generated on 2020-12-23 00:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Such an administrator should exercise his ingenuity in adjusting the curriculum to the individual needs of Franklin H., who threatens to go the road of Robert B. but, being 2½ years younger, gives more promise of being saved.

The grade classification here proposed for the older pupils has been influenced by prevocational and vocational need and not primarily by scholarship. In the more elementary grades classification can with advantage be based entirely upon scholastic achievement, but when a child approaches the end of his formal education, the emphasis should change from that of all-round cultural development and facility with tool subjects to preparation for a specific vocation. The child having an intelligence quotient of 90 is likely to stop school with the eighth grade, which he will complete at about the age of 16, so that the eighth grade certainly, and additional grades if he remains in school, should have for him a strong vocational bent. The 100 IQ child is likely to drop out of school somewhere in the high school, from which he may graduate at about age 19, if he remains. It will be well if for him the ninth grade is in the main prevocational and higher grades vocational in their outlook, as under these conditions he is more likely to stay in school and more likely to find a useful and happy place immediately after leaving. The 110 IQ child is the typical high school graduate whose last one or two years in school should have a vocational bias. The 120 IQ child is college material, out of high school at about age 16 if given a fair chance, and then for the first time called upon to direct his education in view of a vocation ahead. The proportion of children having intelligence quotients of 120 or above is not great, and important as they are for social welfare, they should hardly have the school system administrated for their benefit to the exclusion of that of their less talented brethren. A school system organized for the small fraction who are to go to college will very commonly force

Generated on 2020-12-23 00:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

such a person as Robert B., of IQ about 80, to be in the regular eighth grade at age 17. The ridiculousness of the situation does not change its tragic nature to Robert, and the administrators who have brought it about are culpable. They should be sentenced to teach a year in a reformatory for every child whom they have so grievously offended.

If we glance at the column of Table 14 giving the notes of the principal, we find that he has doubted the propriety of a number of the promotions made by the teacher, in that he has asked for information as to progress after the child has been in the next grade for four weeks. The intention obviously is to check up upon the accuracy of the classification by a review of the case early in the next grade. Such a check-up is scarcely worth attempting if it is to be based upon the judgments of the new teacher of the pupil, for in four weeks he has not learned the capabilities of his pupils. Such a check-up based upon excellent standardized tests may be of great value, for a confirmation of the June test results should be quite sufficient to lead to an immediate change of the grade classification, generally to the decided advantage of the pupil. The autonomy of the high eighth-grade teacher of the preceding year has not been encroached upon (though it may well be if the welfare of pupils is clearly at stake), and the new teacher (low ninth grade) has as yet formed no judgments which he feels in duty bound to fight for.

The problem studied in this chapter has been that of the general classification of pupils into school grades. A less extensive problem, that of classification in a single subject, may commonly be undertaken with value by the teacher of a single grade, acting alone or with others, in the case of departmental instruction. The amount of overlapping (the number of pupils in a given grade making records as good as or better than the median of the grade above, or as poor as or worse than the median of the grade below) was found to be

Generated on 2020-12-23 01:15 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

very great in the eighth grade of the Dewey Junior High School when a battery composed of three tests was used. In general, the amount of overlapping is greater when determined from scores in single tests than when scores derived from batteries are employed.

This is true in this particular grade when dealing with the Woody-McCall Arithmetic Test. Table 15 following, based on published national norms, shows about the same amount of overlapping as would be found using local norms.

TABLE 15

OVERLAPPING OF WOODY-MCCALL TEST MARKS IN THE DEWEY HIGH EIGHTH-GRADE CLASS

	H 4	L 5	H 5	L 6	H 6	L 7	H 7	L 8	H 8	L 9
National June norms	19.5	(22.25)	25	(27.25)	29.5	(30.75)	32	(33)	34	34[1]
Per cent reaching or exceeding national grade norms . .									6	6
Per cent reaching or falling short of grade norms . .			11	19	39	56	83	94		

Certain subjects yield larger measures of overlapping than others. English composition and handwriting yield very large measures — the former, in substantial part at least, because all measures of composition are very unreliable, and the latter because there is actually very wide scatter of ability in handwriting; spelling commonly yields rather large measures of overlapping; while reading and arithmetic yield somewhat smaller measures, though still very appreciable amounts.

[1] The low ninth-grade norm cannot exceed 34, as this is the maximum possible score on this test. The figures given in parentheses are interpolated values.

Generated on 2020-12-23 00:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

A study having as its purpose the classification of pupils in a single subject should proceed just as the scholastic investigation preceding, with the following modifications: (1) there is no problem of weighting of parts; (2) single subject tests are commonly less reliable than batteries of tests, so that greater care must be made in selection and generally longer tests given in a single subject than those which are serviceable as parts of batteries; (3) the results are commonly used for sectioning within a class rather than for determining class groups. As an illustration of this type of study might be mentioned a thoroughgoing examination of the spelling abilities of ninth-grade pupils, possibly with the intention of excusing those above a certain mark from further formal spelling work. As indicated in Chapters IX and X, there are quite a number of spelling tests already available or readily devisable which would be accurate enough to serve this end.

Generated on 2020-12-23 00:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER FIVE

The Determination of Individual Idiosyncrasy

1. The origins of mental peculiarity. The term "idiosyncrasy," as used in this chapter, refers to differences in two abilities of a child as judged by comparison with the age or grade group in which he is located. If a child shows 10-year ability in reading and 12-year ability in computation, he is here considered to possess an idiosyncrasy. There is no moral obliquity attached to this peculiarity. This observation is not uncalled for, in view of the endeavor of teachers and others to eliminate oddity practically wherever and whenever found. The writer has frequently described to his students in courses in education a youngster considerably better in mathematics than in reading or showing unequal development along some other two lines, and has asked for advice as to the guidance and training of such a child. Fully 90 per cent of the replies received stipulate first of all that the teacher should endeavor to bring up the reading, and not uncommonly it is proposed that the child be taken out of arithmetic and given double assignments in reading. In other words, there is something wrong in the situation which will be righted when a dead level of attainment is reached. The writer has elsewhere (1926) considered this question in some detail and will not repeat it here further than to give a plank which he has proposed as a part of a teacher's *credo* (1926, page 25): "I shall respect and endeavor to utilize to a social outcome idiosyncrasy wherever found." This is the point of view underlying the suggestions of this chapter as to the treatment of children possessed of inequalities in mental development. This view has been particularly strengthened by the evidence (cited by the writer in the previous study mentioned) that many idiosyncrasies have their roots in orig-

Generated on 2020-12-23 00:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

inal nature. If they were entirely acquired by training, they could be looked upon as less fundamental. One might then think that the training had been unfortunate in bringing them about and that it should therefore be undone, since man should be at liberty to unmake that which he has made. However, when the oddity is given at birth and grows in later years with no threat to our social structure, it seems presumptuous and audacious to assert that it is a fault. It is to be hoped that we have passed the day when the object of education was to eliminate the original sin with which it was said infants greeted the world. Inequality in individual achievement, though commonly rooted in original nature, can be largely influenced by training, and the teacher can encourage, neglect, or discourage such unevenness of development as is found in his charges.

An idiosyncrasy consisting of inferior respect for the rights of others and superior cunning should certainly be eliminated; if possible, mainly by raising the sense of respect, but it might in such a case as this be defensible to endeavor to decrease the cunning, for the social threat of an individual asymmetrical in this respect is great. Generally, however, no social harm is indicated by inequality of mental development within the individual. Superior memory, coupled with inferior arithmetic, does not menace. It is oddity of this sort that will ordinarily be revealed by achievement tests, and it is with reference to such that the issues herewith deal.

2. Purposes served by a knowledge of idiosyncrasies. What, then, are the purposes in mind when an elementary school teacher studies individual peculiarities? The information clerk at the railroad station needs a certain small amount of arithmetic in addition to a very superior memory, and the cafeteria cashier needs a certain memory ability in addition to much speed and accuracy in computation. If shortcomings are so pronounced as to handicap adult life even

Generated on 2020-12-23 00:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

in connection with vocations not particularly demanding the trait which is weak, then by all means the teacher should attempt to " bring up the weak spot," not, however, at the expense of the " strong spot." Let the forte have full opportunity to flourish and in time blossom into a vocational asset. This attitude may safely serve through the first six grades of school life. Beginning with the junior high school and increasingly with higher grades, idiosyncrasies should be fixed upon as cues for educational and vocational guidance. Not uncommonly an idiosyncrasy of a backward child, catered to, developed still further, and attended to in the choice of a vocation is the only opportunity of the individual leading a life of average social usefulness and economic return. A child of chronological age 14, and of 12 in general scholastic accomplishment in reading, spelling, language usage, and geography, but of accomplishment of average 13-year-olds in arithmetic, has had to face certain peculiar trials in his school life. He has been made aware in many ways, when reading, spelling, reciting history, etc., that he is inferior to classmates of his own age. In arithmetic this is not so true, and it has not been so impressed upon him; in fact, he rather likes arithmetic. This is his opportunity. By special effort he can do average or possibly superior work in arithmetic, and he can get the satisfactions that come from success, which satisfactions every child, no matter how dull, should secure somewhere in his life. Just as soon as this child has a tolerable knowledge of reading, writing, spelling, and history (which will be at about age 13 in the sixth grade), it is well to let these capacities grow as fast as may be possible, but not to let them result in general scholastic retardation. In other words, let the child advance as far as possible and with as much satisfaction as possible in mathematics that he may direct his steps to as important a vocation involving it as is within his power.

Generated on 2020-12-23 00:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

A child who is generally superior, but very markedly superior in some one trait, may become one of the great leaders of the race. This will not be accomplished by his being a general all-round good man. Such all-round geniuses occasionally exist; they are our mathematicians who might have been lawyers, our musicians who might have been scientists, etc. — a Leonardo da Vinci or a Benjamin Franklin. We are now speaking of a less versatile type; our mathematicians who might have been small business men had they not been mathematicians, our musicians who might have been dry-goods salesmen had they not been musicians, etc. — a Raphael or a Beethoven. These latter exist also and in considerable numbers. A 120 IQ man equally developed in all traits will probably lead a useful vocational life characterized by an intelligence quotient of 120, while another man of development of 100 in several traits and of 120 in one trait may, by a judicious selection of a vocation, lead a vocational life ordinarily expected of a man of general all-round development of 120. Idiosyncrasy in a person, characterized, as it must be, by something that is superior as well as by something that is inferior, is like an unpolished gem in a crown of rough stones. It may become tarnished and lost from view so that the crown is always considered common, but if it is cut and polished to the degree that it alone of all the stones permits, it will then lend a dignity to the crown not noticeably excelled by one all of whose stones are brilliant.

Let us then conclude that the main purposes to be served by discovering idiosyncrasies of school children are (1) that weak spots may be strengthened in the elementary school, (2) that bright spots may be further strengthened and tied to vocational intentions in the high school, and (3) that both of these things may be done in the junior high school.

3. Natural predispositions toward idiosyncrasy. Children of the same family differ in their respective talents. Accord-

Generated on 2020-12-23 01:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

ing to one view, the cause is as pictured in the following paragraph:

The twins Amelia and Annabel opened their eyes upon the world at the same hour and were apparently welcomed into the same environment. However, Amelia was born first, and her initial cry mingled in her own consciousness with the passing elevated train. The combination, rather more a feeling than a sound, was soothing. Amelia today leads the church choir and finds a delight which she cannot explain when her small voice joins with the open diapason in hallelujahs to the Creator. Annabel's arrival was not disturbed by any passing train, but the light struck straight into her tiny eyes. The beautiful after-images faded slowly, and Annabel dreamed her first dream. Growing neurons wrote the story, and today she is an artist of note who delights the magazine-cover-gazing public.

Who knows the potency of the initial moment? It is a thing to conjure with. A raindrop falling to the mountain top seems destined to reach the Atlantic, but diverted ever so slightly by the flutter of a bird's wing, it flows to the Pacific. Does it matter? If it reaches the Atlantic it will push another drop around the Horn into the Pacific, and if Amelia turns to music, may she not turn some one else away from it? It may be so, but Amelia is not a person in the abstract if you are Annabel. She is then your twin. If you are her father, she is then your daughter, and it adds but little to life's satisfaction to know that if she does not turn to music, some one else will. The social problems connected with specialization are quite different from the individual problems.

It is with the latter that we are here concerned. The problem of first importance is to know what sorts of specialization are rooted deep in human nature and what are mere incidents. To be an expert upon land values in Florida rather than in California is a mere geographical incident, but to be a musi-

Generated on 2020-12-23 00:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

cian rather than a painter devolves upon very different mechanisms. What are the independent neural constellations? Evidence upon this question can readily be gathered only from adults or from children with sufficient language facility to make their ideas and interests clearly known. If at the age of 4 or 5 Amelia and Annabel betray interests in music and painting, respectively, one does not surely know whether they have been acquired since birth or are inherited. It seems reasonable to the writer to think that traits that may be acquired tend to be acquired in proportion to the trait stimulation and that the amount of such stimulation is, in the case of scholastic achievement, roughly proportional to the emphasis placed by the school upon the different subjects of the curriculum. Arguing from this point of view, reading and arithmetic, in so far as they are acquired, would generally become acquired in the elementary grades. The presumption that such musical and artistic traits as are acquired have been acquired during these same years rather than earlier is not quite so obvious, but even here the attribution of the acquired feature to a nurture factor occurring upon the day of birth seems unreasonable to the writer. He conceives of Amelia and Annabel in the wee small hours of their existence responding to gross bodily sensations — hunger, temperature, respiration, etc. — and not to music or art, reading, writing, history, or arithmetic, nor even to the immediate antecedents of these: sound, color, space, variety in vocalization, small muscle kinæsthesis, or temporal and quantitative relationships. For these reasons the writer largely credits to original nature and not to nurture mental differences found very early in life.

However sound this argument may be, it seems evident that unevenness of mental development is found very early in life. Such unevenness, however, is not random; twenty children in one hundred may be obviously unequally devel-

Generated on 2020-12-23 01:16 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

oped in computation and spelling, while but five are as clearly of unequal capacity in history and geography. It matters not whether we assign the reason for this to (1) greater intrinsic difference between history and geography or (2) greater intrinsic difference in the nervous mechanisms that mediate computation and spelling than in those employed in history and geography. Since we are only dealing with human beings, the two statements have exactly the same meaning, and the importance of the single idea involved lies in the fact that certain idiosyncrasies occur frequently and are of large amount, while others are seldom found and are but trifling when found. If history and geography are such that they do not permit of frequent or great inequality of development between them, while computation and spelling do so permit, then, for the understanding, guidance, and training of childhood, we should seize upon the latter as a feature to be investigated in the case of every child, while the former may be allowed to run its harmless course unprobed.

We shall obtain suggestions as to mental traits which are capable of developing independently of other mental traits by a review of endeavors to determine mental capacities and "psychological types."[1]

Jung classifies individuals upon the basis of their adaptation to situations as introvert or extravert — those who look inward or turn the mind upon itself, and those who look outward. As Jung makes no suggestion that the one type is more intelligent, more gifted in muscular or sense development, etc., than the other, we could, if the classification is sound, have persons unequally developed as judged by the average with respect to an intro-extraversion trait and some second mental trait such as intelligence. Jung, however, carries his classification much further than this. He con-

[1] Three workers have recently reviewed literature bearing upon this point: Klüver (1925), Stead (1926), and Spearman (1927).

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

siders that there are four basic functions, — thinking, feeling, sensation, and intuition, — and that these may be exercised in an introverted or extraverted manner, giving eight different types of mental activity and an equal number of types of persons. The scheme looks rational in several respects, but the writer does not trust a rationalization unaccompanied by a method of objective proof that attempts to fathom non-rational processes such as intuition, feeling, etc. Jung and his school have been peculiarly derelict in that they have failed to utilize well-known techniques of analysis, both experimental and statistical, and to date have no criterion whereby to prove either their own hypothesis or to disprove that of another. In view of the unestablished and uncertain importance of a classification of individuals upon the basis of introversion and extraversion, the counselor of children would do well to consign this classification to the field of investigation and research and not to that of practical application.

The following words from Dr. W. V. Bingham (1926), who is a thorough believer in the reality of intro-extraversion classification, should be taken to heart:

> But for the present, any gesture in the direction of practical utilization of these measures of personality as aids to vocational decisions should be made with the utmost hesitancy, in view not only of their necessarily low reliability, but also of the instability of the very personality characteristics whose share in vocational success is obvious.

Much speculation and considerable experimental investigation took place in the nineteenth century which had as its purpose the determination of mental types: visual, auditory, vocal-motor, tactual and kinæsthetic, and combinations of these. The interest in this problem grew slack largely because of an inability to find pure types. The problem in a somewhat different form has been revived recently by Jaensch

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and his coworkers in their proclamation that certain individuals are *eidetiker* and others are not. The *eidetiker* is one who very readily has projected visual images both of things actually seen before and of things imagined. These images partake more of the nature of objective phenomena than do memory images. They have a definite location in space, a size which is independent of the distance which they are projected, a definiteness, a color toning, and other qualitative differences from the memory image, including a feeling of lack of relationship, at least in part, with the volitional processes of the person sensing them. Jaensch considers that some 30 or 50 per cent of elementary school children are *eidetiker* and are to be clearly differentiated from the rest. He also considers that *eidetiker* school children differ among themselves in the vividness of their images.

The approach of the type psychologists of the last century was quite different from that of Jaensch, but the basic phenomena in which both have shown interest is clearly the same. It would be foolhardy to assert that the new approach is but a repetition of an older one which has proved abortive, but certainly principles and conclusions are as yet far too indefinite to permit their being of service in the routine classification of school children. Let us by all means encourage further investigation of this important subject, but meanwhile not start a classification of children into *eidetiker* and non-*eidetiker* and not attempt vocational counsel and class instruction upon the basis of such a classification.

A very modern attempt endeavors to classify persons upon the basis of internal secretions, giving the thyroid individual the pituitary character, etc. Investigators along these lines have undertaken to establish the relationship between internal secretions and mental traits, but have in all cases found very low correlation. These attempts are but modern versions of the classical endeavor to define character in terms of

Generated on 2020-12-23 00:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the four humors which it was believed that the body secreted. That psychologists no longer seriously attempt to classify men as phlegmatic, sanguinary, choleric, or melancholic is not proof that a classification upon the basis of bodily secretions is impossible, but we may at least bear the earlier debacle in mind and refuse to be converted to the internal-secretion school until much better evidence has been adduced than is the case to date.

There have been two different kinds of attempts made to determine mental traits from racial origins. The one type concerns itself with the relationship of general intellectual level and race. Certain investigators of this problem have concluded that Mediterranean people are on the average inferior to those of Nordic origin. Much evidence bearing upon racial differences of this general all-round sort has been collected and presented, but as an aid in the problem that we are here concerned with, — the discovery of outstanding individual mental traits, — such conclusions as these investigators reach are very nearly worthless, because the differences of individuals within any race are so much greater than the differences between races that knowledge, let us say, that the Jews average higher in general intelligence than the Irish is of little avail in determining whether this particular Jew is superior to this particular Irishman. We cannot look to racial group studies for appreciable aid in the problem of individual classification.

The second type of racial study concerns itself with differences within races of two or more traits. Thus, the Armenian has been described as a sycophant and a tradesman and the Turk as a fanatic and a warrior. The justification for such classifications may lie in the cultural environments of the different races, but as springing from original nature, they certainly are not established, and here again conclusions that have been made in the past are group conclusions rather

Generated on 2020-12-23 00:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

than individual. It should be needless to say that the mental classification of individuals on the basis of racial origin should not be attempted in the practical, everyday segregation of school children in American public schools.

The investigation of Sante Naccarati (1921) had as its purpose the building up of a morphological index which would correlate with mental ability. He found that the ratio of limb length to volume of trunk is an index which correlated in the case of 75 male university students to the extent of .35 with general all-round mental ability as measured by the Thorndike Psychological Examination. This is the first correlation between physical measurements and intelligence of sufficient size to arouse more than a passing interest upon the part of school administrators. The findings of Naccarati are based upon a small population and have as yet never been confirmed by a more extended study. The earlier anthropometric studies of Galton, Pearson (1926), and others, involving brain measurements, cephalic index, various bone measurements, etc., have yielded correlations with mental traits which were much smaller in value and also much more reliably determined than Naccarati's. Certainly we may not yet place confidence in anthropometric measurements or morphological indexes as a means for the mental classification of school children.

There is one type of character analysis that is hoary with age and so universal that very generally there is a presumption in its favor. It is the attempt to read character by facial characteristics. Space does not permit a discussion of the many ramifications that this attempt has taken, nor shall we here discuss graphology — the betrayal of character through handwriting. None of these methods has established itself as having more than the faintest suggestion of validity. The writer finds it hard to believe that this will always be so and in truth expects that some day the analysis of mental ability

Generated on 2020-12-23 00:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and of emotional characteristics will be clearly furthered by quantitative and qualitative measures, facial contours and expressions. It is very possible that he is wrong in this and that it is merely his enjoyment in motion picture close-ups that leads him to look for a contribution in this direction. However that may be, he certainly cannot advise the use of any of the character-reading schemes based upon facial features as aids in the educational or vocational guidance of school children. Dependence of school administrators and employment agencies upon such features is to be found on every hand, but measurement of the efficacy of classification indicates not only that such dependence has not improved classification, but that it has regularly made it worse. Much as we enjoy " playing a hunch," a child's welfare is altogether too serious a matter to the child himself for us to take liberties with it. Let us indulge this type of classification when traveling upon the train: " Yonder man is or ought to be a doctor; and that one a plumber; and the little fellow in the corner a druggist; etc.," and after the trip we can brag about our expertness to our friends. The writer has classified hundreds of persons in this manner and recalls having made but two mistakes: one person classified as a traveling salesman was later discovered to be a minister, and another classified as a department-store manager was discovered to be a university professor of philosophy. The remaining persons classified, not having recrossed his path, may be thought of as enjoying the vocations assigned to them. The credence one gives to his own snap judgments of his fellow men is amusing when it is not serious. The case of the child, since he is more defenseless, more at our mercy, and more trusting, is very likely to be serious. Let us repress our intuitional natures and judge of character by crediting, first, objective mental measurements; second, a child's self-analysis of his abilities and interests, particularly if it is the outcome of a

Generated on 2020-12-23 00:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

period of critical study and self-searching; third, the judgments of teachers who have known the child for a long time and have appraised him with a sympathetic understanding; and fourth, the parents' convictions, which, though based on intimate knowledge, are unfortunately commonly clouded by an affection which is not aware of shortcomings and by an understanding altogether too independent of comparative data regarding other children of the same age and sex. The imposing of a snap judgment — which is merely the adult version of a fairy tale — on a child's uncritical confidence has far-reaching consequences, not only to the child, but to our profession. Elementary and advanced teachers are tied together, and if high school and university teachers find that their charges lack confidence in them, it is well to remember that this lack of confidence has been earned and that ordinarily but little rectification is accomplished in the higher years. The trouble is deep-rooted. The training that would assist one in making a correct appraisal of so knotty a problem as a growing child has been lacking, because we as an organized profession have not appreciated the importance and the complexity of character judgment. It is high time that we look upon it as a difficult and a serious matter.

Most of the inadequate means of mental analysis that have been referred to have assumed a linkage between a readily ascertainable physical feature, or sensory capacity, and a particular mental capacity. Let us now turn to classification schemes which make no such assumption, which state rather that a mind is *sui generis* and to be studied upon its own account. Descriptive words will still need to be used, but they will no longer be correlated with humors, glands, sense, or motor features.

One such approach, drawing its inspiration from the physical idea of the level or of equilibrium, has postulated certain antagonisms or compensations. To quote from Thorndike (1913 educ. psych., pages 360–361):

Generated on 2020-12-23 00:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Such are: — that superiority to the central tendency in vividness and fidelity of imagery of one sort implies inferiority to the central tendency in vividness and fidelity of imagery of other sorts; that superior ability to get impressions through one sense is related to inferiority in getting impressions through other senses; that intensity of attention varies amongst individuals in opposition to breadth of attention, so that a high degree of power to attend to one thing at a time goes with a low degree of power to attend to many things at once; that the quick learner is the poor rememberer; that the man of great artistic gifts, as in music, painting, or literary creativeness, is weak in scientific ability or matter-of-fact wisdom; that divergence above the mode in power of abstract thought goes with divergence below the mode in thought about concrete things; that the man of superior intellect is likely to be of inferior mental health; that the rapid worker is inaccurate; that an agile mind goes with a clumsy body; etc. Not all of these and other supposed antagonisms or inverse relations have been specifically tested by the calculation of the appropriate *r*'s, but those which have been so tested have been found in gross error.

Such common beliefs as those mentioned by Thorndike are perversions of a very simple fact which is characteristic of each individual. If a person is superior to his own average of attainment in one capacity, he will of necessity average inferior to his own average in the sum total of his other capacities. A child cannot be above average weight for his height without at the same time being below average height for his weight. A child cannot be superior to the rest of his mental make-up in mathematics without being inferior to the rest of his mental make-up in something else, perhaps spelling. This, though a mere mathematical necessity, is nevertheless a very important fact to bear in mind in studying human character. Let us then discard entirely any belief in mental antagonism or compensation in the sense that inferiority to the racial average in one trait implies superiority to the racial average in some other trait, but let us keep the concept that inferiority in one trait to an individual aver-

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

age is concomitant with superiority to the individual average in one or more other traits, and let us subsume this concept under the simple proposition, "Idiosyncrasies exist."

Many individual peculiarities are so directly traceable to education that they are more or less uninteresting and, in fact, unimportant in many problems of classification, being of the nature of end products rather than of initial causes. Thus, if one high school graduate knows French and another knows Spanish, each as a result of having studied the language mentioned, we are not interested in this as a measure of capacity but merely as an accomplished fact. It will be important with reference to certain vocations; for example, with reference to two kinds of foreign trade — important, however, as a measure of attainment in these vocations and not of capacity to attain in them. The measurement of acquired idiosyncrasies of this nature is readily accomplished by subject-matter tests.

Another type of idiosyncrasy likewise measured by differences in ability in school subjects is revealed when two children, each having been subjected to the same environmental opportunities, end up with quite different relative abilities. Thus, two children may go to school together year after year and each on the whole do average work, the one, however, being better in reading than in arithmetic, and the other the reverse. A relative superiority of this sort is not a mere incident due to differences in subject matter studied, because the pupils have studied the same things. Because of native differences in capacity for the different subjects or because of an earlier differentiation in interest and effort which has persisted with the years, we discover, perhaps for the first time in the middle or late school years, a genuine difference in relative accomplishment which is, however, more than merely that, for it is a prophecy of differences in capacity to achieve in the future along various related lines.

Generated on 2020-12-23 00:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The commonness of differences in abilities within individuals is, in the first place, dependent upon basic nervous structure and, in the second place, upon environmental impositions. Because of differences in educational stresses, it is possible to develop a child much more keenly aware of the meaning of words than of the meaning of the sentences containing the words, but such a difference when found is not to be attributed to an original nervous structure which gave a predisposition this way. On the other hand, a child subjected to average educational pressure in all of his school subjects may, due to native predisposition, develop greater awareness of the meanings of words than of the numerical relationships between magnitudes. The writer has presented elsewhere (1926) evidence in support of these last two statements. Clearly, it will be advantageous if we can discover what the original predispositions are or, otherwise expressed, if we can discover what are the mental functions which are readily capable of developing more or less independently of the other mental functions which are commonly active. Though these functions have not as yet been established with the certainty which is very clearly needed before routine guidance procedure can be built upon them, nevertheless they seem to the writer far more certainly established than the character types deduced by analogy or by assumption of some physical linkage, such as those discussed in the preceding paragraphs. A further reason for crediting the trait analyses about to be reported lies in the fact that the method of discovery has been inductive, coming out of data studied, and has not grown out of *a priori* assumptions inadequately tested.

In several early studies Spearman (1914) and Hart (1912 and 1914) attempted to determine in an inductive manner the independence or dependence of mental traits. They came to the important conclusion that there is a single gen-

Generated on 2020-12-23 00:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

eral factor running through all sorts of intellectual activity (quite synonymous with " general intelligence ") and that in each separate activity there is also a special factor which is unique to that activity, or to that and other closely allied activities. Spearman considers this central factor to be due to " a central fund of intellective energy " and the special factors to be due to the specific neural mechanisms which mediate the particular activities.

Without committing one's self to the cause of the general factor, it does seem that a general factor (further evidence suggests that there is more than one) does exist or, what amounts to the same thing, that many overlapping factors exist which overlap in such large part that a common factor may be thought to be present in all. To make Spearman's contention clear, let us suppose that we have measures of twenty mental traits, X_1, X_2, . . . X_{20}. It might be supposed that different amounts of the same four mental factors, A, B, C, D, and nothing else, were involved in these twenty traits. Thus, the first trait might be represented by:

$$X_1 = .50\,A + .50\,B;$$

the second trait by:

$$X_2 = .40\,A + .20\,B + .40\,C;$$

etc., for the remaining eighteen. Opposed to this is Spearman's view that each of the twenty may be thought of as due to a common trait G plus a special trait S_1, S_2, S_3, . . . S_{20}. Thus, for example:

$$X_1 = .80\,G + .20\,S_1$$
$$X_2 = .70\,G + .30\,S_2$$
$$X_3 = .80\,G + .20\,S_3$$

etc.

Dr. Godfrey H. Thomson has shown that many situations described by the scores received from pupils in a number of mental tests can be thought of as being the expression of a

Generated on 2020-12-23 00:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

number of general mental factors, *A*, *B*, *C*, *D*, or, as Spearman claims, of a single factor plus specific factors. When both of these are possible explanations of a given situation, which shall be chosen to meet practical needs? Opinions will differ here, but we may at least agree that if the Spearman view is a possible view, it provides a very ready means of cataloguing a person's achievements and capacities. We may pause to note that the Spearman view is assumed by many school people, mental testers, and clinical workers, when they characterize a person, as they so frequently do, in terms of his general intelligence or of his intelligence quotient, and employ no additional mental rubric. This provides a sort of empirical warrant for the view. However, Spearman himself no longer defends it just as here presented. His own students and coworkers have found data which well-nigh conclusively indicate that there is more than one general factor. We must therefore build up a picture of mental life which is more complex than the simple picture first provided by Spearman.

These additional factors may be general, running through all mental activity, or they may be group factors found in a limited number of mental activities. Much remains to be done before this and other points are cleared up, so that we shall merely note some of the more important and far-reaching traits which seem to be entitled to an independent status. In addition to Spearman's general factor *G*, Webb (1915) found a second factor of wide generality which he characterized by the phrase " persistence of motives." To Webb this means constancy of action resulting from deliberate volition. He clearly has here a factor experimentally determined which fits in well with the philosophical concept of " will power." That Webb's factor meets a need in the understanding of character would be agreed to by McDougall, who argues for " purposive strivings " as an essential category

Generated on 2020-12-23 00:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of human life, and probably also by Woodworth, who would find in it much support for his dynamic psychology.

The data collected by Webb have been very carefully studied by Garnett (1919), who finds three general factors running through the data: (1) an intelligence factor, (2) Webb's factor, which he renames "purposefulness," and (3) a factor which he names "cleverness," which is apparently consequent to "association by similarity." Following Garnett's analysis, let us attempt to picture the mental associations of three persons when confronted by the same situations — the first person average in factors (2) and (3), but above average in (1); the second person average in (1) and (3), but above average in (2); and the third person average in (1) and (2), but above average in (3). This is of course a hypothetical illustration, but it may make more explicit the nature of these three general factors. Each of the three subjects listens to the word "play" spoken by another and is asked to state the first thing that comes to mind. Individual (1) replies "work," having analyzed the meaning of the word and having noted that the gamut from play to work and back again completes a cycle, so that the analysis is complete. Individual (2) replies "joy," having partly reasoned and partly felt that the outcome or end of play was happiness; while Individual (3) in the briefest time of all responds "dance." We may call the response of the first individual a more intellectual, that of the second a more purposeful, and that of the third a more apt response. Individuals chronically disposed to think in the first or second or third manner constitute three different mental types. There are, of course, many individuals who lie intermediate between these three, but that there are individuals in considerable numbers of these three sorts is in substance the claim made by Garnett as a result of studying Webb's data.

Why there should be these three types and not such as

Generated on 2020-12-23 00:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

would be represented by the words " anxious," " friendly," " gloomy," etc., cannot be answered. In fact, we are unable to say that these last three words do not represent types. We can, however, assert that they have not been proved so to do, while the other three — " intellectual," " purposeful," and " clever " — are more or less descriptive of three different sorts of school children which have been found to exist.

This subject has just begun to attract the attention of experimentalists, and the first real authoritative chapter on it will be written ten, twenty, or fifty years hence; but meanwhile children are growing, teachers are teaching, guidance counselors are counseling, and Johnny Joneses and Betty Browns are being either neglected or placed in wonderful and fearful classifications. One such unwarranted classification is that upon the basis of goodness or badness. Studies by Voelker (1921), Raubenheimer (1923), Cady (1923), and many others give no indication that it is a unitary or an essential category of human life.

Another unwarranted category is that of general intelligence when used in the sense that those high in this trait are possessed of more ability in all mental traits than those low in it. Certainly if intellect involves the traits already discussed, — namely, purposefulness and association by similarity or other traits which will shortly be considered in more detail (mental manipulation of spatial relationships or of quantitative concepts, ability to think with non-verbal material, etc.), — then it is not unitary and therefore not an essential rubric. It may seem rather presumptuous to imply that " general intelligence " is without experimental warrant. In explanation the writer would say that he believes that it has much warrant if confined within proper limits, if the thing meant thereby is facility in abstract thinking when problems are stated in verbal terms, but as something including this

Generated on 2020-12-23 00:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

along with the other talents mentioned, it is belied
xcellent evidence.
must of course bear in mind that the terms here
ut first approximations to the underlying
cts in the case are that it has experimen-
one single mental function is sufficient
ormance of school children and uni-
ing psychological tests: synonym-
ntence meaning, sentence completion,
y, and still others. The writer prefers to designate
nning through and inclusive of all
abstract relationships when stated
Spearman attributes it to varying
ral fund of intellective energy." In an
we mean exactly the same thing thereby,
statement finds its real meaning in the same
The same situation holds with reference to
ns; facility in the "manipulation of space relation-
" is synonymous with ability to score on form boards,
right- and left-hand tests, on geometrical form tests (cut-
ting a given figure into required parts, or the reverse), etc.
Similarly, Garnett's "cleverness" and "purposefulness"
are but first approximations to the true meanings which are
to be found in the tests employed (and in this case in the
meanings given to a number of other terms by judges). The
reader must not be over-impressed by the particular words
which are used. There is an intrinsic difficulty here which
can be squarely overcome only by the coining of a number of
new terms.

The present meanings of the words of our language are
mere weighted averages; "success" means what people
think it means, and nothing more. In determining the mean-
ing of a word, the opinion of a man whose influence is far-
reaching must be weighted many times that of a hermit, but

Generated on 2020-12-23 00:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

when such allowance is made, we may say that if 90 per cent of individuals attach one meaning to the word "success" and 10 per cent attach another and inharmonious meaning, then the will of the majority prevails and the meaning is that given by the 90 per cent. There are many words — "virtue," "success," "intelligence," "evil," etc. — for which so wide a range of meanings is common that the average meaning is difficult to determine, and when determined, is considered unsatisfactory by a large dissenting minority. This is difficulty enough for any word to carry, but the experimentalist comes along and attempts to redefine the word in terms of performance of a designated group on a designated test or tests.

For him its meaning is no longer a consensus-of-opinion meaning, but a much more objective and explicit thing. He is undoubtedly greatly aided in his thinking processes by the objectivity and explicitness of the word, but he has taken violent liberties with a social concept, — the meaning of a word, — and it is forever after incumbent upon him to iterate and reiterate the meaning in which he uses the term. If he does not do so, he is to be held responsible for any resulting confusion. The acrimonious discussions of recent years hanging upon the nature of intelligence, the intellectual level of adults, racial differences in intelligence, etc., have in the main been between those whose concept of intelligence has come from a social consensus-of-opinion definition and those who have used, though none too explicitly, an entirely new definition based on scores in designated tests. The present writer considers the social warrant for the opinions of the former group to be above reproach and the evidence given by the latter group to be most excellent. In the present text he follows the illogical procedure of the members of the latter group and uses terms already otherwise defined by custom with a meaning ultimately revealed through scores on a test.

Generated on 2020-12-23 00:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

These two meanings of such words and terms as "intelligence," "mechanical ability," "purposefulness," "cleverness," "social interest," "intellectual interest," "mechanical interest," "manipulation of space relationships," and "manipulation of number relationships" have not been sufficiently examined to enable one to assert that they are substantially the same. The author hopes that they are sufficiently similar in meaning to be serviceable, but he would here express his conviction that before long those using terms which find their meanings in objective performances will need to create a lingo of their own, grievous as is this prospect.

As early as 1907 Krueger and Spearman (1907) found a memory factor entering into a number of tests. A memory factor was found also by Abelson (see Bernstein, 1924) and again by Carey (1915–1916) and still again by T. Verner Moore (1915). The author, in a work as yet unpublished, has found evidence in support of a general memory factor.

Before turning to classifications which have arisen in America, we may very briefly note some other unitary factors suggested by the studies of English workers. These are less important both because less universal in their presence and because of less magnitude when found than the four already mentioned.

The persistence of sensory and memory images was studied by Lankes (1915), with the result that "perseveration" seemed to be a general factor and one independent of general intelligence and also of memory. The correlations that Lankes obtained were very low, and his findings accordingly have large probable errors, as his population was only one of 47.

The factor found by Flügel (1913) and called "oscillation" should be reinvestigated, as should also the very important negative result reported by Bernstein, that there is no general speed factor.

Generated on 2020-12-23 00:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

From an intensive study of a very small population, 5, Peak and Boring (1926) conclude that mental speed and general intelligence are much the same. This is, of course, in harmony with Bernstein's findings.

Dr. Cyril Burt (1909–1910) considers that there is an innate central emotional factor and probably also an acquired central moral factor.

A very recent analysis by Stead (1926) attempts a general reconciliation of the views of Spearman, Burt, Webb, and Garnett. He considers that three factors characterize the individual: (1) the amount of total energy at his disposal — that is, the strength of the instincts forming the bases of his activities; (2) the amount of this total energy that is "graded" — that is, the amount that is at his disposal in varying amounts as contrasted with the all-or-none nature of energy when expressed through instinctive behavior; and (3) the "firmness" of the control of the energy represented under (2) — that is, the number of gradations of energy control possessed by the individual and the extent to which such graduated releases of energy are subject to his volitional control. This scheme of Stead's is very interesting, but the writer hardly considers it a direct consequence of his data; it rather seems to be the product of a canny philosophical speculation. We shall therefore await further investigation.

In America the study of the problem has taken a very different turn and in general a more "practical" and a less statistical and analytical trend. The motives for such investigations in America seem to have arisen out of the desire to determine vocational fitness, — this man will make a good salesman, that one a good investigator, etc., — and we accordingly find the salesman type, the research type, the administrator type, etc. Dr. Thorndike (1920) has thought that individuals are possessed of three types of mental ability and that although most people possess all three types in much

Generated on 2020-12-23 01:17 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

the same amount, there are nevertheless certain ones who are outstandingly superior or inferior in one or another of the three. The three kinds of intelligence are: (1) abstract intelligence, or the ability to deal with abstract ideas; (2) social intelligence, or the ability to get on well with one's fellows; and (3) motor intelligence, or the ability to manipulate and understand mechanical contrivances. Though Thorndike has not provided the statistical evidence establishing the independence of these categories, another worker, Dr. Wyman (1924), has studied the interests of children along three lines so similar to the three proposed by Thorndike as to be very pertinent in this connection. She found a very high degree of independence between " intellectual " interest, " social " interest, and " activity " interest.[1]

It also was found that intellectual interest, independent of intelligence as measured by an intelligence test, correlated quite substantially with school achievement. In brief, Wyman's study gives support to the view that there is an interest trait, possibly very similar to Garnett's purpose or Webb's persistence of motives, and further, this trait shows the same lines of cleavage — intellectual, social, mechanical, or activity — postulated by Thorndike when speaking of abilities.

The lines of cleavage in mental structure thus reported are so numerous that it is no little task to attempt to think of all of them in connection with each case studied. However, certain investigations of the writer, which in the main are not as yet published, give much warrant for extending the list still further. The writer's studies (1923 dist. and 1926) in the

[1] When corrected for attenuation, the correlation between intellectual interest and social interest was found to be .36 (P. E. = .11); that between intellectual interest and activity interest, .20 (P. E. = .14); and that between social interest and activity interest, − .08 (P. E. = .19). The corresponding alienation coefficients, or measures of independence between the three interests, are .93, .98, and .99 +, respectively.

Generated on 2020-12-23 00:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

main corroborate earlier conclusions reached by Thorndike (1921). They support the view that the following traits are more or less independent of each other:

(1) Verbal intelligence, or the ability which in the main underlies facility in naming opposites, coördinates, subordinates, supra-ordinates, predicates; and found in tests of mixed relationships, practical judgments, vocabulary, written directions, sentence completion (textual matter of literary content), sentence meaning, paragraph meaning, word meaning, and logical selection.

(2) Quantitative intelligence, or the ability in the main underlying facility in computation and other situations involving numbers as content.

(3) Spatial intelligence, or the ability in the main underlying facility in handling form boards, geometrical forms, and right- and left-hand, Knox cube, and other similar tests.

(4) Memory, or the ability in the main underlying memory for verbal material (as yet the writer does not know whether this factor extends also to non-verbal material).

(5) Drill, or the ability underlying school studies requiring much drill (indicated to exist in connection with computation and spelling and suggested in other connections).

(6) Several traits involving kinæsthetic and motor abilities. Dr. John F. Walker, in a doctor's dissertation on file at the University of California, reports the discovery of a number of different motor abilities, each largely independent of the others. It seems that when considering fineness of control of different muscles, we must think of several abilities and not of one general motor ability.

(7) The work of Seashore and others strongly suggests a musical ability, and it is likely that other sense organs than the ear have concomitant mental phases.

This list, though disconcertingly long, is probably not complete. In particular it may be necessary to add the very

Generated on 2020-12-23 00:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

important category "mental speed." The evidence upon this point is somewhat conflicting. The list, long as it is, is a tremendous abridgment of the adjectives which are variously used to characterize mankind. The writer has seen a list of character traits totaling over 130 which was proposed to be used as a guide in counseling. The vocational counselor would need to be an animated reservoir, filing cabinet, and regression equation in order to collect, arrange, and interpret any such mass of data. In truth, he has to be something of these things as it is, but the task is not the hopeless one it would be if one had one or two hundred character, intellectual, social, and motor traits to appraise.

4. A minimal list of traits to be studied for the understanding of typical school children. We shall now attempt a listing, in the order of importance and availability, of the items of information required for the understanding and school guidance of a child. The same order of importance will in the main hold in connection with vocational guidance.

Items readily secured and of prime importance when they are, for any individual, exceptional:

1. Name; sex (a certain sex becomes exceptional when considered in connection with an activity ordinarily engaged in by the opposite sex); residence; date and place of birth; nationality of parents; vocation of parents; past disciplinary record; recent school transfers; special sensory or motor defects or abilities; past illnesses; present general health condition; bodily size and strength.

Items important for all purposes involving mental classification, education, and guidance:

2. Maturity, or present chronological age. (This is so important that it should be obtained in two independent ways in order to check one against the other.)

Generated on 2020-12-23 00:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

3. Verbal intelligence. (This can be determined by school achievement tests of the reading and vocabulary sorts and by general intelligence tests or such portions of them as are of a verbal nature and do not include numbers and spatial relationships as content.)
4. Social intelligence. (Pending the derivation of an adequate objective test, this trait must be estimated by teachers and others coming into intimate contact with the child.)
5. Activity and mechanical intelligence. (This trait may be estimated by judgments and in part by existing mechanical ability tests. It may eventually be necessary to divide the trait here mentioned into two or more.)
6. Interests along lines (3), (4), and (5) and along other lines specially noted by the child. (These other lines may be correlated with special sensory or motor development, as, for example, are music, drawing, and certain games. Measures of these interests may be estimated by teachers on the basis of replies to a questionnaire and may also be measured by interest tests such as the Wyman test and the Cowdery (1925) test.)
7. Ability with reference to quantitative phenomena — computation, etc. (This can be determined by computation and various other number tests.)
8. Ability with reference to spatial relationships — geometrical forms, etc. (This can be determined by form boards, geometrical form tests, etc.)
9. Memory with reference to verbal material. (Since most tests of memory include element (3) to a large extent, this should at present be estimated by teachers rather than derived from a test score.)

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

10. Special sensory or motor interests and abilities. (Suggestions as to these may be got from a questionnaire, and certain of these may also be readily ascertained by existing objective tests.)

These ten rubrics constitute the essential list. The reader will note that there is no category " general intelligence " included. The writer has of course omitted this intentionally because it has commonly been measured by tests which are a complex of Items (3), (5), (7), (8), and (9).[1] He has, however, placed Item (3), which is the largest single element in this complex, next only to chronological age in importance. Also omitted are such items as " purpose," " cleverness," "intro-" and "extraversion," etc. It seems to the writer that " purpose " should yield to " purposes " and that these are very likely represented in the threefold classification given under (6). The claims of " cleverness," " intro- " and " extraversion " to a place are not as yet sufficiently established.

The writer believes that a less analytical study of a child than that covered by the ten points mentioned will fail many times to ascertain essential peculiarities, while a study involving more points and different points will frequently raise trivial issues or suggest individual oddities which do not correspond to actuality.

[1] Item (2) could also be included here when dealing with mental ages above 14 and IQ's based upon them.

Generated on 2020-12-23 00:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER SIX

Experimental Studies of Certain Inequalities of Development

1. The traits to be studied and an outline of the steps to be followed. The writer wishes that he could here report a study of school children for every one of whom measures from (1) to (10) as listed in the preceding section were available. Such data are not to his hand; so he will present data which are quite rich in information upon some of these points, though inadequate with reference to others.

The Stanford Achievement Test was given near the end of the term to 25 low eighth-grade pupils, with the results tabulated in Table 16.

To aid in sectioning and in determining promotions an educational profile should be drawn up for each child, giving his scores upon all tests which measure disparate capacities. The nine Stanford Achievement Tests are Paragraph Meaning, Sentence Meaning, Word Meaning, Computation, Arithmetic Reasoning, Science Information, History and Literature Information, Language Usage, and Spelling. A study of the first three Stanford Achievement Tests and studies of more or less similar tests found in intelligence test batteries warrant the conclusions that these tests are very similar in terms of the basic, underlying trait which they measure. We shall therefore draw no distinction between them, and in drawing up a profile shall use a chart making provision for "Total Reading Score," but no provision for the three separate tests. The chief function measured by these reading tests is that called in the last section "verbal intelligence." The fourth test is Computation, and it is clearly entitled to consideration separate from the other tests. It has a small bond with verbal intelligence, a large connection with Arithmetic Reasoning, and a small linkage with Spelling, probably

Generated on 2020-12-23 00:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 16

STANFORD ACHIEVEMENT TEST SCORES OF A CLASS OF LOW EIGHTH-GRADE CHILDREN MADE NEAR END OF SEMESTER

NAME	SEX	CHRONOLOGICAL AGE	PARAGRAPH MEANING	SENTENCE MEANING	WORD MEANING	READING TOTAL	COMPUTATION	ARITHMETIC REASONING	SCIENCE INFORMATION	HISTORY AND LITERATURE INFORMATION	LANGUAGE USAGE	SPELLING	TOTAL
A. C.	m.	11–8	108	74	82	264	128	120	93	91	50	208	95.4
G. B.	m.	12–11	88	59	66	213	140	124	71	63	48	172	83.1
C. M.	m.	12–1	82	60	55	197	136	112	65	70	34	154	76.8
A. N.	f.	12–6	110	76	62	248	160	132	86	92	56	194	96.8
P. N.	f.	12–0	86	52	59	197	160	116	62	51	50	178	81.4
M. R.	f.	12–6	98	65	75	238	124	88	84	81	40	206	86.1
H. Z.	m.	12–6	70	56	46	172	132	128	67	57	26	120	70.2
R. A.	f.	13–3	72	51	52	175	148	84	55	43	32	166	70.3
B. H.	f.	13–11	78	63	58	199	140	112	56	20	36	190	75.3
G. M.	f.	13–9	94	59	62	215	144	96	60	47	44	196	80.2
E. M.	m.	13–10	102	78	79	259	160	148	85	84	43	154	93.3
H. P.	f.	13–3	102	68	78	248	120	96	76	76	46	144	80.6
E. B.	f.	13–4	96	62	61	219	148	108	65	61	48	196	84.5
L. B.	m.	14–1	88	72	59	219	164	136	86	64	40	176	88.5
B. C.	m.	14–3	112	78	81	271	116	124	80	84	48	174	89.7
J. D.	f.	14–4	86	72	54	212	140	124	79	56	48	142	80.1
K. E.	m.	14–4	86	67	73	226	160	96	79	82	44	184	87.1
H. N.	m.	14–8	82	64	57	203	112	108	56	56	42	148	72.5
R. A.	m.	14–2	88	74	70	232	140	108	89	76	42	182	86.9
N. W.	f.	14–10	94	58	62	214	140	68	63	60	54	188	78.7
G. F.	m.	15–10	78	56	63	197	124	96	52	60	36	170	73.5
G. J.	f.	15–2	58	35	49	142	116	84	64	27	32	160	62.5
H. W.	m.	15–6	78	36	48	162	120	92	59	41	38	164	67.6
C. C.	f.	16–10	80	68	64	212	120	96	73	51	40	168	76.0
L. G.	m.	16–8	80	57	54	191	144	120	70	63	42	162	79.2
Means	.	13–11	87.8	62.4	62.8	204.2	137.4	108.6	71.0	62.2	42.4	171.8	80.65

due to a drill factor. The fifth Stanford Achievement Test is Arithmetic Reasoning. It is sufficiently different from the reading test, in that it includes " numbers as content," to warrant its standing alone, though it is, in fact, considerably more closely related to reading than it is to computation. The sixth test is Science Information. It is entitled to independent status, for although quite decidedly connected with

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

verbal intelligence, there is another factor, probably interest or purpose, which gives it considerable independence. Upon much the same grounds the next test, History and Literature Information, is entitled to separate treatment. The eighth test is Language Usage. It is somewhat difficult to ascertain just what enters into this, but whatever it is, the aggregation represented by it is quite different from any of the other Stanford Achievement Tests; so it is considered by itself. The last test is Spelling, and here a verbal intelligence, coupled with memory, drill, and interest, give a combination unlike any of the other tests; so we shall consider it separately from the rest. Our profile chart therefore provides, as shown in Chart 2, for Reading, Computation, Arithmetic Reasoning, Science Information, History and Literature Information, Language Usage, and Spelling. The next to the last, or tenth, column of the chart gives age norms.[1] Thus the mean scores made by a random sampling of American white 10–6-year-olds is: Total, 33; Reading, 102; Computation, 73; Arithmetic Reasoning, 39; etc. The 10–6-year-olds constituting this random sampling were found in all grades from the first through the sixth, but the mean school grade of all 10–6-year-olds must not be taken as 4.5 (the middle of the fourth grade) as recorded in the "grade" column immediately opposite 10–6. The entry 4.5 in the last column gives the grade, the mean scores of which, when a random sampling throughout the country is taken, are: Total, 33; Reading, 102; Computation, 73; etc. This total score, 33, is the norm for a random population of 10–6-year-olds, and it is also the norm for a random population of children of all ages found in the middle of the fourth grade. It must not, however, be assumed that grade 4.5 is the mean grade for a random sampling of 10–6-year-olds. This

[1] Age and grade norms here given are those reported in the 1925 Revision of the Manual of Directions for Stanford Achievement Test.

