112
359
THS

THE RELIABILITY AND VALIDITY OF THE COURTIS GENERAL DEVELOPMENT TEST

Thesis for the Degree of M. A.

MICHIGAN STATE COLLEGE

Robert Jay Huyser

1952

THESIS

This is to certify that the

thesis entitled

"The Reliability and Validity of the Courtis Test of Growth, Series G."

presented by

Robert Huyser

has been accepted towards fulfillment
of the requirements for

M.A. degree in Education

H. W. Sundwall
Major professor

Date May 29, 1952

O-169

THE RELIABILITY AND VALIDITY OF THE COURTIS GENERAL DEVELOPMENT TEST

By

Robert Jay Huyser

A THESIS

Submitted to the School of Graduate Studies of Michigan State College of Agriculture and Applied Science in partial fulfillment of the requirements for the degree of

MASTER OF ARTS

Department of Education

1952

THESIS

ACKNOWLEDGMENTS

The writer has received assistance from many persons both in conducting and in reporting this study. Indebtedness is expressed to Dr. Harry Sundwall, under whose helpful guidance this study was conducted, and to Dr. Leonard Luker and Dr. Alfred Dietze, who served as members of the examining committee. Further indebtedness is acknowledged to Dr. Arthur DeLong, who made available much of the data used in this study; to the many students who agreed to act as subjects, several on their own time; and to my wife, Sarah, who assisted in the preparation of both the original and final copies of the this thesis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

THE PROBLEM AND DEFINITIONS OF TERMS USED

Since the beginning of the testing movement, attempts have been made to improve psychological measurement. Many of the tests now in use have been criticized, and rightly so, for their inability to measure a single aspect or factor of the human personality satisfactorily. With few exceptions, the scores of traditional tests are influenced by a great many different factors. As a result, their meaningfulness has been questioned, and a few would claim them to be indices of so much that they cease to have genuine relationship to anything.

Dr. Stuart A. Courtis has proposed a new method of psychological measurement which is said to have promise as a remedy for some of these shortcomings. He has constructed a test using this method but has made no positive claims as to what the test measures. He has, however, provided a few interesting hypotheses which are discussed in a later chapter.

I. THE PROBLEM

Statement of the problem. The purpose of this study was to make certain tests of the reliability and validity of the Courtis General Development Test by means of statistical analysis of appropriate data.

Importance of the study. The Courtis General Development Test was copyrighted in 1930 by Stuart Appleton Courtis, Ann Arbor, Michigan. Since that time it has seen little application, except for a small number

of research projects. It was the contention of the writer that one of the major reasons for this disuse might be the complete lack of acceptable reliability and validity determinations concerning this test.

The Courtis Test, in the writer's opinion, merits investigation because of its unusual construction and because of the promise it is said to offer for the improvement of educational measurement.

This unusual method of test construction is called "ratio" testing --also referred to by Dr. Courtis as differential testing. With this technique, Dr. Courtis has attempted to "cancel out" the many factors which interfere with the precise measurement of individual differences. The theoretical aspects of this technique are reviewed in a later chapter.

Another important aspect of this new method addresses itself to the very nature of measurement. Dr. Courtis seems to have made an honest attempt to advance educational and psychological measurement from the more primitive ordinal type to the ratio type characteristic of the "exact" sciences. He has devised a new unit, called an "isochron", which is claimed to provide for an absolute zero point and for equality of units throughout the scale. These units, it is contended, are capable of being added, subtracted, multiplied, and divided as are the physical units of length, time, mass and so forth.

It is of further interest to note that Dr. Courtis has presented data which seems to indicate that the test possesses considerable cross-cultural fairness. The test has been translated into several European languages and administered to a great many native school children. The average scores of these groups were found to be approximately the same. In fact, Courtis says, "The average ratio for an unselected thousand

school children is independent of sex, age, grade, or language within the limits of ages 9 to 20 or 30."[1] It is concluded, therefore, that the test may be useful for individual comparisons from one age or cultural group to another.

Certainly, if these claims can be supported and the usefulness of the test demonstrated, it seems safe to predict that the test will become accepted and widely used, and that many more tests will be constructed by applying these techniques.

Many of the difficulties met in educational and psychological research result from spurious correlations due to the fact that scores on tests are not determined by single factors. An example of such a difficulty can be seen in the theoretical controversy on the nature of intelligence.[2]

II. DEFINITIONS OF TERMS USED

Differential testing. As used in this study, the term differential testing will be synonymous with the expression ratio testing.

Obtained score. An obtained score is an actual score made by a subject on a given administration of a given test.

Ratio testing. The Courtis measurement technique which involves administering two tests to each individual and expressing the score as

---

[1] Stuart A. Courtis, "Differential Testing as a Method of Psychological Analysis," Address of Retiring Vice-President, Section Q, American Association for the Advancement of Science, Education, Boston, December 29,1933, p. 27.

[2] For discussion refer to Chapter II.

a ratio of the scores on the two tests, thus attempting to "cancel out" factors which interfere with the measurement process.

Reliability. For the purposes of this study, reliability will be defined as the stability of scores on repeated testing under similar conditions.

True score. A true score is a hypothetical score which has been defined as the mean score obtained from an infinite number of administrations of a test to one individual.

Validity. Validity will be considered to mean the relationship between the test scores and the various criteria chosen in this study. The portion of the study concerned with the establishment of the validity of the Courtis Test will be, essentially, a survey of relationships necessary for certain uses of the test.