Generated on 2020-12-23 00:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

may seem rather confusing, and it is in fact quite puzzling. It is due to the fact that age distributions of pupils in elementary grades and also grade distributions of pupils of a given age are not symmetrical. Since this peculiar situation is the one that exists, we must interpret the achievement test norms in connection with the tenth column if we are studying a child in comparison with children of his age, or we must make comparisons with the norms of the last column if we are studying a child in comparison with children of his own school grade.

The only other figures of the table requiring explanation are those of the "sensed-difference-score" column. These are not an essential part of the interpretative procedure unless one wishes to calculate accomplishment quotients or express accomplishment as a fraction of average adult accomplishment. The figures here recorded are simply total scores when expressed in terms of units proportional throughout to sensed differences rather than in the units which happen to be the units of the test. If a first child scores 20 and a second child 21, the second is one raw test unit above the first, and a fourth child scoring 81 is one raw test unit above a third child scoring 80. However, the unit has somewhat changed its significance in these two different parts of the scale. To make them comparable in terms of sensed differences, — that is, differences appreciated by teachers and acquaintances, — the scores must be expressed in the units of the sensed-difference-score column. Doing so, we see that the second child is two sensed units above the first child and that the fourth child is only one sensed unit above the third child. In other words, children (3) and (4) are sensed as being only half as different one from the other as are children (1) and (2). These sensed-difference units have been measured from an estimated zero point, so that quotients in terms of them, except for chance errors and possibly a

Generated on 2020-12-23 00:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

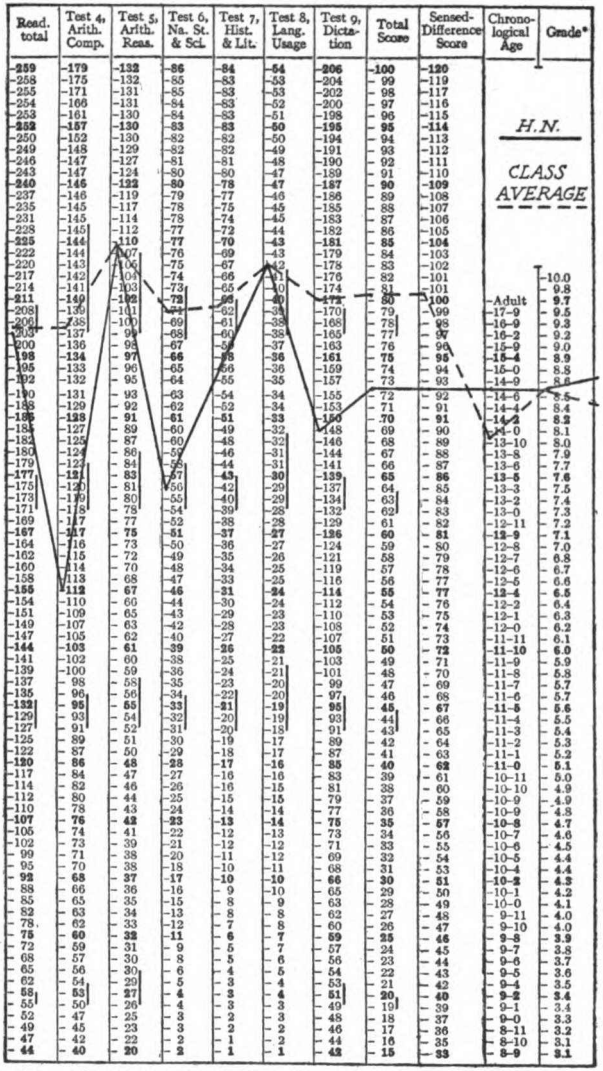
systematic error in the zero point, — which latter is possible, in view of the difficulty of determining zero points, — represent true achievement quotients. Let us designate these values given in parentheses by the letter *s*, meaning by it sensed-difference measures. Then a child's achievement quotient is his *s* score divided by the *s* score which is normal for his age. For example, a 10–6-year-old scoring 43 has an achievement quotient of 65/55, or 118 (as is customary, the decimal point has been dropped). The pupil's score 65 may also be interpreted as stating that his achievement is 65 per cent that of average adults, as the average adult score is 100. This scheme is particularly recommended for use with older children, both because the error in the zero point is less material than with younger children and because with these older children the other achievement-quotient procedures — for example, achievement-age-divided-by-chronological-age — lose the type of significance they possess with children of an earlier age. It is of course not necessary to use a quotient technique at all for a very thorough understanding of a child's accomplishment, for a comparison with age and grade norms is quite sufficient; but if a quotient technique is desired, the writer would recommend, where the data — sense unit scores — are available, that here given, especially for ages above 10.

Chart 2, on the next page, gives sensed-difference scores corresponding to total test scores from 15 to 100 inclusive. Table 17 (page 132) gives sensed-difference scores corresponding to total scores of 15 or less.

In the case of young children the use of the sensed-difference score in calculating quotients is not considered of much interpretative value simply because a small amount of tutelage will here, as with the simpler tasks in a Binet or other intelligence test, double or quadruple the amount of the test material known. Thus, if a 4–0-year-old scores 4

Generated on 2020-12-23 00:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Chart 2. Stanford Achievement Test Educational Profile Chart: Advanced Examination (with Sensed-Difference Scale)



* A grade of 6.5 indicates the middle of the 6th grade, of 7.0 the very beginning of the 7th, etc.

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 17

Stanford Achievement Test Total Score	Sensed-Difference Score	Chronological Age
15	33	8–9
14	32	8–7
13	30	8–5
12	28	8–4
11	26	8–2
10	25	8–0
9	23	7–10
8	21	7–7
7	19	7–5
6	17	7–2
5	14	6–11
4	12	6–7
3	10	6–2
2	8	5–9
1	5	5–1
0	3	4–0

on the Stanford Achievement Test, he has an accomplishment quotient of $\frac{12}{3}$, or 400. This probably fairly represents the case, and though the child does know four times as much reading, arithmetic, etc., as average children of his age, it is not in itself a highly significant fact — not so significant as the statement that he knows as much as average low second-grade pupils. Were we to depress our zero point by adding, say, 30 to each sensed-difference score, we should then obtain a quotient of $(12 + 30)/(3 + 30)$, or 127. This looks more reasonable than the former quotient, but is probably, in fact, very unreasonable, being an understatement of the actual number of times the child is above the average for his age. The facts of achievement are such that the writer does not attribute high value to any quotient, however calculated, for young children, which is based upon subject matter that is as readily influenced by training as the ordinary material used in achievement and intelligence tests.

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

2. The case of H. N. Let us now interpret the scores of certain specific cases. The profile of H. N. — boy, age 14 years, 8 months, just completing the low eighth grade — is shown on Chart 2. His total score is 72.5. The probable error of this score indicated by the vertical bar in the total-score column which is closest to 72.5 (the bar opposite total scores 77–78) is sufficiently small to inform us that we may place considerable confidence in this measure. The average score for the class is 80.7, showing that on the whole H. N.'s achievement is not up to that of the class in which he is located by 8 units, or about three fourths of a school year. His age is quite normal for the grade, so that his total score suggests that he would be better classified if continued in the low eighth grade rather than if promoted to the high eighth grade. Thus far we have considered only his grade, age, and total score. Let us now examine his scores on the separate parts of the achievement test. We should first, however, note that the scores on these parts have larger probable errors, as shown by the lengths of the vertical bars in the respective columns, than the total score, and they must accordingly deviate much more from any point chosen for reference in order to have significance than was necessary in the case of the total score. The standing of H. N. in each of the tests when interpreted in the light of the probable errors of the test scores is not markedly different from his average or total standing, except that he is low in Computation and a little high in Arithmetic Reasoning. The difference in standing in Computation and Arithmetic Reasoning is so large as to be quite significant. In judging of the significance of this difference, it would be very desirable to know its probable error; that is, to have a bar drawn of such length that it was equal to the probable error of the Computation — Arithmetic Reasoning difference. The exact determination of the length of this probable error bar

Generated on 2020-12-23 00:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

is something of an undertaking, but we can secure a very serviceable approximation to it with little labor. Let us note that in the seven different scores — Reading, Computation, Arithmetic Reasoning, Science Information, History and Literature Information, Language Usage, and Spelling — there are no fewer than twenty-one different kinds of differences which may be studied: Reading–Computation; Reading–Arithmetic Reasoning; Reading–Science Information; etc. We should expect just as a matter of chance, since our measures are quite far from being perfectly reliable, that the largest of these twenty-one differences would be very appreciable. In other words, the probable error of a difference, chosen because it is the largest of twenty-one differences, will be much larger than that given by the ordinary formula for the probable error of a difference, since the latter formula is based on the assumption that there has been no choice involved in selecting the difference. As a rather close approximation to the probable error of this largest difference of twenty-one possible differences the writer suggests that the sum of the probable error bars closest to the two scores involved be added and multiplied by $1\frac{1}{2}$. Thus, if we add the P. E. bar in the Computation column, extending from 118 to 123, to that in the Arithmetic Reasoning column, extending from 99 to 107, and draw a bar which is $1\frac{1}{2}$ times this sum, we shall find that the difference between the Computation and the Arithmetic Reasoning scores is about 2 probable errors, so that we are quite safe in considering that H. N. is genuinely inferior in the trait measured in the Computation test to that measured by the Arithmetic Reasoning test. Of course, the cause of this inferiority is not revealed to us, but in view of the rather low Spelling score, it may possibly be due to a dislike for tasks involving drill and memory.

We have examined first of all in the analytical study of

Generated on 2020-12-23 00:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

H. N.'s educational profile the maximum difference found — in this case that between Computation and Arithmetic Reasoning. It is generally serviceable to study the maximum difference first, for if this maximum difference is small, say less than 1½ probable errors, then we shall attach little significance not only to this difference, but also to all other differences, twenty in number, and draw our conclusions as to the best disposition of the child from his total score. If the maximum difference is so great as probably to be significant, then other differences are quite likely to be significant also. Thus, in the case of H. N. we have found the Computation–Arithmetic Reasoning difference to be probably significant. Further, the inferiority of the Science Information and Spelling scores to the Reading and Language Usage scores is probably descriptive of a real difference in the abilities of H. N., though the differences are not great and should not of themselves be the cause of a major alteration in the educational program. These differences might well be contributing causes if other things, especially H. N.'s interests, suggest a specialization or a particular vocational outlook.

As it is important to understand the line of argument here followed, the steps will be summarized: (1) An appraisal of the child's all-round achievement as represented by his total score in comparison with the age and grade norms of the school system in which he is located is first made. (2) An examination of the significance of the major difference found is made. If this major difference could easily have arisen as a matter of chance, then all smaller differences are even more likely to have so arisen. In this case diagnosis is not to be made on the basis of differences between test scores, but on the basis of general level of attainment as given by the total score. (3) If the examination of the major difference warrants the belief that it is descriptive of an achievement or a

Generated on 2020-12-23 00:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

capacity difference in the child, then guidance and future education should take this into account. (4) Lesser differences than this major difference have probable errors which are smaller than the probable error of this major difference, so that these lesser differences should be examined and, if fairly large, — and particularly if they support each other, — they likewise will be considered significant and kept in mind in determining future educational programs and in giving vocational guidance.

It was implied in the last paragraph that certain differences might " support each other." Experience with educational profiles shows that certain differences more or less suggest other differences. This seems to be due to the presence of unmeasured factors; for example, memory ability, spatial relationships ability, purpose, interests. Thus, if Reading and Language Usage are both high or both low, there is suggested an interest in or lack of interest in verbal material. H. N. stands fairly well up in both of these, and as we can attribute this to a single interest in verbal material, we shall say that they support each other. Being high in Reading and low in Language Usage, or vice versa, offers no simple explanation, and thus two such scores do not support each other. Individuals indubitably showing this latter condition are found, though in smaller numbers than those showing the former. We must therefore keep an open mind in the matter and not, before thorough investigation, assume that a difference of some one particular type must be due merely to chance. With a mind ever ready to recognize exceptions, we may nevertheless say that agreements in relative standing in the following pairs of traits " support each other ":

Reading:

Reading–Arithmetic Reasoning: the bond here is probably verbal intelligence.

Generated on 2020-12-23 00:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Reading–History and Literature Information: the bond here is probably a developed literary interest with considerable verbal intelligence.

Reading–Language Usage: the bond here is probably developed social literary interest.

Computation:

Computation–Arithmetic Reasoning: the bond here is probably an interest in mathematics.

Computation–Spelling: the bond here is probably a memory ability and a contentment with drill.

Arithmetic Reasoning:

The connection of Arithmetic Reasoning with Reading and with Computation has already been noted.

Arithmetic Reasoning–Science Information: the bond here is probably a developed interest in science and measurement.

Science Information:

The connection of Science Information and Arithmetic Reasoning has been mentioned.

Science Information–History and Literature Information: the bond here is commonly quite strong and is probably due to a general interest in reading. If the interest is critical as well as broad, Reading, Science Information, and History and Literature Information will all stand fairly close together.

History and Literature Information:

The special connections of History and Literature Information have already been mentioned.

Language Usage:

Special connections of Language Usage have already been mentioned, except for an occasional and not very pronounced connection between Language Usage and Spelling,

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

due probably to a developed interest in the structural phases of language.

Spelling:
The special connections of Spelling have been noted.

The wording used in the preceding statements has suggested some positive trait as the cause of some special bond, but of course low standing may equally well be accounted for by the absence of the positive factor; thus, lack of the usual interest in books, magazines, newspapers, etc., may well result in especially low Science Information and History and Literature Information scores, etc., for other pairs of traits.

We may now sum up the particular case of H. N. The important items are:

He is 9 months older than the average of his class.
He is 8 points in score lower than the average of his class.
He is quite certainly inferior in computation to his other talents.
The Computation–Arithmetic Reasoning difference is some two probable errors in size.
He is probably inferior in Science Information and Spelling to his other talents.
In another year he should probably think seriously of a life vocation and the preparation for it.

Judging by national norms, he shows average achievement for his age and grade, but the community wherein he resides is, as shown by the class profile, apparently quite far from average. The children completing the low eighth grade are 9 months younger than is nationally the case, and they are 8 points in total score higher than the national norm for their grade. They are in all about 1½ years in advance of the national norms. Since H. N. must be classified in this particular city, he will more nearly be with equals if held in the

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

low eighth grade for another half year than if advanced with the class.

We have not needed to consider the national norms at all in reaching this conclusion and have done so only because they are recorded upon our printed profile chart. One need not concern oneself with national norms when classifying for a single school or city, but when children in the upper grades are involved and the question of college or other further training is considered, state or national norms are valuable. The dullest child in a very superior high school might conceivably profit by going to college, whereas the brightest child in a very backward high school might fail most unhappily in college. Since the local norms cannot serve as a guide in such cases, a norm derived from a wider territory is needed.

Our first recommendation is that H. N. be not promoted to the eighth grade. Secondly, since he does not seem to be inferior in arithmetic reasoning, it is possible that an appropriate future vocation would demand talents along this line. If this should be the case, he would clearly be handicapped if of inferior computation ability. We can therefore with reason (*a*) provide his next teacher with his educational profile, pointing out that because of his strength in arithmetic reasoning, special effort should be made to strengthen computation, a talent closely linked socially to the former, but that in doing so there should be no sacrifice of opportunity to work and enjoy the more difficult arithmetic reasoning problems of the grade; (*b*) discuss with H. N. himself the dependence in all vocations of arithmetic reasoning and computation and tell him that if he would profit by his good arithmetic reasoning ability, it is necessary that his computation greatly improve. This covers the recommendations dealing with H. N. They are few in number and directed to persons always available and concerned with the case: H. N. himself, his next teacher, and his principal. Other

Generated on 2020-12-23 00:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

investigations leading to fuller information and perhaps other recommendations would be desirable, but commonly they are not feasible.

In presenting data covering pupils, such as H. N., to his students for report and recommendation, the writer not uncommonly receives a report somewhat as follows: "If H. N. is in good health, I should give him added work in computation"; or, "If his parents are ambitious for him and are willing to coöperate, I should have him do extra home work in arithmetic"; or, "If he did better work on a second test, I should promote him with the class"; etc. Such answers are evasions of the practical issue. Of course, if we had more information about H. N., we could handle his case more intelligently, but H. N., from the school principal's point of view, is just one very prosaic case out of hundreds, and the decision covering him has to be made upon about as much evidence as is here reported — which, by the way, is much more ample than is usually the case where reliable achievement tests are not employed. For reasons presented in Chapter IV, the disciplinary record, when used in connection with promotion and classification, is more commonly misused than otherwise. It is therefore not presented in connection with this case, for the writer would make the recommendation as to grade placement that he has made in regard to H. N., whether he is teacher's pet or the principal's nightmare.

Let us now note H. N.'s achievement quotient. This procedure will serve as an alternative method to that already followed. His total score corresponds to a sensed-difference score of 92.5, so that he is now nearly an average adult in scholastic achievement. The normal sensed-difference score for his age is 92.7, and he has substantially an all-round achievement quotient of 100. It is therefore doubtful if he will ever graduate from high school, and it is certainly to be

Generated on 2020-12-23 00:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

expected that he will not go on to college. The majority of vocations are within his grasp, and he should shortly pursue such school work as will further vocational training.

3. The case of A. C. and that of A. N. Having discussed an average case rather fully, let us now rapidly examine the profile of A. C., probably the brightest member of the class. First, as an exercise, his profile should be plotted on Chart 2. When we have done so, the following facts are obvious:

A. C. is $2\frac{1}{3}$ years younger than the average of his class.
He is 15 points higher in total score than the class average.
He runs above the profile chart in reading, science information, history and literature information, and spelling, and is therefore not well tested in these subjects; i.e., the test is not hard enough for him, as he would presumably score still higher if the tests in these subjects provided room at the top in which to exercise his talents.
He is low in computation and only about 1 year above his class in arithmetic reasoning. The most significant difference is probably that between computation and reading, but since we have no probable error bars near the top of the scale, the number of probable errors represented by this difference is not definitely known. However, following the procedure of adding the probable error bars for Reading and Computation which are given a little lower down on the chart and multiplying by $1\frac{1}{2}$, we see that the Computation–Reading difference is in the neighborhood of 3 probable errors. We may thus place much confidence in it, and when it is supplemented by the Arithmetic Reasoning score, we conclude that A. C. is relatively inferior in arithmetic.

All things considered, including the rapid rate at which A. C. is growing mentally (he is but 11–8 years old), we conclude that if now a misfit, he will be one still more a year

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

hence unless given extra promotions. A good schedule for A. C. would be to cover the next four years of school work in two. He should immediately skip to the low ninth grade, and six months hence skip again, and later repeat the process. He can, of course, at the moment pass any reasonable high school freshman, sophomore, junior, or senior reading test. He can probably pass any reasonable high school American history or general science test and could, after spending two or three weeks upon assigned readings, pass any reasonable freshman or sophomore English literature test. If some one states, "Impossible! Why, the little fellow is a prepubescent; what can he know of the loves of the lords and ladies of the past?" let him reserve his pity for some one else. A. C. should at this stage of his development be fully and scientifically informed of the biological differences and processes of the sexes. With a mind such as his, it should be considered a crime if some big coarse bully is able to poke fun at him because of his intellectual ignorance of certain simple biological facts of life. If informed, as he should be, he can appreciate with a depth and clarity not found in the ruffian or in the sentimental, self-conscious, lovesick youth the great and tender love stories of literature. A. C. should be allowed to skip elementary time-consuming courses in all lines except mathematics. Here he is somewhat backward and probably needs the regular work of a fast-moving section. He is far too young and distant from the terminus of his formal education to decide definitely upon a vocation and should not drop mathematics because he does not expect to follow a vocation involving it. Let him decide that matter four or six years hence, meanwhile taking all the mathematics offered in high school, for he will need it if he follows any of the physical, biological, or social sciences. If A. C. graduates from high school three years hence, with an extra language or two at his command and with an extra science, he

Generated on 2020-12-23 00:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

will be well fixed for foreign travel or college work. He is a superior child in a superior school system, and if not held back, should flourish mightily.

Let us calculate his sensed-achievement quotient. His sensed-achievement score is 114.4, and, as has been noted, this is probably lower than the truth because the test is not hard enough to test this boy adequately. The normal score for his age is 70.0, so his achievement quotient is 163. If he wishes to follow scholastic or research lines, he should be better than average Ph.D. material. Guide and instruct him accordingly.

The case of A. N. is interesting because the record is so uniformly excellent. Her further education should be very similar to that of A. C., with the difference that she can proceed more rapidly in mathematics. A study of the greatest difference given by her profile does not clearly establish the significance of any difference. Her achievement quotient is 148 (115.8/78). Thus four years in college, followed if desired by graduate work, are within her capacity, but both because of her youth and her uniformity of development, a suggestion as to a major field of endeavor to be followed would be premature at this time.

4. The case of G. J. If we now turn to the other end of our achievement distribution, we find an interesting case in the person of G. J. She is 1¼ years older than the class average; 18 points in total score lower; and shows a Reading Dictation difference which is 3½ times its probable error, so that a difference between these two abilities is well established. Other differences found are to be trusted, especially that between Science Information and History and Literature Information.

In all-round achievement G. J. is 18 points, or nearly 2 years, below her class. She would accordingly be with scholastic equals if admitted to the high sixth grade, for,

Generated on 2020-12-23 00:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

although we do not have available the average total score for the 6.5 grade in this particular school system, it is probably as much above the national 6.5 norm as is the 8.5 average for this school above the national 8.5 norm. This amount is 8 points on the total score scale, so that we shall assume the norm for the beginning of the high sixth grade to be, for this city, 55 + 8, or 63. Thus, in ability G. J. belongs with the pupils who are about to start the high sixth grade. However, she has only another year before she reaches the compulsory-school-attendance age limit, and we should hardly preserve school standards to the detriment of the individual. G. J.'s entire future curriculum should revolve around this issue: What can the school give her during the next year or two that will best fit her for a vocation? Let us by all means attempt to do this; otherwise we are tending to force her to an early marriage and motherhood. Consider the cacogenic effect of marrying our dullest pupils at the age of 18 and our brightest ones at the age of 28!

G. J. is not so dull but that she can do many things with pleasure and profit. She has a distaste for reading, history, and literature. It is too late in her life to attempt to change this — rather let this condition alone and search for a vocation not demanding this ability and interest. She seems interested in science and relatively so in mathematics. She is probably a good observer and interested in details, as is suggested by excellence in spelling. Would not her talents find expression as a clerk and assistant in a photograph gallery, in a dentist's office, or perhaps a drug store? With such an outlook in mind, place her probably as a special student in whatever grade will help her to this end. As a special student in junior or senior high school she should be able to select a course involving some or all of the following subjects: general science, elementary bookkeeping, commercial arithmetic, and household arts. If this curriculum fails to open

Generated on 2020-12-23 00:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

her eyes to the beauties of Shakespeare, bear in mind that if it adds two dollars a week to her salary, it will open more vistas than would years of agonizing over Shakespeare. She is nearly an adult and cannot readily read the popular magazines. What chance has Milton with her? Let us not feel sorry for her, for two dollars will buy many yards of amusing cinema, and our pity would be misplaced.

G. J.'s achievement quotient is 88, as given by 83.5/94.5. This is certainly sufficiently high to justify the school in accepting full responsibility for her cultural and vocational training. If there is no place for her to get in school such work as has been mentioned, then the school is at fault, for G. J. herself is quite typical of a substantial portion of the pupils passing through our educational halls.

Generated on 2020-12-23 00:42 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER SEVEN

Elementary Statistical Procedures

Every student who has thumbed the pages of this book to this point is certainly able to follow such directions for giving and scoring as ordinarily accompany a test and obtain a pupil's score. How much credence is to be placed in the score thus obtained is generally very inadequately considered in Manuals of Directions. To overcome this shortcoming to at least a degree is the chief purpose of this chapter. The confidence to be placed in an individual's score depends upon its probable error. The meaning of the probable error of a score is presented in succeeding sections: 1 (Plotting a distribution of scores), 2 (Calculation of the arithmetic average), 3 (Calculation of the standard deviation), 4 (Calculation and meaning of the probable error of a score). A shorter and more usual manner of obtaining the required probable error is given in Sections 5 (Plotting a scatter diagram), 6 (Calculation of a product-moment correlation coefficient), 7 (Expressing means and standard deviations in original test units), and 8 (Probable error of a score via the reliability coefficient). Certain further aids required in a number of important situations are given in the remaining sections.

1. Plotting a distribution of scores. Let us assume that John Doe makes a score of 11 on a certain 20-word spelling test. No sensible teacher would take this score as an infallible index of his spelling ability, but also, lacking further information, no sensible teacher would judge that because his score is 11, his true ability is 13 or 10 or any other number greater or less than 11 that might be stipulated. We do not know whether this obtained score is unduly favorable or otherwise to John, but we do anticipate that it does not exactly represent his ability. In other words, it is probable

Generated on 2020-12-23 00:42 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

that this score of 11 varies from his true ability. This gives us the concept that the score is in error when taken as a measure of his true ability.

If we wish to secure a better measure of his ability, we can give a second similar spelling test. Let us suppose that we do so, and that John makes a score of 10. The 10 is worthy of no more and no less credence than the 11, but if we take the average of the two, 10.5, we may place somewhat more confidence in it than in either of the scores separately. If a third test is given, the average of the three scores is still more trustworthy, etc. Let us give 25 equally difficult spelling tests, and we shall say that we find scores as follows: 11, 10, 16, 9, 13, 12, 15, 10, 12, 11, 11, 13, 9, 14, 12, 17, 10, 11, 11, 9, 12, 10, 15, 13, 12. We desire the mean or arithmetical average of these. We might add them and divide by the number, 25, but generally a simpler way is to make a distribution first, which we shall accordingly proceed to do. Running through the scores rapidly, we find that the smallest is 9 and the largest 17, so we shall draw up a tally sheet as represented by the first two columns of Table 18 on the following page. The other columns will be explained shortly.

The graph of the distribution of frequencies as given in Column *f* is represented in Chart 3 on page 149.

In plotting this distribution, straight lines are drawn connecting the various points. The initial point on the base line immediately above score 8 and the final point immediately above score 18 are correctly located as drawn. Errors are sometimes made in the initial and final points of a graph, but just as the other points are plotted directly above 9, 10, 11, etc., so should the initial and final points be directly on the base line at scores 8 and 18. A glance at this distribution is sufficient to convince one that there is little likelihood of John's true ability being as low as 9 or as high as 16. It probably lies between 10 and 13.

Generated on 2020-12-23 00:42 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 18

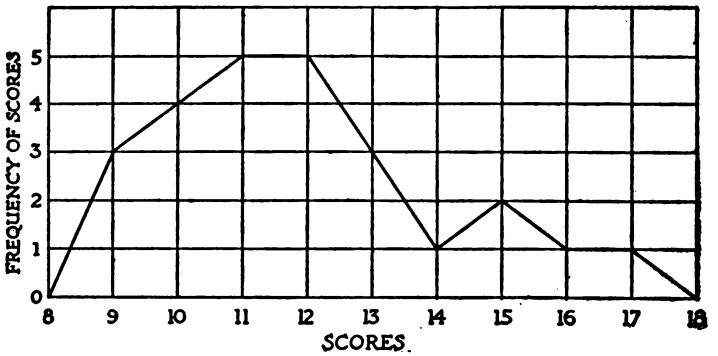
Score Designated X	Tally	Sum of Tallies Designated f	Score as a Deviation from an Arbitrary Origin (in this Case from $X=12.0$) Designated ξ	(Used in Calculation of the Mean) $f\xi$	(Used in Calculation of the Standard Deviation) $f\xi^2$
9	|||	3	−3	−9	27
10	||||	4	−2	−8	16
11	𝍸	5	−1	−5	5
				−22	
12	𝍸	5	0	0	0
13	|||	3	1	3	3
14	|	1	2	2	4
15	||	2	3	6	18
16	|	1	4	4	16
17	|	1	5	5	25
		25		20	114
				−2	
				$\frac{-2}{25}=-.08=M_\xi$	$\frac{114}{25}=4.56$
				12.00	$(-.08)^2=.0064$
				$M=11.92$	4.5536
					$\sqrt{4.5536}=2.134=\sigma_\xi$

2. The calculation of the arithmetic average. To obtain as reasonable an estimate as possible from these 25 scores we need a measure of central tendency — an average. There are two averages which are commonly used in situations such as this. They are the mean or arithmetic average (frequently but not quite accurately called " the average ") and the median. Either one is a good measure, though the mean is somewhat the more reliable and will be used here and is in general to be recommended. The calculation of the median is given in Section 12, and the steps in the calculation of the mean are given in detail in the first five columns of Table 18.

Generated on 2020-12-23 01:18 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHART 3

DISTRIBUTION OF SCORE MADE BY JOHN DOE ON 25 EQUALLY DIFFICULT SPELLING TESTS



The values in the fourth column are labeled ξ (the Greek letter xi). This is the symbol commonly used to indicate scores as deviations from some arbitrary point. Here the point 12 has been taken as the arbitrary origin, simply because it is near the middle of the distribution and using it leads to calculations involving small numbers. Any other point might have been taken to the same final conclusion, but the figures involved would be larger, as the reader can easily verify if he will calculate the mean, using, for example, 30 as an arbitrary origin. With 12 as the point from which deviations are measured, an original or gross score of 9 is represented by a ξ score of -3, a gross score of 10 by ξ equals -2, etc., as given in column 4. In column 5 are recorded the products of the values in the two preceding columns, and they are accordingly labeled $f\xi$. The sum of the values in this column, if we pay proper attention to the algebraic sign, is equal to -2, — that is, the sum of the deviations of the scores from the arbitrary origin is -2, — so that the average deviation, as given by Formula 8, is $\frac{-2}{25}$, or $-.08$. We shall

Generated on 2020-12-23 00:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

designate this by M_ξ, meaning thereby the mean of the series of measures expressed in ξ units :

$$M = \frac{\Sigma\xi}{N} \text{ (Mean in term of } \xi \text{ units)} \quad . \quad . \quad . \quad . \quad . \quad [8]$$

This informs us that the average of the 25 scores is .08 ξ units below the arbitrary origin. Accordingly (12.00 − .08), or 11.92, is the mean in terms of original or X scores. This mean is designated by the letter M if a single series of measures, as here, is being considered. If two series are under consideration, an X (or first) series and a Y (or second) series, the mean of the X's is designated by M_1 and that of the Y's by M_2.

This calculation can very easily be expressed in terms of symbols, as was very briefly explained in Chapter III. Let M equal the value of the mean. Let *Arb. Orig.* equal the value of the arbitrary origin. Let $\Sigma f\xi$ stand for the sum of the $f\xi$ products. The capital Greek letter Σ (sigma) stands for " the sum of " the magnitudes immediately following it. Let i represent the size of the interval in X covered by each unit interval in ξ. In this problem, when one passes from the ξ value of − 3 to a value of − 2, — that is, when one passes over a ξ interval of one unit, — it corresponds to passing from $X = 9$ to $X = 10$, which is one X unit; thus one unit in ξ corresponds to one unit in X, so that in our present problem $i = 1$. If X scores ran 15, 20, 25, 30, 35, . . . and corresponding ξ scores ran − 3, − 2, − 1, 0, 1 . . ., then corresponding to one ξ unit we would have five X units, and i would equal 5. Finally, let N equal the population or number of cases. With this notation the mean is given by the formula:

$$M = \textit{Arb. Orig.} + iM_\xi \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [2]$$

or

$$M = \textit{Arb. Orig.} + i\frac{\Sigma f\xi}{N} \text{ (The mean)} \quad . \quad . \quad . \quad . \quad [2]$$

Generated on 2020-12-23 00:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Utilizing this, we have:

$$M = 12.0 + 1\left(\frac{-2}{25}\right) = 11.92$$

The formula for the mean is very often written:

$$M = Arb.\ Orig. + i\frac{\Sigma\xi}{N} \quad . \ . \ . \ . \ . \ . \ . \quad [2]$$

for it is conventional to understand the same thing by $\Sigma\xi$'s as by $\Sigma f\xi$'s. This method of calculating the mean is called the method of moments, and as must be apparent, it is easier than the ordinary method, in that it involves working with smaller figures, and harder, in that some of the values are negative and algebraic signs must be carefully attended to. One further advantage is that it leads up to the last column in Table 18, which is used in calculating the standard deviation.

The mean of a series of measures is such a value that if deviations of the separate measures from it are listed and added algebraically, they will sum up to zero. Using x to designate measures when expressed as deviations from the mean, we have:

$x = X - M$ (Score as a deviation from the mean) . . [9]

and $\Sigma x = 0$ (The unique property of x scores) . . [10]

The score as a deviation from the mean enters into many formulas. Just to mention one, the standard deviation is defined by the equation:

$$\sigma = \sqrt{\frac{\Sigma x^2}{N}} \quad \text{(The standard deviation)} \quad . \ . \quad [11]$$

Though this constitutes the definition of σ, the standard deviation, Formula 11 is not convenient for computation purposes. Formula 13, given later, involves simple arithmetic computation. Let us here summarize the notation thus far used:

X is the raw or gross score; that is, the score just as yielded by the test.

Generated on 2020-12-23 00:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

M is the mean score for whatever group is worked with.
x is the score as a deviation from the mean.
ξ is the score as a deviation from some point other than the mean.
i is the number of X units corresponding to one ξ unit. It is usually equal to 1, but it must be kept in formulas for the occasional case when it equals some other value.
σ is the standard deviation as given by Formula 11 or 13 and is further defined in the next section.
Σ indicates an operation. It informs us that all measures immediately following it are to be added algebraically.

This notation is simple and well-nigh universal, so the reader would do well to fix it in mind.

Let us now consider more in detail the information given us by mean scores:

John's score on the first spelling test is	11.0
John's mean score on the first 2 tests is	10.5
" " " " " " 3 " "	12.33
" " " " " " 4 " "	11.5
" " " " " " 5 " "	11.8
. . .	
John's mean score on the first 25 tests is	11.92

At each successive step we obtain a more and more reliable estimate of his ability. This progression leads us to the concept "true ability." We shall define an individual's true test score as the average score that he would make if it were possible to test him with an infinite number of similar forms, and we shall designate such a true score of an individual by capital X with the subscript "infinity" — thus, X_∞. It is of course impossible experimentally to obtain a true score, for it is inconceivable that the conditions of the test could be kept constant throughout a long series. Probably John

Generated on 2020-12-23 00:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

is getting a little tired of spelling along toward the one-hundredth test, and when he contemplates the number still ahead, he may not do his best, and furthermore, he has been growing during this test process. No! experimentally we can never get a true score. We can, however, for the purposes of continuing the argument, postulate one. Let us say that John's true spelling score is 11.80 (i.e., $X_\infty = 11.80$). Thus the first test score, when taken as a measure of true ability, is in error by $-.8$ (i.e., $X - X_\infty = -.8$); the second score is an error by -1.8; the third, by 4.2, etc. We call these differences "errors of estimate." All 25 are recorded herewith: $-.8, -1.8, 4.2, -2.8, 1.2, .2, 3.2, -1.8, .2, -.8, -.8, 1.2, -2.8, 2.2, .2, 5.2, -1.8, -.8, -.8, -2.8, .2, -1.8, 3.2, 1.2, .2$.

Irrespective of sign, 12 of these errors are 1.8 or greater, and 13 are 1.2 or smaller. Since original scores are integral, $-9.0, 10.0, 11.0, \ldots,$ — these errors are grouped or bunched at certain values. If we allow somewhat for this in order to get as reasonable an answer as possible, we should find an error value somewhere in the neighborhood of 1.4, such that half of the errors, irrespective of sign, were greater than it and half less. Thus, any single score taken at random is as likely to differ from the true score by less than 1.4 as it is likely to differ from it by more than 1.4. This value 1.4 is called the median error. For any of the single scores chosen by chance we may write $X \pm 1.4$, meaning thereby that the chance that the obtained single score X differs from the true score by an amount less than 1.4 is equal to the chance that it differs by an amount greater than 1.4.

A great many distributions are approximately normal, and if normal, the median error may be calculated by multiplying the standard error (the standard deviation of the errors, as calculated in the next section) by .6745, and when so found it is called the probable error. The probable error

Generated on 2020-12-23 00:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and the median error are identical in a normal distribution, and for most distributions it is entirely safe to think of the probable error and the median error as having similar significance.

We obtained the value 1.4 by taking the deviations from the true score 11.80. Actually, of course, we cannot do this, as we do not know the true score, but we can secure a very good estimate of the value of 1.4 by finding the deviation of single scores from the mean of a number of single scores, as is done in the next section, or still more serviceably, as is done in Section 8.

3. The calculation of the standard deviation. In the last column of Table 18 (page 148) are recorded $f\xi^2$ values. These have been obtained by multiplying the values found in the two preceding columns. The sum of them, 114, is designated by $\Sigma f\xi^2$ or, as is conventional, by $\Sigma\xi^2$. With this notation the standard deviation, universally represented by the small Greek letter sigma, is given by Formula 12 in ξ units and by Formula 13 in original X units:

$$\sigma_\xi = \sqrt{\frac{\Sigma\xi^2}{N} - \left(\frac{\Sigma\xi}{N}\right)^2} \quad \text{(The standard deviation in terms of } \xi \text{ units)} \cdot \cdot \cdot \cdot \cdot \quad [12]$$

Continuing,

$$\sigma = i\sigma_\xi = i\sqrt{\frac{\Sigma\xi^2}{N} - \left(\frac{\Sigma\xi}{N}\right)^2} \quad \text{(The standard deviation)} \quad [13]$$

in which i, N, and $\Sigma\xi$'s, already defined, are, for this problem, equal to 1, 25, and -2, respectively. Substituting the proper values in this formula, we obtain, as shown in Table 18:

$$\sigma = 1.0\sqrt{\tfrac{114}{25} - (-\tfrac{2}{25})^2} = 2.134$$

This is a measure of the spread or scatter of the scores which compose the distribution, and thus it is a measure of the divergence of the scores from the mean of the distribution.

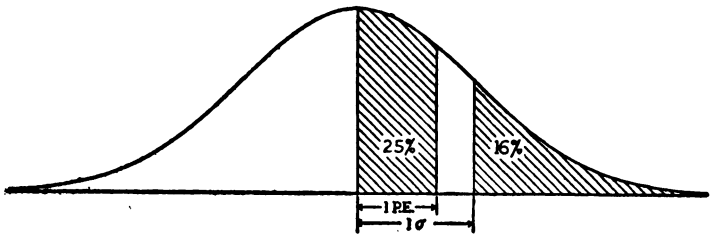
Generated on 2020-12-23 00:44 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

If our population is 25 or greater, the divergence of the measures from the mean of the distribution is very nearly equal to the divergence of the measures from the mean of a distribution of an infinite number of such measures; that is, from the true ability score. This is to say that the divergence of the measures from 11.92, the obtained mean of these 25 measures, is very nearly the same as the divergence from 11.80, the true score. Actually, the standard deviation, which is the square root of the average squared deviation from the mean, equals 2.134, and the square root of the average squared deviation from the true score, 11.80, equals 2.137, as may easily be found by calculation. In this instance we may very serviceably take 2.134, which we can calculate, in place of 2.137, which is unavailable because the true score is unavailable. Now it is known that this can very generally be done, so that we have a procedure for getting a very close estimate of the deviations of single scores from the true mean score, even though this true mean score is unknown.

The standard deviation has involved second-power terms and is a little difficult of interpretation, so it is common to interpret it in connection with a normal distribution, as pictured in Chart 4.

CHART 4

THE NORMAL DISTRIBUTION



Generated on 2020-12-23 00:44 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

4. The calculation and meaning of the probable error of a score. If we go away from the mean either above or below a distance of one σ, we shall find that the proportion of the area still left, the proportion shaded in the upper end, which area of course represents the proportion of the population lying still further up than one standard deviation from the mean, is .16. Thus 16 per cent diverge from the mean upward by more than one standard deviation, 16 per cent diverge downward by more than one standard deviation, and 68 per cent diverge from the mean by less than one standard deviation. These figures are sufficient for a quite accurate interpretation of the meaning of the standard deviation, but Gauss conceived the idea that it would be desirable to have a measure of divergence such that 50 per cent diverge from the mean by less than it and 50 per cent by more than it, and he therefore took just that fraction of the standard deviation which gives this result in a normal distribution. The required fraction is .6745. Gauss called this distance the probable error. Thus by definition the probable error is .6745 of the standard error. (It should be noted that "standard error" and "standard deviation" are identical in meaning — the former being employed when deviations are thought of as errors and the latter when they are not.)

$$\text{P.E.} = .6745\sigma \quad \text{(The probable error)} \quad . \quad . \quad . \quad [14]$$

Accordingly, for practical purposes a probable error may be thought of as being two thirds the size of the standard error. A few other interpretative figures may be given. The relationships of Table 19 hold strictly for the normal distribution and approximately for the majority of distributions:

Generated on 2020-12-23 00:44 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 19

If we go up from the mean 1 σ, the per cent of the population lying beyond is 16.

If we go up from the mean 2 σ, the per cent of the population lying beyond is 2.3.

If we go up from the mean 2 P.E.'s, the per cent of the population lying beyond is 9.

If we go up from the mean 3 P.E.'s, the per cent of the population lying beyond is 2.2.

Within the range $\pm$ P.E. (plus or minus one probable error) lies 50 per cent of the population.

Within the range $\pm$ 2 P.E. lies 82 per cent of the population.

Within the range $\pm$ 3 P.E. lies 95.6 per cent of the population.

Within the range $\pm \sigma$ lies 68 per cent of the population.

Within the range $\pm 2\sigma$ lies 95.4 per cent of the population.

There is only about one chance in twenty-two of a single score lying more than 3 P.E. or 2 σ away from the true score.

There is only about one chance in six of a single score lying more than 2 P.E. away from the true score.

There is only about one chance in three of a single score lying more than 1 σ away from the true score.

There is only about one chance in two of a single score lying more than 1 P.E. away from the true score.

There are about three chances in five of a single score lying more than $\frac{1}{2}\sigma$ away from the true score.

There are about three chances in four of a single score lying more than $\frac{1}{2}$ P.E. away from the true score.

There is much uncertainty among laymen as to the meaning of the probable error. Even in quarters where one would not expect it we find confusion. Thus we find in an otherwise very excellent achievement test study (Powers, 1924) a probable error of 3, with the statement, "This value indicates that the true score of the student on the test will not vary from the obtained score by more than 3 points." This is an egregious blunder, for the accurate statement is, "This value indicates that the true score of the student is as likely to vary from the obtained score by an amount greater than 3 as it

Generated on 2020-12-23 00:44 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

is likely to vary by an amount less than 3." Probably a still better wording, though not in fact differing in meaning, is, "This value indicates that the obtained score is as likely to vary from the individual's true score by an amount greater than 3 as it is likely to vary by an amount less than 3."

We have found a value of 2.134 for the standard error, and by multiplying by .6745, obtain 1.44 for the probable error. This is to be compared with the median error, as roughly determined before, to equal 1.40. The difference between these two values is here, as in general, negligible, so that we shall regularly calculate the probable error by taking .6745 times the standard error and interpret the result as being a median error.

The value 1.44 was calculated from a knowledge of the scores of a single person on 25 similar tests. An equally excellent, or even closer, approximation to the probable error may be obtained by giving two similar tests to 25 individuals. It is entirely feasible to test the members of a class twice, whereas it is generally not feasible to test a single pupil 25 times. We shall thus proceed to calculate the probable error of a score by first finding the correlation between scores on two similar tests. The mathematical analyses involved are too detailed to give here, but the mechanical steps are simple, and the resulting standard error and probable error when obtained are to be interpreted in exactly the same manner as in this present section.

5. Plotting a scatter diagram. The following are the scores received by 36 sixth-grade pupils on two forms of a paragraph-meaning test:

Generated on 2020-12-23 00:44 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 20

SCORES ON THE A B C PARAGRAPH-MEANING TEST

PUPIL	FORM 1	FORM 2
A	56	46
B	74	62
C	62	82
D	74	80
E	48	44
F	74	66
G	78	76
H	78	76
I	78	86
J	26	40
K	72	60
L	68	72
M	76	74
N	92	90
O	58	46
P	80	58
Q	64	54
R	66	70
S	84	78
T	80	60
U	38	40
V	64	74
W	70	62
X	70	58
Y	36	60
Z	68	66
AA	80	80
BB	52	56
CC	62	58
DD	64	58
EE	40	66
FF	92	80
GG	74	66
HH	78	82
II	34	16
JJ	60	64

Generated on 2020-12-23 00:45 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

We wish to draw up a scatter diagram or correlation table indicating for each pupil the two scores received. The correlation chart inserted at the back of this book is a convenient form to use for the calculation of a Pearson product-moment correlation coefficient.[1] There have been a number of forms put out to accomplish this, and any one of them will do equally well with the form here given, provided the arithmetical computations are accurately made. The authors of these various correlation charts, Otis, Ruger, Toops, Thurstone, make various claims as to their respective merits, and the writer also claims certain distinctive features. His chart is undoubtedly the longest of any of them, requiring 50 per cent more labor and time than any of the others. None of the other authors have as yet questioned his claim in this regard, and he therefore has unblushingly characterized his chart as the "long method of calculating r." There is one other unique claim which he makes, and that is that the procedure of his chart provides a more adequate guarantee of arithmetical accuracy than that of the other charts. All the basic quantities needed in the calculation of means, standard deviations, and correlation coefficient are computed in two independent ways, so that there is a check upon each of them. If his chart is as successful in maintaining right to this second claim as to the first, it will continue to serve a need. A third claim, not, however, in any sense unique, as all the correlation-chart makers enlarge upon this point, is that the steps involved are routine and capable of being performed in a mechanical and rule-of-thumb manner. That the steps are of this nature will be apparent as one follows them.

We shall first need to make the requisite entries. It will be noticed that there are 21 intervals in the X and Y scales.

[1] This chart, in packages of 25 or multiples thereof, may be purchased from World Book Company, Yonkers-on-Hudson, New York.

Generated on 2020-12-23 00:45 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

A quick perusal of Form 1 scores shows that the highest score is 92 and the lowest 26. Thus the range of Form 1 scores is $92 - 25$, or 67, and similarly the range covered by Form 2 scores is $90 - 15$, or 75. In order to represent a range of 67 in 21 intervals we must group into classes, having not less than 4 X-units in each class. We may use an interval of 4, 5, 6, . . . but not of 2 or 3. From the standpoint of accuracy it is best to use 4 as the interval, but from the standpoint of simplicity it is a little easier to work with an interval of 5. In the case of Form 2 scores we also must choose an interval of 4 or greater in order to represent a range of 75 in 21 intervals. There is no necessity that the same grouping interval be used for the two forms. An interval of 4 or one of 5 would be quite satisfactory in each case, but for illustrative purposes we shall choose a range of 4 for the interval in Form 1 scores and a range of 5 for the interval in Form 2 scores. If the group has an even number of units per interval, it does not matter much how the interval runs. The interval 26, 27, 28, 29; or 25, 26, 27, 28; or 24, 25, 26, 27; or 23, 24, 25, 26, could be made the first. However, it is desirable to follow some uniform procedure. It is accordingly advised that the intervals be such that the first number in each interval be divisible by the size of the interval. Thus, when grouping in 4's our first interval will be 24, 25, 26, 27, because 24 is divisible by 4. But if there is an odd number of units per interval, it simplifies the procedure somewhat if the middle of each interval is made divisible by the grouping unit. Thus 13, 14, 15, 16, 17 is to be preferred over 14, 15, 16, 17, 18 or 15, 16, 17, 18, 19 or any other arrangement for the first interval.

The X intervals are to be written in by the user in the space provided at the bottom of the chart. It is desirable to do all the work on the chart in red ink, so that there will be no confusion between printed figures and recorded values.

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

In the chart to be found at the back of the book all the figures printed in red correspond to entries and computations made by the worker, while all the rest is part of the printed chart. Small letters in red are simply for reference in the explanation herewith given of the chart. We shall choose for the first X interval 24, 25, 26, 27. These four values might be recorded in the appropriate compartment just above the arrow X, but it will be equally serviceable simply to record the 24 and the 27, understanding of course that 25 and 26 also fall in this interval. We may begin with any compartment, provided only that it gives us room enough at the upper, or right-hand, end of the X scale for the last interval, which in our case is the interval 92–95. Thus we may begin with the first, second, third, or fourth compartment, but not with the fifth. We shall begin with the second, as that will center our values somewhere near the middle of the table, which is convenient. We next label the Y axis. Since there are an odd number (in this case 5) of values entering into each class, a class is completely defined if we simply record its mid value. We therefore designate the classes by 15, 20, 25, etc. The class 15 of course includes all values from 13 to 17 inclusive, the class 20 all values from 18 to 22 inclusive, etc.

Having labeled the X and Y axes, we are to place a tally for each pupil in the appropriate place in the table. Let us do so for Pupil A, who made a score of 56 on Form 1 and 46 on Form 2. The score 56 falls in the interval labeled at the bottom of the sheet 56–59, and 46 falls in the interval labeled 45 on the left margin. Accordingly in the cell which is at the intersection of the 56–59 column and the 45 row we have recorded a tally. This tally is one of the two in the compartment in which occurs the letter *a*. Similar tallies are made for all the other pupils. To facilitate recording these tallies it is convenient to copy the X scale on a sepa-

Generated on 2020-12-23 00:45 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

rate slip which can be moved up and down the chart as required. This saves the care and labor of each time moving the eye from the bottom of the page up the column to the required point. Having recorded these 36 tallies, we are provided with a "scatter diagram," which is the basic table from which the correlation coefficient is computed.

6. The calculation of a product-moment correlation coefficient. For the use of this chart it is not necessary to know the meaning of the symbols employed, but as these meanings are simple and as it gives one a certain confidence in the mechanical operations to know that the symbols stand for very concrete things, they are given herewith. In Section 2 of this chapter, X, x, and ξ are defined. If we have a second variable, we need additional symbols. Exactly corresponding to X, x, and ξ for the first variable we have Y, y, and ζ (zeta) for the second. The symbol d in the chart stands for "difference" and is equal to $\xi - \zeta$, and the symbol s stands for "sum" and is equal to $\xi + \zeta$. The symbols M_ξ, σ_ξ, M_x, and σ_x stand for the mean in the case of the first variable in ξ units, the standard deviation of the first variable in ξ units, the mean of the first variable in X units, and the standard deviation of the first variable in X units, respectively; and M_ξ, σ_ξ, M_y and σ_y have corresponding meanings for the second variable. The symbol M_x is used in place of the more accurate M_X to designate the mean in original test units, simply because x is more readily printed as a subscript than X, and no ambiguity ever arises from so doing. It is also very common to designate the X variable by the subscript 1 and the Y variable by the subscript 2, so that M_1, σ_1, M_2, σ_2 mean the same thing as M_x, σ_x, M_y, σ_y, respectively.

The tallies are added by row and recorded in the column for frequencies at the right, headed f. After recording, this column from top to bottom runs 1, 1, 6, 4, . . . and totals 36. Next to this f column is a column of ζ's and next to it

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

provision for $f\zeta$'s. The f and the adjacent ζ value are multiplied and the answer recorded in the $f\zeta$ column, giving, as shown, 7, 6, 30, . . . Part of these products are positive and part negative. The positive ones total 84 and the negative − 20, and both together yield 64, as recorded at the bottom of the column. The next column provides space for $f\zeta^2$ values. The product of ζ and $f\zeta$ yields $f\zeta^2$, so the product of the two columns preceding the $f\zeta^2$ column yields the desired magnitudes 49, 36, 150, . . . These values are all positive and total 440.

Exactly similar calculations are made dealing with columns instead of rows, yielding the values in the rows headed f, $f\xi$, and $f\xi^2$ at the top of the table. The first check on the accuracy of the work appears here, in that N, the total of the frequencies in the column headed f, should equal the total of the frequencies in the row headed f — in our problem, 36 in each case.

Having 64, which is the $\Sigma\zeta$, we divide by 36, or N, and obtain 1.7778, which is recorded on the far right-hand side of the sheet under the word "Checks" and opposite "$M_\zeta = \frac{\Sigma\zeta}{N} =$," as indicated. This is the value of M_ζ and will shortly be computed in a second and entirely independent way, so as to provide a check on the arithmetic.

It is of course absolutely essential that the sign of M_ζ be recorded. The writer would apologize for so bromidic an observation had not experience shown him that certain computers seem not to concern themselves with this little matter. Similarly, we divide 44, which is the $\Sigma\xi$, by 36 and obtain 1.2222, and record it opposite M_ξ. We also record 440, the $\Sigma\zeta^2$'s, and 626, the $\Sigma\xi^2$'s, in the spaces provided under "Checks." These two values are added and the sum, 1066, placed opposite "$\Sigma\xi^2 + \Sigma\zeta^2$," as shown.

We shall now compute the $\Sigma\xi\zeta$'s, which will complete the

Generated on 2020-12-23 00:46 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

magnitudes required for the computation of r. We shall then proceed to obtain all of them a second way for the purpose of verification. In the lower right-hand corner of each cell of the table is recorded a value which is the product of the ξ values and the ζ values for the column and row at the intersection of which the cell lies. Thus the cell in which the letter *b* is found is in the column having a ξ value of 8 and in the row having a ζ value of 7. The product of 8 and 7 gives the 56 which is recorded in the lower right-hand corner of this cell. Accordingly, 56 is the $\xi\zeta$ product for all the frequencies lying in this cell. There is but a single tally in this cell, so we have $1 \times 56 = 56$ as the product. This 56 is recorded in the upper right-hand corner in the column headed " $+\times+$ " (plus times plus), and similar values are recorded for each other cell in which frequencies are found.

If straight lines are imagined across the row and down the column having zeros in the lower right-hand corners of the cells, the table is divided into four quadrants. For all cells in the upper right-hand quadrant we have positive ξ values and positive ζ values, and thus $\xi\zeta$ products from the cells of this quadrant are termed " $+\times+$ " values and recorded in the " $+\times+$ " column. Similarly values from the cells in the lower left-hand quadrant, for which ξ is minus and ζ is minus, are recorded in the " $-\times-$ " column. Those in the upper left-hand quadrant are recorded in the " $-\times+$ " column and those in the lower right-hand quadrant in the " $+\times-$ " column. Since shortly the $+\times+$ and the $-\times-$ products, which of course are both algebraically positive, are added together, it really does not matter whether or not the $+\times+$ and the $-\times-$ products are kept separate. Likewise the $+\times-$ and the $-\times+$ products may be recorded in the same column if desired. The rule to follow is simply, for every cell, to multiply the cell frequency by the value in the lower right-hand corner of the cell and, keeping the plus

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and minus products separate, record in the ξ' columns. The algebraic sum of these products, 403, designated the $\Sigma\xi'$'s, is copied opposite $\Sigma\xi'$ under "Checks," thus completing the basic constants needed, but we shall not proceed to a calculation of r until we have secured checks for all of these constants.

On the extreme right of the correlation chart is a "Guide to use in summing frequencies along diagonals." This may be cut out and all the windows shaded in the drawing cut out. It can be used in this shape, but it is advantageous to mount it on a light-weight cardboard. If we place this guide at an angle of 45° with the bottom line of the chart, so that it extends from the lower left up toward the upper right (LL–UR) and so that the two cells labeled c are visible through two of the windows, we shall find the value -7 appearing in one of the windows of the guide just beyond the top row of the table. This -7 is the value of d. When the guide is so placed, we are to add all the frequencies appearing in the windows and record the total, 2, in the column headed f immediately next to the column headed d, as shown and indicated by the letter e. The guide may now be slipped to the right a distance of one cell, so that -6 appears in the window at the top of the chart. The sum of the frequencies now visible is 1. This 1 is recorded in the f column opposite the d value -6. The guide is slipped to the right again and the process repeated. Soon the longest diagonal is reached and the frequency 6 found and recorded opposite the d value 0. The guide is then slipped one space farther to the right. There now appears a 1 in a window just below the bottom margin of the chart. The sum of the frequencies now visible is 10, and this number is recorded opposite 1 in the d column. The guide is slipped to the right again and the process repeated until all frequencies are represented in the f column next to the d column. The frequencies in the f column now read 2, 1, 1, blank, 2, 5, . . .

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Turning the guide at right angles, so that it extends from the upper left to the lower right (UL–LR), we proceed just as before, except that now we have s values and they appear in windows just beyond the left and right instead of the top and bottom margins. For illustration, let the guide be so placed from upper left to lower right as to expose the cell lettered *b*. The s value appearing in the window just beyond the right margin is 15. The sum of all the frequencies appearing is 1, and this is recorded, as shown by the letter g, in the f column just opposite the s value 15. The guide is moved one space to the left and the operation repeated, etc., until all the frequencies are represented in the f column next to column s. They run 1 (opposite $s = -15$), blank, blank, 1, blank, blank, 1, . . .

The entries in the fd, fd^2, fs, and fs^2 columns are readily calculated as shown, giving algebraic totals as follows: $\Sigma d = -20$; $\Sigma d^2 = 260$; Σs 108; and $\Sigma s^2 = 1872$. Adding the Σd's and the Σs's and dividing by $2N$, $\left(\frac{-20+108}{72}\right)$, we obtain 1.2222, which we record opposite $\frac{\text{“}\Sigma s + \Sigma d\text{”}}{2N}$ under “Checks,” and note with the proper sense of satisfaction that this agrees with M_t as derived in the other manner. Also we subtract algebraically the Σd's from the Σs's, divide by $2N$, $\left[\frac{108-(-20)}{72}\right]$, and obtain 1.7778 and record opposite $\frac{\text{“}\Sigma s - \Sigma d\text{”}}{2N}$ and again note that we obtain a check. Continuing, we add Σs^2 and Σd^2 and divide by 2, $\left[\frac{1872+260}{2}\right]$, obtain 1066, record in the appropriate place, and note that it checks. Finally, we subtract the Σd^2's from the Σs^2's and divide by 4, $\left[\frac{1872-260}{4}\right]$, obtain 403, record, and note that it checks.

Generated on 2020-12-23 00:46 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

This last check is something of an extravagance, for if $\Sigma\xi^2 + \Sigma\zeta^2$ has been found to equal $(\Sigma s^2 + \Sigma d^2)/2$, then it is a very small demand to place upon a computer to ask that he correctly calculate $(\Sigma s^2 - \Sigma d^2)/4$. If one trusts himself to do this correctly, the first calculation of $\Sigma\xi\zeta$ (that made in the upper right-hand corner of the page) may be forgone.

These checks assure us that all of the basic constants are numerically correct. From here on there is no check, so that the subsequent work involving division and extraction of square roots should be done twice very carefully.

Formula 12 provides us with the standard deviation in ξ units. Accordingly we find σ_ξ by dividing the $\Sigma\xi^2$ by N, subtracting M^2_ξ, and extracting the square root, thus:

$$\sigma_\xi = \sqrt{\tfrac{838}{36} - (1.2222)^2} = 3.987$$

Similarly, $$\sigma_\zeta = \sqrt{\tfrac{440}{36} - (1.7778)^2} = 3.010$$

The formula for the correlation coefficient, when calculated by means of deviations from arbitrary origins, — that is, when ξ and ζ measures are used, — is:

$$r = \frac{\frac{\Sigma\xi\zeta}{N} - M_\xi M_\zeta}{\sigma_\xi \sigma_\zeta}$$ (Product-moment correlation coefficient) . . . [15]

Thus we have:

$$r = \frac{\frac{402}{36} - (1.2222)(1.7778)}{(3.987)(3.010)} = .752$$

This is the reliability coefficient as determined by these particular 36 sixth-grade pupils. We shall shortly utilize this value in obtaining the probable error of a single individual score, but let us first calculate σ_x, σ_y, M_x, and M_y. All these values are readily obtained knowing σ_ξ, σ_ζ, M_ξ, and M_ζ.

The first interval for Form 1 includes scores 24, 25, 26, and 27, so that the midpoint of this interval is 25.5. The

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

ξ value corresponding to this midpoint, as shown on the chart, is -9. Thus an X value of 25.5 corresponds to a ξ value of -9. For the next interval we have X values 28, 29, 30, and 31, with the midpoint 29.5. Thus an X value of 29.5 corresponds to a ξ value of -8. Copying these and certain other data found in the same manner, we obtain:

TABLE 21

X Value		Corresponding ξ Value	
Difference of 4 units	25.5	-9	Difference of 1 unit
	29.5	-8	
	33.5	-7	
	. . .	. . .	
	57.5	-1	
	61.5	0	
	65.5	1	
	69.5	2	
	. . .	. . .	
	93.5	8	

Since 1 unit in ξ corresponds to 4 units in X, we have $i_x = 4$, as recorded on the right of arrow X on the chart. Similarly, $i_y = 5$, as recorded. We shall now use these values.

7. Expressing means and standard deviations in original test units. The standard deviation is a measure of variability or spread, and we immediately see that a certain number of units' variability in ξ implies four times as many units' variability in X. We have:

$\sigma_x = i_x \sigma_\xi$ (The standard deviation in X units, knowing it in ξ units) . . . . . [See Formula 13]

and thus for this problem,

$$\sigma_x = 4(3.987) = 15.948$$

which, if published, should be written, as will be explained later,

$$\sigma_x = 15.9$$

Generated on 2020-12-23 00:47 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

in order not to suggest an unwarranted degree of accuracy in the answer. The value 15.948 is written in the lower right-hand corner of the chart, opposite "$\sigma_x =$."

Similarly, we have $\sigma_y = i_y \sigma_\zeta$, so that:

$$\sigma_y = 5\,\sigma_\zeta = 5(3.010) = 15.050$$

which is to be kept to one decimal place, thus,

$$\sigma_y = 15.0$$

Formula 2, of Chapter III, may be written:

M_x = Arb. Orig.$_x$ + $i_x M_\xi$ . . . . . [See Formula 2]

or again as:

M_y = Arb. Orig.$_y$ + $i_y M_\zeta$ . . . . . [See Formula 2]

The mean in ξ units is 1.2222. By reference to Table 21 we see that $\xi = 0$ corresponds to $X = 61.5$. The mean is 1.2222 units above (since the sign of 1.2222 is plus) zero, which is 4 times 1.2222, or 4.8888 X units above 61.5. Thus by Formula 2

$$M_x = 61.5 + 4(1.2222) = 66.3888$$

This answer is to be recorded opposite "M_x" in the lower right-hand corner of the chart. Similar determination of M_y gives 63.889. This completes all the calculations of the chart except the calculation of the probable errors of r, M_x, M_y, σ_x, and σ_y, which are explained later. These probable errors are the recorded $\pm$'values appearing immediately after these five constants. To one familiar with the meaning of the probable error it is apparent that the answers are not accurate to the number of places to which the work has been carried. The rule to follow in determining the number of figures which should be published is: *Keep to place indicated by the first figure of one half the probable error.* (Kelley, 1924.)

When we divide each of these probable errors by 2, we obtain .024, .90, .85, .64, .60. Accordingly the correlation

Generated on 2020-12-23 00:47 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

coefficient should be kept to the second decimal place and each of the other constants to the first decimal place only. Thus, if our results are to be published, we should write:

The reliability coefficient: $r = .75 \pm .048$

The mean score on Form 1: $M_x = 66.4 \pm 1.8$

The mean score on Form 2: $M_y = 63.9 \pm 1.7$

The standard deviation of Form 1 scores: $\sigma_x = 15.9 \pm 1.3$

The standard deviation of Form 2 scores: $\sigma_y = 15.0 \pm 1.2$

In recording probable errors, two significant figures are always sufficient.

8. The probable error of a score via the reliability coefficient. We have used X to designate the pupil's raw test score and x his score as a deviation from the mean of his group. If we had the scores of a pupil on an infinite number of similar forms and averaged them, we would have his true score. We shall represent this true gross score by the symbol X_∞, and we shall represent his true score as a deviation from his group mean by x_∞. If we take the single score X as evidence of the true score, then $(X - X_\infty)$ is the error involved in the process. It is easily shown that $(x - x_\infty)$ is equal to $(X - X_\infty)$, so $(x - x_\infty)$ is also the error of estimate. In Section 4 of this chapter we found an approximate answer for the standard deviation of such errors of estimate, and now we shall consider a second method leading more briefly to the same result. Let $\sigma_{1\cdot\infty}$ (read, "the standard deviation of single scores for a given fixed value of the true score") represent this standard deviation. We have:

$$\sigma_{1\cdot\infty} = \sigma_1\sqrt{1 - r_{11}}$$ (The standard deviation of errors of estimate when the single score is taken as evidence of the true score) . . . [16]

The correlation coefficient used in this formula is a "reliability coefficient," or correlation coefficient between two similar forms of the same test. To indicate that a certain correlation is that between similar forms, the first subscript

Generated on 2020-12-23 00:47 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of the r is in arabic or lower case and the second subscript in roman or capital type, thus:

r_{1I} is the correlation between a first and second similar form of Test 1.

r_{2II} is the correlation between a first and second similar form of Test 2.

r_{aA} is the correlation between a first and second similar form of Test A.

The probable error is, of course:

$$\text{P.E.}_{1\cdot\infty} = .6745\,\sigma_{1\cdot\infty} = .6745\,\sigma_1\sqrt{1 - r_{1I}}\,. \quad . \quad [17]$$

The standard deviation, σ_1,— identical with σ_x,— is, for this paragraph-meaning test, equal to 15.9, but of course it was just a matter of chance which form was called the first and which the second. Therefore it is well to use for σ_1 the average of the two standard deviations.

$$\sigma_1 = \frac{\sigma_x + \sigma_y}{2} = \frac{15.95 + 15.05}{2} = 15.50$$

Thus, $$\sigma_{1\cdot\infty} = 15.50\sqrt{1 - .752} = 7.72$$

and for the probable error we have:

$$\text{P.E.}_{1\cdot\infty} = (.6745)(7.72) = 5.21$$

These values would be published as

$$\sigma_{1\cdot\infty} = 7.7 \text{ and } \text{P.E.}_{1\cdot\infty} = 5.2$$

The standard error, and consequently the probable error, of a score is a sort of average for the table entire, and thus either may be applied to the score of any individual. Suppose John Doe has a score of 70 on the ABC Paragraph-Meaning Test. How much credence should we place in this score? If we go one standard deviation up we obtain 77.7, and reference to Table 19 shows that there is one chance in six that John's true ability lies above this. There is likewise one chance in six that it lies below 62.3, so that there are four chances in six that it lies between 62.3 and 77.7.

Let us draw, free-hand, a normal curve approximately to

Generated on 2020-12-23 00:48 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

represent this situation. We shall first, according to any convenient scale, record score values along the abscissa or horizontal axis, 50, 51, 52, . . . 90. The middle of our normal curve is to be at 70, John's obtained score, and the height at this point may be made any convenient height. One standard deviation down and up the abscissa scale brings us to 62.3 and 77.7, respectively, and the height of the curve at these points is $\frac{3}{5}$ of the height at 70. The height of the curve at $2\frac{1}{2}$ standard deviations up and down from the mean is practically zero, it being in fact only $\frac{1}{23}$ as high as at the mean; thus we multiply 7.7 by $2\frac{1}{2}$, obtain 19.25, subtract and add this to the mean, and obtain points 50.75 and 89.25. The normal curve to be drawn is thus practically to come down to the base line at $X = 51$ and $X = 89$; to have the height chosen as convenient at the point $X = 70$; and to be $\frac{3}{5}$ as high as this at $X = 62.3$ and $X = 77.7$. With these five points a smooth curve may be drawn free-hand. We now have a graphic aid enabling us to make any sort of judgment desired as to the likelihood of John's true ability being above or below a certain point. If we decide to place pupils of true ability above 60 in one group and those below in another, and if we place John in the upper group, the area under our curve above the point 60 as a fraction of the total area states the chance of our classification being correct. If the reader has drawn a curve according to directions, he will note that approximately 10 per cent of the area falls below 60 and 90 per cent above, so that there are some nine chances in ten that John belongs in the upper group and one chance in ten that he belongs in the lower. Any other point than 60 may be chosen and the chance of correct classification determined in a similar manner. We have used the standard error of the score as our basis for figuring the chance of correct classification. It is fully as common to use the probable error.

$$\text{P.E. of John's score} = .6745(7.7) = 5.2$$

Generated on 2020-12-23 00:48 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Thus there are twenty-five chances in one hundred that John's true ability lies above 75.2, twenty-five in one hundred that it lies below 64.8, and fifty in one hundred that it lies between 64.8 and 75.2. If the reader has made a drawing as directed, these same facts are of course revealed by an examination of it.

With a standard error of a score, $\sigma_{1\cdot\infty}$, which is one half as large as the standard deviation of the group, σ_1, the reader will probably feel that there is a great deal of uncertainty in classifying John on the basis of his test score. This is true, but the ABC Paragraph-Meaning Test was found to have a reliability for the single grade of .75, which is about as high as the majority of educational tests. In plain language, classification upon the basis of the majority of these tests does involve much error and uncertainty. It should not be done except tentatively and with the expectation that the need of changes in classification will soon become apparent and with the opportunity for making such changes an integral part of the administrative machinery.

It is of first importance that the teacher who interprets test scores and classifies pupils should know the error of his technique. Thus, if he classifies on the basis of a test score, he needs to know the standard error of the individual score, $\sigma_{1\cdot\infty}$, or the probable error, $\text{P.E.}_{1\cdot\infty}$. For many of the better tests $\sigma_{1\cdot\infty}$, or $\text{P.E.}_{1\cdot\infty}$, is given by the authors. For others the data from which it may be derived are given; namely, the reliability coefficient and the standard deviation of the group from which the coefficient was calculated. Having these, Formula 16 gives us $\sigma_{1\cdot\infty}$. For a still larger number of tests the reliability coefficient alone is given. This is of little service unless the range or spread of talent for the group from which the coefficient was obtained is also given. Ordinarily, classification problems involve segregating the members of a single class. If promotions are made yearly,

Generated on 2020-12-23 00:48 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

this is a one-grade range of talent. If there are semiannual promotions, we have a one-half-grade range of talent. The distribution of true ability in reading, arithmetic, spelling, etc., in a one-half-grade range is ordinarily almost as great as in an entire grade range, so that no great error is introduced if the standard deviation for a grade range is used also in a system having semiannual promotions. In this text, wherever data permit, the reliability coefficients reported are those obtained in an average one-grade range of talent. The reader will therefore understand that if a reliability coefficient is given in this text without qualification as to the range of talent covered, it is to be assumed to be a one-grade range.

If the teacher has available the reliability coefficients for such a range, and if he calculates the standard deviation of scores for his own grade, he may use this standard deviation, together with the reported reliability coefficient, and secure a standard error of estimate. For example, suppose Miss Black gives the DEF Reading Test to her fifth grade and calculates the standard deviation as was done in Section 3 of this chapter and finds it to equal 3.8, and suppose the entirely trustworthy and capable author of the DEF Reading Test gives the reliability coefficient for a single grade range as being equal to .70. Miss Black may assume that the standard deviation of her class is approximately equal to that of the class used in deriving the correlation coefficient, and thus write:

$$\sigma_{1\cdot\infty} = 3.8\sqrt{1 - .70} = 2.1$$

and

$$\text{P.E.}_{1\cdot\infty} = .6745\,\sigma_{1\cdot\infty} = 1.4$$

obtaining thereby a serviceable estimate of the error of an individual score. The necessity for knowing this is so great, if interpretation is to be sane, that the best possible estimate of it should always be sought. If no reliability coefficients

Generated on 2020-12-23 00:48 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

or standard or probable errors are reported by the author of the test used, a careful examination of the items of the test and a comparison with other somewhat similar tests having known reliabilities will enable one to make an estimate of the reliability and derive an approximate $\sigma_{1\cdot\infty}$. Though this procedure will not give very accurate results, it is always preferable to leaving the matter unconsidered.

9. The probable error under various conditions. There are two ways of bettering the unsatisfactory situation of attempting to classify pupils by means of their raw scores upon a test of low reliability. The first of these is to work with more reliable tests, and if a more reliable, equally valid test is available, this solves the problem. The second way is to use an improved technique of interpretation. No improvement in interpretation can make a genuinely poor test give excellent results, but the technique described in the last section is not the best possible, so that better results than by it are always available if one uses the procedure described in this section. The difference in procedures is very easily explained. In the last section the pupil's score as a deviation from the mean, x, was taken as an indication of his true score, x_∞. It can be easily shown statistically that, in general, a better estimate of the true score is obtained if one takes $r_{11}\,x$ instead of x as the estimate of it. In the last section $(x - x_\infty)$ was the error of estimate, and this led to a standard error of estimate, $\sigma_{1\cdot\infty}$, which was equal to $\sigma_1\sqrt{1 - r_{11}}$. In this section $(r_{11}x - x_\infty)$ is the error of estimate, and the standard error of estimate is given by Formula 18:

$$\sigma_{\infty\cdot 1} = \sigma_1\sqrt{r_{11} - r_{11}^2}$$

(Standard error of estimate of a regressed score which is taken as evidence of the true score) . . . . . [18]

For the probable error we have:

$$\text{P.E.}_{\infty\cdot 1} = .6745\ \sigma_{\infty\cdot 1} = .6745\ \sigma_1\sqrt{r_1 - r_{11}^2} \quad . \quad [19]$$

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Since $\sqrt{r_{1I} - r_{1I}^2}$ equals $\sqrt{r_{1I}}\sqrt{1 - r_{1I}}$, and since the reliability coefficient always has values between 0 and 1, it is obvious that $\sqrt{r_{1I} - r_{1I}^2}$ is less than $\sqrt{1 - r_{1I}}$ and that therefore $\sigma_{\infty \cdot 1}$ is always less than $\sigma_{1 \cdot \infty}$. Accordingly, by the use of this second technique, we shall always have smaller probable errors of estimate than by the use of the first.

Let us write:

$$\bar{x}_\infty = r_{1I}x \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [20]$$

in which the superior bar indicates an estimated value, so that this equation is read, "The estimated true score as a deviation from the mean is equal to the reliability coefficient times the obtained score as a deviation from the mean." If x is 10 units above the mean and $r_{1I} = .6$, then $\bar{x}_\infty$ is 6 units above the mean; and if x is 10 units below the mean, then $\bar{x}_\infty$ is 6 units below the mean. This tendency of the estimated true score to lie closer to the mean than the obtained score is called the principle of regression. It was first discovered by Francis Galton and is a universal phenomenon in correlated data. We may now characterize the procedure of the last and present sections by saying that in the last section regression was not allowed for and in the present it is. If the reliability is very high, then there is little difference between x and $\bar{x}_\infty$, so that this second technique, which is slightly the more laborious, is not demanded, but if the reliability is low, there is much difference in individual outcome, and the refined procedure is always to be used in making individual diagnoses. Roughly, we may consider that individual placement according to the first procedure is excellent if the reliability is .95 (an equivalent excellence is obtained by the second procedure, with a reliability of .947); that it is fair if the reliability is .90 (by second procedure, .887); poor but an improvement over the judgments of single teachers if the reliability is .80 (by second procedure, .72); very poor but about comparable

Generated on 2020-12-23 00:49 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

to a careful teacher's judgment if the reliability is .75 (by second procedure, .50); and so poor as not to be used unless no other means, such as teacher's judgments, are available if the reliability is below .75 (by second procedure, below .50).

We have let $\bar{x}_\infty$ stand for the estimated true score as a deviation from the mean of a group. If $\bar{X}_\infty$ stands for the estimated true raw score, it is directly obtained from a knowledge of $\bar{x}_\infty$ by simply adding the mean for the group, thus:

$$\bar{X}_\infty = \bar{x}_\infty + M \quad \text{(Estimated true raw score)} \quad . \quad . \quad . \quad [21]$$

Since $\bar{x}_\infty = r_{11}x$ and since $x = X - M$, we may substitute and obtain:

$$\bar{X}_\infty = r_{11}X + (1 - r_{11})M \quad \text{(Regression of estimated true score upon raw score)} \quad . \quad . \quad . \quad [22]$$

This is a very simple equation to use. Thus, if Mary Doe's ABC Paragraph-Meaning Test score on the first form is 90 and if the grade mean is 66, the grade standard deviation 15.9, and the reliability .75, we have:

$$\bar{X}_\infty = .75(90) + .25(66) = 84$$

If we take the 84 as our estimate of Mary's true ability, the probable error of our estimate is given by Formula 19, thus:

$$\text{P.E.}_{1\cdot\infty} = .6745(15.9)\sqrt{.75 - (75)^2} = 4.64$$

We thus have 84 ± 4.6 — that is, an estimated true ability of 84, with a probable error of estimate of 4.6 — instead of 90 ± 5.4, as given by the procedure of the preceding section.

In estimating true scores by Formula 22 a table may be built up, thus obviating the necessity of calculation for each pupil. For the ABC Paragraph-Meaning Test such a table may be computed from the equation

$$\bar{X}_\infty = .75\ X + (1 - .75)66.4$$

as given in part in Table 22.

Generated on 2020-12-23 00:49 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 22

Score on Test	Estimated True Score
X	$\bar{X}_\infty$
50	54
51	55
52	56
53	56
54	57
. . .	. . .
63	64
64	65
65	65
66	66
67	67
68	68
69	68
70	69
. . .	. . .
89	83
90	84
91	85
92	86
93	86

This table is rapidly calculated, for the difference between successive values of $\bar{X}_\infty$ is constant and equal to r_{11}. Thus, after calculating the initial fact that corresponding to $X = 50$ we have $\bar{X}_\infty = 54.1$, successive values are obtained by simply adding .75 to each preceding value. We shall follow the rule of Section 7 of this chapter and keep the equivalent score to the nearest integer only, for the one half P.E. of the $\bar{X}_\infty$ score is equal to 2.3.

Certain other errors of estimate are reported in the literature, and in order that there may be no confusion in regard to their meaning, four very common ones are listed here. As before, let r_{11} stand for the reliability coefficient, or the

Generated on 2020-12-23 00:49 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

correlation between X_1 (the score on Form 1) and X_I (the score on a second similar form of the same test), and let σ_1 be the standard deviation of scores on Form 1, or σ_1 may equal the average of the standard deviations on Forms 1 and 2, if both are known.

(1) If x_1 is taken as evidence of x_1, then $(x_1 - x_1)$ is the error of estimate, and the standard deviation of such errors $= \sigma_1\sqrt{2 - 2r_{1I}}$ . . . . . [23]

(2) If $r_{1I}x_1$ is taken as evidence of x_1, then $(x_I - r_{1I}x_1)$ is the error of estimate, and the standard deviation is designated by the symbol $\sigma_{1\cdot I}$ and is given by Formula 24:

$$\sigma_{1\cdot I} = \sigma_1\sqrt{1 - r_{1I}^2}$$

(This procedure is a refinement upon the preceding, in that it allows for regression) . . . . . [24]

(3) If x_1 is taken as evidence of x_∞, then $(x_\infty - x_1)$ is the error of estimate, and the standard deviation of such errors is designated by the symbol $\sigma_{1\cdot\infty}$ and is given by Formula 16:

$$\sigma_{1\cdot\infty} = \sigma_I\sqrt{1 - r_{1I}} \quad . \quad . \quad . \quad . \quad . \quad [16]$$

(4) If $r_{1I}x_1$ is taken as evidence of x_∞, then $(x_\infty - r_{1I}x_1)$ is the error of estimate, and the standard deviation of such errors is designated by the symbol $\sigma_{\infty\cdot 1}$ and is given by Formula 18:

$$\sigma_{\infty\cdot 1} = \sigma_1\sqrt{r_{1I} - r_{1I}^2}$$

(This procedure is a refinement upon the preceding, in that it allows for regression) . . . . . [18]

Each of these is a standard error of estimate. There is no question as to which is "the" right one, for it simply depends upon which process has been followed, as each is right in its proper setting. To secure an estimate of true ability, the fourth process is in all cases the best, but if reliability is high, it is not sufficiently better than the third

Generated on 2020-12-23 00:49 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

to warrant the extra labor. The commonest process has been and will probably continue to be the third, so that $\sigma_1\sqrt{1 - r_{11}}$ will ordinarily be the proper value for the standard error of estimate. We may note that if the first standard error is divided by $\sqrt{2}$, we obtain the third. This way of obtaining the third is common and is found in the literature in connection with standard and probable errors in some one of the three following forms:

$$\text{Standard error of estimate} = \frac{\text{Standard deviation of differences between scores on two similar forms}}{\sqrt{2}}$$

(A second way of writing the relationship given by Formula 16) . . [25]

$$\text{Probable error of estimate} = \frac{.6745\left(\text{Standard deviation of differences between scores on two similar forms}\right)}{\sqrt{2}}$$

(A second way of writing the relationship given by Formula 17) . . [26]

$$\text{Probable error of estimate} = \frac{\text{P.E. of differences between scores on two similar forms}}{\sqrt{2}}$$

(A third way of writing the relationship given by Formula 17) . . [27]

It is recommended that Formulas 16 and 17 be used, as they incorporate σ_1 and r_{11}, each of which it is desirable to know for its own sake. The arithmetic labor involved in calculating standard and probable errors by Formulas 16 and 17 is no greater than in these modified statements, Formulas 25, 26, and 27.

10. Standard scores and their use in calculating idiosyncrasy. If Arthur makes a score of 60 in a paragraph-meaning test and a score of 140 in an arithmetic test, and if these are all the facts that we know about the pupil or the tests, we

Generated on 2020-12-23 00:49 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

do not know which is the better record. If the average for the class is 50 in the first instance and 150 in the second, then we do know that relatively he has done better in paragraph meaning than in arithmetic, as he is above the average in the one and below in the other.

Arthur is 10 paragraph-meaning units above in paragraph meaning and 10 arithmetic units below in arithmetic. We cannot say that he is as far above in the one as he is below in the other, for we do not know that a unit of the one is of equal significance to a unit of the other. If a pupil is one standard deviation above in one test and one standard deviation above in a second, there is much warrant for calling these equally excellent records. We shall use this procedure in general and express deviations from the mean in terms of standard deviations. Such measures of deviation we shall call "standard scores" and designate them by the letter z, with the appropriate subscript. The symbols X, x, M, and σ have been defined. We now add one further symbol to the list.

$$z = \frac{x}{\sigma} = \frac{X - M}{\sigma}$$ (The standard score, or measure of deviation in terms of the standard deviation) . [28]

If paragraph meaning is designated by the subscript 1 and arithmetic by the subscript 2, and if $\sigma_1 = 5$ and $\sigma_2 = 10$, we have for Arthur the following standard scores:

$$z_1 = \frac{60 - 50}{5} = 2.0$$

$$z_2 = \frac{140 - 150}{10} = -1.0$$

and we may now say that his score is twice as far above the mean in paragraph meaning as it is below the mean in arithmetic.

Let d represent the difference between two standard scores, thus:

$$d = (z_1 - z_2)$$ (Measure of idiosyncrasy) . . [29]

Generated on 2020-12-23 00:50 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

This is a measure of relative divergence in one trait from position in a second and, as judged by the group tendency, is a measure of idiosyncrasy. If d is large and if we can place confidence in it, — i.e., that it is due to a difference in the mental make-up of the child and not due to chance, — then it becomes highly significant in determining lines of development that need to be emphasized, lines that should be used in vocational activity, etc.

A second technique for measuring idiosyncrasy which has probably occurred to the reader is to express scores in the two subjects in terms of age norms and divide the one by the other. Thus, if Arthur's score of 60 in paragraph meaning is equivalent to the average 12.0-year score and his score of 140 in arithmetic is equivalent to the average 10.0-year score, we have:

$$\frac{12.0}{10.0} = 1.20$$

as his "paragraph meaning–arithmetic" quotient, and we would think of his paragraph-meaning score as being 1.2 times his arithmetic score. Using this procedure in place of that based upon d involves all the errors present in the d technique plus the added errors due to uncertainty as to the zero points in both paragraph meaning and arithmetic, and therefore this paragraph meaning–arithmetic quotient procedure is not advisable. Even the d measure has a substantial chance error, and it is not recommended unless the user determines his probable error so that he can use it rationally.

11. The probable error of measures of idiosyncrasy. We need to know the standard error of our measure, d, of idiosyncrasy. It is easily determined (Kelley, 1923 new), but before giving the formula for it, we need to define one new symbol. We have let r_{11} stand for the reliability coefficient when dealing with a single test. If we have two tests, we

Generated on 2020-12-23 01:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

shall let r_{2II} stand for the reliability coefficient of the second. Then:

$$\sigma_d = \sqrt{2 - r_{1I} - r_{2II}} \quad \text{(Standard error of the measure of idiosyncrasy)} \quad [30]$$

The reliability coefficients of this formula should be those for the grade in question or at least those obtained from groups of substantially the same range of talent.

We may illustrate the use of d and its standard deviation: If the paragraph-meaning test has a reliability of .75 and the arithmetic test a reliability of .50, then

$\sigma_d = \sqrt{2 - .75 - .50} = .866$, and P.E.$_d$ = .6745 σ_d = .584,

so we have for Arthur:

$$d = [2 - (-1)] = 3 \pm .58$$

The difference is five times its probable error, so a difference of the sort found — namely, paragraph meaning superior to arithmetic — indubitably exists, so that if we wish to eliminate it, utilize it, or augment it, as the case may be, we may proceed with much certainty.

Such a wide difference between abilities as reported for Arthur is not common, so that most differences found with tests of such low reliability, .75 and .50, will be very uncertainly established. In general, r_{1I} and r_{2II} should each be greater than .85 to warrant a general study of idiosyncrasy in pupils.

The standard error of d given by Formula 30 is an average value and applicable to all of a population of d's. If we have the following d values for a class of ten: 3.0, 2.2, − 1.4, 1.8, − 2.1, 3.8, 1.7, − 2.8, −.4, .2, and if the reliability coefficients of the two tests employed are .80 and .71, then, by Formula 30, we find σ_d = .70. If we choose one of these ten d's at random, its standard error is .70, or if we take all of them one after another, .70 is the standard error to be attached to each. If, however, we do not choose one at

Generated on 2020-12-23 00:50 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

random, but because of some feature of the d itself, — for example, if we choose the largest d, the one with the value 3.8, — then .70 is too small a value for its standard error. It is beyond the scope of this text to provide the standard-error formula for such a case as this, but the reader should know that when he exercises choice, based on the differences themselves, as to which of several differences is studied, then the general formula for the standard error applicable to differences chosen at random does not apply. In the illustrative problem of Chapter VI a rule for obtaining the approximate probable error of the largest of twenty-one differences is given. This rule does not apply when the number of differences is other than twenty-one or when the differences are all independent of each other, which in the problem of Chapter VI they are not, since there are but seven original tests from which the twenty-one differences arise.

12. The calculation of the median and of other percentiles. The standard deviation of class scores is required in most of the important formulas here given. It not infrequently happens that an approximate answer will suffice, and such may readily be obtained by calculating the 10th and 90th percentiles, determining the difference between the two, which we shall call D, and using the formula (see Kelley, 1921 new) below:

$$\sigma = .390\,D \quad \text{(Standard deviation determined from the 10–90 percentile range)} \quad . \quad . \quad [31]$$

Let us write $P_{.10}$ for the 10th percentile, $P_{.90}$ for the 90th percentile, and in general P_p for the $(100\,p)$ percentile. Then we have:

$$D = P_{.90} - P_{.10} \quad (D, \text{ the 10–90 percentile range}) \quad . \quad . \quad [32]$$

We now need a formula for the calculation of a percentile. The following, serviceable in calculating the upper and lower quartiles, the median, and all other percentiles, looks rather

Generated on 2020-12-23 00:50 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

formidable, but a numerical example will show that it is very simple to use:

$$P_p = v_p + \frac{pN - F_p}{f_p} i_p \quad \text{(Value of a percentile)} \quad . \quad [33]$$

P_p = the percentile the value of which is to be calculated.
p = the proportion of cases having smaller values than P_p [e.g., if the 15th percentile is being determined, then $p = .15$. Further, $100\,p = 15$, and $P_{.15}$ is the value of the 15th percentile, so that $P_{.15}$ is the value of the (100 p) percentile].
v_p = the value of the lower limit of the class in which the P_p percentile lies.
f_p = the frequency or number of cases in this class.
i_p = the interval or range covered by this class.
F_p = the sum of the frequencies in all the classes below (i.e., classes with smaller X values than) this class.

Let us assume class scores on a geography test, as follows, in which X is the gross score and f the number of pupils receiving each score indicated:

X	f
7	0
8	1
9	7
10	8
11	13
12	5
13	6
14	2
15	0
	42

The lower quartile (*L.Q.*) is such a score that one fourth of the pupils make a score less than it; thus the *L.Q.* is identical with $P_{.25}$, the 25th percentile. Similarly the median (*Med.*) is identical with $P_{.50}$ and the upper quartile (*U.Q.*) with $P_{.75}$.

Generated on 2020-12-23 00:50 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

We shall calculate these three percentiles. For the lower quartile $p = .25$ and $Np = 42(.25) = 10.5$. If we add the frequencies from below (numerically below) up until we obtain a total frequency of 10.5, we find that it brings us into the fourth class listed, thus:

$0 + 1 + 7 + 2.5$ (out of the 8 in the fourth class) $= 10.5$

The class index or mid-value of X for this class is 10.0, and thus the class itself extends from 9.5 to 10.5. The value of the lower limit of this class, v_p, is accordingly 9.5; the range covered by this class, i_p, is 1.0 (as given by $10.5 - 9.5 = 1.0$); the number of frequencies lying in this class, f_p, is 8; and the number lying below this class, F_p, is 8 (as given by $0 + 1 + 7 = 8$). Accordingly we have:

$$P_{.25} = 9.5 + \frac{(.25)42 - 8}{8} 1.0 = 9.81$$

Similarly, $P_{.50}$, or the median, lies in the fifth class, and we have:

$$P_{.50} = 10.5 + \frac{(.50)42 - 16}{13} 1.0 = 10.88$$

Also: $$P_{.75} = 11.5 + \frac{(.75)42 - 29}{5} 1.0 = 12.00$$

These quartiles have been calculated merely for illustrative purposes. To obtain D we need the 10th and 90th percentiles:

$$P_{.10} = 8.5 + \frac{(.10)42 - 1}{6} 1.0 = 8.96$$

$$P_{.90} = 12.5 + \frac{(.90)42 - 36}{6} 1.0 = 12.80$$

and $$D = P_{.90} - P_{.10} = 12.80 - 8.96 = 3.84$$

so that we have for an approximate value for the standard deviation:

$$\sigma = .390(3.84) = 1.50$$

Generated on 2020-12-23 00:51 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

formidable, but a numerical example will show that it is very simple to use:

$$P_p = v_p + \frac{pN - F_p}{f_p} i_p \quad \text{(Value of a percentile)} \quad . \quad [33$$

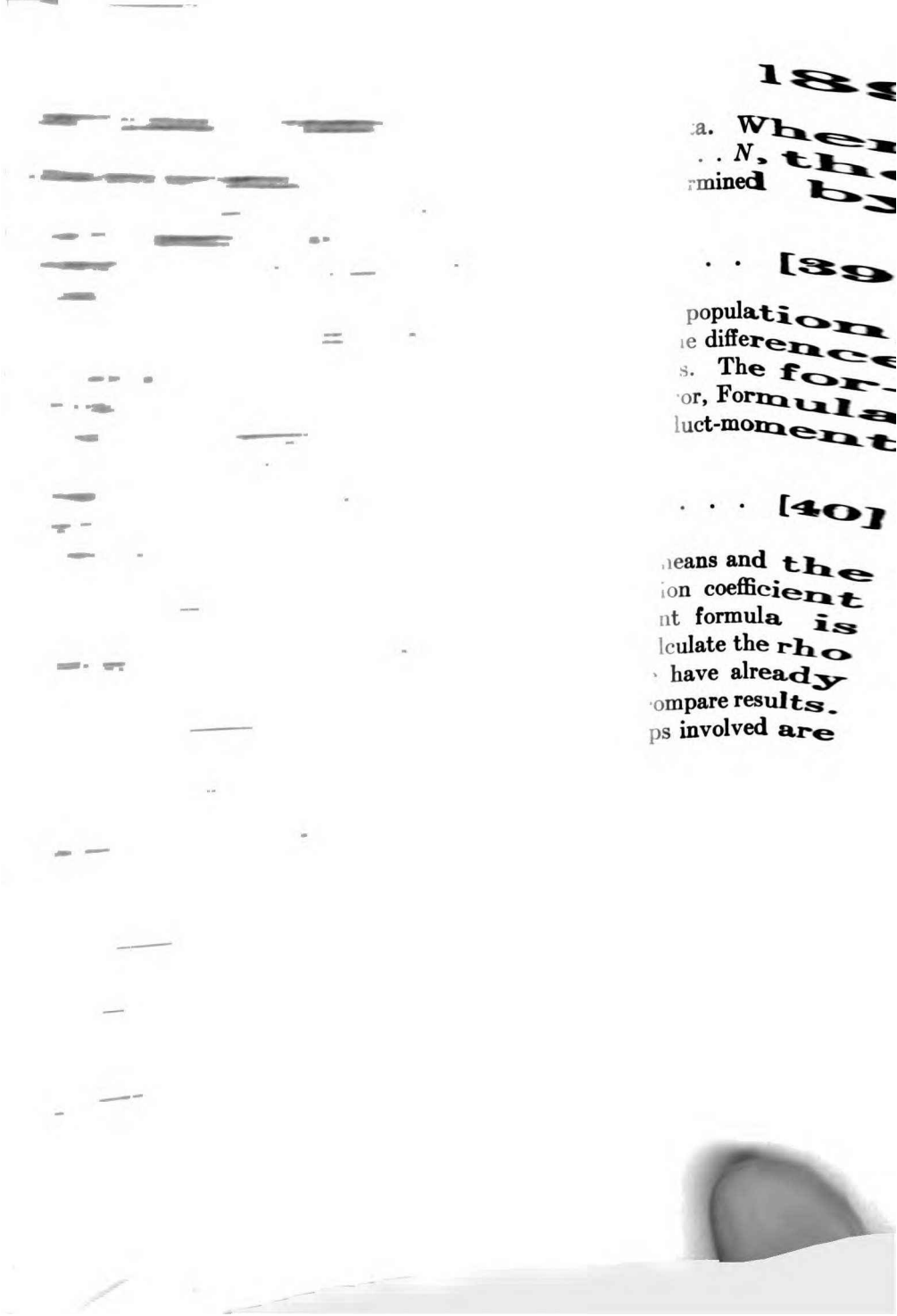
P_p = the percentile the value of which is to be calculated.
p = the proportion of cases having smaller values than P [e.g., if the 15th percentile is being determined, then $p = .15$. Further, $100\,p = 15$, and $P_{.15}$ is the value of the 15th percentile, so that $P_{.15}$ is the value of the $(1($ p) percentile].
v_p = the value of the lower limit of the class in which t P_p percentile lies.
f_p = the frequency or number of cases in this class.
i_p = the interval or range covered by this class.
F_p = the sum of the frequencies in all the classes below (i classes with smaller X values than) this class.

Let us assume class scores on a geography test, as follo in which X is the gross score and f the number of pupils r ceiving each score indicated:

X	f
7	0
8	1
9	7
10	8
11	13
12	5
13	6
14	2
15	0
	42

The lower quartile ($L.Q.$) is such a score that one fou the pupils make a score less than it; thus the $L.Q.$ is ide with $P_{.25}$, the 25th percentile. Similarly the median (is identical with $P_{.50}$ and the upper quartile ($U.Q.$) wit

Generated on 2020-12-23 00:51 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN



Generated on 2020-12-23 00:51 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

As an exercise the student may calculate by Formula 13 the standard deviation for the same series of scores and compare answers.

13. The credence to be placed in measures based on total populations. The most important concept bearing upon reliability for the reader to have is that of standard error, or probable error, of a score of a single individual. Of course every statistical measure has an error, and formulas to obtain these are available in many cases. The standard errors of statistical constants which are based upon the entire population dealt with have $\sqrt{N}$ in the denominators, and therefore the errors in these constants are regularly very much smaller than the errors in individual scores. It is desirable that the sizes of these errors be known, as it is not safe to assume that they are negligibly small. They are given below for the mean, standard deviation, the 10–90 percentile range, and the correlation coefficient. To obtain the probable errors of these four constants it is of course only necessary to multiply the right-hand members in each case by .6745.

$$\sigma_M = \frac{\sigma}{\sqrt{N}}$$ (Standard error of the mean. See also Formula 3, Chapter III) . . . . . [34]

$$\sigma_\sigma = \frac{\sigma}{\sqrt{2N}} = \frac{.7075}{\sqrt{N}}$$ (Standard error of the standard deviation) . . . . . [35]

$$\sigma_\sigma = \frac{.9\,\sigma}{\sqrt{N}}$$ (Approximate standard error of the standard deviation when σ is derived from D) [36]

$$\sigma_D = \frac{.889\,D}{\sqrt{N}}$$ (Standard error of the 10–90 percentile range) . . . . . [37]

$$\sigma_r = \frac{1 - r^2}{\sqrt{N}}$$ (Standard error of the product-moment correlation coefficient) . . . . [38]

All these formulas are based upon the assumption that N is large. If N is less than 25, particularly if less than 10, they all yield too small values for the standard errors.

Generated on 2020-12-23 01:19 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

14. Correlation determined from ranked data. When each of two series of measures is ranked 1, 2, 3, . . . N, the correlation between them may be readily determined by Spearman's rho formula for correlation:

$$\rho = 1 - \frac{6\,\Sigma D^2}{N(N^2 - 1)} \quad . \quad . \quad . \quad . \quad . \quad [39]$$

in which rho is the correlation coefficient, N is the population, and D is a variable, being for each individual the difference between the ranks of his scores for the two series. The formula is very easy to use and has a standard error, Formula 40, but slightly larger than that of the product-moment coefficient of correlation, Formula 38.

$$\sigma_\rho = \frac{1.047(1 - \rho^2)}{\sqrt{N}} \quad . \quad . \quad . \quad . \quad . \quad [40]$$

Since, however, one commonly desires the means and the standard deviations, as well as the correlation coefficient between the two series, the product-moment formula is much the more valuable. Let us, however, calculate the rho correlation for the same data for which we have already calculated the product-moment r, in order to compare results. The data are given in Table 20, and the steps involved are shown in Table 23 on the next page.

Generated on 2020-12-23 00:52 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 23

Pupil	Score on Form 1	Score on Form 2	Rank on Form 1	Rank on Form 2	Difference in Ranks Squared
N	92	90	1½	1	.25
FF	92	80	1½	6	20.25
S	84	78	3	8	25.00
P	80	58	5	26½	462.25
T	80	60	5	23	324.00
AA	80	80	5	6	1.00
G	78	76	8½	9½	1.00
H	78	76	8½	9½	1 00
I	78	86	8½	2	42.25
HH	78	82	8½	3½	25.00
M	76	74	11	11½	.25
B	74	62	13½	20½	49.00
D	74	80	13½	6	56.25
F	74	66	13½	16½	9.00
GG	74	66	13½	16½	9.00
K	72	60	16	23	49.00
W	70	62	17½	20½	9.00
X	70	58	17½	26½	81.00
L	68	72	19½	13	42.25
Z	68	66	19½	16½	9.00
R	66	70	21	14	49.00
Q	64	54	23	30	49.00
V	64	74	23	11½	132.25
DD	64	58	23	26½	12.25
C	62	82	25½	3½	484.00
CC	62	58	25½	26½	1.00
JJ	60	64	27	19	64.00
O	58	46	28	31½	12.25
A	56	46	29	31½	6.25
BB	52	56	30	29	1.00
E	48	44	31	33	4.00
EE	40	66	32	16½	240.25
U	38	40	33	34½	2.25
Y	36	60	34	23	121.00
II	34	16	35	36	1.00
J	26	40	36	34½	2.25
					2397.50 = ΣD^2

Generated on 2020-12-23 00:52 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The entries in the fourth column, $1\frac{1}{2}$, $1\frac{1}{2}$, 3, 5, etc., are rank values assigned to the scores in the second column. As the first two entries in the second column are equal, neither is deserving of the first rank in preference to the other. Therefore ranks 1 and 2, which are to be assigned to these first two entries, are averaged, obtaining $1\frac{1}{2}$, and this average rank is assigned to each of the 92's. A similar procedure is followed throughout wherever there are ties in scores. The ranks recorded in the fifth column are in accordance with the size of the scores in the third column, and here again, wherever there are ties, the average rank is assigned to each of the tied measures. In the sixth column D^2 values are recorded. These are the squares of the differences in ranks as given in the fourth and fifth columns. The sum of the D^2's is recorded at the foot of the sixth column. We thus have for the rho correlation coefficient:

$$\rho = 1 - \frac{6(2397.50)}{36(36^2 - 1)} = .69$$

The trustworthiness of this value is indicated by its standard error:

$$\sigma_\rho = \frac{1.047(1 - .69^2)}{\sqrt{36}} = .091$$

The fact that rho does not exactly equal the product-moment correlation coefficient, which was found to be .75, is in part due to a systematic difference between rho and r, which may be allowed for by the use of Formula 41. This systematic difference is small, so that ordinarily it suffices to report the rho coefficient found without correcting it by using Formula 41.

$$r = 2 \sin \frac{\pi}{6} \rho = 2 \sin (p \times 30°)$$

(Formula for estimating the product-moment correlation, knowing the rank coefficient of correlation) . . . [41]

Generated on 2020-12-23 00:52 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

In the main the difference found between rho and a product-moment *r* is to be attributed to variable and unknown causes, which, however, are quite likely to be present. Thus for the series here studied the discrepancy found between the two values is not surprising.