THE ORGANIZATION OF THE REMAINDER OF THE THESIS

Chapter I has outlined and defined the problem under consideration and presented background material essential to a full understanding of the problem. Chapter II will contain a review of pertinent literature and will include an attempt to fit this problem into its appropriate place with respect to related knowledge and research.

A discussion of the major methodological problems of reliability and validity determinations, as they pertain to this study, is advanced in Chapter III.

Chapter IV will be concerned with a description of the groups studied and materials used. The conditions under which the study was made

will be described in complete form.

A report of the research conducted in this study will be presented in Chapter V. Hypotheses will be advanced and tested; conclusions will be presented where, in the writer's opinion, they are warranted.

Chapter VI will include a summary of procedures and findings of the study. Additional problems in this area which were not considered in this study or which were raised by this study will be discussed.

CHAPTER II

REVIEW OF THE LITERATURE

A good deal of literature concerning theories of intelligence has been written; somewhat less is available with regard to reliability and validity of tests and very little concerning "ratio testing". To the writer's knowledge, the only literature concerned with "ratio testing" has been written by Dr. Courtis.

Literature concerning the Courtis Test. The Courtis Test attempts the use of a basic idea in scientific research, i.e., the law of the single variable. According to the theory presented by Courtis, two tests are administered to each individual. Both tests, like traditional tests, are influenced by a great many variables. Now, by dividing the score on the second test by the score on the first test, all of these extraneous variables are said to "cancel out".[1] The result is an index of the facet of the individual which is being measured. This assumes, of course, that the variables which influence the scores are multiplicative rather than additive and that they influence both tests equally.

Literature on the theoretical validity of the Courtis Test. The validity of a test is the degree of correspondence between scores made on the test and the "true" criterion, i.e., the trait or characteristic which the test was designed to measure. In the case of psychological

---

[1] S. A. Courtis, "Explanations Essential to Understanding," (unpublished folder), Detroit, Michigan, 1951, p. 4.

tests, of course, it is rarely or never possible to measure the "true" criterion directly. A substitute criterion must be chosen.

The problem of finding an acceptable criterion for the Courtis is a difficult one, mainly because the trait being measured, "quality", has been only vaguely defined as to the behaviors expected of a person possessing a certain amount of the trait. Dr. Courtis has referred to quality as:

> ..... the cause within the nature of the organism of differences in growth when all other factors have been held constant .... The causes of differences in the achievements of different individuals OTBE(*) is called quality. It is a nature element.[2]

Dr. Courtis, in classifying the various factors dealt with in measurement, mentions Nature factors, Nurture factors, and Maturity factors. He describes and gives examples of nature factors as follows:

> NATURE FACTORS: age, sex, differences in individual status, health, aptitude, imagination, memory, initiative, etc. The general name that will be used for all nature factors is QUALITY. That is, two children who differ, let us say, in memory for numbers will be described as having memories of so many units of qualtiy.[3]

Dr. Courtis further describes the test as measuring an "element" which, by his definition, does not change.[4]

Besides these more or less general descriptions of "quality," Dr. Courtis makes a few more specific suggestions as to what the test might measure:

(*) Other things being equal.

2 S. A. Courtis, Maturation Units and How to Use Them, Detroit, Michigan, 1950, p. 63

3 S. A. Courtis, Toward a Science of Education, (Explanations and Interpretations to Accompany Maturation Units and How to Use Them), Detroit, Michigan, 1950, p. 17.

4 Ibid., p. 25.

> These [DR's] measure the relative rate at which the individuals learn or develop under uniform conditions. (They correspond somewhat to IQ's. If you use 44.7 as a divisor for the standard value for tests 3/2, 36.4 for tests 5/4, and 41.3 for tests 7/6, you will obtain for the most of the children IQ's comparable with those from intelligence tests.)[5]

In the same general vein is this description of the manner in which one would deal with persons possessing different DR's:

> ..... To a low DR speak slowly, wait between each idea until the individual has mastered it, use concrete illustrations,... ...With high DR's do just the opposite. Talk in terms of abstract principles. Speak quickly and directly. Do not repeat, do not dominate. Let the individual state his needs and give him just what he wants and no more.[6]

and:

> A developmental ratio corresponds roughly to an IQ, or a measure of brightness. . . . . A person who is of average brightness will have a developmental ratio of 100; those who are more gifted by nature will have ratios higher than 100, and those less gifted, will have lower ratios.[7]

The resemblance of these and other descriptions to descriptions of differences in intelligence seems very pronounced. Therefore, a measure of intelligence is proposed as appropriate criterion for assessing the validity of the Courtis Test. However, lack of demonstrable validity using this criterion does not preclude validity for another criterion.

A further test of validity will be made by comparing the scores on each subtest with the scores on each of the other two. This test is

---

5 S. A. Courtis, "Instructions for Giving the General Development Tests," (unpublished paper), p. 4.

6 Ibid., p. 5.

7 S. A. Courtis, "The Interpretation of Scores in the General Development Tests." (unpublished paper), p. 4.

suggested by Dr. Courtis' claim that the three subtests measure the same quality.

Studies on the reliability and validity of the Courtis Test. A recent study concerning reliability of the Courtis Test was conducted by Dr. Arthur R. DeLong.[8] He reported mid-ratio test-retest correlations of .53 (N = 75) and .56 (N = 56) using data collected on college students.