Generated on 2020-12-23 00:52 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

CHAPTER EIGHT

Observations in Support of Certain Principles Used in Preceding Chapters

The five sections of this chapter are more or less technical, so that the reader who is primarily interested in the conclusions and not in the methods of arriving at them is advised to read this first paragraph only, skipping the rest of the chapter. Sections 1, 2, and 3 provide the data and argument leading to the conclusion that some 90 per cent of a general intelligence test and an all-round achievement test measure the same thing. Section 4 provides the argument showing that a reliability of .50 or higher is demanded of a test which is to be used for group-measurement purposes, and a reliability of .96 or higher for a test to be used as a guide in the making of individual diagnoses. Section 5 provides the proof that the proper weighting factor to allow for differences in reliability when a number of tests are combined to build up a single battery is $\sqrt{r_{11}}/(1 - r_{11})$.

1. The proportion of elements in "achievement" and "intelligence" that are identical. Let the achievement test scores be represented by symbols with the subscript 1 and the intelligence test scores by symbols with the subscript 2, and let us suppose that the intercorrelation between the two tests and the reliability coefficients of each and the standard deviation of the scores in each are known for a certain narrow-range group.

Dealing with scores as deviations from means, we have for this narrow-range group:

$$x_1 = x_\infty + e_1$$

in which x_1 is the obtained score; x_∞, the pupil's true score; and e_1, the error of measurement. Let

$$x_2 = x_\infty + e_2$$

Generated on 2020-12-23 00:53 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

be a similar statement for the intelligence measure. The correlation between the true scores and the standard deviations of the true scores can be estimated by the following formulas:

$$r_{\infty\omega} = \frac{r_{12}}{\sqrt{r_{1\text{II}}}\sqrt{r_{2\text{II}}}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [42]$$

$$\sigma_\infty = \sigma_1\sqrt{r_{1\text{II}}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [43]$$

$$\sigma_\omega = \sigma_2\sqrt{r_{2\text{II}}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [44]$$

By these three formulas $r_{\infty\omega}$, σ_∞, and σ_ω may be obtained, so henceforth, in this section, the discussion will pertain to variables x_∞ and x_ω, and not to x_1 and x_2. Let

$$x_\infty = ua + b$$

and

$$x_\omega = wa + c$$

in which u and w are constants for the entire population dealt with; a, the factor making for success in both "achievement" and "intelligence"; b, a factor uncorrelated with a or c, making for success in achievement only; and c, a factor uncorrelated with a or b, making for success in intelligence only. We may thus write:

$$\sigma^2_\infty = u^2\sigma^2_a + \sigma^2_b$$

The left-hand member is the variance [1] of the true achievement scores and is equal to $u^2\sigma^2_a$, the variance due to the factor which is found also in the true intelligence scores, plus σ^2_b, the variance of the factor which is unique to the achievement scores. Similarly,

$$\sigma^2_\omega = w^2\sigma^2_a + \sigma^2_c$$

That is, the variance of the true intelligence scores is equal to $w^2\sigma^2_a$, the variance due to the factor which is found also in the true achievement scores, plus σ^2_c, the variance of the

[1] The term "variance" means σ^2.

Generated on 2020-12-23 00:53 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

factor unique to the intelligence scores. For the correlation between x_∞ and x_ω, we have:

$$r_{\infty\omega} = \frac{\Sigma(ua+b)(wa+c)}{N\sigma_\infty\sigma_\omega} = \frac{uw\sigma^2_a}{\sigma_\infty\sigma_\omega} = \sqrt{\frac{\sigma^2_a}{\sigma^2_a+\frac{\sigma^2_b}{u^2}}}\sqrt{\frac{\sigma^2_a}{\sigma^2_a+\frac{\sigma^2_c}{w^2}}}$$

To simplify this equation, let us assume, relative to the total x_∞ variance, that the variance of the unique portion of x_∞ is equal to that of the unique portion of x_ω, relative to the total variance of x_ω; that is, let us assume that

$$\frac{\sigma^2_b}{\sigma^2_\infty} = \frac{\sigma^2_c}{\sigma^2_\omega} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [45]$$

Then

$$\frac{\sigma^2_b}{u^2} = \frac{\sigma^2_c}{w^2} = \text{(let us say) } \sigma^2_d$$

There are no reasons of which the author is aware for regarding this assumption as extreme, and it greatly simplifies interpretation. Rather freely expressed, it is equivalent to saying that that part of achievement which is not intelligence is as great an amount as that part of intelligence which is not achievement. This matter should be investigated experimentally after the terms " achievement " and " intelligence " have been more objectively defined than at present, but *for the issue here studied* a 10 or 20 or even 50 per cent error in this assumption is of no great moment. We then have:

$$r_{\infty\omega} = \frac{\sigma^2_a}{\sigma^2_a+\sigma^2_d} \quad . \quad . \quad . \quad . \quad . \quad . \quad [46]$$

We thus see that the coefficient of correlation between achievement and intelligence test scores corrected for attenuation is equal to the variance of the common factor divided by the total true variance, or it is that proportion of the total variance which is due to the common factor present in each test.

Generated on 2020-12-23 00:53 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Obviously, if $r_{\infty\omega} = .9$

then $\frac{\sigma^2_a}{\sigma^2_a + \sigma^2_d} = .9$

and $\frac{\sigma^2_d}{\sigma^2_a + \sigma^2_d} = .1$

so we would conclude that 90 per cent of the two traits correlated was identical and 10 per cent was different. From data given in Section 3 of this chapter it is seen that this is approximately the situation that prevails between the Stanford Achievement Test total score and certain well-known intelligence tests.

2. The estimation of the true correlation between general intelligence scores and general achievement scores for a defined range of talent, knowing the correlation in a different range. Before utilizing Formula 46, we must first determine what range of talent is to be employed for determining $r_{\infty\omega}$. Shall we use a one-grade range or a single chronological-age-group range, or shall we use the total range for which the achievement test used, the Stanford Achievement Test, is serviceable — grades 2 to 9? Though it does, in fact, make rather less difference than one might suspect, for $r_{\infty\omega}$ is found to be large even in narrow ranges, still we must not neglect the effect of range upon the correlation between different traits.[1]

It is probably true that certain accomplishments of young children are designated "intelligence," whereas the same

[1] It seems to the writer that Pearson's formula for the effect upon correlation of double selection (Kelley, 1923, stat., Section 64) scarcely applies here because of the assumptions underlying it. Quite different assumptions are employed in the treatment herewith.

Dr. Otis's (1925) formula to accomplish the same purpose is surely theoretically inapplicable. Dr. Otis's formula is, in fact, the well-known formula for the relation between ranges in obtained scores and *reliability* coefficients (see Kelley, 1923, Formula 178). The writer considers it unsound to use this formula when the variables correlated are different; that is, are not equally excellent measures of the same thing.

Generated on 2020-12-23 00:53 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

accomplishments of older children are labeled "achievement," and vice versa — for example, the Stanford-Binet year 9, question 3, which is, "If I were to buy 4 cents' worth of candy and should give the storekeeper 10 cents, how much money would I get back?" would likely appear in an intelligence test for 7-year-olds and in an achievement test for 10-year-olds. Thus the same function is at one time revealed as "intelligence" and at another time as "achievement." The narrower the range of talent considered, the more likely are such situations to be found. Otherwise expressed, such a function as that measured by the Binet question cited would contribute to the measure of "difference" between achievement and intelligence if a one-year range of talent were examined, whereas it would augment the measure of "similarity" if the range included both 7- and 10-year-olds. To do this latter appears logically sound, so it would seem most reasonable to consider the community of function between achievement and intelligence with reference to as wide a range of talent as is biologically homogeneous.

If an age-heterogeneous group of children are included in a correlation study, then the correlation found is due in part to a growth factor affecting both variables. Our thinking as it concerns adults does not involve a difference in growth, for there is negligible growth in mental functions for the range represented by mature (i.e., neither adolescent nor senescent) adults. In order to parallel this situation when dealing with children, we must choose a single age range of talent. The groups for which data are commonly at hand are usually defined in terms of school grades and not of ages, so we must ascertain the grade range that is commensurate with a single age range. For the Stanford Achievement total score we have age and grade standard deviations as in Table 24 on the next page.

Generated on 2020-12-23 00:53 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 24

STANFORD ACHIEVEMENT TOTAL SCORE MEANS AND STANDARD DEVIATIONS

(POPULATIONS OF ABOUT 150 PER GRADE)

RANGE OF TALENT	MEAN SCORE		STANDARD DEVIATION
Unselected 12-Year-Olds	57.2		20.5
Grade 2	9.0		5.4
3	19.6		9.0
4	32.7		10.0
5	47.8		10.8
6	54.7		11.0
7	64.1		11.2
8	72.7		11.4
9	78.9		11.4
Grades 2–9	47.4	(Same number of pupils from each grade)	25.6
3–8	48.6	"	20.9
4–8	54.4	"	17.5

We thus see that a complete sampling of 12-year-olds shows but slightly less variability than do the pupils in six consecutive school grades. The fact that so wide a grade range is required to give us an equivalent variability to that of unselected 12-year-olds may at first sight seem rather surprising. If we will, however, recall that without any particular attempt at homogeneous classification we nevertheless not infrequently find 12-year-olds in the first and second grades and in the ninth and tenth grades, and further, that the 12-year-olds in the first and second grades are generally there by courtesy, belonging properly in the kindergarten, and that the 12-year-olds in the ninth and tenth grades are characteristically very greatly retarded pedagogically, being

Generated on 2020-12-23 00:54 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

commonly of a mental caliber of college freshmen, it no longer appears surprising that in general the variability of six consecutive school grades is approximately equal to that of unselected children of a single age. We shall therefore conclude that we are called upon to reduce all measures of correlation obtained from some other than a six-grade range to the comparable measure for a six-grade range. We shall refer to this range as that of a complete random sampling of children of a single age.

We may investigate analytically the effect of range. We shall let lower-case letters indicate constants determined from a narrow-range group and capital letters those from a wide-range group. Then the problem which concerns us is to estimate $R_{\infty\infty}$, knowing $r_{\infty\infty}$. There are well-known formulas for the estimation of the reliability coefficient to be expected in a wide range when its value in a narrow range is available. We are here, however, dealing with a much more complex problem, for we are correlating measures which at least in part are not measures of the same underlying capacity.

A study of the very extensive correlation data given by Root (1922) shows no systematic change in the size of the correlations found in the various elementary school grades, except for a slight lowering in Grade 9. The reliability coefficients for the Stanford Achievement Test are also very approximately constant for the different grades. The data are not available to inform us if the reliabilities of the well-known intelligence tests are constant from grade to grade, but it seems reasonable to expect them to be approximately so. We shall therefore postulate a situation wherein the true correlation between achievement and intelligence is the same for the various grades, and thus deduce $R_{\infty\infty}$ for a six-grade range, knowing it for a narrower or wider range of talent. We shall assume that the correlations, means, and

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

standard deviations for two single grades separately and for the two combined are as given in the accompanying table:

TABLE 25

	True Correlation	True Standard Deviation		Distance between Lower Half Mean and Other Means		Population
		Achievement	Intelligence	Achievement	Intelligence	
Lower single grade	$r_{\infty\omega}$	σ_∞	σ_ω	0	0	N
Upper single grade	$r_{\infty\omega}$	σ_∞	σ_ω	$k\sigma_\infty$	$k\sigma_\omega$	N
Two grades combined	$R_{\infty\omega}$	Σ_∞	Σ_ω	$\frac{k\sigma_\infty}{2}$	$\frac{k\sigma_\omega}{2}$	$2N$

It is to be noted that in addition to assuming that for the two single grades $r_{\infty\omega} = r_{\infty\omega}$, it is assumed that σ_∞ for the lower single grade equals σ_∞ for the upper single grade, and also that σ_ω is the same for the two grades. It is also assumed that the mean growth in achievement and intelligence from the lower to the upper grade is the same number, k, of standard deviations. These assumptions do little violence to the known facts concerning the Stanford Achievement Test, at least from Grades 3 to 9, but there are a number of tests which rather uniformly have a larger σ_1 (and therefore probably a larger σ_∞) for upper than for lower elementary school grades. The uncertainty of these assumptions necessitates that the conclusions here reached be looked upon merely as first approximations. The writer uses them and offers Table 25 not because this table is experimentally established, but because, pending experimental determination, he believes that its use will give a much truer picture of achievement-intelligence correlations than one in which the effect of the range of talent examined is ignored.

If we accept Table 25, our statistical problem is to relate Σ_∞ with σ_∞; Σ_ω with σ_ω; and $R_{\infty\omega}$ with $r_{\infty\omega}$. Let s stand

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

for a summation covering the cases in the lower grade and S for a summation extending over the upper grade, and let x_∞ and x_ω stand for deviations from the means of the groups dealt with. We then have:

$$\Sigma^2_\infty = \frac{s\left(x_\infty - \frac{k\sigma_\infty}{2}\right)^2 + S\left(x_\infty + \frac{k\sigma_\infty}{2}\right)^2}{2N}$$

$$= \sigma^2_\infty\left(1 + \frac{k^2}{4}\right) \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [47]$$

Similarly, $\Sigma^2_\omega = \sigma^2_\omega\left(1 + \frac{k^2}{4}\right)$ . . . . . [47]

$$R_{\bar{\omega}\omega} = \frac{s\left(x_\infty - \frac{k\sigma_\infty}{2}\right)\left(x_\omega - \frac{k\sigma_\omega}{2}\right) + S\left(x_\infty + \frac{k\sigma_\infty}{2}\right)\left(x_\omega + \frac{k\sigma_\omega}{2}\right)}{2N\Sigma_\infty\Sigma_\omega}$$

$$= 1 - \frac{1 - r_{\infty\omega}}{1 + \frac{k^2}{4}} \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad . \quad [48]$$

We may generalize the preceding solution so that it applies to whatever number of single grades are combined to give the wide-range group. Making the same assumptions as before, — that σ_∞, σ_ω, and $r_{\infty\omega}$ are constant for each single grade entering into the wide-range population and that the difference between each grade mean and that of the grade just above is $k\sigma_\infty$ and $k\sigma_\omega$ for achievement and intelligence, respectively, — we readily obtain the following formulas, in which g is the number of single grades combined in the wide-range population:

$$\Sigma^2_{\bar{\omega}} = \sigma^2_\omega\left\{1 + \frac{k^2}{g}\left[\left(\frac{g-1}{2}\right)^2 + \left(\frac{g-3}{2}\right)^2 + \ldots \left(\frac{g - 2g + 1}{2}\right)^2\right]\right\} \quad [49]$$

Generated on 2020-12-23 00:54 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

If we let ρ represent the $\{\ \}$ term, — i.e., represent $\Sigma^2_\infty/\sigma^2_\infty$, the ratio between the variances in the wide and narrow ranges, — this may be written:

$$\Sigma^2_\infty = \sigma^2_\infty \rho \quad . \quad . \quad . \quad . \quad . \quad . \quad [50]$$

By a similar derivation:

$$\Sigma^2_\omega = \sigma^2_\omega \rho \quad . \quad . \quad . \quad . \quad . \quad . \quad [50]$$

It is also readily found that

$$R_{\infty\omega} = 1 - \frac{1 - r_{\infty\omega}}{\rho} \quad . \quad . \quad . \quad . \quad . \quad [51]$$

A study of Table 24, giving data for the Stanford Achievement Test, shows that for Grades 2 to 8, k is approximately .9. This simply states that neighboring grade means are approximately .9 of a grade (true) standard deviation apart. Assuming this same value for the intelligence test variable, we obtain values of ρ as given in Table 26:

TABLE 26

ESTIMATED RELATIVE VARIANCES IN TRUE ACHIEVEMENT SCORES OR TRUE INTELLIGENCE SCORES FOR DIFFERENT GRADE RANGES

GRADE RANGE	NUMBER OF CONSECUTIVE GRADES COMBINED							
	1	2	3	4	5	6	7	8
ρ, the ratio of the variance in the grade range indicated to that in the single grade	1.000	1.2025	1.5400	2.0125	2.6200	3.3625	4.2400	5.2525

Using these values of ρ in Formula 51, we immediately obtain the desired correlation for wide ranges, knowing it for narrow ranges. These are given in Table 27 and made use of in the next section.

3. The community of function of achievement and intelligence measures. In Section 1 of this chapter it was shown that $r_{\infty\omega}$, the estimated true correlation between true intel-

Generated on 2020-12-23 00:54 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 27

TABLE FOR ESTIMATING THE TRUE CORRELATION BETWEEN ACHIEVEMENT AND INTELLIGENCE TESTS FOR A SIX-GRADE RANGE, BASED UPON DATA FOR NARROWER AND WIDER GRADE RANGES

GRADE RANGE	1	2	3	4	5	6	7	8
	ASSUMED $r_{\infty\omega}$ VALUES	CONSEQUENTIAL $R_{\infty\omega}$ VALUES						
	.50	.58	.68	.75	.81	.85	.88	.90
	.60	.67	.74	.80	.85	.88	.91	.92
	.70	.75	.81	.85	.89	.91	.93	.94
	.75	.79	.84	.88	.90	.93	.94	.95
	.80	.83	.87	.90	.92	.94	.95	.96
	.85	.88	.90	.93	.94	.96	.96	.97
	.90	.92	.94	.95	.96	.97	.98	.98

ligence and achievement, is a reasonable measure of the per cent of intelligence that is achievement and of achievement that is intelligence. In Section 2 it was shown that a complete age population was of about the same variability as the population of six consecutive school grades, and a table was provided for estimating $r_{\infty\omega}$ for a six-grade range, knowing it for narrower and wider ranges. We may now utilize the conclusions of these preceding sections and find the correlation corrected for attenuation — i.e., $r_{\infty\omega}$ values — for certain intelligence and achievement tests and for certain grade ranges, and secondly, estimate what this correlation would be for a six-grade, or complete age group, range, and thus secure a value indicative of the community of function between two tests.

Some very excellent data are provided by Symonds (1924), from which we can secure the measures of correlation corrected for attenuation that we need. Dr. Symonds' variables are expressed in terms of mental or subject ages, thus:

Generated on 2020-12-23 00:55 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

X_1 = mental age on National Intelligence Tests: Scale A
X_2 = mental age on National Intelligence Tests: Scale B
X_3 = reading age on Thorndike-McCall Reading Test
X_4 = arithmetic age on Woody-McCall Arithmetic Test

The needed statistical constants for these variables are reported by Symonds, except the mean mental or subject ages. However, the pupils are from Grades 4 to 8 inclusive, and we shall scarcely be far astray if we take 150 months as the mean mental or subject ages for each of these four variables. Copying from Symonds' article, we have Table 28:

TABLE 28

	N	Standard Deviation in Months	r_{1I}
X_1 = N. I. T. Scale A	232	23.9	.922
X_2 = N. I. T. Scale B	242	22.4	.949
X_3 = Thorndike McCall R. T.	232	22.2	.794
X_4 = Woody-McCall A. T.	229	25.7	.855

Dr. Symonds gives the following accomplishment-ratio statistics:

TABLE 29

Accomplishment–Ratio Statistics

Ratio	N	Standard Deviation of Ratios	Reliability Coefficient
X_3/X_1	226	.119	.344
X_3/X_2	201	.117	.230
X_4/X_1	224	.142	.598
X_4/X_2	196	.123	.487

The formula giving $r_{\infty\omega}$ is:

$$r_{\infty\omega} = \frac{r_{13}}{\sqrt{r_{11}}\sqrt{r_{3III}}}.$$

Generated on 2020-12-23 00:55 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

From Table 28 we may secure r_{1I} and r_{3III} values, and we may get r_{13} by the following process: Let $X_3/X_1 = X_5$. Then from Table 29 we have given r_{5V}. The statistical problem is, then, knowing r_{5V}, r_{1I}, and r_{3III}, to determine r_{13}. It can be shown that if $v_1 = \sigma_1/M_1$ and $v_3 = \sigma_3/M_3$, the following equation holds:

$$r_{5V} = \frac{r_{3III}v^2_3 + r_{1I}v^2_1 - 2\,r_{13}v_1v_3}{\sqrt{v^2_1 - 2\,r_{13}v_1v_3 + v^2_3}} \quad . \quad . \quad . \quad [52]$$

As all of the elements of this equation except r_{13} are known, we may solve the equation for r_{13} and obtain the needed correlation coefficient. Doing so, we secure .793 as the value of r_{13}. Continuing, we immediately obtain:

$$r_{\infty\infty} = \frac{.793}{\sqrt{.922}\sqrt{.794}} = .927$$

This estimated true correlation is for a five-grade range. Referring to Table 27, we find that for a six-grade range we should expect the value .95. Accordingly Symonds' data suggest that no less than 95 per cent of the National Intelligence Tests: Scale A, and the Thorndike-McCall Reading Test are basically measures of the same thing. Further, in view of the size of the population dealt with, this result has but a small chance error. It may seem surprising that there is so much that is common between these two well-known tests, one called an intelligence test and the other a reading test, but such is clearly indicated to be the case.

Dr. Symonds' data enable the determination of three other measures of community of function. Proceeding just as before, we obtain the last three rows of Table 30:

Generated on 2020-12-23 00:55 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 30

Tests Correlated	$r_{80,80}$ Found	$r_{80,80}$ Estimated for a Six-Grade Range
National Intelligence Tests Scale A, and Thorndike-McCall Reading Test . . .	.93	.95
National Intelligence Tests Scale B, and Thorndike-McCall Reading Test . . .	.87	.90
National Intelligence Tests Scale A, and Woody-McCall Arithmetic Test . . .	.81	.85
National Intelligence Tests Scale B, and Woody-McCall Arithmetic Test . . .	.89	.91

Averaging the first two results in the last column of Table 30 and the last two, we obtain:

The community between the reading and intelligence measures is 92.5 per cent.

The community between the arithmetic and intelligence measures is 88 per cent.

The data just cited are the most extensive that the writer has been able to find in the literature, which have been reported in sufficient detail (reliability coefficients, standard deviations, means, and intercorrelations being needed) to determine the amount of community of function between tests. However, there are certain smaller populations which he has been able to use in this connection.

For a population of 22, Whittier State School boys (delinquents), Grades 4–8, the raw correlation between the Stanford-Binet and the Stanford Achievement Test was found to equal .79. The mean Stanford Achievement Test score was 51.4, and the standard deviation of such scores, 14.78. It has been determined, as reported in Chapter X, Section *g*,

Generated on 2020-12-23 00:55 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

that the reliability coefficient of the Stanford Achievement Test for a population yielding the standard deviation 10.6 is .96. Knowing the reliability for this range, we may estimate for a second range by the formula:

$$R_{1\text{II}} = 1 - \frac{\sigma^2_1}{\Sigma^2_1}(1 - r_{1\text{II}})$$

(Formula for estimating the reliability $R_{1\text{I}}$ in a range of standard deviation Σ_1, knowing the reliability $r_{1\text{I}}$ in a range of standard deviation σ_1) [53]

Thus we have $\sigma_1 = 10.6$; $r_{1\text{I}} = .96$; and $\Sigma_1 = 14.78$. From this we obtain $R_{1\text{I}} = .98$. For this same group the standard deviation of the Stanford-Binet scores is 15.6 months. Knowing the reliability of the Stanford-Binet to equal .93 for a group of variability, $\sigma_1 = 18.46$ months, we may use Formula 53 and obtain the Stanford-Binet reliability for this population of 22. Doing so, we find:

$$R_{1\text{I}} = .90$$

Thus for the coefficient corrected for attenuation we have:

$$r_{\infty\omega} = \frac{.79}{\sqrt{.98}\sqrt{.90}} = .84$$

This is for a five-grade range of talent. Referring to Table 27, we secure .87 as the estimated correlation between true scores for a six-grade range; i.e., for a complete age group. Thus we conclude that the community between the Stanford-Binet and the Stanford Achievement Test is 87 per cent.

The writer has gone through the same process for a number of other populations, with the following results:

Population: 25, Neodesha, Kansas, accelerated Grade 4 pupils.
Tests correlated: Stanford Achievement Test and Illinois General Intelligence Test: Mental Age.
Correlations found: Raw correlation, .71. Corrected for attenuation and for range, this yields .97 community of function.

Population: 60, Los Gatos, California, Grades 6–8 pupils.
Tests correlated: Stanford Achievement Test and National Intelligence Tests, Scale B — Form 1.

Generated on 2020-12-23 00:56 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Correlations found: Raw correlation, .66. This leads to a community of function of 89 per cent.

Population: 156, Everett, Massachusetts, Grade 6 pupils.
Tests correlated: Stanford Achievement Test and Otis Self-Administering Tests of Mental Ability: Intermediate Examination.
Correlations found: Raw correlation, .79. This leads to a community of function of 95 per cent.

In addition to the foregoing, the measures of community of function between various mental tests given in Table 31, on the opposite page, are deduced from the published data given by Root (1922), utilizing in addition the reliability data given in Chapter X.

The most exceptional finding here is with reference to the Otis Group Intelligence Scale correlations. Either the reliability coefficient used in the calculations is too high or the test is considerably different from the other intelligence tests. For the other tests we find that the community of function varies from 87 per cent to 99 per cent, the Stanford-Binet and the Mentimeters being slightly less similar to the other intelligence tests than is the case with the rest. It is, however, pertinent to note that the average community between the various intelligence tests mentioned is but slightly higher than that between the intelligence tests and the achievement tests, and in both instances it is very high. To summarize, we have for a complete age population a situation somewhat as follows:

The community between different intelligence tests is about 95 per cent.
The community between intelligence tests and achievement batteries is about 90 per cent.
The community between intelligence tests and reading tests is about 92 per cent.
The community between intelligence tests and arithmetic tests is about 88 per cent.

Generated on 2020-12-23 00:56 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

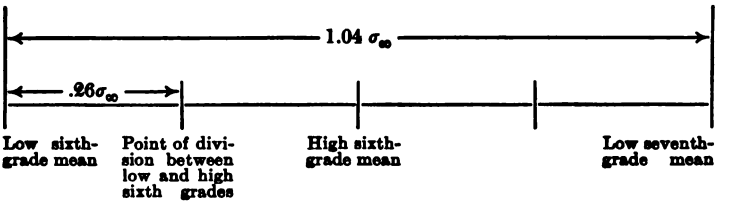
TABLE 31

Tests Correlated	*N*	Grade Range	*r*	Estimated Community of Function for a Complete Age Group
National Intelligence Tests: Scale A — Form 1, and National Intelligence Tests: Scale B — Form 1 .	207	3–8	.94	99%
National Intelligence Tests: Scale A — Form 1, and Mentimeters . .	211	3–8	.93	97%
Stanford-Binet and Mentimeters . .	407	1–12	.88	87%
Terman Group Test of Mental Ability, Form A, and Mentimeters . .	159	7–12	.82	94%
Otis Group Intelligence Scale: Advanced Examination, and Mentimeters	216	5–12	.75	77%
Stanford-Binet and Otis Group Intelgence Scale: Advanced Examination	218	5–12	.80	84%
Stanford-Binet and National Intelligence Tests: Scale A — Form 1 .	211	3–8	.84	94%
Stanford-Binet and National Intelligence Tests: Scale B — Form 1 .	210	3–8	.86	96%
Stanford-Binet and Terman Group Test of Mental Ability: Form A .	160	7–12	.75	93%
National Intelligence Tests: Scale A — Form 1, and Terman Group Test of Mental Ability: Form A . . .	76	7–8	.79	99%

These are the findings that have led to the point of view of Chapter IV, that most of the distinctions drawn between intelligence and achievement are spurious.

Generated on 2020-12-23 00:56 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

4. The reliability requisite for different purposes. Making allowance for the unreliability of the measures of achievement, we can deduce from the study of Kruse (1918) that approximately 15 per cent of sixth-grade children exceed in true all-round scholastic achievement the median of seventh-grade children, and about the same per cent of the seventh grade fall short of the sixth-grade median. If we assume a normal distribution of talent of children in these grades, and if σ_∞ is the standard deviation of the true ability scores of the children in the sixth grade, then by reference to a table of the probability integral we find that it is necessary to go 1.04 standard deviations above the sixth-grade mean to reach the seventh-grade mean. We have, in fact, a situation substantially as diagramed below:



In problems involving group measurement it is ordinarily desirable to distinguish between two mean scores differing by as much as .26σ_∞. To do this with reasonable certainty, this distance should not exceed 1.5 probable errors (the certainty is then represented by chances of five to one). If we take the reliability that leads to this result as the minimal satisfactory reliability, we have a means of determining its numerical value. Let σ_1 be the standard deviation of the obtained scores; σ_∞, the standard deviation of true ability scores; and r_{1I}, the reliability coefficient — all when determined from a one-grade range. Then the equation to be satisfied is:

Generated on 2020-12-23 00:56 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

$$.26\,\sigma_{\infty} = 1.5\left(\frac{.6745\,\sigma_1}{\sqrt{N}}\right) \quad . \quad . \quad . \quad . \quad [54]$$

Let us now take 30 as an average grade population, use Formula 43 of Section 1 of this chapter, and solve equation 54 for r_{1I}:

$$.26\,\sigma_1\sqrt{r_{1I}} = \frac{1.01175\,\sigma_1}{\sqrt{30}}$$

We obtain $r_{1I} = .50$. We shall accordingly conclude that a reliability coefficient, when determined from a single grade range of .50 or higher, is demanded of a test which is to be used for group measurement purposes.

A much higher reliability is needed if individual diagnoses are to be made. In this case the minimal reliability condition to be satisfied is given by Formula 55, where the ( ) term is the probable error of the individual score, whereas in Formula 54 the ( ) term was the probable error of the class mean.

$$.26\,\sigma_{\infty} = 1.5(.6745\,\sigma_1\sqrt{1 - r_{1I}}) \quad . \quad . \quad . \quad [55]$$

from which we obtain $r_{1I} = .94$. This likewise is a reliability coefficient as found from a one-grade range of talent, and since it is a rather high coefficient, it is obvious that relatively few of our intelligence and achievement tests meet this standard of reliability. We are forced to conclude that if they do not, they are of doubtful value in connection with the more important problems involving *individual* classification.

5. Derivation of the weighting factor which is dependent upon the reliability of the test used. The accompanying condensed proof of the weighting procedure of Section 3, Chapter IV, that tests measuring the same thing should be weighted according to the following function of their reliabilities, $\sqrt{r_{1I}}/(1 - r_{1I})$, is given in this text because this proof has not appeared in print in any other place. It is

Generated on 2020-12-23 00:56 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

expected that all readers except mathematicians will pass over this proof.

Let us assume scores x_1, x_2, x_3, . . . x_n, which are all measures of the same thing and which are to be combined to obtain a total score. The appropriate combination is that given by the regression equation relating these to the criterion, x_0, or single thing which it is desired to measure. This equation is:

$$\frac{\bar{x}_0}{\sigma_0} = \beta_{01.23\cdots n}\frac{x_1}{\sigma_1} + \beta_{02.13\cdots n}\frac{x_2}{\sigma_2} + \cdots + \beta_{0n.12\cdots n-1}\frac{x_n}{\sigma_n}$$

in which

$$\beta_{01.23\cdots n} = \frac{\Delta_{01}}{\Delta_{00}}$$

$$\beta_{02.13\cdots n} = \frac{-\Delta_{02}}{\Delta_{00}}$$

etc., where Δ_{01} and Δ_{00} are minors of the major determinant:

$$\Delta = \begin{vmatrix} 1 & r_{01} & r_{02} & \cdots & r_{0n} \\ r_{01} & 1 & r_{12} & \cdots & r_{1n} \\ r_{01} & r_{12} & 1 & \cdots & r_{2n} \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ \cdot & \cdot & \cdot & & \cdot \\ r_{0n} & r_{1n} & r_{2n} & \cdots & 1 \end{vmatrix}$$

Now if x_2, x_3, etc., are all measures of the same thing that x_0 measures, but unequally reliable, then the coefficient of correlation between each and x_0, when corrected for attenuation, will equal 1.00, thus:

$$1 = \frac{r_{01}}{\sqrt{r_{00}}\sqrt{r_{11}}}$$

$$1 = \frac{r_{02}}{\sqrt{r_{00}}\sqrt{r_{22}}}$$

etc.

Generated on 2020-12-23 00:57 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

or

$$\begin{aligned} r_{01} &= \sqrt{r_{00}}\sqrt{r_{1I}} \\ r_{02} &= \sqrt{r_{00}}\sqrt{r_{2II}} \\ &\text{etc.} \\ r_{12} &= \sqrt{r_{1I}}\sqrt{r_{2II}} \\ r_{13} &= \sqrt{r_{1I}}\sqrt{r_{3III}} \\ &\text{etc.} \end{aligned}$$

Making these substitutions in Δ_{01} and evaluating the resulting determinant,[1] we obtain

$$\Delta_{01} = \frac{\sqrt{r_{1I}}}{1 - r_{1I}}[\sqrt{r_{00}}(1 - r_{1I})(1 - r_{2II}) \ . \ . \ . \ (1 - r_{nN})] \quad [56]$$

which, since the [ ] term is constant for all the variables, may be written:

$$\Delta_{01} = \frac{\sqrt{r_{1I}}}{1 - r_{1I}} (\text{constant}).$$

This immediately gives us:

$$\beta_{01 \cdot 23 \ . \ . \ . \ n} = \frac{\sqrt{r_{1I}}}{1 - r_{1I}} (\text{constant}),$$

so that the appropriate weights, in order to allow for differences in reliability, bear the ratios to each other given by the magnitudes $\sqrt{r_{1I}}/(1 - r_{1I})$.

[1] I am indebted to Dr. Harold Hotelling for a suggestion which readily led to the evaluation of this determinant.

Generated on 2020-12-23 00:57 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER NINE

Judgments as to the Excellence of Tests when Used for Individual Measurement and Diagnosis

1. Description of lists and ratings of tests. The number of offerings made to a test-famished generation is sufficient to appease a rapacious appetite. The schoolman of today may satisfy the arithmetic or the reading cravings of his fifth-grade pupils with some thirty "standardized" arithmetic tests and some twenty-five "standardized" reading tests. The salad and dessert courses are not overlooked, for there are some twenty hand-writing scales and at least a dozen Latin tests. In spite of the avidity of the boys and girls and of their solicitous teachers, the time is past when one can give all the standardized tests even to a single subject. Accordingly, in any measurement program the issue, after the purpose has been decided upon, is which of the many available measures to employ. The writer has heard the opinion expressed that it would be officiousness to attempt to answer this question. It naturally cannot be answered to the satisfaction of all test devisers, for there is only one place at the top of each ranking list.

The writer considers that in publishing the ranks of tests, for general excellence for individual measurement, as he here does in this chapter, though errors in ranking are unavoidable and when present very unfortunate, nevertheless, due to the frequency with which it happens, the error of the principal or superintendent who selects a poor measuring device when a good one is available is much more serious. It is a double injustice: first, and perhaps a negligible injustice, to the publisher of the better test, and secondly, an injustice to the pupils whose scholastic futures are affected by the test results. This latter injury is so serious that the

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

writer offers no apology for the subject matter of this chapter. The rankings here given are the consensus of opinion of seven judges, of whom the writer is one. He, as undoubtedly does each of the other six, believes certain rankings to be in error. It is, however, hoped that the rankings will prove of such service that future and improved rankings will be called for, and the writer hopes that these can be made utilizing further judgments; particularly does he hope that authors of tests will present more adequate data as to reliability to facilitate judgment than has been commonly given in the past. The writer would be most happy to receive from test authors data upon this point or references to published sources covering it.

The judges who were asked to serve with him in giving rankings were chosen by the writer because they were known by him to have broad training and experience with either intelligence or achievement tests, or both. These judges were Raymond Franzen, Frank N. Freeman, William A. McCall, Walter S. Monroe, Arthur S. Otis, L. L. Thurstone, Marion R. Trabue, and Martin J. Van Wagenen.

Dr. Thurstone expressed inability to do the task, and Dr. Monroe stated that he had conscientious scruples against such an undertaking because his point of view is that tests should be chosen because of their peculiar adaptability to the specific purposes in mind and that therefore, on the whole, the various tests in a given field should not be compared with each other, but should be retained for the measurement of the specific phases each is peculiarly adapted to measure. The writer regrets that space, as well as a considerable lack of pertinent data, does not permit herewith a thorough investigation of Monroe's point of view. He does, however, question its applicability to the data in hand and would refer the reader to the line of argument and the facts presented in Chapter VIII, Sections 1, 2, and 3, con-

Generated on 2020-12-23 00:57 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

cerning the community of function between general intelligence and achievement.

Two of the other judges expressed unfamiliarity with a large portion of the field and therefore gave rankings in but certain of the classifications. Accordingly, the rankings given in this chapter are based in the main upon reports of five judges.

Fallible as these rankings are, the writer believes that they constitute a radical improvement upon the judgments of teachers, principals, and superintendents generally. These are men and women who are busy with other things and who incidentally find themselves called upon to make a test selection, which can be wisely made only by one of technical knowledge and wide experience in the test field. A study of the individual rankings as recorded in subsequent sections shows that on the whole there is a very fair degree of agreement among the judges. Part of this agreement may reasonably be attributed to advertising, for there is reason to believe that a well-advertised test will be ranked higher than an equally good one which is less widely used and less well known, but the greater part of the agreement may surely be attributed to actual differences in merit which are recognized by these judges working entirely independently of one another. All judgments, including those of the writer, were made without a knowledge of the judgments of the other judges. Some half-dozen revisions of these original rankings have been made when the judge in question has stated that he was in error in his first ranking, due to ignorance of certain facts. In one classification the writer called the attention of certain judges to their failure to rank certain tests. (These appeared on a second sheet which had been overlooked.) On the whole, however, it may be said that the basic rankings are entirely independent of each other.

Generated on 2020-12-23 00:58 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

After rankings had been summarized, it developed that certain tests had been overlooked. The judges were requested to insert these in the lists, giving the average rankings. This was done as recorded in Table *sss*. Though the procedure here is different from that employed in making the initial rankings, it should, however, be entirely fair and in no sense interfere with the independence of judgment.

Just before publication another search revealed a number of tests which should be in the lists. These were not presented to the judges both because time did not permit and because most of these tests were so new that the judges would have had little opportunity to become acquainted with them. These tests are listed at the ends of the ranked lists in Chapter X. They are presented without recommendation. Undoubtedly they range in general merit from very poor to excellent.

In addition to ranking the tests of a given classification, the judges were asked to indicate the number of tests which they considered sufficiently excellent and reliable to be used for individual measurement and classification; the number not sufficiently reliable for this, but satisfactory for group measurement and classification; and finally, the number of such merit that they were of doubtful value for either purpose. The median number considered of " individual value " or of " group value " is recorded in connection with each classification. Let us note in detail how this might work out, and has done so, in certain instances.

Generated on 2020-12-23 00:58 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 32

CLASSIFICATION *X*

TESTS RANKED FOR GENERAL EXCELLENCE FOR THE PURPOSE OF INDIVIDUAL MEASUREMENT AND CLASSIFICATION	JUDGES							MEDIAN RATINGS
	1	2	3	4	5	6	7	
A	*	*	5	3	1	2½	4	3
B	*	*	1	2	2	1	3	2
C	*	*	4	6	3	*	1	3½
D	*	*	3	1	5	2½	2	2½
E	*	*	6	5	6	*	6	6
F	*	*	2	4	4	4(½)†	5	4
No. having individual value			1	3	3	1	4	3
No. having group value			4	1	2	2	2	2
No. of doubtful value or of value unknown to the judge			1	2	1	1	0	1
No. not reported upon	6	6	0	0	0	2	0	0

* Not reported upon by the judge (generally because not known by him) and therefore not ranked.

† The ½ in parentheses indicates that the judge is but slightly familiar with the test and considers that his judgment upon this test should receive but half weight. It was originally planned to give this fractional weight, but fractional weightings occurred so infrequently that this was not done, and full weight was given to the ranking when calculating median rankings.

The alphabetical order of the tests need no longer be maintained, and the data of Table 32 may be organized to appear as in Table 33, on the opposite page.

One might conclude upon reading this table that the consensus of opinion was that Tests B, D, and A had individual value; Tests C and F, group value; and that Test E was of doubtful value. This is probably not far from the mark, but it should be noted that the judges did not express their opinions upon this specific matter. Reference to Table 32 shows that Judge 4 would say that these three tests, B, D, and A, have individual value, but Judge 5, who likewise considered three tests to have such value, attributed it to Tests A, B, and C. Though it seems fairly reasonable to

Generated on 2020-12-23 00:58 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TABLE 33

CLASSIFICATION *X*

Individual value, 3. Group value, 2

TEST	NO. OF JUDGES RATING IF LESS THAN 5	JUDGES					MEDIAN RATINGS
		3	4	5	6	7	
B		1	2	2	1	3	2
D		3	1	5	2½	2	2½
A		5	3	1	2½	4	3
C	4	4	6	3		1	3½
F		2	4	4	4(½)	5	4
E	4	6	5	6		6	6

conclude that if there are three tests having individual value, they are Tests B, D, and A, nevertheless only two judges, 4 and 7, expressed the judgment that these particular tests had this value.

2. The detailed classifications and ratings of the various tests. Tables *a* to *sss* are given herewith, because the writer considers that in so important a matter as this it is incumbent upon him to publish rankings in such a manner that he can be checked up upon them. Each judge has been informed as to his individual number and can thus verify his individual rankings. The reader may use these tables to ascertain the variability of judgments of the excellence of any particular test, while the tables of Chapter X will prove most serviceable in obtaining an average judgment as well as other information about a test.

Generated on 2020-12-23 00:58 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

TABLES GIVING RANKINGS OF TESTS[2]

(*a*) PRIMARY GENERAL INTELLIGENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Pintner, R. — Cunningham, B. V., — Primary Mental Test		2(½)	4	1	3	2	1	1[1]	2
Park, B. — Franzen, R., — Primary Test		3(½)	b7	2	2	1		2	2
Dearborn, W. F., — Group Test of Intelligence		5(½)	2	6	1	7	2	4	4
Rhode Island Intelligence Test (Bird, G. E., and Craig, C. E.)			5	4	5	6(½)		3	5
Haggerty, M. E., — Intelligence Examination δ 1		7(½)	3	3	4	5	6	b7	5
Otis, A. S., — Group Intelligence Scale; Primary Examination		1(½)	1	5	7	b7	5	b7	5
Detroit First Grade Intelligence Test (Engel, A. M.)		5(½)	6	b7	6	3	4	7	6
Kingsbury, F. A., — Primary Group Intelligence Test		b7	7	7	b7	4		5	7
[2] Army Beta		b7	b7	b7	b7	b7	b7	b7	b7
[2] Cole, L. W. and Vincent, L. — Group Intelligence Test		5(½)	b7	b7	b7	b7	3	b7	b7

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:59 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

[2] Holley, C. E., — Picture Completion Test		b7	b7	b7	b7	b7	b7(½)	b7	b7
[2] Meyers, G. C., — Mental Measure		b7	b7	b7	b7	b7	7	b7	b7
[2] Pressey, S. L., — Mental Survey Tests, Primer Scale		b7	b7	b7	b7	b7	b7	6	b7
Test known by a single judge									
Gunnison's Primary Test A	1								
No. having individual value		8	9	2	1	3	4	2	3
No. having group but not individual value		3	3	7	12	6	6	7	7
No. of doubtful value		1	1	4	0	4	0	4	3
No. not reported upon		2	1	1	1	1	4	1	1

[1] Judge G. states: "For middle of range only. Extreme cases should be checked up by individual mental examinations."

[2] Here and in subsequent sections of this table the tests below the heavy horisontal rule are not in order of median ratings. In this section in each case the rating is "below rank 7," hence the notation "b7."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:59 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*b*) ELEMENTARY GENERAL INTELLIGENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
National Intelligence Test (Parts A and B)		2	1	1	1	2	1	3	1
National Intelligence Test (Part A only)		5	5½	8½	2½	3	6	4	5
Haggerty, M. E., — Intelligence Examination δ 2		b7	2	5	5	1	7	b9	5
Dearborn, W. F., — Group Test of Intelligence	4		7	4		9½	4		5½
Otis, A. S., — Self-Administering Tests of Mental Ability: Intermediate Examination		3	3	6	6(½)	9½	b9	2	6
National Intelligence Test (Part B only)		6	5½	8½	2½	4	8	7	6
Otis, A. S., — Group Intelligence Scale		4	b9	7	b9	6	b9	1	7
Illinois General Intelligence Scale. (Buckingham, B. R.)		7	8½	2	8	b10	2	b9	8
Mentimeters School Group 2 A (Trabue, M. R.)			b9	3	b9	8	3	b9	9
Army Alpha		b7	b9	b9	b9	b10	9	b9	b9
Army Beta		b7	b9	b9	b9	7	b9	b9	b9
Chicago Group Intelligence Test (Freeman, F. N. and Rugg, H. O.)	3		b9	b9		b10			b9
Morgan, J. J. B., — Mental Test			b9	b9	b9	b10		6	b9
Meyers, G. C., — Mental Measure			b9	b9	b9	b10	b9	8	b9
New Jersey Composite Test, — Intelligence Section			b9	b9	b9		b9	b9	b9

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:59 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Pintner, R., — Non-Language Mental and Educational Survey Tests			4	b9[1]	9	b10		b9	b9
Otis, A. S., — Classification Test		1	b9	b9	4	b10		b9	b9
Peters, C. C., — Test of General Information with Sociologically Determined Weightings	4		b9	b9		b10	b9		b9
Pressey, S. L., — Mental Survey Scale			8½	b9	7	b10	b9	b9	b9
Theisen, W. W., — Fleming, — Classification Test	3		b9			b10		b9	b9
Thorndike, E. L., — Standard Group Examination of Intelligence Independent of Language	4		b9	b9[2]		b10		b9	b9
Thorndike, E. L., — Visual Vocabulary			b9	b9	b9	5[3]	b9	b9	b9
Completion Exercises Alpha and Beta (Trabue, M. R. and Kelley, T. L.)			b9	b9	b9	b10	5	5	b9
Completion Test Language Scales (Trabue, M. R.)			b9	b9	b9	b10	b9	b9	b9
Wylie, A. T., — Opposites Test			b9(½)	b9	b9	b10[3]	b9	b9	b9
Tests known by a single judge									
Chelsea Mental Tests (Ballard, P. B.)	1								
Northumberlin Mental Tests (Thomson, G. H.)	1								
No. having individual value		8	12	5	3	13	7	2	7
No. having group but not individual value		5	5	9	17	6	7	9	7
No. of doubtful value		9	8	10	0	5	5	11	8
No. not reported upon		6	3	4	8	4	9	6	6

[1] Judge C states: "Would rate high in a non-verbal classification."
[2] Judge C states: "Would rate high in a non-verbal classification."
[3] Judge E states: "I believe that the Wylie and the Thorndike Visual Vocabulary should be judged on a basis of three or four forms, since that many may be given in the same time and at the same expense as one form of the tests with which they are competing."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 00:59 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*c*) JUNIOR HIGH SCHOOL GENERAL INTELLIGENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Terman, L. M., — Group Test of Mental Ability		1½	3	1	1	1	1	1	1
Otis, A. S., — Group Intelligence Scale		1½	1	3	2	6	7	2	2
National Intelligence Test (Parts A and B)		4	2	7	5½	5	4	4	4
Miller, W. S., — Mental Ability Test		6	4	2	b9	4(½)	3	5	5
Haggerty, M. E., — Intelligence Examination δ 2		7	5	5	3	2	6	b9	5
Otis, A. S., — Self-Administering Tests of Mental Ability: Intermediate Examination		3	4	6	5½	7	b9	3	5½
Completion Exercises Alpha and Beta. (Trabue, M. R. and Kelley, T. L.)			b8	8	8	b8	5	6	8
Mentimeters School Group 2 A (Trabue, M. R.)			b8	4	7	b8	2	b9	8
Army Alpha		5	7	b9	b9	3	9	9	9

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:00 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Junior and Senior High School Classification Test. (Chapman and Wells)	3	b7	b8				8($\frac{1}{2}$)		b8
Chicago Group Intelligence Test (Freeman, F. N. and Rugg, H. O.)	3		b8	b9		b8			b8
Morgan, J. J. B., — Mental Test			b8	b9	b9	b8		7	b8
Meyers, G. C., — Mental Measure		b7	b8	b9	b9	b8	b9	b9	b8
Peters, C. C., — Test of General Information with Sociologically Determined Weightings	4		b8	b9		b8	b9		b8
Pressey, S. L., — Mental Survey Scale		b7	8	b9	4	b8	b9($\frac{1}{2}$)	b9	b8
Theisen, W. W., — Fleming, — Classification Test	3		b8			b8		b9	b8
Thorndike, E. L., — Standard Group Examination of Intelligence Independent of Language	4		b8	9		b8		8	b8
Completion Test Language Scales (Trabue, M. R.)			b8	b9	9	b8	b9	b9	b8
Wylie, A. T., — Opposites Test				b9($\frac{1}{2}$)	b9	8	b9	b9	b8
No. having individual value		7	9	4	4	8	5	2	5
No. having group but not individual value		3	4	8	10	2	9	9	7
No. of doubtful value		0	5	5	0	8	1	5	4
No. not reported upon		9	1	2	5	1	4	3	3

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:00 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*d*) HIGH SCHOOL GENERAL INTELLIGENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Terman, L. M., — Group Test of Mental Ability		3½	4	1	1	2	5	1	2
Otis, A. S., — Group Intelligence Scale		3½	b7	3	2	b7	8	2	3½
Thorndike, E. L., — Intelligence Examination for High School Graduates		1	1	4	b8	b7	1	b8	4
Otis, A. S., — Self-Administering Tests of Mental Ability: Advanced Examination		3½	3	6	4	b7(½)	b8	3	4
Thurstone, L. L., — Psychological Examination		3½	2	5	b8	b7	3	b8	5
Miller, W. S., — Mental Ability Test		8	6	2	6	5	7	4	6
Haggerty, M. E., — Intelligence Examination δ 2		6½	b7	b8	8	3	6	b8	8
Completion Exercises Alpha and Beta (Trabue, M. R. and Kelley, T. L.)			b7	7	3	b7	b8	5	8
Mentimeters School Group 2 A (Trabue, M. R.)			b7	8	b8	7	4	b8	8
Army Alpha		6½	b7	b8	b8	4	b8	8	b7
Junior and Senior High School Classification Test (Chapman and Wells)	2		b7				b8		b7

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:00 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Brown University Psychological Examination (Colvin, S. S.)			5	b8	b8	b7	2	b8	b7
Morgan, J. J. B., — Mental Test	4		b7	b8	b8			6	b7
Meyers, G. C., — Mental Measure		b8	b7	b8	b8	b7	b8	7	b8
Otis, A. S., — General Intelligence Examination for Business Institutions		b8		b8	b8	6		b8	b8
Peters, C. C., — Test of General Information with Sociologically Determined Weightings	4		b7	b8		b7	b8($\frac{1}{2}$)		b7
Pressey, S. L.,— Mental Survey Scale			b7	b8	b8	b7	b8	b8	b8
Theisen, W. W.,— Fleming — Classification Test	3		b7			b7		b8	b7
Institute of Educational Research Tests of Selective Rational Thinking. (Thorndike, E. L.)			7	b8	b8	1		b8	b8
Completion Test Language Scale. (Trabue, M. R.)			b7	b8	5	b7	b8	b8	b7
Wylie, A. T., — Opposites Test				b8($\frac{1}{2}$)	7	b7	b8	b8	b8
Test known by a single judge									
General Intelligence Test (Spearman, C.)	1								
No. having individnal value		9	8	4	4	9	7	3	7
No. having group but not individual value		1	7	10	14	5	7	4	7
No. of doubtful value		0	4	5	0	5	3	12	5
No. not reported upon		12	3	3	4	3	5	3	3

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:00 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(e) COLLEGE GENERAL INTELLIGENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thorndike, E. L., — Intelligence Examination for High School Graduates		1	1	1	1	1	1	1	1
Psychological Examination for College Freshmen (Thurstone, L. L.)		3	3	2	2	3	3(½)	2	3
Brown University Psychological Examination (Colvin, S. S.)		b7	2	3	3	2	2	4	3
Terman, L. M., — Group Test of Mental Ability		6	b8	4	5	4	6	5	5
Otis, A. S., — Self-Administering Tests of Mental Ability: Advanced Examination		4	6	6	4	b7	b8	3	6
Otis, A. S., — Group Intelligence Scale		5	4	7½	6	6	8	7	6
Roback, A. A., — Mentality Tests	3	2	7					b8	7
Completion Exercises Alpha and Beta (Trabue, M. R. and Kelley, T. L.)			b8	7½	7	b7	5	8	7½
Army Alpha		7	8	b8	b8	5	b8	b8	b8
Miller, W. S., — Mental Ability Test		b7	b8	5	b8	b7	7	6	b7
Meyers, G. C., — Mental Measurement		b7	b8	b8	b8	b7	b8	b8	b8
Thorndike E. L., — Standard Group Examination of Intelligence Independent of Language		b7	5	b8		b7		b8	b7

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Mentimeters School Group 2 A (Trabue M. R.)			b8	b8	b8	b7	4	b8	b8
Wylie, A. T., — Opposites Test . .				b8($\frac{1}{2}$)	8	7[1]	b8	b8	b8
Test known by a single judge . . .									
General Intelligence Examination (Spearman, C.)	1								
No. having individual value . . .		8	7	1	8	8	5	5	5
No. having group but not individual value		2	4	10	10	2	5	3	5
No. of doubtful value		2	3	3	0	4	3	6	4
No. not reported upon		3	1	1	2	1	2	1	1

[1] Judge E states: "I believe that the Wylie and the Thorndike Visual Vocabulary should be judged on a basis of three or four forms, since that many may be given in the same time and at the same expense as one form of the tests with which they are competing."