The validity of the Courtis Test was investigated recently in a study by Rusch,[9] using the rank-difference technique on data obtained with high school freshmen (N = 140), a correlation between Courtis Mid DR and Kuhlman-Anderson IQ of .37 was found. This would provide for predictions of one score from the other which were seven per cent better than chance. It may be assumed that this sample was relatively unbiased.

In a more recent unpublished study, Jacobs[10] compared scores on the Courtis Test with Wechsler-Bellevue IQ. These data, obtained on 44 residents of the Lansing Boys Vocational School, yielded the following correlations.

Courtis score	Wechsler score	correlation
median DR	Full-scale IQ	.47
median DR	Verbal IQ	.48
median DR	Performance IQ	.50
Mid DR	Full-scale IQ	.37

---

8 Arthur R. DeLong, "How Does a Constant Disturbance Factor Affect the Developmental Ratios on the Courtis General Development Test," East Lansing, Michigan: Michigan State College Department of Elementary Education, 1952. p. 6.

9 R. Rusch, "Psychology Seminar," (unpublished paper), Naperville, Michigan, 1951, p. 3.

10 J. Jacobs, "Correlation Between the Courtis Test and the Wechsler-Bellevue Intelligence Scale," (verbal report of findings), East Lansing, Michigan, 1952.

Literature on theories of intelligence. Since the time intelligence was first measured objectively by Binet in 1904, three distinct theories of intelligence have emerged. Spearman postulated a two-factor theory, stating the intelligence was made up of a general factor, "G" and specific "s" factors. He recognized elusive group factors resulting from overlapping specific factors. Thorndike theorized that thought consisted of associations and that the number of these associations or bonds that an individual had, or could have, determined his intelligence. Thurstone, on the other hand, proposed that factors could be isolated into "primary abilities," each of which would be unrelated to other "primary abilities."

Experimentation has failed to prove or disprove any of these theories. The data have been inconclusive. In administering intelligence tests made up of subtests, each designed to measure separate aspects of intelligence, the subtests are found to correlate from about .17 to about .50. Proponents of the "G" factor theory explain the "low" intercorrelations as due to differing experience backgrounds among testees. Proponents of the primary factor theory explain the "high" intercorrelations on the basis of "impurity" in the tests.

"Ratio" testing and the theories of intelligence. If "ratio" testing should prove to be an answer to "impurities" in testing, the way would be open for research to determine whether specific aptitudes are related and whether the theory of the general factor is tenable. This study will not be concerned with these questions directly; rather it will attempt only to determine whether the specific "ratio" tests being investigated

are reliable and valid in some instances where validity would be assumed on the basis of theory underlying the test. If the tests were demonstrated to be reliable and valid, the way would be open for the construction of tests of separate abilities which may be empirically shown to be pure. Tests of this type might prove fruitful for research concerning the organization of mental factors.

CHAPTER III

METHODOLOGY

This chapter contains a discussion of the important methodological problems pertinent to this study. The approach selected will be outlined and the reasons for its use presented.

On the Assessment of Reliability. The reliability of a test refers to the consistency or stability attained in its measurements. As such, an index of reliability reveals the degree of confidence which may be placed in scores obtained with the test; i.e., it tells how closely a score may be expected to approximate some "true" score.

Statistically, the numerical value of a reliability coefficient corresponds exactly to the proportion of the score variance[1] that is due to real differences in individuals in the trait measured by the test. The remainder of the variance is due to errors of measurement.[2]

The experimental and statistical procedures used to determine reliability determine what is to be considered true variance and what is to be called error variance. There are four more or less distinct methods of assessing reliability, each having variations. These are: (1) equivalent forms, (2) test-retest, (3) split-halves, and (4) analysis of variance among individual items.

---

[1] Standard deviation squared.

[2] E. F. Lindquist (Ed.), Educational Measurement, American Council on Education, Washington, D. C., 1951, p. 561.

The test-retest procedure is, in a sense, the most conservative of the above methods. The reason for this is that any real change in the trait measured between the two administrations of the test, or in the manner in which the individual responds to the test, is considered by this method to be error variance. This is quite important in some cases, while in others it is relatively unimportant. In the measurement of attitudes and other traits which may be comparatively ephemeral and unstable, reliability may be decidedly underestimated if much time elapses between the test and retest administrations.

However, in a test which measures a hereditarily determined trait, or any other trait that is highly stable and not subject to fluctuation, this procedure is considered to be quite acceptable and, in fact, avoids certain disadvantages in the other methods.

Another consideration is the nature of the test itself. Certain tests lend themselves to certain kinds of reliability determinations.[3] In the writer's opinion, the most appropriate method for the Courtis Test is the test-retest approach.

On the Assessment of Validity. The validity of a test is the degree to which a test measures whatever it was designed to measure. There are essentially two aspects of validity; namely, reliability and relevance. The reliability of a test can be thought of as placing a ceiling on the possible validity of a test. The other phase of validity, relevance, concerns the relationship between scores on the test and the actual trait which the test was designed to measure. It follows,

---

3 Lindquist, op. cit., p. 577.

then, that to assess the validity of a test, it is necessary to have some independent measure of the trait in question. This measure is referred to as a criterion.

The criteria chosen for this study will be various measures of intelligence. Independent estimates of intelligence are provided by the Otis Quick scoring Mental Ability Test and the California Short-Form Test of Mental Maturity.

It is, of course, possible for a test to be valid for measuring one trait and not valid for another. Therefore, to demonstrate lack of validity for one purpose does not, a priori, demonstrate lack of validity for some other purpose.

CHAPTER IV

THE MATERIALS USED AND GROUPS STUDIED

The materials used in this study include, in addition to the Courtis Test, the California Short-Form Test of Mental Maturity, Advanced '50 S-Form, and the Otis Quick-scoring Mental Ability Test, Advanced, Gamma, Form Bm. The groups studied were composed of students in an undergraduate course in Child Growth and Development at Michigan State College. These groups, as well as the materials used to test them, will be described in greater length in this chapter.

I. MATERIALS USED

The Courtis General Development Test. The Courtis Test, as previously stated, is unique in its construction, the differential technique of measurement being its outstanding feature. Excerpts from the Courtis Test are presented in Figure 1 to illustrate the manner in which this idea has been applied. This test consists of three subtests, each of which contains two separately timed tests.

The first of these subtests is referred to as the "Cat and Dog" test. It is administered as follows: The subject's attention is directed to the first part of the test (part "a" in Figure 1). He is instructed to identify the animal which is like the key animal by underlining the appropriate choice and placing its identifying number in the parentheses following the four choices. The individual is then given a signal to

start work on the test. After thirty seconds, and at thirty-second intervals thereafter during the test, he is instructed to circle the choice at which he is looking and to place an appropriate number beside the circled choice. The second portion of the test is administered in the same way except that the individual is instructed to identify the animal which is the opposite of the key animal.

A procedure similar to that described above is used in the two remaining subtests. In the first of these, this idea is applied to words. (Figure 1, Parts "c" and "d") The testee responds to the first portion of the word test by selecting the word which is the same as the key word. Again he locates his progress each thirty seconds when the examiner signals, "Mark 1," and so on. The scoring is the same as in the "Cat and Dog" test. In the second "thinking" portion of the word test the procedure is the same except that the testee selects an antonym of the key word.

Again the same procedure is used in the third test where the subject matter is numbers. (Parts "e" and "f", Figure 1) In the first part of the number test, the subject underlines the number which is identical to the key number. The second portion requires that he pick the number which is the reverse of the key number. The testee responds to the signals, "Mark 1," etc., as he did in the first two tests.

The test is scored by counting the number of responses in each thirty-second interval for each test. In order to increase reliability, the highest and lowest scores are crossed out.[1] Then the remaining scores

---

[1] S. A. Courtis, "Instructions for Giving the General Development Tests," (unpublished paper), p. 3.

a b

Cat and Dog Test

c d

1. low	earth	high	down	low	23. long	short	long	extended	length
2. well	well	play	ill	healthy	24. laugh	happiness	cry	smile	laugh
3. lost	found	wasted	sorry	lost	25. early	beginning	late	soon	early
4. win	gain	success	win	lose	26. stand	stand	manners	sit	rise
5. first	number	leader	last	first	27. evening	morning	eve	twilight	evening
6. cheap	cheap	poor	inexpensive	dear	28. strong	powerful	strength	strong	weak
7. front	back	leader	before	front	29. near	known	close	far	near

Word Test

e f

1	42	57	24	75	51	42	23	32	67	41	23	32	76
2	53	64	53	35	46	71	24	62	37	26	71	62	73
3	17	62	28	17	71	82	25	14	41	23	14	58	85
4	14	14	41	23	85	58	26	71	62	71	82	28	17
5	16	83	52	61	38	16	27	24	42	57	51	24	75
6	24	24	75	42	57	51	28	61	83	52	38	16	61
7	35	35	46	71	53	64	29	43	56	34	61	65	43

Number Test

Figure 1

EXCERPTS FROM EACH PART
OF THE COURTIS TEST

are added and the score on the second test is expressed as a per cent of the score on the first test. This percentage is referred to as a percentage of development.

These percentages are then transmuted by means of a table into linear units called "isochrons." The table of isochronic values was prepared by Dr. Courtis and contains loglog values obtained from the Gompertz growth curve.[2] These units, according to Dr. Courtis, may be added, subtracted, multiplied, and divided as are the units used in the physical sciences. However, it seems appropriate to point out that the derivation and use of these units requires an assumption to the effect that the Gompertz curve adequately describes all growth and learning.[3]

After converting the percentage scores to isochronic units, each person's isochronic score is divided by the average isochronic score for the group of which he is a member. The ratio thus obtained is multiplied by 100 providing a number of the order of an IQ.[4] Thus, if a person's isochronic score is average, he will have a differential ratio of 100, while scores below average will provide differential ratios below 100 and scores above average will be transmuted to differential ratios above 100.

---

[2] For a discussion of the Gompertz curve see Croxton and Cowden, Applied General Statistics, p. 447.

[3] John C. Flanagan, "Units, Scores, and Norms," Educational Measurement, (E. F. Lindquist, editor), Washington, D. C.: American Council on Education, 1951, p. 722.

[4] S. A. Courtis, "The Interpretation of Scores in the General Development tests," (unpublished paper), p. 4.