(f) PRIMARY ACHIEVEMENT BATTERIES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Pressey, L. C., — Scale of Attainment No. 1	4	ind		doubt		gr		gr	gr[1]

[1] "Ind." in this column indicates that the median opinion of the judges was that the test has "individual value"; "gr" that it has "group value"; and "doubt," that its value is doubtful.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*g*) ELEMENTARY ACHIEVEMENT BATTERIES [1]

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Achievement Test (Kelley, T.L., Ruch, G. M., Terman, L. M.)		1	1	1	1	1	1	1	1
Otis, A. S., — Classification Test . .		2	2	3	2	b3	b4	2	2
Illinois Examination (Buckingham, B. R. and Monroe, W. S.) . . .		3	4	2	3	b3	2	3	3
Classroom Products Survey Test (Chapman, J. C.)		4		4	4	2	4	4	4
Courtis, S. A., — Standard Tests, Reading, Writing, and Arithmetic			b4	b4	b4	b3	3	b4	b4
New Jersey Composite Test, Achievement Section			3	b4	b4		b4	b4	b4
Pressey, L. C., — 2d Grade Attainment Scale	4			b4	b4	b3		b4	b4
Pressey, L. C., — 3d Grade Attainment Scale	4			b4	b4	b3		b4	b4
Mentimeters School Group 2 A (Trabue, M. R.)			b4	b4	b4	b3	b4	b4	b4
No. having individual value . . .		4	4	3	1	2	1	2	2
No. having group but not individual value		0	2	3	3	1	4	7	4
No. of doubtful value		0	0	3	0	5	2	0	2
No. not reported upon		5	3	0	0	1	2	0	1

[1] One of the judges constructed a battery composed of Thorndike-McCall Reading, Woody-McCall Arithmetic, and Morrison-McCall Spelling and ranked it 1. He then stated: "I ranked the Stanford Achievement Test 2 because it is correlated too highly with intelligence, no matter what the level of instruction has been. These 'achievement batteries' are not usefully distinguished from 'general intelligence.' Of course I would need to and can present evidence to show what I mean by this."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:01 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*h*) JUNIOR HIGH SCHOOL ACHIEVEMENT BATTERIES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Achievement Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)		1	1	1	1	1	1	1	1
Otis, A. S., — Classification Test . .		2		3	2		2	2	2
Mentimeters School Group 2 A (Trabue, M. R.)		[1]	2	2	3	2	3	3	$2\frac{1}{2}$
No. having individual value . . .		1	1	3	2	1	1	1	1
No. having group but not individual value		1	1	0	1	0	1	1	1
No. of doubtful value		0	0	0	0	1	1	1	1
No. not reported upon		1	1	0	0	1	0	0	0

[1] Judge A states: "Trabue Mentimeters are not achievement."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:01 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*i*) HIGH SCHOOL ACHIEVEMENT BATTERIES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A[1]	B	C	D	E	F	G	
High School Content Examination (Iowa Entrance Examination, Ruch, G. M.)			1	1	1	1	1	1	1
Vocational Guidance Tests (Consists of High School Arithmetic, Geometry, Physics, and Technical Information Tests.) (Thurstone, L. L.)			b1	2	2	2	1	2	2
Mentimeters School Group 2 A (Trabue, M. R.)			b1	3	3	3	2	3	3
No. having individual value		2		3	2	2	0	1	2
No. having group but not individual value		0		0	1	0	1	0	0
No. of doubtful value		1		0	0	1	2	2	1
No. not reported upon		0		0	0	0	0	0	0

[1] Judge A reports inability to compare high school achievement batteries.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:01 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(7) PRIMARY READING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Haggerty Reading Examination Sigma 1. (Haggerty, M. E. and Noonan, M.)		1		1	1	1[1]	2	3	1
Detroit Group Test in Word Recognition (Oglesby, E.)		2		2	2	2	1	1	2
Pressey, L. W., — First Grade Attainment Scale in Reading		3		3	3	3	3(½)	2	3
No. having individual value		3		1	0	3	1	1	1
No. having group but not individual value		0		2	3	0	2	2	2
No. of doubtful value		0		0	0	0	0	0	0
No. not reported upon		0		0	0	0	0	0	0

[1] Judge E states: "Attempts at validating these tests have convinced me that they measure different combinations of qualities which have been grouped under the abstraction 'reading.' The Thorndike-McCall is the only one that measures what I mean by reading."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:01 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*k*) ELEMENTARY READING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thorndike, E. L., and McCall, W. A., — Reading Scales				4	1	1[1]	7	2	2
Stanford Reading Test (Kelley, T. L., Ruch, G. M., and Terman, L. M.)				1	2	2	2	5	2
Thorndike, E. L., — Test of Word Knowledge				2	9	5	1	7	5
Haggerty, M. E., — Reading Examination Sigma 3				5	3	4	8	8	5
Chapman, J. C., — Cook, S. A., — Speed of Reading Test	3				5		6	3	5
Monroe, W. S., — Standardized Silent Reading Tests, Revised				3	4	b9	b9	6	6
Thorndike, E. L., — Visual Vocabulary				6	b9	6	5	b9	6
Gray, W. S., — Silent Reading Test				9	b9	7	4	1	7
Burgess, M. A., — Reading Test				b9	6	8	b9	4	8
Completion Test Language Scales Alpha and Beta. (Trabue, M. R. and Kelley, T. L.)	4			7		b9	3	b9	8½
Courtis, S. A., — Silent Reading Test No. 2				b9	7	3	9	b9	9
Fordyce, C., — Scale for Measuring Ability in Silent Reading	3				b9	9		9	9

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Brown, H. A., — Silent Reading Test	3				b9	b9		b9	b9
Holley, C. E., — Sentence Vocabulary Test	4			b9(½)	b9	b9		b9	b9
Kansas Silent Reading Test. (Kelly, F. J.)				b9	8	b9	b9	b9	b9
Completion Test Language Scales (Trabue, M. R.)	4			8		b9	b9	b9	b9
Witham, E. C., — Silent Reading Tests 1 and 2	2			b9				b9	b9
Witham, E. C., — English Vocabulary Test 1 and 2	2			b9				b9	b9
No. having individual value		11		3	2	6	4	4	4
No. having group but not individual value		3		8	12	3	7	9	8
No. of doubtful value		4		4	0	6	2	5	3
No. not reported upon		0		3	4	3	5	0	3

[1] See comment of Judge E in footnote of preceding section.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:02 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*l*) JUNIOR HIGH SCHOOL READING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thorndike, E. L. and McCall, W. A., — Reading Scales				4	1	1	5	1	1
Stanford Reading Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)				1	2	3	2	2	2
Van Wagenen, M. J., — Reading Scales A, B, and C				3	4	2	4(½)	3	3
Haggerty, M. E., — Reading Examination Sigma 3				5	3	4	b5	4	4
Holley, C. E., — Sentence Vocabulary Test	4			b5(½)	b4	b4		b4	b4
Thorndike, E. L., — Test of Word Knowledge				2	b4	b4	1	b4	b4
Completion Test Language Scales (Trabue, M. R.)				b5	b4	b4	b5	b4	b4
Completion Test Language Scales Alpha and Beta (Trabue, M. R. and Kelley, T. L.)				b5	b4	b4	3	b4	b4
Witham, E. C., — Silent Reading Test 1 and 2	2			b5				b4	b4
Witham, E.C., — English Vocabulary Test 1 and 2	2			b5				b4	b4
No. having individual value		5		3	2	4	3	3	3
No. having group but not individual value		2		4	6	1	3	5	4
No. of doubtful value		3		3	0	3	1	2	2
No. not reported upon		0		0	2	2	3	0	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:02 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:02 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*m*) HIGH SCHOOL READING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Van Wagenen, M. J., — Reading Scales A, B, and C				1	2	2	4(½)	2	2
Thorndike, E. L., and McCall, W. A., — Reading Scales				2	3	1	5	1	2
Inglis, A., — Vocabulary Test				3	b5	3[1]	1	b4	3
Haggerty, M. E., — Reading Examination Sigma 3				4	1	5	b5	3	4
Thorndike, E. L., — Test of Word Knowledge				b5	5	4	2	b4	5
Monroe, W. S., — Standard Silent Reading Test				5	4	b5	b5	4	5
Holley, C. E., — Sentence Vocabulary Test	4			b5(½)	b5	b5		b4	b5
Completion Test Language Scales (Trabue, M. R.)				b5	b5	b5	b5	b4	b5
Completion Test Language Scales Alpha and Beta (Trabue, M. R. and Kelley, T. L.)				b5	b5	b5	3	b4	b5
Witham, E. C., — Silent Reading Test 1 and 2	2			b5				b4	b4
Witham E., C., — English Vocabulary Test 1 and 2	2			b5				b4	b4

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:03 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

No. having individual value . . .		5		1	0	5	4	3	3½
No. having group but not individual value		3		7	9	0	4	4	4½
No. of doubtful value		3		3	0	4	0	4	2
No. not reported upon		0		0	2	2	3	0	1

[1] Judge E states: "I find it hard to rank vocabulary and reading tests together. The Inglis stands higher in my regard than this rank would show, but not as a measure of reading."

(*n*) COLLEGE READING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Inglis, A., — Vocabulary Test . . .				1	3	2	1	1	1
Completion Test Language Scales Alpha and Beta (Trabue, M. R. and Kelley, T. L.)				2	1	1	3	2	2
Thorndike, E. L., — Test of Word Knowledge				3	2	3	2	3	3
No. having individual value . . .				0	1	3	2	0	1
No. having group but not individual value				3	2	0	1	0	2
No. of doubtful value				0	0	0	0	3	0
No. not reported upon				0	0	0	0	0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:03 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*o*) ELEMENTARY READING TESTS, ORAL

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Gates, A. I., — Graded Word Knowledge Test	4			1	2	1		2	1½
New Standardized Oral Reading Check Test (Gray, W. S.)	4			2	1	2		1	1½
No. having individual value				1	2	2		1	1½
No. having group but not individual value				1	0	0		1	½
No. of doubtful value				0	0	0		0	0
No. not reported upon				0	0	0		0	0

(*p*) ELEMENTARY LITERATURE APPRECIATION TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford History and Literature Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)		ind		ind	gr	ind		ind	ind

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:03 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*q*) JUNIOR HIGH SCHOOL LITERATURE APPRECIATION TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford History and Literature Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.) . .		ind		ind	gr	ind		ind	ind
Van Wagenen, M. J., — Reading Scales A, B, and C, English Literature		ind		gr	gr	ind	gr	ind	gr to ind

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:20 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*r*) HIGH SCHOOL LITERATURE APPRECIATION TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford History and Literature Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.) . .	3			ind	1	1			1
Van Wagenen, M. J., — Reading Scales A, B, and C, English Literature	3			gr	2	2			2
Abbott, A., and Trabue, M. R., — Exercises in Judging English Poetry		gr		gr	3	3[1]	gr	gr	3
No. having individual value . . .		2		1		2	0	2	2
No. having group but not individual value		1		2		1	1	1	1
No. of doubtful value		0		0		0	2	0	0
No. not reported upon		0		0		0	0	0	0

[1] Judge E states: "I marked the Abbott-Trabue 3 because of its unreliability in its present form. I believe it has very great possibilities."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:03 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(s) ELEMENTARY AND JUNIOR HIGH SCHOOL COMPOSITION SCALES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Hudelson, E., — English Comp. Scale				2	1	b4	1	1	1
Nassau County Supplement to Hillegas Scale (Trabue, M. R.)				3	3	1	4	2	3
Lewis, E. E., — English Comp. Scales				1	4	3	5	4	4
Hudelson, E., — Typical Composition Ability Scale				4	2(½)	b4	2	b5	4
Thorndike, E. L., — Extension of the Hillegas Scale				5	b5	2	3	5	5
Van Wagenen, M. J., — English Composition Scales	4			b5	5		b5	3	5½
Harvard-Newton Composition Scale (Ballou, F. W.)				b5	b5	4	b5	b5	b5
Breed, F. S., and Frostic, F. W., — Composition Scales	4			b5	b5		b5	b5	b5
Hillegas, M. B., — Composition Scale				b5	b5	b4	b5	b5	b5
Willing, M. H., — Scale for Measuring Composition				b5	b5	b4	b5	b5	b5
No. having individual value [1]		7		0	0	0	0	0	0
No. having group but not ind. value		1		8	10	8	7	5 [2]	8
No. of doubtful value		2		2	0	0	3	5	2
No. not reported upon		0		0	0	2	0	0	0

[1] Judges were asked to give opinion as to "ind" or "gr" value upon assumption that one composition (written in 20 or 30 minutes) is rated by one teacher.

[2] Judge G states: "Of value only when teacher has had practice or training in use of scales and is free from constant error (constant error would render group value of little consequence). Not less than two samples should be graded for each pupil and by not less than two judges."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:04 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*t*) HIGH SCHOOL COMPOSITION SCALES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Hudelson, E., — English Composition Scale				2	1	4	1	4	2
Nassau County Supplement to the Hillegas Scale (Trabue, M. R.)				3	3	1	4	b5	3
Thorndike, E. L., — Extension of the Hillegas Scale				b5	5	2	3	1	3
Lewis, E. E., — English Composition Scales				1	4	3	5	3	3
Hudelson, E., — Typical Composition Ability Scale				4	2(¼)	b4	2	5	4
Van Wagenen, M. J., — English Composition Scales	4			5	b5		b5	2	5¼
Hillegas, M. B., — Composition Scale				b5	b5	b4	b5	b5	b5
Willing, M. H., — Scale for Measuring Composition				b5	b5	b4	b5	b5	b5
No. having individual value[1]		8		0	0	0	0	0	0
No. having group but not individual value		0		8	8	7	7	5	7½
No. of doubtful value		0		0	0	1	1	3	½
No. not reported upon		0		0	0	0	0	0	0

[1] See footnotes of preceding section.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:04 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*u*) ELEMENTARY SPELLING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Morrison, J. C., — McCall, W. A., — Spelling Scale				2	1	1	5	1	1
Iowa Spelling Scales (Ashbaugh, E. J.)				1	3	5	2	3	3
Stanford Dictation Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)				3	2	3	4	5	3
Sixteen Spelling Scales (Briggs, T. H., et al.)				4	7	2	6	2	4
Buckingham, B. R., — Extension of Ayres Scale				5	4	4	1	3	4
Ayres, L. P., — Spelling Scale				6	5[1]	6	3	4	5
Tidyman, W.F., — Standard Spel. Tests	2			b8				6	7½
Courtis, S. A., — Standard Supervisory Tests in Spelling				b8	6	8	8(½)	b8	8
Monroe, W. S., — Timed Sentence Spelling Test				7	b8	9	b8	7	9
Courtis, S. A., — Standard Research Test in Spelling				b8	b8	7	7	b8	b8
Nebraska Spelling Scale (Fordyce, C.)	3				b8		b8(½)	b8	b8
100 Spelling Demons (Jones, N. F.)				b8	b8	b9	b8	b8	b8
Starch, D., — Spelling Lists				8	8	b9	b8	b8	b8
No. having individual value		3		5	6	6	6	5	5½
No. having group but not ind. value		5		6	6	5	4	5	5
No. of doubtful value		5		1	0	1	2	3	1½
No. not reported upon		0		1	1	1	1	0	1

[1] Judge D states: "Not in appropriate form for testing."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*v*) JUNIOR HIGH SCHOOL SPELLING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Sixteen Spelling Scales (Briggs, T. H., et al.)				1	1	1	1	2	1
Stanford Dictation Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)				2	3	2	2	3	2
Morrison, J. C., — McCall, W. A., — Spelling Scale				3	2	3	4½	1	3
Monroe, W. S., — Timed Sentence Spelling Test				4	4	4	3	4	4
Nebraska Spelling Scales (Fordyce, C.)	3				b4		4½	b5	b4
Starch, D., — Spelling Lists				5	b4	5	b5	b5	b4
Tidyman, W. F., — Standard Spelling Tests	2			b5				5	b4
No. having individual value		2		3	4	3	2	3	3
No. having group but not individual value		2		3	2	2	2	3	2½
No. of doubtful value		3		0	0	0	2	1	½
No. not reported upon		0		1	1	2	1	0	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:04 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*w*) HIGH SCHOOL SPELLING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Sixteen Spelling Scales (Briggs, T. H., et al.)				1	1	1	1	1	1
Stanford Dictation Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)				2	2	2	2	2	2
Monroe, W. S., — Timed Sentence Spelling Test				b2	b2	b2	b2	b2	b2
Starch, D., — Spelling Lists				b2	b2	b2	b2	b2	b2
No. having individual value		2		2	1	2	2	0	2
No. having group but not individual value		0		2	3	2	1	0	$1\frac{1}{2}$
No. of doubtful value		2		0	0	0	1	4	$\frac{1}{4}$
No. not reported upon		0		0	0	0	0	0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*x*) ELEMENTARY LANGUAGE USAGE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Language Usage (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			1	1	1(½)[1]		3	1
Charters, W. W., — Diagnostic Language Test				3	2	2(½)	3	1	2
Charters, W. W., — Diagnostic Language and Grammar Test				2	3	3(½)	2	2	2
Kirby, T. J., — Grammar Test	2					4(½)		4	4
Wilson, G. M., — Language Error Test	4			4	4(½)		1	b7	4
Pressey, S. L., — Ruhlen, H., — Diagnostic Tests in English Composition (punctuation)	3			6	6			6[2]	6
Pressey, S. L., — Bowers, E. V., — Diagnostic Tests in English Composition (capitalization)	3			6	6			6[2]	6
Pressey, S. L., — Conkling, F. R., — Diagnostic Tests in English Composition (grammar or inflected forms)	3			6	6			6[2]	6

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:21 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Pressey, S. L., — Diagnostic Tests in English Composition (vocabulary, grammar, and punctuation). Early forms				b7	b7		b6	b7	b7
Pressey, S. L., — Conkling, F. R., — Diagnostic Tests in English Composition (sentence structure)				b7	5	b5		b7	b6
Starch, D., — Grammatical Scale A				b7	b7	b5	4	b7	b7
Starch, D., — Punctuation Scale				b7	b7	5(½)	5	b7	b7
Clapp, F. L., — Standardized School Tests in Correct English	1								
No. having individual value		9		3	3	0	3	2	3
No. having group but not individual value		3		6	8	5	0	3	6½
No. of doubtful value		1		2	0	3	3	2	1½
No. not reported upon		0		2	2	5	7	1	2

[1] Judge E states: "My familiarity with these tests has been limited by a strong conviction that we do not wish to test proof-reading ability."

[2] Judge G states: "Of individual value if used as a group of tests."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:05 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*y*) JUNIOR HIGH SCHOOL LANGUAGE USAGE AND GRAMMAR TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Charters, W. W., — Diagnostic Language and Grammar Test				2	1	3(½)	3	1	2
Stanford Language Usage Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			1	4	2(½)		3	2½
Wilson, G. M., — Language Error Test				4	3	b5	1	5	4
Charters, W. W., — Diagnostic Language Test				5	2	b5	4	2	4
Briggs, T. H., — English Form Test				b5	5	1	2	b5	5
Briggs, T. H., — Analogies Test	4			3	b5		5 [1]	b5	5½
Diagnostic Tests in English Composition (Pressey, S. L., Ruhlen, H., Conkling, F. R., Bowers, E. V.)	3			b5	b5			4	b5
Starch, D., — English Grammar Test	3				b5	5(½)		b5	b5
Starch, D., — Grammatical Scale A				b5	b5	b5	b5	b5	b5
Starch, D., — Punctuation Scale				b5	b5	4(½)	b5	b5	b5
No. having individual value		5		3	3	2	5	2	3
No. having group but not individual value		4		6	7	3	0	3	4
No. of doubtful value		1		0	0	3	2	5	2½
No. not reported upon		0		1	0	2	3	0	½

[1] Judge F states: "The results of this test are doubtless more significant than the results of the others, but it is not called 'Language and Grammar.'"

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:05 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(z) HIGH SCHOOL LANGUAGE USAGE AND GRAMMAR TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Wilson, G. M., — Language Error Test	4			1	1		1	1	1
Starch, D., — English Gramar Test	3				2	2		b1	2
Starch, D., — Grammatical Scale A				2	3	3	2	b1	2½
Starch, D., — Punctuation Scale				3	b3	1	3	b1	3
No. having individual value		3		0	1	0	1	0	½
No. having group but not individual value		0		3	3	1	0	1	1½
No. of doubtful value		1		0	0	2	2	3	1½
No. not reported upon		0		1	0	1	1	0	½

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:05 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*aa*) ELEMENTARY ENGLISH FORM TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Pressey, S. L., Conkling, F. R., — Diagnostic Tests in English Composition (sentence structure) . .	2			gr				ind [1]	

[1] Judge G states: "Individual value if used as a group of tests."

(*bb*) JUNIOR HIGH SCHOOL ENGLISH FORM TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Briggs, T. H., — English Form Test				1	1	1	1	2	1
Diagnostic Tests in English Composition, — Language Usage and Grammar (Pressey, S. L., Ruhlen, H., Conkling, F. R., Bowers, E. V.)	3			2	2			1	2
Starch, D., — Punctuation Scale . .	4			b2		2	2	b2	b2
No. having individual value . . .	2			0	1	1	1	1	1
No. having group but not individual value	0			3	1	1	0	1	1
No. of doubtful value	1			0	0	0	1	1	1
No. not reported upon	0			0	1	1	1	0	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*cc*) HIGH SCHOOL ENGLISH FORM TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Briggs, T. H., — English Form Test				1	1	1	1	1	1
Starch, D., — Punctuation Scale . .				2	2	2	2	2	2
No. having individual value . . .		1		0	1	1	1	0	1
No. having group but not individual value		0		2	1	1	0	0	½
No. of doubtful value		1		0	0	0	1	2	½
No. not reported upon		0		0	0	0	0	0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*dd*) ELEMENTARY ARITHMETIC TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Arithmetic Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4	1		2	2	3		7	2½
Buckingham, B. R., — Scale for Problems in Arithmetic				5	3	5	2	2	3
Woody, C., — Arithmetic Scales				7½	1	2	3	6	3
Woody, C., — McCall, W. A., — Mixed Fundamentals				4	4	1	5	b11	4
Woody, C., — Van Wagenen, M. J., — Arithmetic Scales	3			9		4		4	4
Monroe, W. S., — Diagnostic Arithmetic Test				3	5	7	6	9	6
Spencer, P. L., — Diagnostic Arithmetic Test	2			1				b11	6½
Cleveland Survey Arithmetic Test (Judd, C. H. et al.)				b11	7	8	1	1	7
Stevenson, P. R., — Arithmetic Problem Analysis Test	3				b11		4	8	8
Otis, A. S., — Arithmetic Reasoning Test				b11	9	b11	7	3	9
Monroe, W. S., — Standardized Reasoning Tests in Arithmetic				6	6	b11	9	10	9
Courtis, S. A., — Standard Research Tests, Arithmetic, Series B				b11	b11	9	8	5	9
Monroe, W. S., — General Survey Arithmetic Test	4			10	8	6		b11	9
Peet, H. E., — Dearborn, W. F., — Progress Tests in Arithmetic	3			11			11	11	11

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Clapp, F. L., — Number Combinations	2			b11				b11	b11
Courtis, S. A., — Standard Practice Tests in Arithmetic	4			b11	b11[2]	11		b11	b11
Rochester Attainments in Arithmetic Chart	3			b11		b11	b11[3]		b11
Starch, D., — Arithmetical Scale A				b11	b11	b11	b11	b11	b11
Stone, C. W., — Reasoning Test				b11	11	b11	10	b11	b11
Theisen, W. W., — Woody, C., — Parallal Tests	3			7½		b11		b11	b11
Thompson, T. E., — Minimal Essentials in Arithmetic				b11	b11	b11	b11	b11	b11
Witham, E. C., — Standardized Arithmetic Tests	2						b11	b11	b11
Pittsbury Arithmetic Scale (Guy, J. F.)	1								
Ruch, G. M., — Knight, F. B., — Problem Scales	1								
No. having individual value		16		3	4	5	8	5	5
No. having group but not individual value		2		14	11	9	6	16	11
No. of doubtful value		6		3	0	3	1	0	2½
No. not reported upon		0		4	9	7	9	3	5½

[1] Judge A states: "You have reasoning and fundamentals mixed here. Cannot well be compared."
[2] Judge D states: "These are the best teaching tests of all. Should have a division on practice tests."
[3] Judge F states: "This is no measure in the sense the others are."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*ee*) JUNIOR HIGH SCHOOL ARITHMETIC TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Arithmetic Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			1	2	1		3	1½
Buckingham, B. R., — Scale for Problems in Arithmetic				3	1	3	2	1	2
Otis, A. S., — Arithmetic Reasoning Test				2	3	2	3	2	2
Stevenson, P. R., — Arithmetic Problem Analysis Test	3				b3		1	b3	b3
Witham, E. C., — Standardized Arithmetic Tests	2						b3	b3	b3
No. having individual value	5			1	1	1	3	3	2
No. having group but not individual value	0			1	3	2	1	2	2
No. of doubtful value	0			1	0	0	0	0	0
No. not reported upon	0			2	1	2	1	0	1

(*ff*) HIGH SCHOOL AND COLLEGE ARITHMETIC TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thurstone, L. L., — Arithmetic Test	3			gr	gr	ind			gr

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:06 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*gg*) JUNIOR HIGH SCHOOL ALGEBRA TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Rogers, A. L., — Test of Mathemetical Ability	4			ind	ind	doubt		ind	ind

(*hh*) HIGH SCHOOL ALGEBRA TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Hotz, H. G., — Algebra Scales . .				2	1	1	1	2	1
Douglas, H. R., — Diagnostic Tests for 1st year Algebra				1	2	2	2	3	2
Thurstone, L. L., — Algebra Test . .	3			3		3	3		3
Rugg, H. A., — Clark, J. R., — Standardized Tests in 1st year Algebra				5	4	5	4	1	4
Kelley, T. L., — Mathematical Values Test	4			[1]	5	4	5	b5	5
Illinois Standardized Algebra Tests (Monroe, W. S., and Williams, L. W.)				4	3	b5	b5	5	5
Coleman, W. H., — Scale in Algebra	1								
No. having individual value . . .		5		1	3	3	1	5	3
No. having group but not individual value		0		5	2	3	5	0	3
No. of doubtful value		2		0	0	0	0	0	0
No. not reported upon		0		1	2	1	1	2	1

[1] Judge C states: "Considered not comparable to rest in this classification and therefore not ranked."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*ii*) COLLEGE ALGEBRA TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thurstone, L. L., — Algebra Test . .	3			gr	gr	ind			gr

(*jj*) HIGH SCHOOL GEOMETRY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Columbia Geometry Test (Hawkes, H. E., and Wood, B. D.)				1	2($\frac{1}{2}$)	1	1	3	1
Minnick, J. H., — Geometry Test .				2	1	3	b3($\frac{1}{2}$)	1	2
Thurstone, L. L., — Geometry Test (Part of Vocational Guidance Test)	3			3	3($\frac{1}{2}$)	2			3
Schorling, R., — Plane Geometry Test	4				b3	b3	2	2	3
Stockard, L. V., — Bell, J. C., — Geometry Test					b3	b3	b3	3($\frac{1}{2}$)	b3	b3
No. having individual value . . .		4		1	2	2	2	0	2
No. having group but not individual value		0		2	3	2	2	3	2
No. of doubtful value		1		1	0	1	0	1	$\frac{1}{2}$
No. not reported upon		0		1	0	0	1	1	$\frac{1}{2}$

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*kk*) COLLEGE GEOMETRY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Columbia Geometry Test (Hawkes, H. E., and Wood, B. D.)				1	1(½)	1	1	1	1
Thurstone, L. L., — Geometry Test	3			2	2(½)	2			2
No. having individual value		2		1	1	2		1	1
No. having group but not individual value		0		1	1	0		0	1
No. of doubtful value		0		0	0	0		0	0
No. not reported upon		0		0	0	0	1	1	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*ll*) ELEMENTARY AND JUNIOR HIGH SCHOOL GEOGRAPHY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Posey, C. J., — Van Wagenen, M. J., — Geography Scales	4			1	2	4		1	1½
Spencer, P. L., — Gregory, C. A., — Geography Test	4			2	1	1		4	1½
Buckingham, B. R., — Stevenson, P. R., — U. S. Geography Information and Problems				4	3	3	1	2	3
Buckingham, B. R., — Stevenson, P. R., — Place Geography Test				5	4	2	2	3	3
Hahn, H. H., Lackey, E. E., — Geography Scale				6	6	5	3	5	5
New York Standard Geography Tests (Nifenecker, E. A.)	3				5	6		b6	6
Witham, E. C., — Geography Tests	3				b6		5	6	6
Courtis, S. A., — Geography Test	4			b6	b6		4	b6	b6
Whittier Geography Scale A	3			3	b6			b6	b6
Wisconsin Geography Test (Hinterberg, E.)	1								
Olmsted, M. C., — Diagnostic Geography Tests	1								
No. having individual value		6		3	3	1	2	4	3
No. having group but not individual value		2		4	6	3	1	3	4
No. of doubtful value		3		0	0	2	2	2	1
No. not reported upon		0		4	2	5	6	2	3

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:07 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*mm*) ELEMENTARY GENERAL SCIENCE TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford Science Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)		ind		ind	ind	ind		ind	ind

(*nn*) JUNIOR HIGH SCHOOL GENERAL SCIENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Ruch, G. M., — Popenoe, H. F., — General Science Test				1	2(½)	1	1	2	1
Stanford Science Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			2	1	2		3	2
Dvorak, A. — General Science Scales	2			3				1	2
No. having individual value . . .		2		2	2	2	1	3	2
No. having group but not individual value		1		0	0	0	0	0	0
No. of doubtful value		0		1	0	0	0	0	½
No. not reported upon		0		0	1	1	2	0	½

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*oo*) HIGH SCHOOL GENERAL SCIENCE TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Ruch, G. M., Popenoe, H. F., — General Science Test	4			1	1	1		1	1
Dvorak, A. — General Science Scales	2			2				2	2
No. having individual value		1		1	1	1		2	1
No. having group but not individual value		1		1	0	0		0	1
No. of doubtful value		0		0	0	0		0	0
No. not reported upon		0		0	1	1		0	0

(*pp*) BIOLOGY TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Ruch, G. M., — Cossmann, L., — Biology Test	4			ind	ind	ind		ind	ind

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*qq*) HIGH SCHOOL CHEMISTRY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Powers, S. R., — Test for General Chemistry		2($\frac{1}{2}$)		1($\frac{1}{2}$)	1	1	2	1	1
Glenn, E. R., — Welton, L. E., — New Type of High School Chemistry Tests for Instructional Purposes		1($\frac{1}{2}$)		3($\frac{1}{2}$)	3		1	2	2
Rich, S. G., — Chemistry Test		b3		2($\frac{1}{2}$)	2	2		3	2
Rivett, B. J., — Time Limit Test in Chemistry	3	3($\frac{1}{2}$)				b2		b3	b2
No. having individual value		4		1	1	1	?	1	1
No. having group but not individual value		0		1	2	1	?	2	2
No. of doubtful value		0		1	0	1	?	1	0
No. not reported upon		0		1	1	1	2	0	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Glenn, E. R., — Obourn, E. L., — New Type of High School Physics Tests for Instructional Purposes	4				4	1	1	2	1½
Thurstone, L. L., — Physics Test (Part of Vocational Guidance Tests)				1	2	2	3	b3	2
Iowa Physics Test (Camp, H. L.)	4			2	1	4		3	2½
Chapman, J. C., — Test in Electricity, Magnetism, Sound, Light, Heat, Mechanics				3	3	3	2	1	3
Starch, D., — Physics Test				4	b4	b4	4	b3	b3
No. having individual value		3		0	1	5	2	0	1½
No. having group but not individual value		0		3	4	0	1	4	2
No. of doubtful value		2		1	0	0	1	1	1½
No. not reported upon		0		1	0	0	1	0	0

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stanford History and Literature Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			1	2	1		1	1
Hahn, H. H., — History Scales				2	1	2	1	b2	2
Harlan, C. L., — Information Test in American History				b2	b2		2	2	2½
Boston Research Tests in U. S. History (Penell, O. C.)	1								
No. having individual value		2		1	0	1	0	1	1
No. having group but not individual value		1		1	3	1	2	2	1½
No. of doubtful value		1		1	0	0	0	0	½
No. not reported upon		0		1	1	2	2	1	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:08 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(tt) JUNIOR HIGH SCHOOL AMERICAN HISTORY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Van Wagenen, M. J., — American History Scales				1	1	2	1	1	1
Stanford History and Literature Information Test (Kelley, T. L., Ruch, G. M., Terman, L. M.)	4			2	b3	1		3	2½
Barr, A. S., — Diagnostic Tests in American History	4			b3	2(½)	b3		2	3
Pressey, L. W., — Richards, R. C., — American History Test				b3	3(½)	3	3(½)	b3	3
Bell, J. C., — McCollum, D. F., — U. S. History Test	4			b3	b3	b3		b3	b3
Hahn, H. H., — History Scales				3	b3	b3	2	b3	b3
No. having individual value		5		3	1	1		3	3
No. having group but not individual value		0		2	5	2		3	2
No. of doubtful value		1		1	0	3		0	1
No. not reported upon		0		0	0	0		0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:27 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Barr, A. S., — Diagnostic Tests in American History	4			1	1	2		1	1
Pressey, L. W., — Richards, R. C., — American History Test				2	2	1	1	2	2
No. having individual value		1		0	0	0		1	0
No. having group but not individual value		1		2	2	1		1	2
No. of doubtful value		0		0	0	1		0	0
No. not reported upon		0		0	0	0		0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*vv*) HIGH SCHOOL ANCIENT HISTORY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Institute of Educational Research College Entrance Examination in Ancient History (Wood, B. D.)				1	1	1	1	1	1
Sackett, L. W., — Ancient History Test	4			2	2		2	2	2
Davis, S. B., — Hicks, E. E., — True False Test in Roman History	1								
No. having individual value				1	0	1	1	1	1
No. having group but not individual value				1	2	0	1	1	1
No. of doubtful value				0	0	0	0	0	0
No. not reported upon				1	1	2	1	1	1

(*ww*) HIGH SCHOOL MODERN EUROPEAN HISTORY TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Vannest, C. G., — Diagnostic Test in Modern European History	1								

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*xx*) COLLEGE ANCIENT HISTORY TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Institute of Educational Research College Entrance Examination in Ancient History (Wood, B. D.) .				ind	gr	ind	ind	ind	ind

(*yy*) CITIZENSHIP SCALE

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Chassell, C. F., Upton, S. M., — Citizenship Scales				ind	gr	doubt	gr	ind	gr

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(22) CHARACTER TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Cady, V. M., — Tests of Incorrigibility	4			1	3(½)	2		1	1½
Voelker, P. E., — Character Tests				2	1	b3	1	b1	1½
Downey, J. E., — Will-Temperament Test				b2	2	1	2	b1	2
Pressey, S. L. — X-O Tests for Investigating the Emotions				b2	b3	3	3	b1	b2
No. having individual value				0	0	0	2	1	0
No. having group but not individual value				2	4	2	1	0	2
No. of doubtful value				2	0	2	0	3	2
No. not reported upon				0	0	0	1	0	0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:09 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*aaa*) ELEMENTARY DRAWING SCALES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Carey-Kline Drawing Scales . . .	2				1			1[1]	1
Thorndike, E. L., — Drawing Scale				1($\frac{1}{2}$)	2	1	1	3[1]	1
Child, H. G., — Drawing Scale . .	2			2($\frac{1}{2}$)				2[1]	2
No. having individual value . . .				2	2	0	1	3	2
No. having group but not individual value				0	1	1	0	0	0
No. of doubtful value				0	1	0	0	0	0
No. not reported upon				1	1	2	2	0	1

[1] Judge G states: "Individual value only in case teachers can use scales with small error, without a serious constant error, when several specimens of drawing are collected from each pupil and rating is done by at least two judges."

(*bbb*) JUNIOR HIGH SCHOOL DRAWING SCALE

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thorndike, E. L., — Drawing Scale .	4			ind	ind	gr		ind[1]	ind

[1] See previous footnote.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Ayres, L. P., — Handwriting Scale, — Gettysburg Edition			1	4	3	2	2	1[1]	2
Thorndike, E. L., — Handwriting Scale			3	3	2	1	3	3[1]	3
Freeman, F. N., — Chart for diagnosing faults in handwriting			5	1	b6	3	6	2[2]	4
Kansas City Scale for Measuring Handwriting	2		4				4		4
Frasier, G. W., — Handwriting Test	3			b6			5	4[1]	5
Ayres, L. P. — Handwriting, — Three Slant Edition			2	5	4		b6	b6[1]	5
Starch, D., — Handwriting Scale			b6	6	1	b6	1	5[1]	5½
Handwriting and Measuring Tablets (Clark, F. L., Wells, J. B., and Freeman, F. N.)	2		b6			b6			b6
Courtis, S. A., — Standard Practice Tests in Handwriting			b6	b6	5	b6	b6	b6	b6
Courtis, S. A., — English Test No. 1, — Handwriting	4			b6		5	b6	b6	b6
Gray, C. T., — Standard Score Card for Measuring Handwriting			b6	2	b6	4		b6	b6
New York City Penmanship Scale (Lister, C. C., and Meyers, G. C.)	4		b6		6		b6	6[1]	b6

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Palmer, A. N., — Standard for the Evaluation of Efficiency	3		b6			b6		b6	b6
Zaner Handwriting Scale	4		6		b6	6	b6		b6
No. having individual value . . .		5	11	8	0	0	5	8	5
No. having group but not individual value		0	1	0	9	5	2	0	3
No. of doubtful value		10	0	1	0	5	4	3	3
No. not reported upon		0	2	5	5	4	3	3	3

[1] Judge G states: "Individual value only when teachers can use the scales with small errors, without a serious constant error, when at least two specimens are collected from each pupil, and when ratings are made by at least two judges."
[2] Judge G states: "For diagnostic work only."

(*ddd*) COLLEGE HANDWRITING SCALES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Freeman, F. N., — Chart for Diagnosing Faults in Handwriting . .			3	1	2	1	1	1[2]	1
Ayres, L. P., — Adult Handwriting Scale			1	2	1		2	3[1]	2
Frasier, G. W., — Handwriting Scale			2	3	3(½)		3	2[2]	3
No. having individual value . . .		2	3	3	0	0		2	2
No. having group but not ind. value		0	0	0	3	1		0	½
No. of doubtful value		1	0	0	0	0		1	½
No. not reported upon		0	0	0	0	2		0	0

[1] See preceding note 1.
[2] See preceding note 2.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*eee*) TYPING TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Blackstone, E. G., — Stenographic Proficiency Tests	2			1	2(¼)				1¼
Thurstone, L. L., — Typist Test	2			2	1				1½
Rogers, H. W., — Stenographic and Typist Tests	2			3(¼)	3(¼)				3
No. having individual value		2			2	0			1
No. having group but not individual value		0		1	3	0			1½
No. of doubtful value		1		0	0	3			¼
No. not reported upon		0		0	0	0			0

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*fff*) GENERAL CLERICAL TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Thurstone, L. L., — Clerical Examination	4			1	1	1	1		1
National Business Ability Tests (Cody, S.)	3			2	2		2		2
Ruggles Diagnostic Test of Aptitude for Clerical Office Work	1								
No. having individual value		1		1	0	1	0		1
No. having group but not individual value		0		1	2	0	1		0
No. of doubtful value		1		0	0	0	1		1
No. not reported upon		0		0	0	1	0		0

(*ggg*) JUNIOR HIGH AND HIGH SCHOOL MECHANICAL ABILITY TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stenquist, J. L., — Mechanical Aptitudes Tests	3			gr	gr		ind		gr

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:10 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*hhh*) ELEMENTARY JUNIOR HIGH AND HIGH SCHOOL MUSIC TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Kwalwasser, J., Ruch, G. M., — Test of Musical Accomplishment	2			4	1		[1]		2½
Mosher, R. M., — Sight Reading Music Test	2			3	2				2½
Seashore, C. E., — Rhythm Test	4			1	6½	4	3		3½
Seashore, C. E., — Tonal Memory Test	4			2	6½	1	6		4
Seashore, C. E., — Sense of Intensity Test	4			7	6½	2	2		4½
Seashore, C. E., — Sense of Pitch Test	4			5	6½	3	5		5
Beach, F. A., — Standardized Music Tests	2				3	7			5
Seashore, C. E., — Sense of Time Test	4			6	6½	5	1		5½
Seashore, C. E., — Sense of Consonance Test	4			8	6½	6	4		6½
Hillbrand, E. K., — Sight Singing Test	1								
No. having individual value				3		4		6	4
No. having group but not individual value				5		1		0	2
No. of doubtful value				0		2		4	2
No. not reported upon				2		3		0	2

[1] Judge F states: "What do we know about this anyhow? Better get musicians to judge this."

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:11 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*iii*) SUNDRY: ELEMENTARY SCHOOL TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
The Teachers Word Book (Thorndike, E. L.).	4			ind	ind	ind	ind		ind
Home Economics Information Test (Teachers College)[1]	3			gr	ind		ind		ind
Information Test on Foods (Illinois Home Economic Association) . .	1								

[1] Dr. W. A. McCall states that this test is still in the formative stage.

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:22 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*iii*) SUNDRY: HIGH SCHOOL TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Blackstone, E. G., — Stenographic Proficiency Tests	3	ind		ind	gr($\frac{1}{4}$)				ind
Bureau of Personnel Research, Carnegie Institute of Technology, Vocational Tests: Will Profile, social relations, business information, meeting objections, interest analysis	2			gr	ind($\frac{1}{4}$)				gr
Goodspeed, H., — Dodge, B., — Preliminary Judgment Test in Home-Making	3			ind	gr		ind		ind
Hoke, E., — Prognostic Test of Stenographic Ability	2			ind($\frac{1}{4}$)	gr				gr
Hoopingarner, N. L., — Analysis of Work Interests Questionnaire	2			doubt	gr				doubt – gr
Mathematical Values Test (Kelley, T. L.)				ind	ind	gr	gr	ind	ind
Miner, J. B., — Analysis of Work Interests Test	2			doubt	gr				doubt – gr
Murdock, K., — Sewing Scale		ind		ind	ind	ind	ind	ind	ind
Murdock, K., — Analytic Sewing Scale	3	ind		ind	ind				ind
Rogers, A. L., — Test of Mathematical Ability				ind	ind	doubt	ind	ind	ind
Thurstone, L. L., — Vocational Guidance Tests		ind		ind	ind	ind	gr		ind

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Whittier Scale for Grading Home Conditions	4	ind		ind($\frac{1}{2}$)	ind	ind[1]			ind
Wilkins, L. A., — Prognosis Test in Modern Languages	4	gr		gr	gr	gr			gr

[1] Judge E states: "If enough judges used."

(*kkk*) ELEMENTARY PHYSICAL DEVELOPMENT MEASURES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
The judges were not sufficiently familiar with these to warrant ranking									
Athletic badge tests for boys and girls (Playground and Recreation Assn. of America.)	2								
Baldwin, B. T., — Physical Development Scale	1								
Rapeer, L. W., — Scale for Measuring Physical Education, Health, Physical Development	1								
Reilley, F. J., — Standards in Physical Training	0								

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*lll*) JUNIOR HIGH SCHOOL AND HIGH SCHOOL PHYSICAL DEVELOPMENT MEASURES

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Athletic badge tests for boys and girls (Playground and Recreation Association of America)	1								

(*mmm*) HIGH SCHOOL AND COLLEGE FRENCH TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Columbia French Test (Meras, A. M., Roth, S., Wood, B. D.)	4			1	2	1	1		1
Henmon, V. A. C., — French Test .				3	1	2	2	1	2
Twigg, A. M., — French Vocabulary Test	2			2($\frac{1}{2}$)	4				3
Handschin, C. H., — Modern Language Tests, — French				4	3	4	3	2	3
Starch, D., — French Test	4			b4		3	4	3	b3
No. having individual value . . .		3		2	2	2	2	2	2
No. having group but not individual value		0		1	2	1	2	0	1
No. of doubtful value		2		2	0	1	0	1	1
No. not reported upon		0		0	1	1	1	2	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:28 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*nnn*) HIGH SCHOOL AND COLLEGE GERMAN TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Columbia College Placement Examination in German (Betz, F., Betz, G. A., Wendt, H. G., Wood, B. D.).	4			1	1	1		1	1
Whipple, G. M., — German Vocabulary Test				3	2	3	1	2	2
Starch, D. — German Test				2	3	2	2	3	2
No. having individual value		1		1	1	1	0	1	1
No. having group but not individual value		0		2	2	0	2	1	1
No. of doubtful value		2		0	0	2	0	1	1
No. not reported upon		0		0	0	0	1	0	0

(*ooo*) HIGH SCHOOL AND COLLEGE SPANISH TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Handschin, C. H., — Modern Language Tests, — Spanish	3			gr	gr	doubt			gr

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:23 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Henmon, V. A. C., — Latin Test	4			2	2		1	1	1½
Brown, H. A., — Latin Test	4			1	1		2	b4	1½
Ullman, B. L., — Kirby, T. J., — Latin Comprehension Test	3			3(½)	3			2	3
Stevenson, P. R., — Latin Vocabulary Test	3			5(½)	4			3	4
Starch, D., — Latin Test	2						3	b4	4
Pressey, L. W., — Latin Syntax Test	2			5(½)	b4				b4
Tyler, C., — Pressey, S. L., — Test in Latin Verb Forms	3			5(½)	b4			4	b4
White, D. S., — Latin Test	1								
No. having individual value		4		2	3	0	2	5	2½
No. having group but not individual value		0		4	3	0	1	1	2½
No. of doubtful value		4		0	0	8	0	0	1
No. not reported upon		0		2	2	0	5	2	2

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*qqq*) HIGH SCHOOL LATIN COMPOSITION TEST

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Godsey, E., — Diagnostic Latin Composition Test	2			gr($\frac{1}{2}$)	gr				gr

(*rrr*) HIGH SCHOOL LATIN-DERIVATIVE VOCABULARY TESTS

	No. of Judges Rating if Less Than 5	Judges							Median Ratings
		A	B	C	D	E	F	G	
Stevenson, P. R., — Coxe, W. W., — Latin Derivative Test	2			1($\frac{1}{2}$)	1				1
Kansas Latin Derivative Test (Holtz, W. L., Godsey, E.)	2			2($\frac{1}{2}$)	2($\frac{1}{2}$)				2
Wentworth, M. M., — Latin Test .	1								
No. having individual value . . .		0		0	1	0		0	0
No. having group but not individual value		0		2	1	0		0	0
No. of doubtful value		1		0	0	3		3	2
No. not reported upon		2		1	1	0		0	1

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:24 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Classification and Test	Individual or Group Value as Judged by Median Rating	Interpolated Rank Values — Judges							Median Ratings
		A	B	C	D	E	F	G	
Primary General Intelligence Test Mentimeters School Group 2 A. (Trabue, M. R.)	gr	b7		4½	8		4½	8½	8
Elementary General Intelligence Test Multi-Mental Scale (Elementary School Form) (McCall, W. A., et al.)	ind	5½	6½	b9	2½		9½	1½	6
Junior High School General Intelligence Test Multi-Mental Scale (Elementary School Form) (McCall, W. A., et al.)	gr	6½	7½	b9	3½		8½	3½	7
High School General Intelligence Test Psychological Examination. (Thurstone, L. L., — Prepared for Committee on Personnel Research, National Research Council, 1925.)	ind	4½	2½	1½(½)	6½(?)		5	6½	4¾
Scholastic Aptitude Tests. (College Entrance Examination Board, 1925.)	gr			7½(½)	6¾(?)		8½	6¾	7⅛
Multi-Mental Scale (Elementary School Form) (McCall, W. A., et al.)	gr to ind	5½	7½	b9	6¼		9½	6½	7

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

College General Intelligence Tests									
Psychological Examination. (Thurstone, L. L., — Prepared for Committee on Personnel Research, National Research Council, 1925.)	ind	$1\frac{1}{2}$	$1\frac{1}{2}$	$1\frac{1}{2}(\frac{1}{2})$	$3\frac{1}{3}$(?)		$3\frac{1}{2}$	$6\frac{1}{2}$	$2\frac{5}{12}$
Scholastic Aptitude Tests. (College Entrance Examination Board, 1925.)	gr to ind			$5\frac{1}{2}(\frac{1}{2})$	$3\frac{3}{4}$(?)		$4\frac{1}{2}$	$7\frac{1}{2}$	5
Primary Reading Test									
Gates, A. I., — Test of Reading Vocabulary for the Primary Grades, 1926	gr			$2\frac{1}{2}$	4		$2\frac{1}{2}$	$\frac{1}{2}$	$2\frac{1}{2}$
Elementary Reading Tests									
Ohio Literacy Tests (Foster, —, — Goddard, H. H.)	gr			$10\frac{1}{2}$	b 12[1]		$8\frac{1}{2}(\frac{1}{10})$	$10\frac{1}{2}$	$10\frac{5}{12}$
New York State Regents Literacy Test (Morrison, J. C.)	doubt to gr			b13	b 12[1]		$8\frac{1}{2}(\frac{1}{10})$	$10\frac{3}{4}$	b11
High School Reading Tests									
Iowa Reading Comprehension Test (Ruch, G. M.)	gr to ind			$3\frac{1}{2}(\frac{1}{2})$	4		$3\frac{1}{2}$	$3\frac{1}{2}$	$3\frac{1}{2}$
Whipple, G. M., — High School and College Reading Test	gr	$6\frac{1}{2}$		$1\frac{1}{2}$	7		$5\frac{1}{2}$	$3\frac{3}{4}$	$5\frac{1}{2}$

[1] Judge D states: "These should not be considered for use in schools except for some civic purpose."

Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

(*333*) GIVING DATA UPON TESTS INTERPOLATED IN PRECEDING RANKINGS (*continued*)

Classification and Test	Individual or Group Value as Judged by Median Rating	Interpolated Rank Values — Judges							Median Ratings
		A	B	C	D	E	F	G	
College Reading Tests									
Whipple, G. M., — High School and College Reading Test . . .	gr	1/2		2 1/2	1 1/2		5	1 1/2	1 1/2
Iowa Reading Comprehension Test (Ruch, G. M.)	gr			2 3/4 (1/2)	1		4	1 3/4	2 1/2
Elementary and Junior High School Composition Scale									
Hudelson, E., — Maximal Composition Ability Scale	gr	3 1/2		4 1/2	4		5 1/2	6 1/2	5
High School Composition Scale									
Hudelson, E., — Maximal Composition Ability Scale	gr	4 1/2		b6	5		5 1/2	6 1/2	5 1/2
High School Geometry Test									
Schorling, R., Sanford, V., — Geometry Test	gr	1/2		1 1/2	2		1 1/2	1/2	1 1/2
High School Latin Test									
Lohr, L., Latshaw, H., — Latin Form Test	gr	4 1/2		2 1/2 (1/2)			4 1/2		4 1/2

Original from UNIVERSITY OF MICHIGAN

Sundry										
Cross English Test	gr	gr		gr	gr		ind	ind	gr	gr
Henman, V. A. C., — French Word List	ind			ind			ind		ind	ind
Seven Tests for Appreciation of Literature. (Logasa and McCoy)	gr			gr				gr	gr	gr
Spink Grading Chart for Mechanical Drawing	ind			ind(½)				ind	ind	ind

Generated on 2020-12-23 01:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

CHAPTER TEN

Classified and Graded Lists of Tests, Giving Reliability and Other Information

1. Description of lists and ratings of tests. In order to obtain as authoritative information as possible about intelligence and educational tests, the following letter was sent to the authors of all tests less than ten years old (or older, if known to be still in use) which, as far as the writer could ascertain, had ever been used after their first presentation to the public:

Questionnaire Sent to Authors of Tests

I am preparing a text upon the Interpretation of Educational Measurements and feel that it would be of great value could I include statements from the authors covering certain salient features of a select list of important educational tests. The information desired is indicated below under 11 headings. I have supplied this information myself so far as possible. May I ask you to correct any items which I have put down which are incorrect and to fill in the items which are lacking? I shall greatly value your aid in this matter and I am sure the future readers of the text will be equally appreciative.

Very sincerely yours,

1. Author
2. Name of test
3. Date first issued — 3 *r*. If revised, date of revision
4. No. of divisions, sections, or forms
5. Publisher

5 *d*. Source for Directions for giving and scoring

5 *n*. Source or sources for norms

6. Reliability coefficient of test

6 *g*. Population and grade or grades used in determining this reliability

6 σ. Standard dev. of the scores of this group upon a single form

6 *s*. Source, if published, of information given in 6, 6 *g*, and 6 σ

If items 6, 6*g*, and 6σ are not available, will the author express his opinion as to the reliability of the test by checking in the appropriate blanks provided below?

6 *r*. I consider the test, in the function which it measures, to be, in comparison with the average teacher's judgment, more reliable . . ., about as reliable . . ., less reliable . . .

I recommend it for { group . . . / individual as well as group . . . } placement and diagnosis.

7. Grades for which test is recommended by author
8. Time required to give test
9. Talent required to score test
10. Cost
11. Function measured: For the grades for which applicable the important phases of . . . which are not measured by this test are

Generated on 2020-12-23 01:29 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The order in which tests are listed in the following tables is that given in Chapter IX, Sections *a–rrr*, as modified by including the information given in Section *sss* of Chapter IX.

The data contained in the replies to this letter are recorded in ordinary type in the subsequent lists of this chapter. Thus, all statements appearing in ordinary type have been subscribed to by the authors of the tests concerned. In the case of joint authorship, the statements in ordinary type have been subscribed to by at least one of the authors. In certain cases authors did not reply, but statements directly emanating from them, in that they have been taken from manuals of directions and publishers' announcements, periodical articles written by the authors, have been the source of information here published. In such instances information is published in ordinary type as being directly attributable to the author. All statements appearing in italic type come from a source not directly attributable to the author. These statements are to be credited to the writer, unless otherwise noted.

In order to condense the space required in making available the very voluminous data which are to be had, the following abbreviations are used in the subsequent lists:

When an approximate value is indicated, the Latin *circa*, meaning " about," is used and abbreviated " ca."

Item 1 of Questionnaire. No abbreviation of the word " author " has been found necessary. The surname of the author, followed by the given name or initials, is the first item recorded.

Item 2. No abbreviation for " name of test " has been necessary. If the name of the test does not include that of the author, it is given in full, but if the name of the author is a part of the name of the test, a single dash indicates that fact, thus: " (1) Hotz, H. G. (2) — Algebra Scales."

Item 3. The word " date " stands for the date first issued. It has been intended to report here the date of first publica-

Generated on 2020-12-23 01:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

tion, but the reticence of publishing houses in attaching a date to their products has made it very difficult to be certain of the correctness of some of these dates. Not uncommonly the date of copyright of the Manual of Directions has been here recorded, but this may be too late a date. "Rev." stands for the date of revision, if there has been a revision.

Item 4. The abbreviations here used are "f" for "forms," "div." for "divisions," "pts." for "parts," "ser." for "series," "sca." for "scales," and "sec." for "sections." The term "forms" is here restricted to refer to comparable or equally difficult and equally excellent duplications of a test of a given type. Two forms are correlated to obtain a reliability coefficient, not two divisions, parts, series, or sections.

Item 5. "Pub." is the abbreviation for "publisher."

Items 5 d and 5 n. It has become so common to find both directions for giving and scoring and norms in the Manual of Directions that data upon these two items are not listed in the following tables.

Item 6. "Reliab." is the abbreviation for "reliability." Two kinds of information bearing upon reliability are here reported. Under "Reliab. j-a" (meaning reliability according to the judgment of the author) are recorded the judgments made by the author of the test in answering question 6 *r*. If the author stated that he considered the test more reliable than the average teacher's judgment, a "+" is recorded; if less reliability, a "−"; and if about as reliable, an "=" is recorded. Further, if he recommended his test for group placement and diagnoses, "gr." is recorded; and if for individual as well as group placement and diagnoses, "ind." is entered. The other sort of data bearing upon reliability reported under "Reliab." are such as have resulted in statistically determined measures of it. This information is available when for a given test either (*a*), (*b*), or (*c*) following are known: (*a*) the standard deviation, σ_1, and the relia-

Generated on 2020-12-23 01:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

bility coefficient, r_{11}, for a given group; (*b*) the standard deviation, σ_1, and the standard error of a score, $\sigma_{1.\infty}$ (or the probable error of a score, P. $E._{.1.\infty}$) for a given group; or (*c*) the standard error of a score (or the probable error) and the reliability coefficient for a given group. Having either the information (*a*), (*b*), or (*c*), we have at hand all the necessary facts because of the following relationships:

$$\sigma_{1.\infty} = \sigma_1\sqrt{1 - r_{11}} \qquad \text{[Formula 16, Chapter VII]}$$

$$\text{P.E.}_{.1.\infty} = .6745\,\sigma_1\sqrt{1 - r_{11}}$$

In this equation σ_1 and r_{11} are, of course, values derived from the same data. Unfortunately the majority of authors have not presented data (*a*), (*b*), or (*c*). Some of them have, however, given r_{11} and the age or grade range involved in its determination. This is of much assistance in estimating the reliability, for the change of reliability with change in range follows much the same lines as given for achievement-intelligence correlations in Table 27 (Section 2, Chapter VIII), and thus reliability for certain ranges may be estimated, knowing them for other ranges. In this connection the abbreviation "Cr." is quite frequently found, followed by data upon reliability. When this occurs, the coefficients given are those determined by Miss M. Alice Cronin and reported in a master's thesis at Stanford University.

A still richer source of data bearing upon phases, both of reliability and validity of high school tests, is the work of Ruch and Stoddard, as listed in the bibliography (1927). No one has as yet done for the elementary field what these authors have done for the secondary field in making available the information which is necessary for a full and precise utilization of test scores.

Item 7. "Gra." stands for "grades for which test is recommended by the author." "Age" stands for "ages for which test is recommended by the author."

Generated on 2020-12-23 01:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Item 8. "Time" stands for "time required to give the test."

Item 9. "Talent" stands for "talent required to score the test." "Good cler." stands for "good clerical help required to score the test."

Item 10. "Cost" stands for the "cost of the test when purchased in bulk," lots of 25 or 100 being usually quoted.

Item 11. In the blank space in Question 11 was recorded the field covered by the test before the question was sent to the author. Thus, if a reading test is being considered, the question would read: "For the grades for which applicable the important phases of *reading* which are not measured by this test are. . . ." The author's reply to this question is recorded following "Function measured: reading, except." Where Item 11 is omitted, it indicates that the author did not answer this question.

The reader must not come to the conclusion that tests for which data as just described are not recorded are less excellent tests than those for which such data are given. The writer would say that it has been very difficult to collect these data, and their presence or absence is largely contingent upon his success in this undertaking and more or less unconnected with questions of general excellence of the test.

The general scheme of classification has been to list tests under the following headings:

Primary (kindergarten, first grade, and low second)
Elementary (Grades 2 to 8 inclusive)
Junior High School (Grades 7 to 9 inclusive)
High School (Grades 9 to 12 inclusive)
College (Grades 12 to 16 inclusive)

Many a test belongs in more than one of these classifications, and in such case it is to be found listed in each classification to which it is applicable, but the information as to publisher, reliability, etc., is given only in the first, or the most important, classification in which listed. Cross references in other classifications to this one are given.

Generated on 2020-12-23 01:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

The order in which all tests preceding "tests not rated" are listed in each classification is that of their median rankings, except in the "Sundry" classification, which is alphabetical. Fewer tests are listed in a classification in this chapter than in the same classification in the preceding chapter. All tests of the preceding chapter which were judged to be so low that they fell below the line where ranking was attempted have been omitted from the lists of this chapter. Any injustice in this procedure probably affects one or more of the following tests:

Elementary General Intelligence Test. Wylie, A. T.: Opposites Test. (Some of the judgments upon this test may have been of preliminary forms.)

Elementary General Intelligence Test. Ballard, P. B.: Chelsea Mental Tests. (Known by but one judge.)

Elementary General Intelligence Test. Thomson, G. H.: Northumberland Mental Test. (Known by but one judge.)

Junior High School General Intelligence Test. Dearborn, W. F.: Group Test of Intelligence, Intermediate, Ser. 2. (Omitted from this classification by oversight.)

High School and College General Intelligence Test. Spearman, C.: — General Intelligence Test. (Known by but one judge.)

Junior High School Reading Test. Thorndike, E. L.: — Word Knowledge Test. (Ranked very high by two judges and very low by three.)

Junior High School Arithmetic Test. Stevenson, P. R.: — Arithmetic Problem Analysis Test. (Ranked very high by one judge and very low by two.)

Junior High School American History Test. Hahn, H. H.: — American History Scales. (Ranked fairly high by two judges and low by three.)

Writing Tests. Gray, C. T.: — Standard Score Card for Measuring Handwriting. (Ranked fairly high by two judges and low by three.)

Generated on 2020-12-23 01:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

2. The detailed classifications and ratings of the various tests. Since this text has concerned itself with problems of measurement and classification involving large populations, little attention has been given to individual tests. One exception to this rule is made herewith in connection with the Stanford-Binet test. This test, though individual, has proved of such value that it has been and is now being used upon groups which are quite as extensive as those to which the better group tests have been applied. The accompanying data have been kindly supplied by Dr. Terman:

(1) Terman, L. M. (2) Stanford–Binet. (3) Date; Mimeographed and distributed to about 25 persons, 1914; printed, 1916. (4) 1 f. (5) Pub.: Houghton Mifflin Company.
(6) Reliab. coef. of test: .90 to .95
Reliab. coef. for chron. age group: 8.0–9.0, approx. .92
Reliab. coef. for chron. age group: 12.0–13.0, approx. .93
Reliab. coef. for adults approx. .93

Population used in determining this reliability:
Population, 8-year-old group: 108
Population, 12-year-old group: 57
Population, adults: 180

Standard deviation of the scores of this group upon a single form:
Standard deviation of 8-year-old group: 12.4 mo.
Standard deviation of 12-year-old group: 18.46 mo.
Standard deviation of an adult group: 24.6 mo.

For Dickson's 149 1st-grade pupils:
Mean age: 7 yr., 0.2 mo.
σ of age: 13.14 mo. (or 15.6 IQ)
Mean mental age: 6 yr., 2.11 mo.
σ mental age: 17.48 mo.
$r_{\frac{1}{2}\frac{I}{II}} = .85$. Brown's formula gives .92.

For Knollin's 180 adults (140 prisoners and 40 business men):
Mean mental age: 14.1
σ mental age: 24.6 mo. (or about 12.8 IQ)
$r_{\frac{1}{2}\frac{I}{II}} = .87$. Brown's formula gives .93.

Generated on 2020-12-23 01:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(7) Age: 3 and up. (8) Time: about 45′. (9) Talent: One experienced in giving individual intelligence tests. (10) Aims to measure general intelligence and not specifically school training.

It would have been desirable to incorporate a classification "Non-Verbal General Intelligence Tests." The number of these, both individual and group, is quite extensive, particularly since the appearance of Army Beta. The most recent test of this sort, as well as the one having the highest reliability (in fact, so high as to seem unreasonable) as reported by the author, is briefly described, from published sources, below:

(1) Dodd, Stuart C. (2) International Group Mental Tests. (3) Date: 1926. (4) 1f rotater edition and 1f paper-and-pencil edition. (5) Pub.: Princeton University Press, Princeton, New Jersey. (6) Reliab.: $N = 112$ 6th-grade orphans; retesting coef.: 78; reliab. coef. (split-half method): .97. (7) Age: kindergarten to adult. (8) Time: 170′ to 235′.

The tests rated by the judges, *in the order of their median ratings*, are given for the various classifications in the following tables.

Tables Giving Data Covering Selected Tests

(*a*) **Primary General Intelligence Tests: Ind. (Number of tests having Individual Value) 3. Gr. (Additional number having group value) 8.**

(1) Pintner, R. and Cunningham, B. V. (2) — Primary Mental Test (3) Date: 1923 (4) 1f (5) Pub: WBC (6) Reliab: *No reliab given:* Retest. coef. = .88 for N = 22 kgtn children, Retest. coef. = .93 for N = 23 kgtn children having $\sigma = 8$, Reliab. j-a: + ind (7) Gra: kgtn to Gra 2 (8) Time: 30′–50′ (9) Talent: good prim. teacher (10) Cost: $1.25 per 25.

(1) Park, B. and Franzen, R. (2) — Primary Test (3) Date: 1923 (4) 1f (5) Pub: Privately, Miss

Generated on 2020-12-23 01:30 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Bessie Park, Primary Supervisor, Des Moines Public Schools (6) Reliab: 15 low 1st grade classes, N in each varying from 30 to 40, r_{11} = .80, σ = 8.50 (7) Gra: h. kgtn to low 1st (8) Time: 40′ (9) Talent: good cler. (10) Cost: $1.25 per 25 (11) Function measured: Measures ability to do first grade work, — a measure of persistence as well as intelligence.

(1) Dearborn, W. F. (2) — Group Test of Intelligence. Series 1. (3) Date: 1920 (4) 1f of 3 pts Exam 1, 2 and 3 (5) Pub: Lippincott (7) Gra 1–3.

---

(1) Bird, G. E. and Craig, C. E. (2) Rhode Island Intelligence Test. (3) Date: 1923 (4) 2f (5) Pub: PSPC (6) Reliab: r = .92 Given in Jour. Ed. Res. 8 — '23; N = 330,–3–6 yr olds; σ = 3.9. Reliab. j-a: + ind (7) Gra: kgtn. Age 3–6 (8) Time: no limit, — ca. 15′ (9) Talent: good kgtn teacher (10) 50¢ per 25 (11) Function measured: General Intelligence, except, Speed.

(1) Haggerty, M. E. (2) — Intelligence Examination δ1 (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab: Va. Survey v. 8, p. 148; in reliab. the test is not quite so satisfactory as δ2 (7) Gra: 1–3 (8) Time 30′ (9) Talent: good cler. (10) Cost: $1.25 per 25 (11) Function measured: General Intelligence.

(1) Otis, A. S. (2) — Group Intelligence Scale; Primary Examination.

(1) Engel, A. M. (2) Detroit First Grade Intelligence Test (3) Date: 1920 Rev. 1921 (4) 1f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 1 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.10 per 25.

(1) Kingsbury, F. A. (2) — Primary Group Intelligence Test.

(1) Trabue, M. R. (2) — Mentimeters School Group 2 A. See (*b*).

Generated on 2020-12-23 01:25 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(*b*) Elementary General Intelligence Tests: Ind. 8, Gr. 7.

(1) Haggerty, M. E., Terman, L. M., Thorndike, E. L., Whipple, G. M., Yerkes, R. M. (2) National Intelligence Test (3) Date: 1920 (4) 2 pts. of 2f each (5) Pub: WBC (6) Reliab: *Scale A, Gra. 4–8 combined,* $r_{11} = .922$; $\sigma = 29$ *(derived from P.E.* $_{score} = 5.4$*): Scale B, Gra. 4–8 combined* $r_{11} = .949$; $\sigma = 32$ *(derived from P.E.* $_{score} = 4.9$*);* $N = 232$: *See P. M. Symonds, Jour. Ed. Res., Apr. '24 and Jour. Ed. Psych., Oct. '24.*
Terman and Whitmire, in an unpublished study, found the correlation between Scales A and B for 1073 children, gra. 3–8, to be .928.
A. I. Gates, Jour. Ed. Psych. Dec. '23, gives the reliability of the composite score on Scales A and B, for 75 pupils, gra. 3–6, as .93. From the above the following are estimated by T. L. K: Scale A, — for a single gra. range r_{11} *= ca. .70 and* σ *= ca. .15. Scale B, — for a single gra. range* r_{11} *= ca. .75 and* σ *= ca. .14. Scales A and B, — for a single gra. range* r_{11} *= ca. .85 and* σ *= ca. .27.*

(1) Haggerty, M. E., Terman, L. M., Thorndike, E. L., Whipple, G. M., Yerkes, R. M. (2) National Intelligence Test (Part A only).

(1) Haggerty, M. E. (2) — Intelligence Examination $\delta 2$ (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab: *No reliab. coef. given.* Author reports that Stenquist found r [$\delta 2$ with NIT–A] = .81, gra. 4–8, N = 500. That Miller found r [$\delta 2$ with Miller] = .79, gra. 9, N = 55. Author found r [$\delta 2$ with Miller] = .61, gra. 9, N = 442. *Est. reliab. for single gra.* = .6 (7) Gra: 3–9 (8) Time: 30′ (9) Talent: Good cler. (10) Cost: $1.10 per 25.

(1) Dearborn, W. F. (2) — Group Test of Intelligence, Series 2 (3) Date: 1920 (4) 1f of 2 pts Exam. 4 and Exam. 5 (5) Pub: Lippincott (7) Gra: 4–9 (8) Time: 50′ for each pt.

Generated on 2020-12-23 01:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Otis, A. S. (2) — Self-administering Tests of Mental Ability (Intermediate Examination) (3) Date: 1922 (4) 2f (5) Pub: WBC (6) Reliab: r = .95, N = 427, gra. = 4–9, σ = *16.76; — Deduced from P.E. of score which is given on directions sheet, using formula* [16] *of Chapter VII*, Reliab. j-a: + ind (7) Gra: 4–9 (8) Time: 40′ (9) Talent: clerk (10) Cost: 80¢ per 25.

(1) Haggerty, M. E., Terman, L. M., Thorndike, E. L., Whipple, G. M., Yerkes, R. M. (2) National Intelligence Test (Part B only).

(1) McCall, W. A., et al. (2) Multi-Mental Scale, (Elementary School Form) (3) Date: 1925 (4) 1f (5) Pub: T. C. Bur. Pub. (6) Reliab: $r_{\frac{I}{II}}$ = .89, r_{II} = .94, Pop. gra. 3–9 inclusive. *Pop. used in determining reliab. same as used in construction of test, therefore reliab. reported may be expected to be spuriously high by a small amount.* (7) Gra: 2–9 (8) Time: 25′ (10) Cost: $1.00 per 100.

(1) Otis, A. S. (2) — Group Intelligence Scale. See (*c*).

---

(1) Buckingham, B. R. (2) Illinois General Intelligence Scale (3) Date: 1919 Rev. 1920 (4) 2f (5) Pub: PSPC (6) Reliab: r = .92, gra. 3–8, N = 958, average r_{11} per gra. for gra. 3, 4, 5, = .90; average r_{11} per gra. for gra. 6, 7, 8, = .76. Reliab. j-a: ind (7) Gra: 3–8 (8) Time: ca. 30′ (9) Talent: good cler. (10) Cost: $2.00 per 100 when separate from rest of Ill. Exam.

(1) Trabue, M. R. (2) — Mentimeters School Group 2A (3) Date: 1920 (4) 1f (5) Pub: Doubleday, Page and Co. (6) Reliab: No reliab. given. r_{12} with Stanford-Binet = .88, gra. 1–12, N = 407. Average r_{12} per g = .6. Reliab. j-a: + ind (7) Gra: 1–12 (8) Time: varies (9) Talent: good intelligent help (11) Function measured: Academic intelligence, except social reactions, mechanical skills, and artistic judgments.

Generated on 2020-12-23 01:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Tests not rated: (1) . . . (2) Army Alpha Revised. See (*d*).

(1) Baker, H. J. (2) — Detroit Intelligence Test (3) Date: 1927 (4) 3 pts., C, M, and W of 1f each (5) Pub: PSPC (7) Gra: C for gra. 2–4; M for gra. 5–9; W for high school and college (10) Cost: Each pt $3.00 per 100.

(1) Whipple, G. M. and Whipple, H. D. (2) — Illinois General Intelligence Scale (3) Date: 1926 (4) 2f (5) Pub: PSPC (7) Gra: 3–8 (10) Cost: $2.00 per 100.

(*c*) Junior High General Intelligence Tests: Ind. 5, Gr. 8.

(1) Terman, L. M. (2) — Group Test of Mental Ability (3) Date: 1920 (4) 2f (5) Pub: WBC (6) Reliab: r = .89; N = 132; g = 9th; σ = 24.2. Reliab. j-a + ind (7) Gra: 7–13 (8) Time: 35′ (9) Talent: good cler. (10) Cost: $1.20 per 25.

(1) Otis, A. S. (2) — Group Intelligence Scale (3) Date: 1918 (4) 2f (5) Pub: WBC (6) Reliab: r = .967; g = 4–8; σ = 31.3. Reliab. j-a: + ind (7) Gra: 5–16 (8) Time: 65′ (9) Talent: good cler. (10) Cost: $1.25 per 25 (11) Function measured: General Intelligence, except, — None, unless you refer to such special abilities as musical ability, etc.

(1) Haggerty, M. E., Terman, L. M., Thorndike, E. L., Whipple, G. M., Yerkes, R. M. (2) National Intelligence Test (Parts A and B). See (*b*).

(1) Miller, W. S. (2) — Mental Ability Test (3) Date: 1922 (4) 2f (5) Pub: WBC (6) Reliab: retesting coef. = .91; N = 109; gra. = 10; σ = 14.3 (7) Gra. 7–16 (8) Time: 30′ (9) Talent: good cler. (10) Cost: 80¢ per 25.

(1) Haggerty, M. E. (2) — Intelligence Examination δ2. See (*b*).

---

(1) Otis, A. S. (2) — Self-Administering Tests of Mental Ability (Intermediate Examination). See (*b*).

Generated on 2020-12-23 01:26 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Trabue, M. R. and Kelley, T. L. (2) — Completion Exercises Alpha and Beta (3) Date: 1917 (4) 2f (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{11} = .90; gra. 2–9; N = ca. 100. From preceding (by T. L. K.) it is est. r_{11} = ca. .55 and σ = ca. 1.0 for single gra. Reliab. j-a: + ind (7) Gra. 2–16 (8) Time: 25′ in gra. 2, 60′ in college (9) Talent: superior cler. help (10) Cost: $1.25 per 100 (11) Function measured: General Intelligence, except memory, number concepts, strictly non-verbal capacities, social attitudes, political sagacity, mechanical knowledge and skill, and appreciations.

(1) McCall, W. A., et al. (2) Multi-Mental Scale (Elementary School Form). See (*b*).

(1) Trabue, M. R. (2) — Mentimeters School Group 2A. See (*b*).

(1) — (2) Army Alpha.

Tests not rated: (1) — (2) Army Alpha Revised. See (*d*).
(1) Baker. (2) — Detroit Intelligence Test. See (*b*).
(1) Whipple and Whipple. (2) — Illinois General Intelligence Scale. See (*b*).

(*d*) High School General Intelligence: Ind. 8, Gr. 9.

(1) Terman, L. M. (2) — Group Test of Mental Ability. See (*c*).

(1) Otis, A. S. (2) — Group Intelligence Scale. See (*c*).

(1) Thorndike, E. L. (2) — Intelligence Examination. See (*e*).

(1) Otis, A. S. (2) — Self-Administering Tests of Mental Ability (Higher Examination) (3) Date: 1922 (4) 1f (5) Pub: WBC (6) Reliab: r = .92; N = 253; gra. = 7–12; σ = *13.82; — Deduced from P. E. of score which is given on directions sheet, using formula* [16] *of Chapter VII.* Reliab. j-a: + ind (7) Gra: 7–12 (8) Time: 40′ (9) Talent: clerk (10) Cost: 80¢ per 25 (11) Function measured: Mental Ability.

Generated on 2020-12-23 01:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Thurstone, L. L. (2) Psychological Examination (Prepared for Committee on Personnel Research, National Research Council, 1925). See (*e*).

(1) Thurstone, L. L. (2) — Psychological Examination. See (*e*).

(1) Miller, W. S. (2) — Mental Ability Test. See (*c*).

(1) Haggerty, M. E. (2) — Intelligence Examination δ2. See (*b*).

(1) McCall, W. A., et al. (2) Multi-Mental Scale. See (*b*).

---

(1) College Entrance Examination Board (2) Scholastic Aptitude Tests. See (*e*).

(1) Trabue, M. R. and Kelley, T. L. (2) — Completion Exercises Alpha and Beta. See (*c*).

(1) Trabue, M. R. (2) — Mentimeters School Group 2A. See (*b*).

Tests not rated:

(1) Bregman, E. O., with coöperation of Cattell, J. McK. (2) Army Alpha Revised (3) Date: 1925 (4) If of 8 pts. (5) Pub: Psychological Corporation, 3939 Grand Central Terminal, N. Y. City (6) Reliab: Not less than Army alpha, and probably more (7) Gra: Same as Army alpha (8) Time: About same as Army alpha (9) Talent: Same as for Army alpha (10) Cost: $5.00 per 100 to psychologists associated with the Psychological Corporation (11) Function measured: Same as Army alpha.

(1) Baker (2) — Detroit Intelligence Test. See (*b*).

(*e*) College General Intelligence Tests: Ind. 6, Gr. 6.

(1) Thorndike, E. L. (2) — Intelligence Examination (3) Date: 1918 (4) 3f each year (5) Pub: T. C. Bur. Pub. (6) Reliab: $r = .85$; $N = 171$; gra. = normal school: $\sigma = 12.5$. Reliab. j-a: + ind (7) Gra: 13 (8) Time: 3½ hours (9) Talent: super. cler. (11) Function measured: General Intelligence, except intelligence in dealing with 3

Generated on 2020-12-23 01:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

dimensional objs. (as in biology or in engineering), with people and their passions (as in the ministry, business, or politics), and with esthetic or perceptual matters.

(1) Thurstone, L. L. (2) Psychological Examination (Prepared for Committee on Personnel Research, National Research Council, 1925) (3) Date: 1925 (5) Pub: American Council on Education (6) Reliab: By Spearman-Brown formula $r_{11} = .959$, $N = 250$. Reliab. on separate pts varies from .71 to .98.

(1) Thurstone, L. L. (2) — Psychological Examination for College Freshmen (3) Date: 1919, Rev. 1922–23–24 (4) 1f (5) Pub: C. H. Stoelting Co. (6) Reliab: j-a: + (7) Gra: = college freshmen (8) Time: 30′ (9) Talent: good cler. (10) Cost: $16.50 per 100.

(1) Colvin, S. S. (2) Brown University Psychological Examination.

(1) Terman, L. M. (2) — Group Test of Mental Ability. See (*c*).

(1) College Entrance Examination Board (2) Scholastic Aptitude Tests (3) Date: 1925 and later (4) New f's each year; 10 sub-tests (5) Pub: Released by C.E.E.B. only (7) Gra: 12–13 (8) Time: 2 hrs. 30′.

---

(1) Otis, A. S. (2) — Self-Administering Tests of Mental Ability (Higher Examination). See (*d*).

(1) Otis, A. S. (2) — Group Intelligence Scale. See (*c*).

(1) Roback, A. A. (2) — Mentality Tests.

(1) Trabue, M. R. and Kelley, T. L. (2) — Completion Exercises Alpha and Beta. See (*c*).

Tests not rated: (1) . . . (2) Army Alpha Revised. See (*d*).

(1) Carpenter, M. F. and Stoddard, G. D., under direction of Seashore, C. E. and Ruch, G. M. (2) Iowa Placement examination (3) Date: 1925 (4) 2

Generated on 2020-12-23 01:31 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

pts, — an aptitude and a training pt, each covering English, chemistry, foreign languages, mathematics, and physics (5) Pub: Extension Division, University of Iowa (6) Reliab: On each of 6 pts varies from .87 to .93 for a population of 100 *of undesignated grade range — See Stoddard, Iowa Placement Examination, University of Iowa Studies, Vol. III, No. 2, 1925* (7) Gra: 12–13 (8) Time: ca 8 hrs.

(*f*) Primary Achievement Batteries: Ind. 0, Gr. 1.

(1) Pressey, L. C. (2) — Scale of Attainment No. 1.

(*g*) Elementary Achievement Batteries: Ind. 2, Gr. 4.

(1) Kelley, T. L., Ruch, G. M., Terman, L. M. (2) Stanford Achievement Test. See Reading (*k*), Arithmetic (*dd*), Gen. Science (*mm*), History (*ss*), Language Usage and Grammar (*x*), and Spelling (*u*), for various parts. (3) Date: 1923 (4) 2f Elem. Exam., 2f Advanced Exam. (5) Pub: WBC (6) Reliab: .95 and .96. Average r_{11} per gra. for gra. 2–3 is .95; average r_{11} per gra. for gra. 4–9 is .96; N = 1204 in gra. 2–9; σ Total Score (h.2 and low 3 combined) = 5.7; σ Total Score (low 8 and h.8 combined) = 10.6. Above derived from data given in Manual. Reliab. j-a: + ind (7) Gra: 2–9 (8) Time: gra. 2–3 ca. 75′, gra. 4–9 ca. 135′ (9) Talent: good cler. (10) Cost: gra. 2–3 $1.10 per 25, gra. 4–9 $1.90 per 25 (11) Function measured: Elementary school studies, except, — Mechanical studies, home economics, art, music, and citizenship habits and attitudes.

(1) Otis, A. S. (2) — Classification Test (3) Date: 1923 (4) 2f (5) Pub: WBC (6) Reliab: r_{11} = .95; N = 253; gra. 4–8. Reliab. j-a: + ind (7) Gra: 4–8 (8) Time: 70′ (9) Talent: clerk (10) Cost: $1.10 per 25 (11) Function measured: Mental ability and general achievement.

---

(1) Buckingham, B. R. and Monroe, W. S. (2) Illinois Examination. See Illinois General Intelligence

Generated on 2020-12-23 01:26 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Scale (*b*), Monroe Elementary Reading (*k*), and Monroe Elementary Arithmetic General Survey Scale (*dd*) (3) Date: 1919, Rev. 1920 (4) 2f and two exam. (5) Pub: PSPC (7) Gra: Exam. 1, 3–5, Exam. 2, 6–8 (8) Time: ca. 60′ (10) Cost: $4.00 per 100 (11) Function measured: Elementary school work.

(1) Chapman, J. C. (2) — Classroom Products Survey Test (3) Date: 1920, Rev. 1921 (4) 1f (5) Pub: Lippincott (6) Reliab: r_{11} = .6 to .85 chiefly .6 and .7 (for individual tests and in single gra. ranges); gra. = 6, 7, or 8. Reliab. j-a: + ind (Cum grano salis) (7) Gra: 5–8 (8) Time: 90′ (9) Talent: good cler. (10) Cost: $3.50 per 100 (11) Function measured: Elementary school work, except, — Informational content in general science and the humanities.

(*h*) **Junior High School Achievement Batteries: Ind. 1, Gr. 1.**

(1) Kelley, T. L., Ruch, G. M., Terman, L. M. (2) Stanford Achievement Test. See (*g*).

---

(1) Otis, A. S. (2) — Classification Test. See (*g*).

(1) Trabue, M. R. (2) — Mentimeters School Group 2A. See (*b*).

(*i*) **High School Achievement Batteries: Ind. 2, Gr. 0.**

(1) Ruch, G. M. (2) — High School Content Examination (Iowa Entrance Examination) (3) Date: 1923 (4) new form each year (5) Pub: Univ. of Iowa, Iowa City (6) Reliab: r_{11} = .90 to .95; σ (for r_{11} = .95) = 46.6. A random sample of 100 from 1400 applicants for entrance to Univ. of Iowa gave an r_{11} of .90; σ of 1400 scores = 48.8. The reliabilities reported were found by correlating the sum of the scores on sections 1 and 2 with the sum of the scores on sections 3 and 4. In so far as these are not strictly comparable halves the reliabilities reported are probably a trifle lower than the true values. In each case the Spearman-Brown Formula was

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

applied. (7) Gra: 12–13 (8) Time: 80′ (9) Talent: good cler. (11) Function measured: (1) General mastery of basic high school subjects; (2) prediction of college success.

(1) Thurstone, L. L. (2) — Vocational Guidance Test. See High School Algebra (*hh*), High School Arithmetic (*ff*), High School Geometry (*jj*), and High School Physics (*rr*). The vocational guidance test consists of the preceding plus a Technical Information Test. Reliab. coef. has not been determined.

Tests not rated: (1) Carpenter et al. (2) Iowa Placement Examination. See (*e*).

(1) Trabue, M. R., et al. (2) North Carolina High School Senior Examination, 1927 Edition (3) Date: 1927 (4) 1f 9 Sec. (5) Pub: Bureau of Educational Research, University of North Carolina (8) Time: Omitting foreign language 65′; time for foreign language 20′ (11) Function measured: A: English, Literature and forms, B: Comprehension of reading, C: Mental agility (verbal), D: History, American and general, E: Modern times and civics, F: General science, G: Mathematics, H: Latin and French.

(*j*) **Primary Reading Tests: Ind. 1, Gr. 3.**

(1) Haggerty, M. E. and Noonan, M. E. (2) — Reading Examination, Sigma 1 (3) Date: 1919, Rev. 1921–22 (4) 1f, 2 pts (5) Pub: WBC (6) Reliab: *No. reliab. given;* $r = .84$ (Retest after 6 weeks); N = 200; gra. = 1–3; *Cr: Deduced the following from a population of 94 in gra. h.2–h.3: Gra. range low 3–h.3, Part I:* $r_{11} = .79$; $\sigma = 3.31$. *Part II:* $r_{11} = .81$; $\sigma = 3.11$. *Total* $r_{11} = .88$; $\sigma = 6.1$ (7) Gra: 1–3 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 25.

---

(1) Oglesby, E. M. (2) Detroit Group Test in Word Recognition (3) Date: 1924 (4) 10f (5) Pub: WBC (6) Reliab: From Jour. Ed. Res., June, 1924, low 1, h.1, low 2, h.2: Average r_{11} ($\frac{1}{2}$ gra. range) = .68 (Similar forms correlated). Average

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

r_{11} (½ gra. range) = .96 (Split test method: odds vs. evens); N = 554. *σ for ½ gra. range estimated = 8.* (7) Gra: low 2 (8) Time: 4′ (9) Talent: *good cler.* Cost: 90¢ per 25 (11) Function measured: *Mental Ability.*

(1) Gates, A. I. (2) Reading Vocabulary Test for Primary Grades (3) Date: 1926 (4) 2f (5) Pub: T. C. Bur. Pub: (7) Gra: 1–2 (8) Time: 15′ (10) Cost: $3.00 per 100.

(1) Pressey, L. W. (2) — First Grade Attainment Scale in Reading (3) Date: 1923 (4) 2f (5) Pub: PSPC (6) Reliab: Average per ½ gra. for gra. 1B and 1A = .80 to .85; N = ca. 150 per ½ gra. Reliab. j-a: + gr. (7) Gra: 1 (8) Time: 15′ (9) Talent: good cler. (10) Cost: $1.00 per 100 (11) Function measured: Reading, except, — Getting meaning from passages of any length. Test measures only the recognition of most common words.

Tests not rated: (1) Gates, A. I (2) — Primary Reading Tests, — Reading of words, phrases, and sentences (3) Date: 1926 (4) 2f (5) Pub: T. C. Bur. Pub. (7) Gra: 1–2 (8) Time: 15′ (10) Cost: $3.00 per 100

(1) Gates, A. I (2) — Primary Reading Tests, — Reading of paragraphs of directions (3) Date: 1926 (4) 2f (5) Pub: T. C. Bur. Pub: (7) Gra: 1–2 (8) Time: 20′ (10) Cost: $3.00 per 100.

(k) Elementary Reading Tests: Ind. 4. Gr. 9.

(1) Thorndike, E. L. and McCall, W. A. (2) — Reading Scales (3) Date: 1920 (4) 10f (5) Pub: T. C. Bur. Pub. (6) Reliab: Thorndike reports for a group of constant age 10–15, r_{11} = ca. .70. McCall reports for a random sampling of 12-yr-olds, r_{11} = .8; N = 500; σ = 10T (The "T" refers to McCall T-Scores): *Cr: Deduced from a population of 75, gra. h.7–h.8, f.II vs f.IV, r = .57; σ = 9.1; gra. low 7–h.7. Current and Ruch*[1]*: r_{11} = 75; σ = 4.45. Ruch reports (in a personal letter) that C. L. Cushman found: (a) r_{11} = .54; σ = 3.19; N = 73 in gra. 3. (b) r_{11} = .71; σ = 3.61; N = 93 in gra. 4. (c) r_{11} = .56;*

[1] See footnote, page 308.

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

σ = 2.89; N = 63 in gra. 5. (d) r_{11} = .60; σ = 3.15; N = 100 in gra. 6. Standard deviations given by Current and Ruch and by Cushman are in raw test scores and those given by McCall and Cronin are in T-scores. Reliab. j-a; McCall reports + ind; Thorndike reports +, using 2 or more forms; ind., using preferably 4 forms. If using 1f gr. value. (7) Gra: 2–12 (8) Time: 30′ (9) Talent: super. cler. (10) Cost: $2.00 per 100 (11) Function measured: Reading, except, — An exact measure of speed, comprehension on a single level of difficulty, emotional appreciation of what is read.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Reading Test (3) Date: 1922 (4) 2f of 3 pts each (5) Pub: WBC (6) Reliab: (Par Mean) r_{11} = .78; (Sen Mean) r_{11} = .80; (Wd Mean) r_{11} = .90; (Total) r_{11} = ca. .91; N = 1204 in gra. 2–9; average N per gra. = 150. These r_{11}'s are average r_{11}'s for one grade in grades 2–9. The σ's for each part and each grade are given in the manual. The σ of the Reading Total Score (h.2nd and low 3rd combined) = 24.9; (low and h.8th combined) = 34.4. *Current and Ruch*[1]: *r_{11} = 93; σ = 41.4.* Reliab. j-a: + ind (7) Gra. 2–9 (8) Time: gra. 2–3, 25′, gra. 4–9, 40′ (9) Talent: good cler. (10) Cost: $1.10 per 25 (11) Function measured: Reading, except, — Pronunciation and speed of reading.

(1) Thorndike, E. L. (2) — Test of Word Knowledge (3) Date: 1921 (4) 8f (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{11} = .83, *9th gra. pupils* (7) Gra: 4–10 (8) Time: 20′ (9) Talent: good cler. (10) Cost: $1.50 per 100 (11) Function measured: Vocabulary, except, — The "active" or speaking and writing vocabulary. This test is for the reading or "passive" vocabulary.

(1) Haggerty, M. E. and Haggerty, L. C. (2) — Reading Examination Sigma 3 (3) Date: 1919, Rev. 1921–22 (4) 2f (5) Pub: WBC (6) Reliab: *No reliab. given.* *r = .885 (Retest after 2 days) N = 126;*

[1] See footnote, page 308.

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

gra. = 5c–8a Current and Ruch[1]*: r_{11} = .83; σ = 27.7 on f1 and 19.7 on f2* (7) Gra: 6–12 (8) Time: 45′ (9) Talent: good cler. (10) Cost: $1.10 per 25.

---

(1) Chapman, J. C. and Cook, S. A. (2) — Speed of Reading Test: Lippincott.

(1) Monroe, W. S. (2) — Standardized Silent Reading Tests, Revised (3) Date: 1920 (4) 3f of 2 tests each (5) Pub: PSPC (6) Reliab: *Cr: N = 41; gra. low 7–h.7; f1 vs f2. Rate: r_{11} = .79. Comprehension: r_{11} = .65. Comprehension: average r_{11} per gra. for gra. 3–5 = .66; average r_{11} per gra. for gra. 6–8 = .73. Rate: average r_{11} per gra. for gra. 3–5 = .75; average r_{11} per gra. for gra. 6–8 = .83. Current and Ruch*[1]*: r_{11} = .76; σ = 3.25* Reliab. j-a: ind (7) Gra: Test 1, 3–5, Test 2, 6–8 (8) Time: 4′ (9) Talent: good cler. (10) Cost: 80¢ per 100 (11) Function measured: Silent reading rate and comprehension.

(1) Thorndike, E. L. (2) — Visual Vocabulary Test.

(1) Gray, W. S. (2) — Silent Reading Test.

(1) Burgess, M. A. (2) —Reading Test.

(1) Trabue, M. R. and Kelley, T. L. (2) — Completion Test Language Scales Alpha and Beta. See (*c*).

(1) Foster, —, and Goddard, H. H. (2) — Ohio Literacy Tests.

(1) Courtis, S. A. (2) — Silent Reading Test No. 2. (6) *Current and Ruch*[1]*: r_{11} = .77; σ = 57.8 on f1 and 66.3 on f2.*

(1) Fordyce, C. (2) — Scale for Measuring Ability in Silent Reading.

Tests not rated: (1) Stone, C. R. (2) — Narrative Reading Tests (3) Date: 1922, Gra. 7 test 1923 (4) 1f of

[1] For reliability coefficients credited to Current, W. F. and Ruch, G. M., see reference (1925). As all of these were determined from the same 154 children in grades 4–8, they should be highly comparable. Form A was correlated with Form B except in case of Chapman Test.

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

5 pts (5) Pub: PSPC (7) Gra: 3–4; 5–6; 6; 7; Jr. High Sch. (8) Time: 40′–60′ including scoring (10) $4.00 per 100 for each pt plus $.75 for time cards for each pt (11) Function measured: Rate and comprehension.

(1) Chapman, J. C. (2) — Unspeeded Reading Comprehension Test (3) Date: 1925 (4) 1f (5) Pub: Lippincott (6) Reliab: *Current and Ruch*[1]: $r_{1I} = .89$ *and* $\sigma_1 = 7.3$. r_{1I} *was derived by Spearman-Brown formula.* (5) Gra: 5–12 (8) Time: 30′ (10) Cost: $1.00 per 50.

(1) Gates, A. I. (2) — Silent Reading Tests, Grade 3–8 (3) Date: 1926 (4) 4 pts of 2f each; A reading to appreciate general significance, B reading to predict outcome of given events, C reading to understand precise directions, D reading to note details (5) Pub: T. C. Bur. Pub. (7) Gra: 3–8 (8) Time: 30′ (10) Cost: $8.00 per 100. Separate pts sold at $3.00 per 100.

(*l*) **Junior High School Reading Tests: Ind. 3, Gr. 4.**

(1) Thorndike, E. L. and McCall, W. A. (2) — Reading Scales. See (*k*).

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Reading Test. See (*k*).

(1) Van Wagenen, M. J. (2) — Reading Scales A, B and C. History (2f), General Science (2f), English Literature (3f), and English Literature Interpretation (2f). (3) Date: 1921 (4) 4 pts (5) Pub: PSPC (6) Reliab: r_{1I} and σ not given, but P.E. of score = 3.5 scale points or approximately $\frac{1}{3}$ stand. dev. of high school freshmen (7) Gra: 7–12 (8) Time: 50′ each part (9) Talent: good cler. (10) Cost: $3.00 per 100 (11) Function measured: Ability to pick out the topic of the paragraph, the ability to summarize what has been read, the ability to evaluate the accuracy of the data or its relative value.

---

[1] See footnote, page 308.

Generated on 2020-12-23 01:32 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Haggerty, M. E. (2) — Reading Examination, Sigma 3. See (*k*).

Tests not rated: (1) Van Wagenen, M. J. (2) — English literature interpretative reading scale alpha and beta. (3) Date: 1927 (4) 2f (5) Pub: PSPC (6) Reliab: P.E. *score* = 2, which is approximately one-third of normal gain of '6 scale points made during a grade in the elementary school. N = 600. The σ for 8th grade pupils lies between 8 and 9 scale points. Reliab. reported is equivalent to $r_{11} = .94$ for an 8th gra. group (7) Gra: Jun. and Sr. high school (8) Time: 40′.

(1) Gates (2) — Silent Reading Tests. See (*k*).

(*m*) **High School Reading Tests: Ind. 4. Gr. 5.**

(1) Van Wagenen, M. J. (2) — Reading Scales A, B, and C. See (*l*).

(1) Thorndike, E. L. and McCall, W. A. (2) — Reading Scales. See (*k*).

(1) Inglis, A. (2) — Vocabulary Test, 1923.

(1) Monroe, W. S. (2) — Standardized Silent Reading Test (3) Date: 1919 (4) 2f (5) Pub: PSPC (6) Reliab: Test is similar in form to Monroe Standardized Silent Reading Test, II for gra. 6–8 (7) Gra: 9–12 (8) Time: 8′ (9) Talent: good cler. (10) Cost: $1.00 per 100 (11) Function measured: Speed and comprehension in silent reading.

(1) Ruch, G. M. (2) Iowa Reading Comprehension Test. (6) Reliab: Ruch and Stoddard: $r_{11} = .88$; $\sigma = 6.6$; N = 100 in gra. 12.

---

(1) Haggerty, M. E. (2) — Reading Examination, Sigma 3. See (*k*).

(1) Thorndike, E. L. (2) — Test of Word Knowledge. See (*k*).

(1) Monroe, W. S. (2) — Standard Silent Reading Test. See (*k*).

(1) Whipple, G. M. (2) — High School and College Reading Test.

Generated on 2020-12-23 01:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Test not rated: (1) Van Wagenen. (2) — Reading scales *A*, *B*, and *C*. See (*l*).

(*n*) **College Reading Tests: Ind. 1, Gr. 4.**

(1) Inglis, A. (2) — Vocabulary Test.

(1) Whipple, G. M. (2) — High School and College Reading Test.

(1) Trabue, M. R. and Kelley, T. L. (2) — Completion Test Language Scales Alpha and Beta. See (*c*).

(1) Ruch, G. M. (2) Iowa Reading Comprehension Test. See (*m*).

(1) Thorndike, E. L. (2) — Test of Word Knowledge. See (*k*).

Test not rated: (1) Steeves et al. (2) Columbia Research Bureau English test. See (*r*).

(*o*) **Elementary Reading Tests, Oral: Ind. $1\frac{1}{2}$, Gr. $1\frac{1}{2}$.**

(1) Gates, A. I. (2) — Graded Word Knowledge Test (3) Date: 1924 (4) 4f (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{11} *per gra group* (*estimated from data on pp 216–217, T. C. Record, Vol. 26, No. 3*) = *ca. .80* (7) Gra: 1–6 (8) Time: no time limit (9) Talent: good elementary teacher (11) Function measured: Pronunciation.

(1) Gray, W. S. (2) New Standardized Oral Reading Check Test (3) Date: 1923 (4) 5f (5) Pub: PSPC (6) Reliab. j-a: + ind (7) Gra: 1–8. Set I: gra. 1–2. Set II: gra. 2–4. Set III: gra. 4–6. Set IV: gra. 6–8. (8) Time: to read 150 words (9) Talent: good teacher of reading (10) Cost: $1.50 per set, 20 of each of 5f (11) Function measured: Speed and accuracy, except comprehension or quality of expression.

(*p*) **Elementary Literature Appreciation Test: Ind. 1, Gr. 0.**

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford History and Literature Information Test. See (*ss*).

Generated on 2020-12-23 01:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(q) Junior High School Literature Appreciation Tests: Ind. 1½, Gr. ½.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford History and Literature Information Test. See (*ss*).

---

(1) Van Wagenen, M. J. (2) — Reading Scales A, B, and C. English Literature. See (*l*).

Tests not rated: (1) Burch, (2) — Comprehension in Literature Test See (*r*).

(1) McDade, J. E. (2) Plymouth Educational Tests No. 130A and No. 132A (4) 2 tests, 1f each (5) Pub: Plymouth Press (7) Gra: 3–8 (10) Cost: Each test 60¢ per 100 (11) Function measured: Test 130A measures pupils familiarity with English literary classics, Test 132A measures familiarity with authorship of classics in English.

(r) High School Literature Appreciation Tests: Ind. 2, Gr. 1.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. Stanford History and Literature Information Test. See (*ss*).

(1) Van Wagenen, M. J. (2) — Reading Scales *A*, *B*, and *C*, English Literature. See (*l*).

---

(1) Abbott, A., and Trabue, M. R. (2) — Exercises in Judging English Poetry (3) Date: 1921 (4) 2 f (5) Pub: T. C. Bur. Pub. (6) Reliab: (Elementary School) $r_{11} = 0$; (High School) $r_{11} = .44$; (College) $r_{11} = .66$; (Graduate English Students) $r_{11} = .72$. Reliab. j-a: + gr. (7) Gra: 12 and college (8) Time: ca. 45′ (9) Talent: good cler. (10) Cost: 5¢ per copy (11) Function measured: Appreciation of poetry, except the analysis of moods and bases for judgments expressed.

Tests not rated: (1) Steeves, H. R., Abbott, Allan, and Wood, B. D. (2) Columbia Research Bureau English Test (3) Date: 1925 (4) 2f of 4 pts (5) Pub: WBC (6) Reliab: By Spearman-Brown formula $r_{11} = .965$. N = 100 entering freshman in two universities. $\sigma_1 = 30.5$. Reliab. of pt 1, spelling = .80;

Generated on 2020-12-23 01:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of pt 3, vocabulary = .94; of pt 4, literary knowledge = .90 (7) Gra: 12–13 (8) Time: 2 hrs. (9) Talent: clerks (10) Cost: $1.40 per 25 (11) Function measured: Spelling; the mechanics of English, including punctuation; vocabulary; and literary knowledge.

(1) Burch, Dr. Mary C. (2) — Comprehension in Literature Test (3) Date: 1927 (4) 2f each of Tests 1, 2, 3 (5) Pub: Author, 15 So. 13th St. San Jose, California. (6) Reliab: Gra. 7–12 combined, Test 1 .936, Test 2 .939, Test 3 .929, Three tests combined .976 (7) Gra: 7–12 (8) Time: ca. 46′ for 3 tests (11) Function measured: Comprehension of English literature at different levels of difficulty.

(1) Van Wagenen (2) — English literature interpretative reading scale alpha and beta. See (*l*).

(*s*) Elementary and Junior High School Composition Scales: Ind. o, Gr. 9.

(1) Hudelson, E. (2) — English Composition Scale (3) Date: 1921, Rev. 1923 (4) 1f (5) Pub: WBC (6) Reliab: r_{1I} = .40 (*estimated reliab. of rating of one judge*) Reliab. j-a: gr (7) Gra: 4–12 (8) Time: irrelevant (9) Talent: Teacher or supervisor of judgment — preferably a teacher of English trained in the use of composition scales (10) Cost: manual and scale, 25¢ (11) Function measured: General merit of a composition, except that the specific qualities of composition (such as spelling, coherence, etc.) are not measured separately.

(1) Trabue, M. R. (2) Nassau County Supplement to Hillegas Scale (3) Date: 1915 (4) 1f. For all practical purposes this scale is one form of the Hillegas Scale. (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{1I} = .82 (Median judgments of 4 teachers against those of 4 other teachers *when rating a single composition*) Reliab. j-a: ind. if repeated trials are used as a basis. (7) Gra: 4–12 (8) Time: irrelevant (9) Talent: Teacher or supervisor of good judgment (10) Cost: 8¢ per copy

Generated on 2020-12-23 01:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(11) Function measured: English composition, except speed of composition and detailed analysis of faults.

(1) Lewis, E. E. (2) — English Composition Scales (3) Date: 1921 (4) 5 pts of 1f each, 4 letter writing scales and one narrative composition scale. (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 4–12 (8) Time: irrelevant (9) Talent: no specific talent (10) Cost: 25¢ each (11) Function measured: Letter writing.

(1) Hudelson, E. (2) — Typical Composition Ability Scale. See (*t*).

(1) Hudelson, E. (2) — Maximal Composition Ability Scale.

(1) Thorndike, E. L. (2) — Extension of the Hillegas Scale.

(1) Van Wagenen, M. J. (2) — English Composition Scales (3) Date: 1923 (4) 1f (5) Pub: WBC (6) Reliab: depends upon capacity and training of the teacher evaluating the work. Reliab. j-a: much more reliable (7) Gra: 4–12 (8) Time: irrelevant (9) Talent: Exceptionally capable teacher of English Composition (10) Cost: 25¢ each (11) Function measured: English Composition, except more detailed elements in composition writing, as spelling, punctuation and grammar as such.

Test not rated: (1) Clark, F. L. (2) — Letter writing test (3) Date: 1926 (4) 1f of 3 pts (5) Pub: PSPC (7) Gra: 5–12 (10) Cost: $3.00 per 100.

(*t*) **High School Composition Scales: Ind. o, Gr. $8\frac{1}{2}$.**

(1) Hudelson, E. (2) — English Composition Scale. See (*s*).

(1) Trabue, M. R. (2) Nassau County Supplement to the Hillegas Scale. See (*s*).

(1) Thorndike, E. L. (2) — Extension of the Hillegas Scale.

Generated on 2020-12-23 01:33 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Lewis, E. E. (2) — English Composition Scales. See (*s*).

(1) Hudelson, E. (2) — Typical Composition Ability Scale (3) Date: 1923 (4) 1f (5) Pub: PSPC (6) Reliab. j-a: + ind (7) Gra: 1–16 (8) Time: irrelevant (10) Cost: 25¢ each (11) Function measured: General composition merit; not specific merits.

(1) Hudelson, E. (2) — Maximal Composition Ability Scale.

(1) Van Wagenen, M. J. (2) — English Composition Scales. See (*s*).

Test not rated: (1) Clark. (2) — Letter writing test. See (*s*).

(*u*) **Elementary Spelling Tests: Ind. $5\frac{1}{2}$, Gr. 5.**

(1) Morrison, J. C., McCall, W. A. (2) — Spelling Scale (3) Date: 1923 (4) 8f (5) Pub: WBC (6) Reliab: r_{11} = .931; N = 577; gra. range is 2–8 inclusive. *Ruch reports (in a personal letter) that C. L. Cushman correlated Test I with Test XIII and found: (a) r_{11} = .84; σ = 6.32; N = 66 in gra. 3. (b) r_{11} = .90; σ = 5.98; N = 70 in gra. 4. (c) r_{11} = .75; σ = 5.06; N = 54 in gra. 5. (d) r_{11} = .86; σ = 6.72; N = 55 in gra. 6.* Reliab. j-a: + ind. McCall qualifies this by: "With caution as to reliability of .7 to .9" (7) Gra: 2–8 (8) Time: ca. 25′ (9) Talent: good cler. (10) Cost: 25¢ (11) Function measured: Morrison: Spelling, except words involving use of capital letters. McCall: Spelling, except non-conscious spelling.

(1) Ashbaugh, E. J. (2) Iowa Spelling Scales (3) Date: 1922 (4) Many comparable lists may be built up from scaled words given (5) Pub: PSPC (6) Reliab. j-a: + gr (7) Gra: 2–8 (8) Time: varies (9) Talent: good speller (10) Cost: all 7 scales 50¢. Single grades 6¢ each in quantity (11) Function measured: Spelling.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Dictation Test (3) Date: 1923 (4) 2f

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(5) Pub; WBC (6) Reliab: Average r_{11} per gra. for gra. 2–9 = .86; N = 1204 in gra. 2–9; Average N = 150; Average σ per gra. = 21.3. Reliab. j-a: + ind (7) Gra: 2–9 (8) Time: ca. 20′ (9) Talent; good cler. (11) Function measured: Spelling ability, except ranges of ability considerably above or below the average for the school gra. tested, e.g. a very superior speller in the 9th gra. is inadequately measured by the 9th gra. test.

(1) Briggs, T. H., et al. (2) Sixteen Spelling Scales. See (*v*).

(1) Buckingham, B. R. (2) — Extension of Ayres Scale.

(1) Ayres, L. P. (2) — Spelling Scale (3) Date: 1915 (4) 1 list 1000 words (5) Pub: Russell Sage Found. (6) Reliab. j-a: + ind (7) Gra: 3–8 (8) Time: varies (no. of words used) (9) Talent: Ability to read scale (10) Cost: 10¢ for scale (11) Function measured: Spelling, except ability to spell words which are not among the 1000 most commonly used.

---

(1) Tidyman, W. F. (2) — Standard Spelling Test.

(1) Courtis, S. A. (2) — Standard Supervisory Tests in Spelling.

(1) Monroe, W. S. (2) — Timed Sentence Spelling Test (3) Date: 1918 (4) 1f. Test I, gra. 3–4: Test II, gra. 5–6: Test III, gra. 7–12 (5) Pub: PSPC (6) Reliab. j-a: ind (7) Gra: 3–12 (8) Time: 12′ (9) Talent: good speller (10) Cost: each test 4¢.

Test not rated: (1) Van Wagenen, M. J. (2) — Spelling Scales (3) Date: 1926 (4) 5 sca (5) Pub: PSPC (6) Reliab: P. E. of individual score = 1.3 scale points, or approximately $\frac{1}{3}$ of the normal gain of 4 scale points made during a grade in the elementary school. Based on N = 500. The σ of 1200 8th gra. pupils at St. Paul was 5.7. *The reliab. reported is equivalent to* r_{11} = .95 *for this 8th gra. group* (10) Cost: Single copy 20¢.

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(*v*) **Junior High School Spelling Tests: Ind. 3, Gr. $2\frac{1}{2}$.**

(1) Briggs, T. H., Hudelson, E., Kelley, T. L., Stetson, E. L., and Woodyard, E. (2) Sixteen Spelling Scales Standardized in Sentences for Secondary Schools (3) Date: 1920 (4) 12f easy words and 4f hard words (5) Pub: T. C. Bur. Pub. (6) Reliab: .65–.70 per gra. group for a single scale (estimated by T. L. K.) Reliab. j-a: gr. (if 1 scale is used): ind. (if 3 are used) (7) Gra: 7–12 (8) Time: ca. 10′ per scale (9) Talent: good cler. (10) Cost: 40¢ (11) Function measured: Spelling, except spelling when the attention is otherwise engaged.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Dictation Test. See (*u*).

(1) Morrison, J. C. and McCall, W. A. (2) — Spelling Scale. See (*u*).

---

(1) Monroe, W. S. (2) — Timed Sentence Spelling Test. See (*u*).

Test not rated: (1) Van Wagenen. (2) — Spelling scales. See (*u*).

(*w*) **High School Spelling Tests: Ind. 2, Gr. $1\frac{1}{2}$.**

(1) Briggs, T. H., Hudelson, E., Kelley, T. L., Stetson, E. L., and Woodyard, E. (2) Sixteen Spelling Scales Standardized in Sentences for Secondary Schools. See (*v*).

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Dictation Test. See (*u*).

Test not rated: (1) Van Wagenen. (2) — Spelling scales. See (*u*).

(*x*) **Elementary Language Usage Tests: Ind. 3, Gr. $6\frac{1}{2}$.**

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Language Usage Test (3) Date: 1923 (4) 2f (5) Pub: WBC (6) Reliab: Average r_{11}

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

per gra. for gra. 4–9 = .67; N = 839 in gra. 4–9; Average N = 140; Average σ per gra. = 10.4. Reliab. j-a: = gr. (7) Gra: 4–9 (8) Time: 8′ (9) Talent: good cler. (11) Function measured: Language usage, except language habits when attention is otherwise engaged.

(1) Charters, W. W. (2) — Diagnostic Language Test. See (*y*).

(1) Charters, W. W. (2) — Diagnostic Language and Grammar Test. See (*y*).

---

(1) Kirby, T. J. (2) — Grammar Test. (6) Reliab: *Ruch and Stoddard: Principles* r_{11} = *.91;* σ = *9.1; N = 128 in gra. 7–12. Sentences:* r_{11} = *.70;* σ = *4.3; N = 136 in gra. 7–12. Additional reliab. coef's given by Ruch and Stoddard.*

(1) Wilson, G. M. (2) — Language Error Test. See (*y*).

(1) Pressey, S. L. and Ruhlen, H. (2) — Diagnostic Tests in English Composition (Punctuation). See (*bb*).

(1) Pressey, S. L. and Bowers, E. V. (2) — Diagnostic Tests in English Composition (Capitalization). See (*bb*).

(1) Pressey, S. L. and Conkling, F. R. (2) — Diagnostic Tests in English Composition (Grammar or inflected forms). See (*bb*).

Tests not rated: (1) Coxe, W. W., Cornell, Ethel L., Orleans, J. S., and Richards, E. B. (2) New York English Survey Tests (3) Date: 1925 (4) 1f of 4 pts (5) Pub: PSPC (7) Gra: Language usage 4–8; sentence structure 4–8; grammar 7–8; literature information 7–8 (10) Cost: $1.00 per 100 for each pt.

(1) Franzeen, C. E. (2) — Diagnostic Tests in Language (3) Date: 1924 (4) 2f of 3 pts (5) Pub: Bureau of Administrative Research, University of Cincinnati (7) Gra: 3–8 (8) Time: 20′–45′

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(10) Cost: $2.00 per 100 for each pt (11) Function measured: Pt 1, pronouns; Pt 2, verbs; Pt 3, varied constructions.

(*y*) **Junior High School Language Usage and Grammar Tests: Ind. 3, Gr. 4.**

(1) Charters, W. W. (2) — Diagnostic Language and Grammar Test (3) Date: 1918, Rev. 1922 (4) 2f of 3 pts each (5) Pub: PSPC (6) Reliab: *G. M. Ruch reports (in a personal letter): (a) Language, r_{11} = .78; σ = 6.85; N = 80 in gra. 9: (b) Grammar, r_{11} = .78; σ = 7.40; N = 80 in gra. 9.* (7) Gra: 7–8 (8) Time: no time limit (9) Talent: good cler. (10) Cost: $1.50 per 100 (11) Function measured: Language usage, except initiative; i. e. original composition.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Language Usage Test. See (*x*).

(1) Wilson, G. M. (2) — Language Error Test (3) Date: 1923 (4) 3f (5) Pub: WBC (6) Reliab: r_{11} = .90; gra. = 3–8 combined; N = 103; σ = 7.4; *Estimated r = ca. .65 for single grade.* Reliab. j-a: + ind: For placement, helpful; for diagnosis, very helpful. (7) Gra: 3–12 (8) Time: 5′ to 15′ (9) Talent: good cler. (10) Cost: 80¢ per 25 (11) Function measured: Language forms, except technical grammar, composition ability.

---

(1) Charters, W. W. (2) — Diagnostic Language Test (3) Date: 1918, Rev. 1922 (4) 2f of 5 pts each (5) Pub: PSPC (6) Reliab: See (*y*) *Charters Diagnostic Language and Grammar Test.* (7) Gra: 4–8 (8) Time: no time limit (9) Talent: good cler. (10) Cost: 80¢ per 100 (11) Function measured: Language usage, except initiative of expression.

(1) Briggs, T. H. (2) — English Form Test. See (*bb*).

(1) Briggs, T. H. (2) — Analogies Test.

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Tests not rated: (1) Cox et al. (2) N. Y. English Survey Tests. See (*x*).

(1) Leonard, S. A. (2) — Test of Grammatical Correctness (3) Date: 1923 (4) 1f (5) Pub: National Council of Teachers of English, 506 W. 69th St., Chicago (6) Reliab: r_{11} = .67. N = 756 pupils in gra. low 5 — high 11 (7) Gra: 7–12 (8) Time: ca. 12′ (10) Cost: 90¢ per 100.

(1) Leonard, S. A. (2) — Sentence Recognition Test (3) Date: 1923 (4) 1f (5) Pub: National Council of Teachers of English, Chicago (6) Reliab: r_{11} = .75 N = 582 pupils in gra. low 5 — high 11 (7) Gra: 7–12 (8) Time: ca 12′ (10) Cost: 90¢ per 100.

(1) Witham, E. C. (2) — Grammar Test (pronouns) (3) Date: 1924 (5) Pub: J. L. Hammett Co. (7) Gra: 6–8 (10) Cost: $1.00 per 50.

(1) Franzeen. (2) — Diagnostic tests in language. See (*x*).

(1) McDade, J. E. (2) — Language-Grammar Test (3) Date: 1924 (4) 1f (5) Pub: Plymouth Press (7) Gra: 4–12 (8) Time: 20′ (10) Cost: $3.00 per 12 (folders may be re-used).

(*z*) **High School Language Usage and Grammar Tests: Ind. $\frac{1}{2}$, Gr. 1$\frac{1}{2}$.**

(1) Wilson, G. M. (2) — Language Error Test. See (*y*).

---

(1) Starch, D. (2) — English Grammar Test.

Tests not rated: (1) Steeves et al. (2) Columbia Research Bureau English Test. See (*r*).

(1) Tressler, J. C. (2) — English Minimum Essentials Test (3) Date: 1925 (4) 3f of 7 pts each (5) Pub: PSPC (6) Reliab: Forms A and B correlated. r_{11} = .78. σ of Form A scores = 11.9. N = 123, all in low 12th gra. (7) Gra: 8–12 (8) Time: 40′–55′ (10) Cost: 75¢ per 25 per form

Generated on 2020-12-23 01:34 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(11) Function measured: Good usage in grammatical correctness, vocabulary, punctuation and capitalization, sentence structure, sentence sense, inflection and accent, and spelling.

(1) Leonard (2) — Grammatical Correctness. See (*y*).
(1) Leonard (2) — Sentence Recognition. See (*y*).
(1) McDade (2) — Language-Grammar test. See (*y*).

(*aa*) **Elementary English Form Test: Ind. $\frac{1}{2}$, Gr. $\frac{1}{2}$.**

(1) Pressey, S. L. and Conkling, F. R. (2) — Diagnostic Tests in English Composition (Sentence Structure). See (*bb*).

(*bb*) **Junior High School English Form Test: Ind. 1, Gr. 1.**

(1) Briggs, T. H. (2) — English Form Test (3) Date: 1921 (4) 2f of 7 pts each (5) Pub: T. C. Bur. Pub. (6) Reliab: $r_{11} = .76$; N = 100; gra. = 7 and 8 combined. *Cr: Deduced from a population of 88 in gra. low 7–low 8. Total:* $r_{11} = .79$; $\sigma = 4.1$. *Average of the reliabilities for the 7 separate pts:* $r_{11} = .35$; *gra. low 7–h.7.* Reliab. j-a: + ind only diag. (7) Gra: 7–9 (8) Time: no time limit (9) Talent: good cler. (10) Cost: 80¢ per 100 (11) Function measured: 7 fundamental details of punctuation or capitalization — or better, of composition form.

---

(1) Pressey, S. L., Ruhlen, H., Conkling, F. R., and Bowers, E. V. (2) — Diagnostic Tests in English Composition, — Language Usage and Grammar (3) Date: 1923 (4) 1f of 4 pts (5) Pub: PSPC (6) Reliab: *Ruch and Stoddard: By Spearman-Brown formula: Capitalization,* $r_{11} = .79$; $\sigma = 3.9$. *Punctuation,* $r_{11} = .64$; $\sigma = 5.4$. *Grammar,* $r_{11} = .90$; $\sigma = 6.0$. *Sentence Structure,* $r_{11} = .73$; $\sigma = 3.8$. *N = 99 in gra. 9* Reliab. j-a: + ind (7) Gra: 7 and above (8) Time; varies with part 10′–20′ (9) Talent: good cler. (10) Cost: 100 for 75¢ for "Capitalization" or "Punctuation": 100 for $1.50 of "Inflected Forms" or "Sentence Structure" (11) Function

Generated on 2020-12-23 01:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

measured: English Composition, except obviously, paragraphing, word choice, and rhetorical factors.

(*cc*) High School English Form Tests: Ind. 1, Gr. $\frac{1}{2}$.

(1) Briggs, T. H. (2) — English Form Test. See (*bb*).

---

(1) Starch, D. (2) — Punctuation Scale (3) Date: 1916 (4) 1f (5) Pub: University Coöperative Company, 506 State Street, Madison, Wisconsin. (6) Reliab. j-a: + ind (7) Gra: 5–12 (8) Time: no time limit (9) Talent: good cler. (10) Cost: 80¢ per 100.

Test not rated (1) Steeves et al. (2) Columbia Research Bureau English Test See (*r*).

(*dd*) Elementary Arithmetic Tests: Ind. 5, Gr. 11.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Arithmetic Test (3) Date: 1923 (4) 2f of 2 pts each (5) Pub: WBC (6) Pt 1 (Ar. Comp.) Average r_{1I} per gra. for gra. 2–9 = .74. Pt 2 (Ar. Reasoning) Average r_{1I} per gra. for gra. 2–9 = .77. Average r_{1I} for total score (same conditions) ca. .85; N = 1204 in gra. 2–9; Average N per gra. = 150; the σ's for each part and each grade are given in the manual. The σ of the arithmetic total score (h.2 and low 3 combined) = 22.1; (low 8 and h.8 combined) = 37.7. Reliab. j-a: + ind (7) Gra: 2–9 (8) Time: gra. 2–3, 20′, gra. 4–8, 40′ (9) Talent: good cler. (10) Cost: $1.00 per 25 (11) Function measured: Arithmetic, except speed in arithmetic computation and speed in arithmetic reasoning.

(1) Buckingham, B. R. (2) — Scale for Problems in Arithmetic (3) Date: 1919 (4) 2f 3 div. each (5) Pub: PSPC (6) Reliab: *Cr: found on Div. II, f1 vs f2; N = 38; gra. low 6–h.6;* r_{1I} = .72. Reliab. j-a: + ind (7) Gra: Div. 1, gra. 3–4; Div. 2, gra. 5–6; Div. 3, gra. 7–8 (8) Time: not over 1 hour (9) Talent: good cler. (10) Cost:

Generated on 2020-12-23 01:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

80¢ per 100 (11) Function measured: Arithmetic, except ability in fundamentals of any but the simplest character.

(1) Woody, C. (2) — Arithmetic Scales (3) Date: 1916, Rev. 1920 (4) 1f Ser A, 2f Ser B (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{11} = ca. .75 on the average, depending on teachers, scores, amount of testing, etc. Reliab. j-a: + (7) Gra: 3–8 (8) Time A–20′, B–10′ (9) Talent: good cler. (10) Cost: A $1.00 per 100, B $1.50 per 100.

(1) Woody, C. and McCall, W. A. (2) — Mixed Fundamentals (3) Date: 1917 (4) 4f (5) Pub: T. C. Bur. Pub. (6) Reliab: *Kelley findings: f1 vs f2; r_{11} = .70; N = 70; gra. = 8; σ = 3.34. Cr. findings: f1 vs f2; N = 57; gra. low 8–h.8; r_{11} = .70. Ruch reports (in a personal letter) that C. L. Cushman found:* (a) *r_{11} = .71; σ = 2.71; N = 87 in gra. 3.* (b) *r_{11} = .55; σ = 3.52; N = 85 in gra. 4.* (c) *r_{11} = .81; σ = 2.80; N = 65 in gra. 5.* (d) *r_{11} = .50; σ = 2.48; N = 74 in gra. 6.* Reliab. j-a: + ind with caution (7) Gra: 3–8 (8) Time: 20′ (9) Talent: good cler. (10) Cost: 60¢ per 100 (11) Function measured: McCall states: "Arithmetic; except arithmetic of problem variety, arithmetic beyond fundamentals in integers, fractions and decimals, exact measure of weight."

(1) Woody, C. and Van Wagenen, M. J. (2) — Arithmetic Scales.

---

(1) Monroe, W. S. (2) — Diagnostic Arithmetic Test (3) Date: 1917 (4) 1f of 4 pts, total of 21 tests (5) Pub: PSPC (6) Reliab: *Cr. findings on 1st 11 tests: Addition, 3 tests, average r_{11} = .45; subtraction, 2 tests, r_{11} = .84; multiplication, 3 tests, average r_{11} = .70; division, 3 tests, average r_{11} = .64; N = 56; gra. low 6–h.6. Total of 11 tests (total score not used in diagnosis) r_{11} = ca. .93.* Reliab. j-a: + ind (7) Gra: Pt I, Integers gra. 4–8; Pt II, Integers, gra. 5–8; Pt III, Common Frac-

Generated on 2020-12-23 01:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

tions, gra. 6–8; Pt IV, Decimal Fractions, gra. 6–8 (8) Time: Pt 1–8′, Pt 2–12′, Pt 3–11′, Pt 4–2.5′ (9) Talent: good cler. (10) Cost: 80¢ per 100.

(1) Spencer, P. L. (2) — Diagnostic Arithmetic Test (3) Date: 1923 (4) 2f, 3 ser each (5) Pub: Bureau of Administrative Research, University of Cincinnati (7) Gra: Test I, gra. 3–4; Test II, gra. 4–6; Test III, gra. 7–8 (8) Time: 2 hours (9) Talent: good cler. (10) Cost: $2.00 per 100 (11) Function measured: Arithmetic, except speed and the degree of difficulty that a child can master.

(1) Judd, C. H., Courtis, G. H., Courtis, S. A., and Ayres, L. P. (2) Cleveland Survey Arithmetic Test (3) Date: 1916 (4) 1f (5) Pub: Courtis, Detroit (6) Reliab. j-a: + ind (7) Gra: 3–8 (8) Time: 22′ (9) Talent: good cler. (10) Cost: $2.20 per 100 (11) Function measured: Arithmetic, except decimals, percentage, mensuration, and reasoning.

(1) Stevenson, P. R. (2) — Arithmetic Problem Analysis Test (3) Date: 1923 (4) 2f (5) Pub: PSPC (6) Reliab. j-a: + ind (7) Gra: 4–6 and 7–9 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 100 (11) Function measured: Arithmetic, except fundamentals.

(1) Otis, A. S. (2) — Arithmetic Reasoning Test (3) Date: 1918, Rev. 1920 (4) 2f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 4–9 (8) Time: 6′ (9) Talent: good cler. (10) Cost: 40¢ per 25 (11) Function measured: Arithmetic reasoning; exception, none.

(1) Monroe, W. S. (2) — Standardized Reasoning Tests in Arithmetic (3) Date: 1918 (4) 2f of 3 tests each (5) Pub: PSPC (6) Reliab: *Cr. findings on test II: Correct principle,* $r_{11} = .60$; *correct answer,* $r_{11} = .56$; $N = 52$; *gra. low 7–h.7.*

Generated on 2020-12-23 01:35 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Reliab. j-a: + ind (7) Gra: Test I, gra. 4–5: Test II, gra. 6–7: Test III, gra. 8 (8) Time: 25′ (9) Talent: good cler. (10) Cost: 80¢ per 100.

(1) Courtis, S. A. (2) — Standard Research Tests, Arithmetic, Series B (3) Date: 1914 (4) 4f (5) Pub: Courtis, Detroit (6) Reliab. j-a: + ind (7) Gra: 4–8 (8) Time: 26′ (9) Talent: good cler. (10) Cost: 1½¢ per copy (11) Function measured: Arithmetic, except 4 processes with whole numbers.

(1) Monroe, W. S. (2) — General Survey Arithmetic Tests (3) Date: 1920, Rev. 1920–21 (4) 3f, 2 scales each (5) Pub: PSPC (6) Reliab. j-a: + ind (7) Gra: Sca 1, gra. 3–5: Sca 2, gra. 6–8 (8) Time: Sca 1, 7′; Sca 2, 17.5′ (9) Talent: good cler. (10) Cost $1.00 per 100.

(1) Peet, H. E. and Dearborn, W. F. (2) — Progress Tests in Arithmetic.

Tests not rated: (1) Ruch, G. M., Knight, F. B., Greene, H. A., and Studebaker, J. W. (2) Compass Diagnostic Tests in Arithmetic (3) Date: 1925 (4) 20 tests (5) Pub: Scott, Foresman and Co. (7) Gra: 5–8 (8) Time: Varies on different tests from 16′–61′ (10) Cost: Varies from $.20 to $1.20 per 25.

(1) Jones, F. D. (2) — Self Correcting problems (3) Date: 1918, revised 1922 and 1925 (4) 5 sets of cards, 1 ser. for each gra. 2, 3, 4, 5, and 1 ser. for gra. 5–8 combined (5) Pub: Jones Mfg. Co., Alhambra, Calif. (7) Gra: 2, 3, 4, 5, and 5–8 (9) Talent: Pupils, as tests are "self correcting."

(1) Wildeman, Edw. (2) — Test in common fractions (3) Date: 1922 (4) 1f (5) Pub: Plymouth Press, Chicago (7) Gra: 5–8 (8) Time: 15′ (10) Cost: 90¢ per 100.

(*ee*) Junior High School Arithmetic Tests: Ind. 2, Gr. 2.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Arithmetic Test. See (*dd*).

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Buckingham, B. R. (2) — Scale for Problems in Arithmetic. See (*dd*).

---

(1) Otis, A. S. (2) — Arithmetic Reasoning Test. See (*dd*).

Tests not rated: (1) Ruch et al. (2) Compass diagnostic tests in arithmetic. See (*dd*).

(1) Wildeman. (2) — Test in common fractions. See (*dd*).

(*ff*) High School and College Arithmetic Test: Ind. 0, Gr. 1.

(1) Thurstone, L. L. (2) — Arithmetic Test (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 12–13 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 25 (11) Function measured: Arithmetic reasoning, except speed in calculation.

(*gg*) Junior High School Algebra Test: Ind. 1, Gr. 0.

(1) Rogers, A. L. (2) — Test of Mathematical Ability. (6) Reliab: *Ruch and Stoddard:* $r_{11} = .82$; $\sigma = 34$; $N = 28$ *in gra. 9.*

(*hh*) High School Algebra Tests: Ind. 3, Gr. 3.

(1) Hotz, H. G. (2) — Algebra Scales (3) Date: 1918 (4) 1f, 5 sca: addition and subtraction, multiplication and division, equation and formula, problem, and graph. (5) Pub: T. C. Bur. Pub. (6) Reliab: *Ruch and Stoddard:* $r_{11} = .92$; $\sigma = 8.70$; $N = 175$ *pupils in gra. 9.* Reliab. j-a: + ind (7) Gra: first year algebra (8) Time: each scale 20′ (9) Talent: good cler. (10) Cost: 70¢ per 100.

(1) Douglas, H. R. (2) — Diagnostic Tests for 1st year Algebra (3) Date: 1921, Rev. 23 (4) 2f, 2 ser each (5) Pub: Bureau of Administrative Research, University of Cincinnati (6) Reliab: *Ruch and Stoddard:* $r_{11} = .80$; $\sigma = 5.20$; $N = 175$ *first year pupils. Again,* $r_{11} = .84$; $\sigma = 4.89$; $N = 43$ *first year pupils.* Ser A, $r_{11} = .63$; gra. = 9; N = 104; average $\sigma = 4.7$ (*No* r_{11} *for Ser B*) Re-

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

liab. j-a: = ind (7) Gra: Series A for end of first semester in algebra, and Series B for end of second semester. (8) Time: A 40′, B 105′ (9) Talent: good cler. (10) Cost: A $1.60, B $4.00 per 100 (11) Function measured: Algebra, except definition of terms, ability to state rules, axioms, theorems, verbal problems (separately), many minor processes.

(1) Thurstone, L. L. (2) — Algebra Test (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 12–13 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 25.

---

(1) Rugg, H. A. and Clark, J. R. (2) — Standardized Tests in 1st year Algebra.

(1) Kelley, T. L. (2) — Mathematical Values Test. See (*jjj*).

(1) Monroe, W. S. and Williams, L. W. (2) Illinois Standardized Algebra Tests (3) Date: 1920 (4) 1f (5) Pub: PSPC (6) Reliab: *Ruch and Stoddard:* $r_{11} = .88$; $\sigma = 9.32$; $N = 38$ *gra. 9 pupils.* Reliab. j-a: Monroe: + ind. Williams: =, perhaps a little better; ind, test needs perfecting, however. (7) Gra: high school classes (8) Time: ca. 30′ (9) Talent: good cler. (10) Cost: $2.50 per 100 (11) Function measured: Algebra 1st year processes.

Test not rated: (1) Otis, A. S. and Wood, B. D. (2) Columbia Research Bureau Algebra Test (3) Date: 1927 (4) 2f (5) Pub: WBC (6) Reliab: $r_{11} = .86$, N = 322 Columbia freshman at entrance, $\sigma = 10.4$ (8) Time: 90′.

(*ii*) **College Algebra Test: Ind. 0, Gr. 1.**

(1) Thurstone, L. L. (2) — Algebra Test. See (*hh*).

(*jj*) **High School Geometry Test: Ind. 2, Gr. 3.**

(1) Hawkes, H. E. and Wood, B. D. (2) Columbia Research Bureau Plane Geometry Test (3) Date: 1923 (4) 2f, (5) Pub: WBC (6) Reliab: By

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Spearman-Brown formula $r_{11} = .93$, N = 1349 high school pupils at end of geometry course; $\sigma = 54$. Reliab. j-a: + ind (7) Gra: For pupils having completed one or more half years of plane geometry (8) Time: 1 hour (9) Talent: good cler. (10) Cost: $1.20 per 25 (11) Function measured: Plane Geometry except original applied problems, but it does include ordinary "originals."

(1) Minnick, J. H. (2) — Geometry Test (3) Date: 1919 (4) 1f (5) Pub: Houston Club Book Store, Univ. of Penn. (6) Reliab: *Ruch and Stoddard derive, just how it is not fully explained,* $r_{11} = .63$; $\sigma = 19.6$; $N = 61$. Reliab. j-a: = ind (7) Gra: high school classes (8) Time: 2 hours. There is a test, Test W, which is a 20 minute adaptation of the long one and is for group use only (9) Talent: good geometry teacher (10) Cost: $1.50 per 100 (11) Function measured: Geometry, but not the full content of geometry and not an appreciation of place of geometry in society. (*This is not an exact quotation but a paraphrasing of a longer statement from author of test.*)

---

(1) Thurstone, L. L. (2) — Geometry Test (Part of Vocational Guidance Test) (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 12 and 13 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 25 (11) Function measured: Only ability to apply principles of geometry to original problems. Does not test memory of rules, definitions, or theorem proofs.

(1) Schorling, R. and Sanford, V. (2) — Geometry Test.

(1) Schorling, R. (2) — Plane Geometry Test (3) Date: 1921.

(*kk*) College Geometry Tests: Ind. 1, Gr. 1.

(1) Hawkes, H. E. and Wood, B. D. (2) Columbia Research Bureau Plane Geometry Test. See (*jj*).

---

(1) Thurstone, L. L. (2) — Geometry Test. See (*jj*).

(*ll*) Elementary and Junior High School Geography Tests: Ind. 3, Gr. 4.

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Posey, C. J. and Van Wagenen, M. J. (2) — Geography Scales (3) Date: 1922 (4) 1f 2 div. 14 pts. All parts yield equivalent scores within the limits of the P.E. Thought: S, General, Div. 1; R, Gen, Div. 2. Information: R, Gen, Divs. 1 and 2; S, Gen, Divs. 1 and 2; T, Gen, Divs. 1 and 2; U, Gen, Div. 2; V, Gen, Div. 2; W, Gen, Div. 2; A, U.S. and North America, Divs. 1 and 2; B, U.S. and North America, Divs. 1 and 2; F, Europe, Div. 2; G, Europe, Div. 2; K, South America, Asia, Africa, Div. 2; L. South America, Asia, Africa, Div. 2. (5) Pub: PSPC (6) Reliab: P.E. of scale score is 2.1 scale points or approximately $\frac{1}{3}$ of a grade difference. *Ruch and Stoddard: Thought R, $r_{11} = .58$: $\sigma = 9.1$, $N = 169$ in gra. 5–7* (7) Gra: 5–8 (Div. 1, 5–6, Div. 2, 7–8) (8) Time: 40′ (9) Talent: super. cler. (10) Cost: $1.50 per 100 (11) Function measured: Geography, except location of places on maps, special kinds of geog. such as physical or commercial geog. as units, ability to acquire detailed geog. information from pictures, ability to read geog. stories or treatises.

(1) Spencer, P. L. and Gregory, C. A. (2) — Geography Test (3) Date: 1922 (4) 3f (5) Pub: Bureau of Administrative Research, University of Cincinnati (6) Reliab: Univ. of Ore. Bureau Research Price List gives average $r_{11} = .88$ and P.E. of Meas. = 1.62 (from which $\sigma = 6.4$). Gra. range and N not stipulated. *Ruch and Stoddard: $r_{11} = .81$; $\sigma = 19.4$; $N = 168$ in gra. 5–7* (7) Gra: 6–8 (8) Time: 45′ (9) Talent: good cler. (10) Cost: $4.00 per 100.

(1) Buckingham, B. R. and Stevenson, P.R. (2) — U. S. Geography Information and Problems (3) Date: 1923 (4) 1f (5) Pub: PSPC (6) Reliab: *Ruch and Stoddard: By Spearman-Brown formula $r_{11} = .87$; $\sigma = 11.6$; $N = 195$ in gra. 5–7* (7) Gra: 6–9 (8) Time: 14′ (9) Talent: good cler. (10) Cost: $2.00 per 100 (11) Function measured: Geography, except place geography.

Generated on 2020-12-23 01:43 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Buckingham, B. R. and Stevenson, P.R. (2) — Place Geography Test (3) Date: 1922 (4) 3f (5) Pub: PSPC (6) Reliab: *Ruch and Stoddard: av. r_{11} = .86; av. σ = 44.8; N = 82 in gra. 5–7.* Reliab. j-a: + ind (7) Gra: 4–8 (8) Time: no time limit (9) Talent: teacher of geography (10) Cost: Teacher's booklet, 20¢. (Pupil requires no material.) (11) Function measured: Geography, except general information and ability to apply geographical principles to concrete situations.

(1) Hahn, H. H. and Lackey, E. E. (2) — Geography Scale (3) Date: 1918, Rev. 1923 (4) 1f (5) Pub: PSPC (Distributors) (6) Reliab: *Ruch and Stoddard: For 30 items r_{11} = .81; σ = 5.6; N = 175 in gra. 5–7.* Reliab. j-a: = ind (7) Gra: 4–8 (8) Time: varies (9) Talent: good geog. teacher (10) Cost: Teacher's copy 20¢. (Pupil requires no material.) (11) Function measured: It is a complete geog. test, testing the entire body of subject matter treated in common, by six authors of modern textbooks on geography. The scale is arranged for diagnostic testing.

(1) Nifenecker, E. A. (2) New York Standard Geography Tests.

(1) Witham, E. C. (2) — Geography Tests.

Tests not rated: (1) Buckingham, B. R., Stevenson, P. R., Ridgley, D. C., and Shipman, Julia M. (2) Information Problems Test in Geography (3) Date: 1926 (4) 3f of Europe test; 2f of U. S. 2f of So. Am., and 2f of Asia: 2 pts to each test (*a*) information and (*b*) problems (5) Pub: PSPC (7) Gra: 5–8 (8) Time: ca. 15′ (9) Talent: Easy to score (10) Cost: $2.00 per 100 for each test.

(1) Courtis, S. A. (2) — Supervisory Geography Test (6) Reliab: *Ruch and Stoddard: By Spearman-Brown formula, r_{11} = .95; N = 166 in gra. 5–7.*

(1) McDade, J. E. (2) Plymouth Educational Tests, Nos. 60A, 63A and 64A. (5) Pub: Plymouth

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Press (10) Cost: 60¢ per 100 (11) Function measured: 60A, ability to define geographical terms; 63A, ability to locate places on the map of the world; 64A, ability to locate places on the map of the U. S.

(*mm*) **Elementary General Science Tests: Ind. 1, Gr. 0.**

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Science Information Test (3) Date: 1923 (4) 2f (5) Pub: WBC (6) Reliab: Average per gra. for gra. 4–9 = .82; average σ per gra. = 13.3; N = 839 in gra. 4–9; average N per gra. = 140. Reliab. j-a: + ind (7) Gra: 4–9 (8) Time: 12′ (9) Talent: good cler. (11) Function measured: Science information, except specialized science abilities, i.e., a marked specialization in science interest and information.

(*nn*) **Junior High School General Science Tests: Ind. 2, Gr. 0.**

(1) Ruch, G. M. and Popenoe, H. F. (2) — General Science Test (3) Date: 1922 (4) 2f (5) Pub: WBC (6) Reliab: fA vs fB; σ = 9.6; r_{11} = .79; N = 25; derived from ½ fA vs ½ fB. σ total = 14.25; r_{11} (total) = .92; N = 23. Ruch and Stoddard give additional reliab. coefs. (7) Gra: 7–9 (8) Time: 45′ (9) Talent: good cler. (10) Cost: $1.30 per 25.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford Science Information Test. See (*mm*).

(*oo*) **High School General Science Tests: Ind. 1, Gr. 1.**

(1) Ruch, G. M. and Popenoe, H. F. (2) — General Science Test. See (*nn*).

---

(1) Dvorak, A. (2) — General Science Scales (3) Date: 1924 (4) 1 easy form, 2 harder comparable forms (5) Pub: PSPC (6) Reliab: P.E. of estimate = 2 (*This is equivalent to a* r_{11} = .96 *for a single gra. range.*) P.E. of distrib. (9th gra.) = 10. *From which* σ = *ca. 15* (7) Gra: 1st year General Science (8) Time: 20′ (9) Talent: super. cler.

Test not rated: (1) Toops, H. A. (2) — General Science Test (3) Date: 1919 (4) 1f (5) Pub: Test given

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

in School Science and Mathematics, November, 1925 (7) Gra: 12–13 (8) Time: 16′ (11) Function measured: Test based on Caldwell and Elkenberry's General Science.

(*pp*) **Biology Test: Ind. 1, Gr. 0.**

(1) Ruch, G. M. and Cossmann, L. (2) — Biology Test (3) Date: 1924 (4) 2f (5) Pub: WBC (6) Reliab: figured from 5 high school classes, — average r_{11} = .82; average σ = ca. 11.5 (7) Gra: Biology classes, usually grade 10, or more elementary classes in any high school grade, or perhaps college freshmen (8) Time: 45′ (9) Talent: good cler. (10) Cost: ca. 6¢ per blank (11) Function measured: Biology, general knowledge.

Tests not rated: (1) Laidlaw, O. W. and Woody, Clifford (2) Michigan Botany Test (3) Date: 1925 (4) 1f of 4 pts (5) Pub: PSPC (6) Reliab: By Spearman Brown formula r_{11} = .87. N = 272, pupils just finishing first year botany in 11 high schools. σ_1 = 11.87 (7) Gra: Where botany is given (10) Cost: $1.00 per 25.

(1) Cooprider, J. L. (2) — Information Exercises in Biology (3) Date: 1925 (4) 1f (5) Pub: PSPC (10) Cost: 50¢ per 25.

(*qq*) **High School Chemistry Tests: Ind. 1, Gr. 2.**

(1) Powers, S. R. (2) — Test for General Chemistry (3) Date: 1924 (4) 2f (5) Pub: WBC (6) Reliab: r_{11} = .796; σ = 9; *gra. range and N not given. Ruch and Stoddard:* r_{11} = .84; σ = 6.1; N = 101 *in gra. 11–12. Additional reliab. coefs. given in Ruch and Stoddard.* (7) Gra: First two years of chemistry, whether taken in high school or in college (8) Time: 35′ (9) Talent: good cler. (10) Cost: $1.10 per 25.

---

(1) Glenn, E. R. and Welton, L. E. (2) — New Type of High School Chemistry Tests for Instructional Purposes.

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Rich, S. G. (2) — Chemistry Test (3) Date: 1923 (4) 2f (5) Pub: PSPC (6) Reliab: r_{11} = .60; N = 66; average σ = 11.8; gra. = 12 and 13 combined Reliab. j-a: + ind (7) Gra: First two years of chemistry, whether taken in high school or college (8) Time: ca. 35′ (9) Talent: good cler. (10) Cost: $1.00 per 25 (11) Function measured: Chemistry except those phases that are taught today, but are not validated by the 7 social aims of education as adopted by the N.E.A. — especially details of technical information.

(77) **High School Physics Tests: Ind. $1\frac{1}{2}$, Gr. 2.**

(1) Glenn, E. R. and Obourn, E. L. (2) — New Type of High School Physics Tests for Instructional Purposes.

(1) Thurstone, L. L. (2) — Physics Test (Part of Vocational Guidance Tests) (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab. j-a: + ind (7) Gra: 12–13 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $1.00 per 25.

---

(1) Camp, H. L. (2) Iowa Physics Test (3) Date: 1920 (4) 2f, 3 ser; Series A, Mechanics; Series B, Heat; Series C, Electricity and Magnetism (5) Pub: PSPC (6) Reliab. j-a: + ind (7) Gra: High School classes in physics (8) Time: 30′ to 40′ (9) Talent: good physics teacher (10) 50¢ per 25 (11) Function measured: Physics, except light and sound.

(1) Chapman, J. C. (2) — Test in Electricity, Magnetism, Sound, Light, Heat, Mechanics (3) Date: 1919 (4) 1f (6) Reliab. j-a: + ind (7) Gra: High School classes in physics (8) Time: ca. 10′ (9) Talent: good physics teacher (10) Cost: single copy 25¢, only 1 necessary (11) Function measured: Physics, except any "physics sense," i.e., power to use elementary physics knowledge in situations which are not stereotyped.

Test not rated: (1) Farwell, H. W. and Wood, B. D. (2) Columbia Research Bureau Physics Test (3)

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Date: 1925 (4) 2f (5) Pub: WBC (6) Reliab: By Spearman-Brown formula r_{11} = .863. N = 575 high school pupils. σ_1 = 25.5 (7) Gra: High school, and college freshmen (8) Time: 75′ (10) Cost: $1.30 per 25 (11) Function measured: Topics of physics in following proportions; mechanics, 16 per cent; heat, 16 per cent; sound, 8 per cent; light, 16 per cent; electricity, 32 per cent; miscellaneous, 12 per cent.

(*ss*) **Elementary American History Tests: Ind. 1, Gr. 1½.**

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford History and Literature Information Test (3) Date: 1923 (4) 2f (5) Pub: WBC (6) Reliab: average r_{11} per gra. for gra. 4–9 = .82; average σ per gra. = 14.4; N = 839 in gra. 4–9; average N per gra. = 140. Reliab. j-a: + ind (7) Gra: 4–9 (8) Time: 12′ (9) Talent: good cler. (11) Function measured: History and literature information, except specialized abilities in history and literature.

---

(1) Hahn, H. H. (2) — History Scales (3) Date: 1920, Rev. 1923 (4) 1f (5) Pub: PSPC (Distributors) (6) Reliab. j-a: = ind (7) Gra: 7–8 (8) Time: varies (9) Talent: history teacher (10) Cost: 1 copy 25¢, only 1 necessary (11) Function measured: A complete test testing the subject matter found in each of six modern texts. It is arranged especially for diagnostic testing.

(1) Harlan, C. L. (2) — Information Test in American History.

(*tt*) **Junior High School American History Tests:[1] Ind. 3, Gr. 2.**

(1) Van Wagenen, M. J. (2) — American History Scales (3) Date: 1919, Rev. 1924 (4) 1f, 2 or 3 div. 4 pts. All parts yield equivalent scores within the limits of the P.E. Information: R1 and R2,

[1] Ruch and Stoddard give comparative reliabilities and intercorrelations of 6 U. S. history tests.

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

General; S1, S2, and S3, Gen.; T1, T2 and T3, Gen.; U1, U2 and U3, Gen.; V1 and V2, Gen.; C1 and C2, Discovery to Revolutionary War; F1 and F2, Revolutionary to Civil War; K1 and K2, Civil War to Present. The only Thought Scale is R2. (5) Pub: T. C. Bur. Pub. (6) Reliab: P.E. of scale score is 2.1 scale pts or approximately ⅓ of a grade difference (7) Gra: Div 1, 5–6; Div 2, 7–8; Div 3, 9–12 (8) Time: 40′ (9) Talent: Good teacher of American History for Thought Scale. Superior Clerical Help for Information Scales (10) Cost: Infor. $2.00 per 100, Thought $2.50 per 100 (11) Function measured: American History, except historical judgment or evaluation of statements and inferences, ability to read for the thought content.

(1) Kelley, T. L., Ruch, G. M., and Terman, L. M. (2) Stanford History and Literature Information Test. See (*ss*).

(1) Barr, A. S. (2) — Diagnostic Tests in American History. (See *uu*).

---

(1) Pressey, L. W., and Richards, R. C. (2) — American History Test (3) Date: 1922 (4) 1f (5) Pub: PSPC (6) Reliab. j-a: Pressey reports: + ind. Richards reports: + gr. (7) Gra: 6–12 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $2.00 per 100 (11) Function measured: American History, except ability to read texts, judge of comparative importance of events, see relationships of past events to present conditions.

Tests not rated: (1) Gregory, C. A. (2) — Tests in American History (3) Date 1923 (4) 2f of 3 tests each (5) Pub: Bureau of Administrative Research, University of Cincinnati (7) Gra: 7–12 (10) Cost: $3.50 per 100 for each test.

(1) McDade, J. E. (2) Plymouth educational tests Nos. 80A, 81A and 82A (4) 3 tests (5) Pub: Plymouth Press, (7) Gra: 3–8 (10) Cost: Each

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

test 60¢ per 100 (11) Function measured: 80A, — U. S. history, events-dates; 81A, — U. S. history, events-names; 82A, — U. S. history, names-events.

(1) Witham, E. C. (2) — Comprehensive 7th and 8th Grade History Tests, Nos. 1, 2, 3, and 4. (3) Date: 1924 (4) 2f of 2 tests (5) Pub: J. L. Hammett Co. (7) Gra: Tests 1 and 2, which are equivalent forms, are for gra. 7; tests 3 and 4, likewise equivalent, are for gra. 8 (8) Time: ca. 20′ (10) Cost: Each test $1.00 per 50. (11) Function measured: To measure interest in history and to stimulate teachers and pupils in this subject.