Three DR's (or QI's[5] as Dr. Courtis has more recently called them) are obtained as a result of the scoring operations described above. These three values should be quite close together.[6] If they are, the middle one is chosen; if one differs markedly, it is rejected and the other two averaged; and if the two extreme scores differ from the middle one by more than 10 points, the test is to be given a second time and the need for constant effort explained.[7]

The California Short-Form Test of Mental Maturity. The California Short-Form Test of Mental Maturity is constructed on the basis of the multiple factor theory of intelligence. It is composed of seven subtests, each designed to measure some aspect of intelligence. The following scores are provided: 1. Total Mental Factors, 2. Verbal, 3. Non-verbal, 4. Spatial Relations, 5. Logical Reasoning, 6. Numerical Reasoning, 7. Verbal Concepts. Reliability coefficients for each of the above are presented in Table I. These reliabilities were obtained by the split-half method (corrected by use of the Spearman-Brown formula) using data obtained on 250 college freshmen. The standard deviation of the derived IQ's is given in the manual as 16 IQ points.[8]

---

5 Quality Index

6 This statement assumes that each of the three tests measures the same thing.

7 S. A. Courtis, "Instructions for Giving the General Development Tests," (unpublished paper), p. 4.

8 Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs, "Manual, California Short-Form Test of Mental Maturity, Advanced, Grades 9 to Adult, 1950, S-Form." Los Angeles, California: California Test Bureau. 1950. p. 4.

TABLE I

RELIABILITY OF CALIFORNIA SHORT-FORM TEST OF MENTAL MATURITY[9]

Score	Reliability Coefficient
Total mental factors	.94
Language	.92
Non-language	.88
Spatial relations	.87
Logical reasoning	.85
Numerical reasoning	.88
Verbal concepts	.92

[9] loc. cit.

The Otis Quick-scoring Mental Ability Test. The Otis Quick-scoring Mental Ability Test is constructed in accordance with the concept of general intelligence. As such, it combines a variety of items, all designed to measure general intelligence, to provide a single score. The split-half reliability of the Otis, Quick-scoring Mental Ability Test, corrected by the Spearman-Brown formula, is presented separately for each grade. These values are as follows: grade 10, .90; grade 11, .91; grade 12, .85.

No exact information concerning variability in the standardization group is provided. However, the manual states, "'Gamma IQ's' found by this method tend to be somewhat less variable than ordinary IQ's."[10]

Nature of the groups studied. Two groups of college students were studied. The subjects were students in a course in Child Growth and Development given in the Division of Education at Michigan State College during winter term of 1952. No attempt at random selection was made, Most of the students were in their junior year and were majoring in elementary education.

These groups appeared to be somewhat homogeneous and selected in IQ but not significantly so with regard to scores on the Courtis Test; at least this was found to be true where comparisons with less selected populations[11] were possible. Also, the information available concerning the performance of less selected groups on the tests used in this study, makes possible some fairly reasonable predictions as to what might be expected had the study been done using other groups.

---

10 Otis, "Manual for Administering and Scoring the Otis Quick-Scoring Intelligence Scale," 1937, p. 4.

11 Anonymous, "Tabulations for Norms Based on Groups of Children Alike in SEX-AGE-GRADE," p. 3.

The Courtis Test was administered to this entire group twice during the winter term. The interval between the two administrations was about two months. On the last day of the term, the group was given the Otis test.

The experimental groups. Group I was made up of all the students on whom the above scores were available. This group contained seventy-four subjects. The scores made by these subjects were used in all reliability determinations.

The following term, spring, 1952, as many members of group I as could be contacted were requested to take the California Short-Form Test of Mental Maturity. Fifty-seven persons responded to the request. These subjects constituted group II. The means and standard deviations on the criterion tests for this group were as follows:

California Test of Mental Maturity

Mean 112.7 S.D. 8.3 (IQ points)

Otis Intelligence Scale

Mean 112.4 S.D. 8.0 (IQ points).

The distributions of the Otis and the California Mental Maturity IQ's for the experimental groups are presented in tabular form in Table II and graphically in Figure 2.

TABLE II

FREQUENCY DISTRIBUTIONS OF CALIFORNIA MENTAL MATURITY AND OTIS IQ'S FOR THE EXPERIMENTAL GROUPS

IQ Mid-points	California Mental Maturity, Group II	Otis Group I	Otis Group II
137	1	0	0
134	0	0	0
131	1	0	0
128	2	1	0
125	3	4	3
122	2	11	8
119	2	7	6
116	5	7	6
113	18	9	6
110	5	12	10
107	6	10	9
104	7	5	3
101	2	0	0
98	3	4	4
95	0	2	1
92	0	0	0
89	0	1	1
N	57	73	57

AMERICAN PAD & PAPER CO., HOLYOKE, MASS.

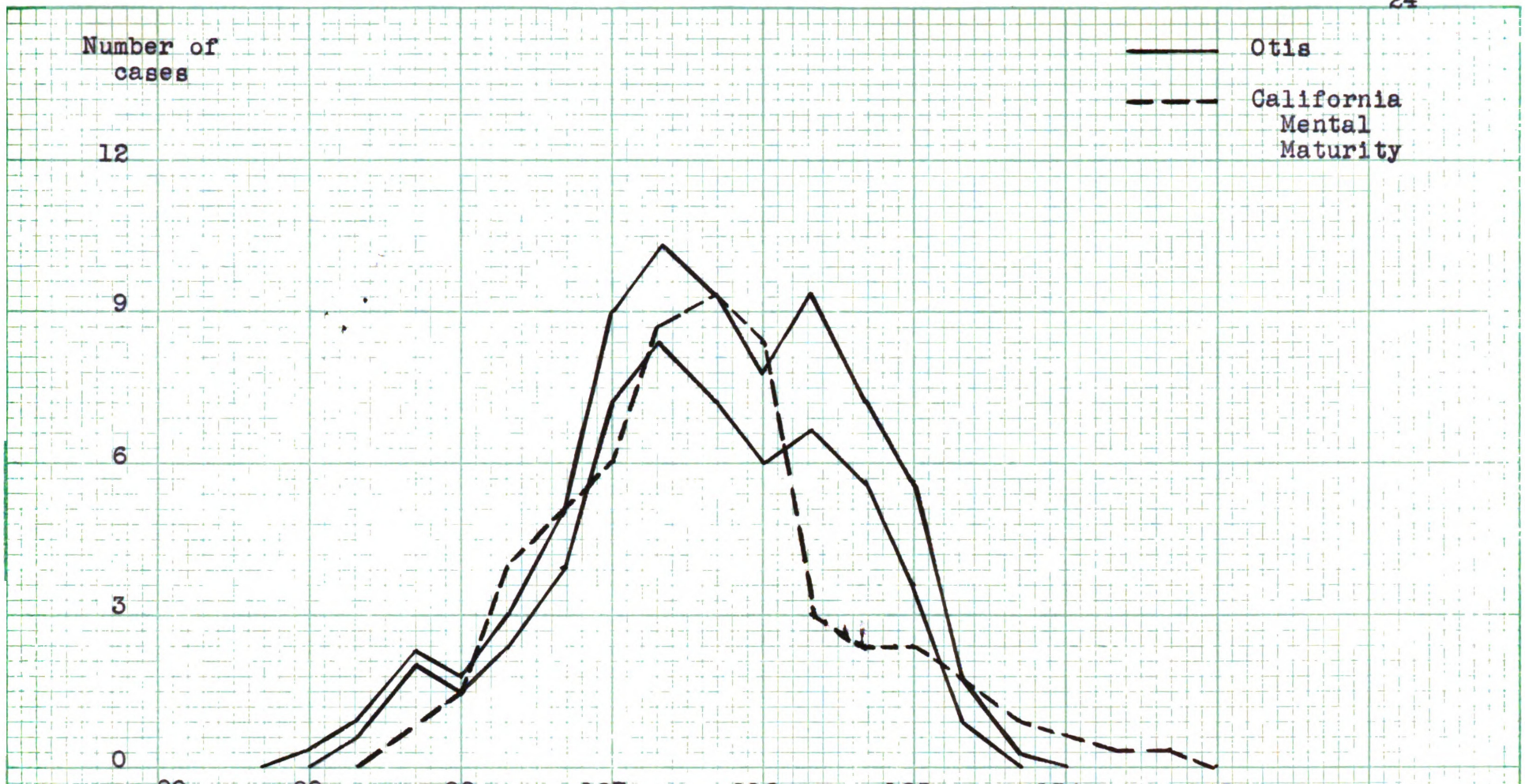

Figure 2

SMOOTHED DISTRIBUTION OF CALIFORNIA MENTAL MATURITY AND OTIS IQ'S FOR THE EXPERIMENTAL GROUPS

CHAPTER V

PRESENTATION OF DATA

This chapter will be devoted to a presentation of a statistical treatment of the data which summarizes the findings of the study. A brief discussion will parallel the presentation of these data.

Reliability. The assessment of reliability was approached with two major purposes in mind; first, to evaluate the reliability of the test for use with college groups such as the experimental group, and second, to estimate the reliability which might be obtained if the test were used with an unselected population.

The obtained reliability coefficients for group I are presented in Table III. These coefficients are estimates of the test-retest correlation for groups similar to the experimental group. An estimate of the reliability of the test for unselected groups was obtained by the use of data believed to have been provided by Courtis.[1] These estimates were obtained by adjusting the correlation to correct for curtailment of variance in the experimental group.[2] Unfortunately, this procedure could not be applied to the reliability coefficients obtained for the number test or the Mid DR's since no data were available from which to

---

[1] Anonymous, "Tabulations for Norms Based on Groups of Children Alike in SEX-AGE-GRADE," p. 3.

[2] Lindquist, E. F., (Ed.), Educational Measurement, Washington, D. C.: American Council on Education, p. 595.

TABLE III

TOTAL AND SUBTEST RELIABILITIES
FOR THE COURTIS GENERAL DEVELOPMENT TEST,
AND ESTIMATED POPULATION RELIABILITIES

Score	Reliability Coefficient	Population Estimate*
Total		
Mid DR	.460	....
Subtest		
Cat and Dog DR	.485	.618
Word DR	.454	.566
Number DR	.286	....

*Estimates of the population variance were unavailable for Mid and Number DR's.

estimate the population variance on these scores. However, it would seem reasonable to expect that each value would be increased in similar proportion.

Reliabilities of this order are very low by comparison with the reliabilities of other available tests and are definitely below the level claimed to be desirable for all but the most crude comparisons.[3]

Reliability of timed tests. It was hypothesized that the procedure of discarding the "high" and "low" scores obtained during the five 30 second intervals of each timed test, rather than increasing the reliability of the test as claimed by Dr. Courtis, actually reduced it. It was believed that a further major cause of this unreliability might be the use of the procedure of dividing the score on one test by the score on another. That is, if the "true score" on test 2 were 60 and the "true score" on test 3 were 40, an error of no more than 5 points in both might throw the ratio anywhere from .54 to .82, depending on where the errors occurred.

For these reasons it was decided to determine the reliability of each timed test using both scoring procedures. These values are presented in Table IV. These findings seem to offer an explanation for the unreliability of the ratio scores; i.e., the scores from which they are derived lack adequate stability.