(*uu*) **High School American History Tests: Ind. 0, Gr. 2.**

(1) Barr, A. S. (2) — Diagnostic Tests in American History (3) Date: 1918 (4) 1f (5) Pub: PSPC (6) Reliab: *Ruch and Stoddard:* $r_{11} = .77$; $\sigma = 9.7$; $N = 50$ *in gra. 9 and 12* (7) Gra: 11–12 (8) Time: 30′ (9) Talent: good cler. (10) Cost: $4.00 per 100.

(1) Pressey, L. W. and Richards, R. C. (2) — American History Test. See (*tt*).

(*vv*) **High School Ancient History Tests: Ind. 1, Gr. 1.**

(1) Wood, B. D. (2) Institute of Educational Research College Entrance Examination in Ancient History (3) Date: 1922.

---

(1) Sackett, L. W. (2) — Ancient History Test.

(*ww*) **High School Modern European History Test:**

Test not rated: (1) Vannest, C. G. (2) — Diagnostic Test in Modern European History.

(*xx*) **College Ancient History Test: Ind. 1, Gr. 0.**

(1) Wood, B. D. (2) Institute of Educational Research College Entrance Examination in Ancient History (3) Date: 1922.

(*yy*) **Citizenship Scale: Ind. 0, Gr. 1.**

(1) Chassell, C. F., Upton, S. N., and Chassell, L. M., (2) Citizenship Scales (3) Date: 1922 (4) 8 equiv-

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

alent sca (5) Pub: T. C. Bur. Pub. (6) Reliab: *No reliab. given.* Average of 10 r's from paired scales rated by 1 teacher is .895 (7) Gra: 1–12 (8) Time: Scores based on observations extending over weeks or months (9) Talent: capable grade teacher (10) Cost: 50¢ per 100 (11) Function measured: Conduct, except the motive which lies back of a given act. Moreover, when the scales are marked by the teacher, the ratings are based on observations, more or less remote, of the behavior of the pupils rather than upon the actual practice of the pupil at the time the ratings are assigned. When the scales are used for self-measurement the ratings are subject to errors resulting from the rater's failure to represent his own practice accurately.

(zz) **Character Tests: Ind. 0, Gr. 2.**

(1) Cady, V. M. (2) — Tests of Incorrigibility (3) Date: 1923 (4) 5 tests with substantially duplicate forms (5) Pub: Tests used in "The Estimation of Juvenile Incorrigibility," Jour. of Delinquency, Monograph 2. (6) Reliab: r_{11} = .746; N = 150: boys 12.5–14.5 yrs old. Reliab. j-a; + ind (7) Gra: 4–12 (8) Time: *ca. 2 hours* (9) Talent: good cler. (11) Function measured: Moral development, except those not represented by moral reliability, social judgments, and mental complexes and inversions.

(1) Voelker, P. E. (2) — Character Tests. Tests used in a study of "The Functions of Ideals in Education." (3) Date: 1921 (4) 3 ser (5) Pub: T. C. Bur. Pub. (6) Reliab: r_{11} = 83; N = 150; gra. = 5–12 inclusive. Reliab. j-a: + ind (7) Gra: elementary and high school (8) Time: A few hours spread over a number of weeks (9) Talent: A superior teacher or scout leader with inscrutable facial expression (10) Cost: book $1.35 (11) Function measured: Reliability, except individual's self-control and purpose.

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(*aaa*) Elementary Drawing: Ind. 2, Gr. 0.

(1) Carey-Kline (2) — Drawing Scales.

(1) Thorndike, E. L. (2) — Drawing Scale (3) Date: 1913, Rev. 1924 (4) 1f (5) Pub: T. C. Bur. Pub. (7) Gra: All. (8) Time: irrelevant (9) Talent: Teacher or supervisor of good judgment; preferably good drawing teacher (11) Function: To measure drawing. Its main purpose is to reduce constant errors. It should reduce variable errors somewhat.

(*bbb*) Junior High School Drawing Scale: Ind. 1, Gr. 0.

(1) Thorndike, E. L. (2) — Drawing Scale. See (*aaa*).

(*ccc*) Elementary to High School Writing Scales: Ind. 5, Gr. 3.

(1) Ayres, L. P. (2) — Handwriting Scale, — Gettysburg Edition (3) Date: 1917 (4) 1f (5) Pub: Russell Sage Foundation, 130 E. 22 St., New York City (6) Reliab. j-a: + ind (7) Gra: 4–8 (8) Time: 5′ (9) Talent: A little practice in use of scale (10) Cost: 10¢ (11) Function measured: Legibility and speed, except beauty.

(1) Thorndike, E. L. (2) — Handwriting Scale (3) Date: 1910, Rev. 1912 (4) 1f (5) Pub: T. C. Bur. Pub. (6) Reliab. j-a: Main service is to reduce constant errors. It probably reduces variable errors somewhat. (7) Gra: 2–12 (8) Time: irrelevant (9) Talent: good judge of handwriting (10) Cost: 12¢.

(1) Freeman, F. N. (2) — Chart for diagnosing faults in Handwriting (3) Date: 1914 (4) Provision for measuring five different features of handwriting (5) Pub: Houghton, Mifflin Co. (6) Reliab. j-a: + ind (7) Gra: All (8) Time: irrelevant (9) Talent: good judge of handwriting (10) Cost: 40¢.

(2) Kansas City Scale of Handwriting.

(1) Frasier, G. W. (2) — Handwriting Test.

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Ayres, L. P. (2) — Handwriting, — Three Slant Edition.

(1) Starch, D. (2) — Handwriting Scale.

Scale not rated: (1) Leamer, E. W. (2) — Diagnostic Practice Sentences in Handwriting (3) 1924, Rev. 1925 (4) 5 sets of 15 cards each (5) Pub: PSPC (7) Gra: 2–8 (8) Time: 10′ per day (10) Cost: 28¢ per set. (11) Function measured: The directions include information which makes it possible to analyze a child's writing in terms of slant, letter-formation, spacing, alignment and quality of line.

(1) Connor, Bertha A. (2) Muscular Movement Penmanship Gradient (3) Date: 1922 (4) 1 sca. for each gra. (5) Pub: Houghton Mifflin Co. (7) Gra: 1–8 (10) Cost: $1.20 per gra.

(*ddd*) **College Handwriting Scales: Ind. 2, Gr. $\frac{1}{2}$.**

(1) Freeman, F. N. (2) — Chart for Diagnosing Faults in Handwriting. See (*ccc*).

(1) Ayres, L. P. (2) — Adult Handwriting Scale.

---

(1) Frasier, G. W. (2) — Handwriting Scale.

(*eee*) **Typing Tests: Ind. 1, Gr. 1$\frac{1}{2}$.**

(1) Blackstone, E. G. (2) — Stenographic Efficiency Test (3) Date: 1923 (4) 5f (5) Pub: WBC (6) Reliab: r_{11} = .92; all high school and intermediate gra.; σ = 14 pts (7) Gra: For use in commercial schools or business colleges (8) Time: 3′ (9) Talent: good cler. (10) Cost: $1.00 per 25 (11) Function measured: Typing, except arrangement, punctuation.

---

(1) Thurstone, L. L. (2) — Typist Test (3) Date: 1920 (4) 1f (5) Pub: WBC (6) Reliab. j-a: ind (7) Gra: For applicants for stenographic typing positions (8) Time: 30–45′ (9) Talent: good office superintendent (10) Cost: $1.50 per 25.

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Rogers, H. W. (2) — Stenographic and Typist Tests.

(*fff*) **General Clerical Tests: Ind. 1. Gr. 0.**

(1) Thurstone, L. L. (2) — Clerical Examination (3) Date: 1919 (4) 1f (5) Pub: WBC (6) Reliab. j-a: ind (7) Gra: For applicants for office positions (8) Time: 30′ to 45′ (9) Talent: good office superintendent (10) Cost: $1.50 per 25 (11) Function measured: Clerical Ability.

Test not rated: (1) Benge, E. J. (2) — Clerical Test (3) Date: 1923 (5) Pub: C. H. Stoelting Co. (10) Cost: $5.00 per 25.

(*ggg*) **Junior High and High School Mechanical Ability Test: Ind. 0, Gr. 1.**

(1) Stenquist, J. L. (2) — Mechanical Aptitudes Tests (3) Date: 1921 (4) 1f (5) Pub: WBC (6) Reliab: r_{11} = .6 − .7; N = 200; gra. = 6, 7, and 8 combined. *See Ruch and Stoddard for further reliab. coefs.* Reliab. j-a: + ind (7) Gra: 6–12 (8) Time: 95′ (9) Talent: good cler. (10) Cost: $1.50 per 25 (11) Function measured: Mechanical aptitude, except manipulative skill.

Test not rated: (1) MacQuarrie, T. W. (2) — Test for Mechanical Ability (3) Date: 1925 (4) 7 pts. (5) Pub: Author, Teachers College, San Jose, Calif. (6) Reliab: r_{11} = .90 for total battery (7) Gra: 6–12, ages 14 up (8) Time: ca. 25′ (10) Cost: $1.50 per 25.

(*hhh*) **Elementary, Junior High and High School Music Tests: Ind. 4, Gr. 2.**

(1) Kwalwasser, J., Ruch, G. M. (2) — Test of Musical Accomplishment (3) Date: 1924 (4) 1f (5) Pub: Extension Division, University of Iowa, Iowa City, Iowa (6) Reliab: Retest after 1 month interval, r = .88; σ = 42.4; N = 49 in gra. 8, 10, and 12. By Spearman-Brown formula: r =

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

.97; $\sigma = 51.5$; N = 167 in gra. 6, 8, 10, and 12. For this same population and via Spearman-Brown formula, the reliabilities of the parts of the test are: (*a*) Knowledge of musical symbols and terms, $r_{11} = .92$; (*b*) Recognition of syllable names, $r_{11} = .87$; (*c*) Detection of pitch errors in melody, $r_{11} = .77$; (*d*) Detection of time errors in melody, $r_{11} = .72$; (*e*) Recognition of pitch names, $r_{11} = .89$; (*f*) Knowledge of time signatures, $r_{11} = .80$; (*g*) Knowledge of key signatures, $r_{11} = .95$; (*h*) Knowledge of note values $r_{11} = .70$; (*i*) Knowledge of rest values, $r_{11} = .72$; (*j*) Recognition of familiar melodies from notation, $r_{11} = .77$. (7) Gra: 4–12.

(1) Mosher, R. M. (2) — Sight Reading Music Test.

(1) Seashore, C. E. (2) — Sense of Rhythm. This and other Seashore Records are for sale by Columbia Graphophone Company, New York City, at $1.50 per record. Users should not be encouraged to use average of the showing of these tests as an index to musicality as a whole. They are tests of specific capacities, some of which have little or no relationship to one another. They are ways of finding specific information about certain elements of musical capacity. The measurement is far more accurate than any direct judgment without measurement. I (Seashore) always want it understood that these tests should not be used by themselves, but to supplement and elucidate the judgment of musical observers of the children, unless the interest in making the test is specific; for example, a survey of the sense of pitch, of the sense of rhythm, in which case the test is adequate in itself. Each test requires about 20 minutes. (6) Reliab: *Ruch and Stoddard: Pitch*, $r_{11} = .70$; $\sigma = 11.95$. *Intensity*, $r_{11} = .66$; $\sigma = 8.12$. *Time*, $r_{11} = .53$; $\sigma = 7.86$. *Consonance*, $r_{11} = .35$; $\sigma = 7.71$. *Memory*, $r_{11} = .66$; $\sigma = 15.30$. $N = 100$. *Rhythm*, $r_{11} = .50$; $\sigma = 7.22$, $N = 50$.

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Seashore, C. E. (2) — Tonal Memory Test.

---

(1) Seashore, C. E. (2) — Sense of Intensity.

(1) Seashore, C. E. (2) — Sense of Pitch.

Tests not rated: (1) Hutchinson, H. E. (2) — Music Test (3) Date: 1924 (4) 1f (5) Pub: PSPC (7) Gra: 7–12 (10) Cost: 50¢ per 25.

(1) Torgersen-Fahnstock (2) — Music Test (3) Date: 1926 (4) 1f of 2 pts (5) Pub: PSPC (7) Gra: 4–9 (10) Cost: 75¢ per 25 (11) Function measured: Pt. A; theoretical knowledge, Pt. B; ear training.

(iii) Sundry: Elementary School Tests: Ind: 2, Gr. 0.

(1) Thorndike, E. L. (2) The Teachers Word Book (3) Date: 1921 (5) Pub: T. C. Bur. Pub. (7) *All gra: This book contains an alphabetical list of the 10,000 most frequently used words in a count of over four million words. The frequency value of each word is given so that vocabulary tests, spelling tests, reading tests, etc., may be built up by the teacher.*

(1) Graduate Students of Household Arts Education Department, Teachers College, under the direction of Professors McCall, Cooley and others (2) Home Economics Information Test (3) Date: 1921, Rev. 1923 (4) 1f of 3 pts (5) Pub: T. C. Bur. Pub. (6) Reliab. j-a: + (7) Gra: 8 (8) Time: 3 hours (9) Talent: good cler. (10) Cost: 15¢ per set and 1 directions sheet with every 25 sets or 35¢ per set with directions sheet (11) Function measured: Household Arts, except skills, appreciation.

Test not rated: (1) Horn, Ernest (2) A Basic Writing Vocabulary (3) Date: 1920 (5) Pub: Univ. of Iowa (7) All Gra: *This book contains an alphabetical list of the 10,000 most frequently used words in writing (letters) determined from a count*

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

of over five million words. Frequency values in the Thorndike notation are given. May be used to build up tests.

(fff) Sundry: High School, College, and Vocational Tests: Ind. 10, Gr. 5.

(1) Blackstone, E. G. (2) — Stenographic Proficiency Tests. See (*eee*).

(1) Goodspeed, H. and Dodge, B. (2) — Preliminary Judgment Test in Home-making.

(1) Henman, V. A. C. (2) — French Word List.

(1) Kelley, T. L. (2) — Mathematical Values (3) Date: 1920 (4) If yielding 13 different scores (5) Pub: T. C. | Bur. Pub. (6) Reliab. j-a: + gr. (for the 13 different scores. The total score is not employed) (7) Gra: 8–12 (8) Time: ca. 90′ (9) Talent: super. high school teacher (10) Cost: 1 copy 5¢ (11) Function measured: High school mathematics, except the mechanical phases. The test yields separate scores on 13 different fundamental mathematical values.

(1) Murdock, K. (2) — Sewing Scale.

(1) Murdock, K. (2) — Analytic Sewing Scale.

(1) Rogers, A. L. (2) — Test of Mathematical Ability.

(1) Spink, — (2) — Grading Chart for Mechanical Drawing.

(1) Thurstone, L. L. (2) — Vocational Guidance Test. This is composed of five parts: Arithmetic, Algebra, Geometry, Physics, and Technical Information. Information concerning the first four may be found under their respective headings.

(1) Whittier Scale for Grading Home Conditions.

---

(1) Bureau of Personnel Research, Carnegie, Institute of Technology (2) — Vocational Tests. Will profile, social relations, business information, meeting objectives, interest analysis.

(1) Cross, —— (2) — English Test.

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Hoke, E. (2) — Prognostic Test of Stenographic Ability.

(1) Logasa and McCoy (2) — Seven Tests for Appreciation of Literature.

(1) Wilkins, L. A. (2) — Prognosis Test in Modern Languages.

Tests not rated: (1) Frasier, G.W. and Armentrout, W. D. (2) — Standard Achievement Test on an Introduction to Education (3) Date: 1924 (4) 1f promised each year (5) Pub: Scott Foresman and Co., Chicago, Ill. (7) Gra: College classes in Education (8) Time: 50′ (10) Cost: 5¢ per copy (11) Function Measured: Covers material presented in Frasier and Armentrout's text "An Introduction to Education."

(1) Kehner, Tyler (2) — Background Test in Social Science (3) Date: 1924 (4) 2f (5) Pub: Harvard University Press (7) Gra: 9–12 (8) Time: 40′–50′ (10) Cost: $1.25 per 25 (11) Function measured: Factual background of social science.

(1) King, Florance B. and King, H. F. (2) — Food Tests (3) Date: 1924 (4) 1f (5) Pub: Indiana University Book Store, Bloomington, Ind. (7) Gra: 6–12 (8) Time: 30′ (10) Cost: 10¢ per 10.

(1) Moss, F. A., Hunt, T., Omwake, K. T. and Ronning, M. M. (2) George Washington Series Social Intelligence Test (3) Date: 1927 (4) 1f (5) Pub: Center for Psychological Service, 2024 Q St., N. W., Washington, D. C. (7) Gra: *High school, college, and industry* (10) Cost: $12.00 per 100 (11) Function measured: Test designed to measure one's ability to get along with others.

(1) Patrick, —— (2) — Industrial Arts Test (5) Pub: PSPC (10) Cost: 50¢ per 25.

(1) Thurstone, L. L. (2) — Spatial Relations Test (5) Pub: C. H. Stoelting Co. (10) Cost: $2.50 per 25.

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

(1) Weber, J. J. (2) — Standard Achievement Test on Aims, Purposes, Objectives, Attributes, and Functions in Secondary Education (3) Date: 1926 (4) 1f (5) Pub: PSPC (10) Cost: $1.00 per 25.

(1) Witham, E. C. (2) Hall of Fame Test (3) Date: 1924 (4) 1f (5) Pub: J. L. Hammett Co. (7) Gra: 7–16 (8) Time: 25′.

(*kkk*) **Elementary Physical Development Measures.** See (*kkk*) of Chapter IX.

(*lll*) **Junior High and High School Physical Development Measures.** See (*lll*) of Chapter IX.

(*mmm*) **High School and College French Tests: Ind. 2, Gr. 1.**

(1) Méras, A. M., Roth, Suzanne, and Wood, B. D. (2) Columbia Research Bureau French Test (3) Date: 1923, Rev. 1924 (4) 2f. (5) Pub: WBC (6) Reliab: r_{11} = .96; N = 1353 high school 2nd, 3rd and 4th year pupils; σ = 41.4. Reliab. j-a: + ind (7) Gra: Those with 1–4 years French (8) Time: 90′ (9) Talent: good cler. (10) Cost: $1.30 per 25 (11) Function measured: French, except cultural and "spiritual" gains, oral and aural skills, except as these are correlated with ability to read and write the language.

(1) Henmon, V. A. C. (2) — French Test (3) Date: 1921 (4) 4f (5) Pub: WBC (6) Reliab: *Ruch and Stoddard: r_{11} = .61; σ = 51.3 on f1 and σ = 60.3 on f2; N = 60.* Reliab. j-a: + (7) Gra: high school classes (8) Time: 20′ (9) Talent: good cler. (10) Cost: 50¢ per 25 (11) Function measured: French, except knowledge of grammar.

---

(1) Twigg, A. M. (2) — French Vocabulary Test.

(*nnn*) **High School and College German Tests: Ind. 1, Gr. 1.**

(1) Betz, F., Betz, G. A., Wendt, H. G., and Wood, B. D. (2) Columbia College Placement Exami-

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

nation in German (3) Date: 1923, Rev. 1924 (4) 10f (5) Pub. — (6) Reliab. j-a: + ind (7) Gra: those with 1–4 years German (9) Talent: good cler. (11) Function measured: German, except cultural and "spiritual" gains, oral and aural skills, except as these are correlated with ability to read and write the language. *This test has been replaced by the Columbia Research Bureau German Test.*

---

(1) Whipple, G. M. (2) — German Vocabulary Test.

Tests not rated: (1) Purin, C. M. and Wood, B. D. (2) Columbia Research Bureau German Test (3) Date: 1925 (4) 2f (5) Pub: WBC (6) Reliab: By Spearman-Brown formula, r_{11} = .962, N = 1067 high school 2nd, 3rd and 4th year pupils. σ_1 = 39.5 (7) Gra: Those with 1–4 years of German (8) Time: 90′ (10) Cost: $1.30 per 25 (11) Function measured: Vocabulary, comprehension, and grammar. Test does not measure oral and aural skills and cultural content except as dependent upon a knowledge of the written language.

(1) Van Wagenen, M. J. and Patterson, —— (2) — Reading Scales in German (3) Date: 1927 (4) 4f of 2 divisions (5) Pub: PSPC (7) Gra: Division 1 for 1st year, Division 2 for 2nd and 3rd years.

(*ooo*) High School and College Spanish Test: Ind. o, Gr. 1.

(1) Handschin, C. H. (2) — Modern Language Tests, — Spanish.

Test not rated: (1) Callcott, Frank, and Wood, B. D. (2) Columbia Research Bureau Spanish Test (3) Date: 1925 (4) 2f 3 pts (5) Pub: WBC (6) Reliab: By Spearman-Brown formula, r_{11} = .965, N = 1061 high school 2nd, 3rd and 4th year pupils. σ_1 = 37.8 (7) Gra: Those with 1–4 years Spanish (10) Cost: $1.30 per 25 (11) Function measured: Vocabulary, comprehension,

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

and grammar. Test does not measure oral or aural skills and cultural content except as dependent upon a knowledge of the written language.

(*ppp*) **High School Latin Tests: Ind. $2\frac{1}{2}$, Gr. $3\frac{1}{2}$.**

(1) Henmon, V. A. C. (2) — Latin Test (3) Date: 1917 (4) 4f. In addition to the four forms, each containing a vocabulary and a sentence test, there is a form (Test X) for research use in school surveys. (5) Pub: WBC (6) Reliab: Vocabulary, r_{11} = .96; N = 348. Sentence, r_{11} = .80; N = 275. Each 1st year pupils. *Ruch and Stoddard give several reliab. coef's on each pt. They vary from .66 to .80 on Vocabulary and from .50 to .71 on Sentences, for gra. 9–12 pupils.* Reliab. j-a: + (7) Gra: high school classes (8) Time: 20′ (9) Talent: good Latin teacher (10) Cost: 50¢ per 25 (11) Function measured: Latin, except knowledge of grammar.

(1) Brown, H. A. (2) — Latin Test (3) Date: 1919 (4) 1f 5 pts (5) Pub: The Parker Company, Madison, Wisconsin (6) Reliab. j-a: + (7) Gra: high school classes (8) Time: L. connected, 15′; L. sentence a, 40′; L. sentence b, 30′; L. grammar, 30′; L. vocabulary, 30′. (9) Talent: good Latin teacher (10) Cost: L. connected, L. sentence a, L. sentence b, 100 each for $1.25; L. vocabulary and L. grammar, 100 each for 75¢ (11) Function measured: High school Latin.

(1) Ullman, B. L. and Kirby, T. J. (2) — Latin Comprehension Test (3) Date: 1922 (4) 1f (5) Pub: Extension Division, University of Iowa, (6) Reliab: r_{11} = .85; σ = 6.59; gra. = 2–8 semesters of Latin. Reliab. j-a: + ind (7) Gra: All high school classes in Latin except first year (8) Time: 30′ (9) Talent: good Latin teacher.

---

(1) Stevenson, P. R. (2) Latin Vocabulary Test (3) Date: 1923 (4) 3f (5) Pub: PSPC (6) Reliab.

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

j-a: ind (7) Gra: 8–12 (8) Time: 30′ (9) Talent: good cler. (10) Cost: 50¢ per 25 (11) Function measured: Latin vocabulary, except syntax, translation, etc.

(1) Lohr, L. and Latshaw, H. (2) — Latin Form Test.

(1) Starch, D. (2) — Latin Test.

Test not rated: (1) Inglis, Alex. (2) — Latin Tests (3) Date: 1923 (4) Several f; separate pts covering general vocabulary, morphology, and syntax. (5) Pub: Harvard University Press (7) Gra: All in which Latin is studied (8) Time: 30′ for each pt (10) Cost: $1.25 per 25 for each pt.

(*qqq*) **High School Latin Composition Test: Ind. o, Gr. 1.**

(1) Godsey, E. (2) — Diagnostic Latin Composition Test (3) Date: 1922 (4) 2f (5) Pub: American Classical League, °/₀ Mason D. Gray, East High School, Rochester, New York (6) Reliab. j-a: + ind (7) Gra: high school classes (8) Time: 30′ (9) Talent: good Latin teacher (10) Cost: $1.00 per 100 (11) Function measured: Latin Composition.

(*rrr*) **High School Latin-Derivative Vocabulary Test: See (*rrr*) of Chapter IX.**

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

BIBLIOGRAPHY

BERNSTEIN, C. E. "Quickness and Intelligence: An Inquiry concerning the Existence of a General Speed Factor." *British Journal of Psychology, Monograph No. 7* (1924), page 2.

BINET, A., and SIMON, T. "Le développement de l'intelligence chez les enfants." *L'Année Psychologique*, Vol. XIV (1908).

BINGHAM, W. V. "Personality and Vocation: A Note on Effects of Introversion on Dominant Interests." *British Journal of Psychology* (April, 1926).

BOBERTAG, O. "Ueber Intelligenzprüfungen — nach der Methode von Binet und Simon." *Zeitschrift für angewandte Psychologie*, Vol. VI (1912).

BOLTON, T. L. "The Growth of Memory in School Children." *American Journal of Psychology*, Vol. IV (1892).

BURT, CYRIL. "Experimental Tests of General Intelligence." *British Journal of Psychology*, Vol. III (1909–1910).

CADY, VERNON M. *The Estimation of Juvenile Incorrigibility. Journal of Delinquency, Monograph No. 2* (April, 1923).

CAREY, N. "Factors in the Mental Processes of School Children." *British Journal of Psychology*, Vol. VII (March–October, 1915); Vol. VIII (May, 1916).

CATTELL, J. MCKEEN. "Mental Tests and Measurements." *Mind*, Vol. XV (1890).

CHADDOCK, A. E. *Principles and Methods of Statistics*. 1925.

CHAMBERS, G. G. *An Introduction to Statistical Analysis*. 1925.

COWDERY, K. M. *An Evaluation of the Expressed Attitudes of Members of Three Professions: Medical, Engineering, and Legal.* (Dissertation on file at Stanford University; 1925–1926.)

CURRENT, W. F., and RUCH, G. M. "Further Studies on the Reliability of Reading Tests." *Journal of Educational Psychology* (September, 1925).

DE MOIVRE, A. A paper printed in 1733 and bound in as an Appendix to certain editions of A. de Moivre, *Miscellanea Analytica* (1730), containing a derivation of the normal probability curve, is reported by Pearson (1924 hist.).

EBBINGHAUS, H. "Ueber eine neue Methode zur Prüfung geistiger Fähigkeiten und ihre Anwendung bei Schulkindern." *Zeitschrift für Psychologie und Physiologie der Sinnesorgane*, Vol. XIII (1897).

FECHNER, G. T. *Psychophysik*. 1860.

FLÜGEL, J. C. "The Influence of Attention in Illusions of Reversible Perspective." *British Journal of Psychology*, Vol. V (June, 1913).

—— "Freudian Mechanism Factors in Moral Development." *British Journal of Psychology* (June, 1917).

—— and MCDOUGALL, WILLIAM. "Some Observations on Psychological Contrast." *British Journal of Psychology* (March, 1915).

FRANZEN, RAYMOND. "The Accomplishment Quotient." *Teachers College Record* (November, 1920).

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

FRANZEN, RAYMOND. "Statistical Issues." *Journal of Educational Psychology* (September, 1924).

FULLERTON, G. S., and CATTELL, J. McKEEN. *On the Perception of Small Differences.* University of Pennsylvania Press, Philadelphia; 1892.

GALTON, FRANCIS. *Hereditary Genius: An Inquiry into Its Laws and Consequences.* 1869.

—— *Natural Inheritance.* 1889.

—— "Grades and Deviates." *Biometrika*, Vol. V, Part IV (1907).

GARNETT, MAXWELL. "On Certain Independent Factors in Mental Measurements." *Proceedings of the Royal Society*, No. A675, Series A, Vol. XCVI (September, 1919).

GARRETT, HENRY E. *Statistics in Psychology and Education.* 1926.

GODDARD, H. H. "Two Thousand Normal Children Measured by the Binet Measuring Scale of Intelligence." *Pedagogical Seminary*, Vol. XVIII (1911).

HART, B., and SPEARMAN, C. "General Ability: Its Existence and Nature." *British Journal of Psychology*, Vol. V (1912).

—— and SPEARMAN, C. "Mental Tests of Dementia." *Journal of Abnormal Psychology* (October–November, 1914).

JONES, D. C. *A First Course in Statistics.* London; 1921.

KELLEY, TRUMAN L. *Educational Guidance.* Teachers College, Columbia University, New York; 1914.

—— "Comparable Measures." *Journal of Educational Psychology* (1914).

—— "Principles Underlying the Classification of Men." *Journal of Applied Psychology*, Vol. III (March, 1919).

—— "A New Measure of Dispersion." *American Statistical Association Quarterly* (June, 1921).

—— *Statistical Method.* 1923.

—— "The Principles and Technique of Mental Measurement." *American Journal of Psychology*, Vol. XXXIV (July, 1923).

—— "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores." *Journal of Educational Psychology*, Vol. XIV (September, 1923).

—— "Distinctive Ability." *School and Society* (October 13, 1923).

—— "How Many Figures Are Significant?" *Science* (December 5, 1924).

—— *The Influence of Nurture upon Native Differences.* 1926.

KLEMM, OTTO. *A History of Psychology* (tr. by E. C. Wilson and Rudolf Pintner). 1914.

KLÜVER, H. "An Analysis of Recent Work on Psychological Types." *Journal of Nervous and Mental Diseases* (December, 1925).

KRUEGER, F., and SPEARMAN, C. "Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten." *Zeitschrift für Psychologie*, Vol. XLIV (1907).

KRUSE, PAUL J. "The Overlapping of Attainments in Certain Grades." *Teachers College Contributions to Education*, No. 92 (1918). Teachers College, Columbia University, New York.

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

KUHLMANN, F. "Degree of Mental Deficiency in Children as Expressed by the Relation of Age to Mental Age." *Journal of Psychoasthenics*, Vol. XVII (1913).

LANKES, W. "Perseveration." *British Journal of Psychology*, Vol. VII (1915).

MCCALL, W. A. "A Proposed Uniform Method of Scale Construction." *Teachers College Record*, Vol. XXII (January, 1921).

—— *How to Measure in Education*. 1922.

MOORE, T. V. "The Temporal Relations of Meaning and Imagery." *Psychological Review Monograph Supplement*, Vol. XXII, No. 3 (May, 1915).

MUELLER, G. E. *Grundlegung der Psychophysik*. 1878.

NACCARATI, SANTE. "The Morphologic Aspect of Intelligence." *Archives of Psychology*, No. 45 (1921).

NATIONAL SOCIETY FOR THE STUDY OF EDUCATION. "The Measurement of Educational Products." *The Seventeenth Yearbook*, Part II. 1918. (Includes bibliography of 606 titles.)

—— "Intelligence Tests and Their Use." *The Twenty-first Yearbook*. 1922.

NORSWORTHY, NAOMI. "The Psychology of Mentally Deficient Children." *Archives of Psychology*, No. 1 (1906).

ODELL, C. W. *Educational Statistics*. 1925.

OTIS, ARTHUR S. Otis Group Intelligence Scale. (Edition of 1920 describing test of 1917.)

—— "A Method of Inferring the Change in a Coefficient of Correlation Resulting from a Change in the Heterogeneity of the Group." *Journal of Educational Psychology* (May, 1925).

PEAK, HELEN, and BORING, EDWIN G. "The Factor of Speed in Intelligence." *Journal of Experimental Psychology* (April, 1926).

PEARL, RAYMOND. *Medical Biometry and Statistics*, 1923.

PEARSON, KARL. "Historical Note on the Origin of the Normal Curve of Errors." *Biometrika*, Vol. XVI, Part IV (December, 1924).

—— "On Our Present Knowledge of the Relationship of Mind and Body." *Annals of Eugenics*, Vol. I, Parts III–IV (April, 1926).

PINTNER, RUDOLF. *Intelligence Testing: Methods and Results*. 1923.

POWERS, S. W. "Tests of Achievement in Chemistry." *Journal of Chemical Education*, Vol. I, No. 7 (September, 1924).

RAUBENHEIMER, ALBERT S. *Experimental Study of Some Behavior Traits of the Potentially Delinquent*. (Dissertation on file at Stanford University; 1923.)

RICE, J. M. "The Futility of the Spelling Grind." *Forum*, Vol. XXIII (1897).

RIETZ, H. L., et al. *Handbook of Mathematical Statistics*. 1924.

ROOT, W. T. "Correlation between Binet Tests and Group Tests." *Journal of Educational Psychology* (May, 1922).

Generated on 2020-12-23 01:36 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

RUCH, G. M., and STODDARD, G. D. *Tests and Measurements in High School Instruction.* 1927.

RUGG, HAROLD O. *Statistical Methods Applied to Education.* 1917.

SPEARMAN, C. "The Proof and Measurement of Association between Two Things." *American Journal of Psychology,* Vol. XV (1904).

—— "Demonstration of Formulæ for True Measurement of Correlation." *American Journal of Psychology,* Vol. XVIII (1907).

—— "Theory of Two Factors." *Psychological Review,* Vol. XXI (1914).

—— *The Nature of Intelligence and the Principles of Cognition.* 1923.

—— *The Abilities of Man.* 1927.

—— *See also* Hart and Spearman.

—— *See also* Krueger and Spearman.

STEAD, H. G. "Factors in Mental and Scholastic Ability." *British Journal of Psychology* (General Section) (January, 1926).

STERN, W. *The Psychological Methods of Testing Intelligence* (tr. by G. M. Whipple). 1914.

SYMONDS, P. M. "The Accuracy of Certain Standard Tests for School Sectioning and Marking." *Journal of Educational Psychology* (October, 1924).

—— *Measurement in Secondary Education.* 1927.

TERMAN, L. M. *The Measurement of Intelligence.* 1916.

—— *The Intelligence of School Children.* 1919.

—— "The Mental Test as a Psychological Method." *Psychological Review,* Vol. XXXI, No. 2 (1924).

THOMSON, G. H. "A Direct Deduction of the Constant Process Used in the Method of Right and Wrong Cases." *Psychological Review,* Vol. XXVI, No. 6 (1919).

THORNDIKE, EDWARD L. *Mental and Social Measurements.* Teachers College, Columbia University, New York; 1904. (Revised Edition, 1913.)

—— "Scale for Measuring the Handwriting of Children." *Teachers College Record* (March, 1910).

—— *Educational Psychology, Vol. III,* with subtitle *Work and Fatigue, Individual Differences.* (1913).

—— "Educational Measurements Fifty Years Ago." *Journal of Educational Psychology,* Vol. IV (November, 1913).

—— *A Study of Engineering Education.* Bulletin No. 11. Carnegie Foundation for Advancement of Teaching, New York; 1918.

—— "Intelligence and Its Uses." *Harper's Monthly Magazine* (January, 1920).

—— "On the Organization of Intellect." *Psychological Review,* Vol. XXVIII, No. 2 (1921).

THURSTONE, L. L. *The Fundamentals of Statistics.* 1925.

—— "The Mental-Age Concept." *Psychological Review,* Vol. XXXIII (July, 1926).

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

TILLINGHAST, C. C., et al. "Bibliography of Tests for Use in High Schools." *Teachers College Record*, Vol. XXIII, No. 4 (September, 1922). (In part annotated; 80 titles.)

URBAN, F. M. "Die psychophysischen Massmethoden als Grundlagen empirischer Messungen." *Archive für die gesammte Psychologie*, Vol. XVI (1909).

VOELKER, PAUL F. "The Function of Ideals in Social Education." *Teachers College Contributions to Education*, No. 112 (1921). Teachers College, Columbia University, New York.

WEBB, EDWARD. "Character and Intelligence." *British Journal of Psychology, Monograph Supplement No. 3* (1915).

WOODWORTH, R. S. "Combining the Results of Several Tests." *Psychological Review*, Vol. XIX (1912).

WOODYARD, ELLA. "The Effect of Time upon Variability." *Teachers College Contributions to Education*, No. 216 (1926). Teachers College, Columbia University, New York.

WYLIE, A. T. "A Brief History of Mental Tests." *Teachers College Record*, Vol. XXIII, No. 1 (January, 1922).

WYMAN, J. B. *On the Influence of Interest on Relative Success*. (Dissertation on file at Stanford University; 1924.)

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

DIRECTORY, HOME OFFICES, BRANCHES, AND AGENCIES OF HOUSES PUBLISHING TEST MATERIALS

(Publishers handling but a single test are not included in this list.)

American Council on Education, 26 Jackson Place, Washington, D. C.

Chicago, University of, Press, 5750 Ellis Avenue, Chicago, Illinois
Agents: Baker & Taylor Company, 55 Fifth Avenue, New York City
Agents in Canada: The Macmillan Company
Agents in England: Cambridge University Press, Fetter Lane, London, E. C. 4

Cincinnati, Bureau of Administrative Research, College of Education, University of Cincinnati, Cincinnati, Ohio

Courtis, S. A., 9110 Dwight Avenue, Detroit, Michigan

Ginn & Company, 15 Ashburton Place, Boston, Massachusetts
70 Fifth Avenue, New York City
2301–2311 Prairie Avenue, Chicago, Illinois
95 Luckie Street, Atlanta, Georgia
1913 Bryan Street, Dallas, Texas
199 East Gay Street, Columbus, Ohio
45 Second Street, San Francisco, California
7 Queen Square, Southampton Row, London, W. C. 1, England

Gregg Publishing Company, 20 West 47th Street, New York City
623–633 South Wabash Avenue, Chicago, Illinois
80 Boylston Street, Boston, Massachusetts
Phelan Building, San Francisco, California
Kern House, 36–38 Kingsway, London, W. C. 2, England

Hammett, J. L., Company, Kendall Square, Cambridge, Massachusetts
Newark, New Jersey

Harvard University Press, Cambridge, Massachusetts

Houghton Mifflin Company, 2 Park Street, Boston, Massachusetts
16 East 40th Street, New York City
2451–2459 Prairie Avenue, Chicago, Illinois
612 Howard Street, San Francisco, California
Agents: Thomas Allen, 366–378 Adelaide Street, West, Toronto, Canada
Agents: Constable & Company, Ltd., 10 Orange Street, Leicester Square, London, W. C. 2, England

Iowa, University of, (University Editor), Iowa City, Iowa
(Extension Division), Iowa City, Iowa

Johns Hopkins Press, Baltimore, Maryland

Lippincott, J. B., Company, East Washington Square, Philadelphia, Pennsylvania

North Carolina Bureau of Educational Research, University of North Carolina, Chapel Hill, North Carolina

Plymouth Press, 7850 Lowe Avenue, Chicago, Illinois

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Psychological Corporation, 3939 Grand Central Terminal, New York City
Public School Publishing Company, Bloomington, Illinois
Russell Sage Foundation, 130 East 22d Street, New York City
Scott, Foresman & Company, 623 South Wabash Avenue, Chicago, Illinois
68 North Pryor Street, Atlanta, Georgia
5 West 19th Street, New York City
Stoelting, C. H., Company, 424 North Homan Avenue, Chicago, Illinois
Teachers College, Bureau of Publications, Teachers College, Columbia University, New York City
World Book Company, Yonkers-on-Hudson, New York (Cable address of home office is "Foresta, Yonkers")
2126 Prairie Avenue, Chicago, Illinois
14 Beacon Street, Boston, Massachusetts
1307 Pacific Avenue, Dallas, Texas
424 West Peachtree Street, Atlanta, Georgia
149 New Montgomery Street, San Francisco, California.
100–104 Fifth Street, Portland, Oregon
101 Escolta, Manila, Philippine Islands

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:37 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

INDEX

Chapter IX includes pages 214–287. Chapter X includes pages 288–348.

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:38 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

MEASUREMENT AND ADJUSTMENT SERIES
Edited by Lewis M. Terman

TESTS & MEASUREMENTS IN HIGH SCHOOL INSTRUCTION

BY G. M. RUCH
Professor of Education, University of California
AND GEORGE D. STODDARD
Assistant Professor of Psychology and Education
University of Iowa

A BOOK that is designed as a handbook and guide for principals and teachers on the preparation, selection, and use of high school tests. It summarizes and interprets the widely scattered contributions in educational magazines and monographs on the measurement of achievement and intelligence in the secondary schools. The material is up-to-date in every respect.

It points out the values to be derived from the use of standard tests as well as the limitations which should be recognized in the use of measuring instruments. Attention is given to the most outstanding problems of measurement as it applies to high school instruction without neglecting the important details with which the test administrator should be familiar. Complete instructions are given for the development and use of the new-type objective examinations.

The wide experience of the authors in the derivation and use of tests in high school enables them to present in an unusually clear and definite manner information on high school testing which is essential to successful use of test materials. This book is admirably adapted as a textbook for use in courses in measurement.

Cloth. xxii+382 pages. Price $2.20

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Generated on 2020-12-23 01:39 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

MEASUREMENT AND ADJUSTMENT SERIES
Edited by Lewis M. Terman

STATISTICAL METHOD IN EDUCATIONAL MEASUREMENT

BY ARTHUR S. OTIS
Author of Otis Group Intelligence Scale and other tests

THE statistical methods that the school administrator or active researcher needs most to know and use are explained in this book in such a way that they can be understood by those who have had no previous introduction to the subject. Any one can obtain from the book a working knowledge of the subject that will make clear the reasons for the various kinds of statistical procedures and the meaning of the results.

Throughout the book, the practical application of methods has been kept in mind. Ample material is given for applying what is studied in the text and many diagrams and drawings are used to make explanations clearer. Several useful devices such as a percentile graph, an age calculator, and an I. Q. slide rule are included.

The conciseness and clearness with which the subject is presented makes this book an admirable guide for all schoolmen and a teachable textbook for beginning classes in statistics or in educational measurement.

Cloth. xii+339 pages. Illustrated. Price $2.16

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

MEASUREMENT AND ADJUSTMENT SERIES
Edited by Lewis M. Terman

MENTAL TESTS IN CLINICAL PRACTICE

By F. L. Wells
Chief of Psychological Laboratory
Boston Psychopathic Hospital

This book is an authoritative guide to the study of individual mentality and personality. It describes the technique of examination methods and suggests improvements of procedure in the administration of language and non-language mental tests.

The significance of test reactions is clearly explained and other factors to be taken into account in psychometric measurement are enumerated. The book also covers the free association experiment, the rôle of the clinical psychologist in solving vocational problems, details of office practice, and personality traits. Sufficient case material has been presented to show clearly how testing technique may be used to the greatest advantage.

The author has been actively engaged in effective clinical work for some years. He writes from extensive experience. The volume which he has produced is an indispensable handbook for clinical examiners of every grade of expertness, and it is also suitable for use as a textbook in normal schools, colleges, and universities.

Cloth. $x+315$ *pages.* *Price* $2.16

WORLD BOOK COMPANY
Yonkers-on-Hudson, New York
2126 Prairie Avenue, Chicago

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

MEASUREMENT AND ADJUSTMENT SERIES
Edited by Lewis M. Terman

EARLY CONCEPTIONS AND TESTS OF INTELLIGENCE

BY JOSEPH PETERSON
Professor of Psychology
George Peabody College for Teachers

THE issues involved in the use of intelligence tests are in many cases best clarified by making known the experiments and conceptions which led to their development. It is to supply this needed historical background of modern mental testing that this book has been written.

The book is a concise but comprehensive treatment of the history of intelligence testing leading up to the methods of today. It tells of the experiments and procedure of early workers in this field and gives their conceptions of what they were attempting to do. It includes also a discussion of problems regarding the use of tests and a bibliography of books on intelligence.

The author's clarity of exposition, his freedom from bias in the treatment of unsettled questions, and the inclusion of exercises and suggestions for further work make the book suitable as a text in a course in testing, in the history of psychology, and in advanced psychology in colleges and normal schools. The book will also give to educators in service an understanding of mental measurement which will be most helpful to them in the practical use of tests.

Cloth. xiv+320 pages. Price $2.16

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

MEASUREMENT AND ADJUSTMENT SERIES
Edited by Lewis M. Terman

Mental Tests and the Classroom Teacher

By Virgil E. Dickson
Director, Bureau of Research and Guidance
Oakland, California

WRITTEN primarily for teachers, from kindergarten to university, so that they may know how to use tests as an aid to better teaching and the adjustment of classroom methods to the needs of all types of students.

In a simple, straightforward way it shows why mental tests are needed, what they are like, and how they can be made most useful in the schoolroom. It points out the safe and sensible path for the teacher to follow, cautioning against the dangers of misusing tests as well as proving what practical good can be accomplished with them.

The author's unequalled experience with tests enables him to view the subject from every angle. He knows the teacher's problems and writes in terms of everyday classroom practice. His book is a most helpful guide, for administrators as well as teachers, in the practical use of tests.

Cloth. xvi + 231 *pages. Price* $1.80.

WORLD BOOK COMPANY
Yonkers-on-Hudson, New York
2126 Prairie Avenue, Chicago

Generated on 2020-12-23 01:40 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN

Stanford Achievement Test

BY TRUMAN L. KELLEY, GILES M. RUCH, LEWIS M. TERMAN
Leland Stanford Junior University

THIS new battery of achievement tests is designed to measure very thoroughly the knowledge and ability of pupils in the school subjects in grades two to eight. It covers all the ground necessary to cover for ordinary purposes of educational testing.

The score in any subject is immediately comparable with the score in any other subject, and valid composite scores for any number of subjects taken together are readily obtainable. Age norms as well as grade norms make possible the derivation of a satisfactory Educational Quotient.

Scoring is easy and objective and the directions are easily mastered. The complete examination may be given in two or three sittings in one day. Both money and time cost for a complete survey of educational achievement have been greatly reduced.

Primary Examination, for grades 2 and 3, contains tests in arithmetic, reading and spelling. *Advanced Examination,* for grades 4 to 8, contains tests in arithmetic, reading, spelling, science information, and history and literature. *Arithmetic Examination* and *Reading Examination* are each for grades 2 to 8.

PRIMARY EXAMINATION: FORM A or FORM B. Price per package of 25 tests, including Key and Class Record, $1.10 net.

ADVANCED EXAMINATION: FORM A or FORM B. Price per package of 25 tests, including Key and Class Record, $1.90 net.

ARITHMETIC EXAMINATION: FORM A or FORM B. Price per package of 25 tests, including Key and Class Record, $1.00 net.

READING EXAMINATION: FORM A or FORM B. Price per package of 25 tests, including Key and Class Record, $1.00 net.

MANUAL OF DIRECTIONS. Price 30 cents net.

SPECIMEN SET. Price 60 cents postpaid.

WORLD BOOK COMPANY
YONKERS-ON-HUDSON, NEW YORK
2126 PRAIRIE AVENUE, CHICAGO

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from
UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

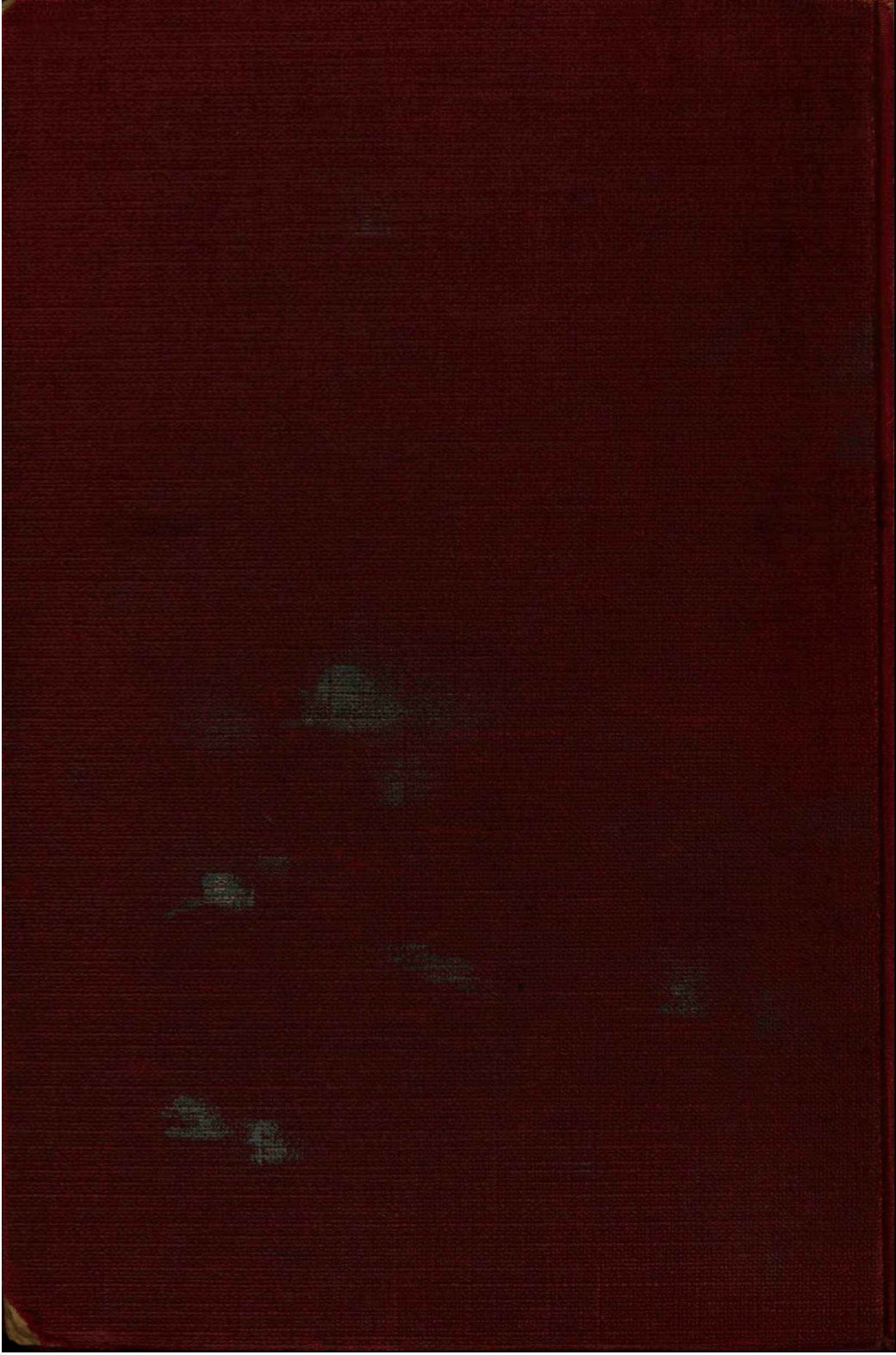
Digitized by Google

Original from
UNIVERSITY OF MICHIGAN

Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google

Digitized by Google

Original from
UNIVERSITY OF MICHIGAN



Generated on 2020-12-23 01:41 GMT / https://hdl.handle.net/2027/mdp.39015001994071
Public Domain, Google-digitized / http://www.hathitrust.org/access_use#pd-google
Digitized by Google
Original from UNIVERSITY OF MICHIGAN