The comparison of scoring methods favored, in each case, the procedure of retaining the "high" and "low" scores. The increase in reliability was not significant in every case, but the combined probability was significantly in favor of the method in which the extreme scores were retained.

---

3 Ibid., p. 609

TABLE IV

COMPARISON OF THE RELIABILITIES OF THE TIMED TESTS WHEN HIGH AND LOW SCORES WERE DISCARDED AND WHEN THESE SCORES WERE INCLUDED

Test	High and Low Discarded	High and Low Included
2	.56	.70
3	.64	.71
4	.68	.73
5	.74	.78
6	.63	.65
7	.49	.75

Ratio scores were recomputed from the scores obtained by retaining extreme scores in the timed tests. However, a comparison of the reliability of these scores with those found by the original method of scoring failed to support the hypothesis of increased reliability. The data, presented in Table V, appear somewhat contradictory and do not significantly favor either method over the other.

Intercorrelations among the subtests. It is of interest to note that the reliability of the Mid Dr (Table I) is no greater than the average of the subtest reliabilities, as would be expected if the three subtests actually measured one "element".[4] The intercorrelations among the subtests were obtained for each administration of the test. These correlations were found to be extremely low. However, it was recognized that this could have resulted from the unreliability of the subtests rather than from lack of similarity in the functions measured. Therefore, in order to maximize reliability, the "high" and "low" scores were included and the ratios for both administrations averaged. This procedure provided scores which were estimated to be somewhat more reliable (cat and dog test, .48; word test, .73; number test, .64). The intercorrelations among these sets of scores were then found and corrected for attenuation to provide an estimate of the relationships existing among the actual "traits" measured by the subtests. These values, presented in Table VI, are low enough to cast considerable doubt on the claimed similarity of

---

[4] S. A. Courtis, "Differential Testing as a Method of Psychological Analysis," (address of retiring Vice-President Section Q, American Association for the Advancement of Science, Education), Boston, 1933, p. 26.

TABLE V

COMPARISON OF RELIABILITIES OBTAINED WHEN VARIOUS SCORING METHODS WERE USED ON THE SUBTESTS

Subtest	DR (Using Isochrons)	Ratio (High and Low Scores Discarded)	Ratio (High and Low Scores Retained)
Cat and Dog Test	.49	.47	.32
Word Test	.45	.57	.58
Number Test	.29	.38	.47

TABLE VI

SUBTEST INTERCORRELATIONS FOR THE COURTIS TEST

Tests	DR's Administration I	DR's Administration II	Average Ratio Administration I & II	Corrected for Attenuation*
Cat and Dog With Words	.21	.04	.30	.42
Cat and Dog With Numbers	.07	.04	.25	.35
Words With Numbers	.04	.25	.26	.41

In this sample r must equal .23 to be significantly different from zero at the 5 per cent level, .30 for significance at the 1 per cent level.

*In correcting for attenuation, the reliability of the average ratios were estimated by the formula $r_{(1+2)} = \frac{4 r_{12} \sigma_1 \sigma_2}{\sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 r_{12}}$.

the functions measured. A relatively small portion of the "true" variance in any one of the tests seems to be accounted for, or accompanied by, variation in either of the other two sets of scores.

A similar procedure was used to estimate the true relationship among the functions measured by the individual timed tests. These findings are presented in Table VII and seem to indicate that there is somewhat more homogeneity among the timed tests than among the subtests.

Validity of the Courtis Test. Validity coefficients were obtained with group II using all scores on the California Short-Form Test of Mental Maturity as well as the Otis Quick-scoring Mental Ability Test as criteria. Validity coefficients for Mid DR and subtests using all criteria are presented in Table VIII. None of the DR validity coefficients were significantly different from zero. Of the five subtest coefficients which were significant, two were negative.

The correlation of each of the timed tests with the criteria was found and is presented in Table IX. Of these 42 coefficients, seven were significantly different from zero. The scores from both administrations were averaged and "high" and "low" scores were retained to obtain improved reliability in the timed tests.

The effect of practice on the Courtis Test. It was hypothesized that if practice effects actually "canceled out" in the differential test, there would be no difference in mean scores if the Courtis Test were administered to the same group twice. This hypothesis was tested statistically by the application of a "t" test for significance of difference in means. The results of this operation are presented in Table X.

TABLE VII

INTERCORRELATIONS
AMONG INDIVIDUAL TIMED TESTS
CORRECTED FOR ATTENUATION

	2	3	4	5	6
3	.58				
4	.68	.48			
5	.42	.48	.70		
6	.73	.59	.97	.70	
7	.54	.54	.76	.65	.90

TABLE VIII

DR AND SUBTEST VALIDITY COEFFICIENTS
OBTAINED FOR THE COURTIS TEST

Test	Otis	Total Mental Factors	Language	Nonlanguage	Spatial Relations	Logical Reasoning	Numerical Reasoning	Verbal Concepts
Administration I								
Mid DR	.07	.04	.07	.02	.09	.06	.18	-.06
Administration II								
Mid DR	.06	.06	.16	-.11	-.13	.20	.03	.14
Average of two administrations								
Cat & Dog ratio	.16	-.15	.06	-.10	-.09	.11	.08	-.11
Word ratio	.36*	-.21	.19	-.29*	-.32*	.06	.14	.13
Number ratio	.12	.28*	.18	.27*	.16	.20	.10	.10

NOTE: A correlation of .26 is required for significance at the 5 per cent level, .34 for significance at the 1 per cent level.

*Significantly different from zero.

TABLE IX

CORRELATION OF TIMED TESTS FROM THE COURTIS WITH THE CRITERIA

Test	Otis	Total Mental Factors	Language	Nonlanguage	Spatial Relations	Logical Reasoning	Numerical Reasoning	Verbal Concepts
Mean Test 2	-.02	.31*	.14	.24	.30*	-.02	.18	.09
Mean Test 3	.01	.17	.14	.13	.11	.06	.28*	.29*
Mean Test 4	.04	.11	.08	.08	.17	-.09	.08	.04
Mean Test 5	.28*	.07	.21	-.16	-.14	.01	.22	.15
Mean Test 6	.13	.13	.11	.09	.14	-.15	.18	.07
Mean Test 7	.10	.29*	.27*	.25	.24	-.00	.21	.12

NOTE: A correlation of .26 is required for significance at the 5 per cent level, .34 for significance at the 1 per cent level.

*Significantly different from zero.

TABLE X

SIGNIFICANCE OF DIFFERENCES IN
MEAN DEVELOPMENTAL RATIOS ON RETEST

Developmental Ratio	Mean Administration I	Mean Administration II	Difference	t
Cat and Dog	73.5	78.8	5.3	4.4
Word	62.0	69.9	7.9	7.9
Number	75.0	81.1	6.1	8.2

"t" (d.f. = 73) must exceed 1.96 to be significant at the 5 per cent level and 2.58 for significance at the 1 per cent level.

In all subtests the mean scores were higher on retest, and all differences were found to be highly significant (beyond the 1% level of significance). In the light of this evidence it seems reasonable to reject the hypothesis that practice effects "cancel out".

CHAPTER VI

SUMMARY AND CONCLUSIONS

I. SUMMARY

This study has had as its major focus the problem of determining the reliability and validity of the Courtis General Development Test. It was believed that this test merited study because of its new method of construction and because of its claimed culture-fairness.

Dr. Courtis' theory of differential measurement was reviewed in order to draw attention to its main features and to provide a background for the hypotheses which were to be presented and tested later in the study.

The problems of assessing reliability and validity were reviewed and the methods chosen for this study were discussed in this context.

The Courtis General Development Test was fully described, as was the recommended procedure for administering and scoring the test. The Otis Quick-Scoring Mental Ability Test and the California Test of Mental Maturity, which were chosen as criteria for the validity study, were reviewed.

All of the above tests were administered to two groups of college students, all of whom were enrolled in a course in Child Growth and Development given in the Division of Education of Michigan State College winter term, 1952. Group I contained seventy-four persons while group II was composed of fifty-seven. No attempt was made to select the subjects

randomly. The performance of both groups on both intelligence tests was presented as a part of the description of the groups.

Test-retest reliabilities were obtained for the Courtis Mid DR, each subtest, and for the individual timed tests. These are presented in Tables III and IV. Possible reasons for the surprisingly low reliabilities found for the test were discussed.

Validity coefficients found for DR's and subtests on the Courtis Test were presented in Table VIII. Out of forty coefficients, five were significantly different from zero.

The hypothesis that there would be no practice effect; i.e., no difference in means on retest, was tested. This null hypothesis was rejected on the basis of a "t" test showing all differences to be significant beyond the 1 per cent level of significance.

II. CONCLUSIONS

The major conclusions which seem to follow from this study are:

1. The reliability of the Courtis Test is too low for individual comparisons of any kind on the college level, and probably too low for comparisons of this type at any educational level.
2. Only in group comparisons could score differences be meaningful (with the probable exception of the number test which fails, with the present scoring procedure, to show enough stability for even the crudest of comparisons).
3. A major reason for the low reliability of the ratio scores seems to have been found in the low reliabilities of the timed tests.

4. The hypothesis that discarding "high" and "low" scores would increase the reliability of the timed tests was not supported. Evidence was presented which favored the inclusion of extreme scores when computing the scores on the timed tests.

5. The hypothesis that all three subtests measure the same trait was not supported to the degree necessary for comparison with other measures.

6. The validity of the test was found to be in serious doubt. An important reason for this, undoubtedly, is the low reliability of the test.

7. The portion of this study dealing with the relevance aspect of validity has tended to show very little, if any, evidence of validity. However, in view of the unreliability of the test, the select nature of the group, and the possible questionability of the choice of criterion, these results are held to be inconclusive. An unequivocal answer to this question should wait, in the writer's opinion, until the reliability of the test is improved.

BIBLIOGRAPHY

A. BOOKS

Anastasi, Anne, Differential Psychology. New York: Macmillan Company, 1937. 615 pp.

Courtis, Stuart A., Maturation Units and How to Use Them. Detroit, Michigan: Stuart A. Courtis, 9110 Dwight Ave., Detroit 14, Michigan, 1950. 148 pp.

_____, Toward a Science of Education, (Explanations and Interpretations to Accompany Maturation Units and How to Use Them). Detroit, Michigan: Stuart A. Courtis, 9110 Dwight Ave., Detroit 14, Michigan, 1950.

Croxton, Frederick E., and Dudley J. Cowden, Applied General Statistics. New York: Prentice-Hall, Inc., 1939. 944 pp.

Lindquist, E. F., editor, Educational Measurement. Washington, D. C.: American Council on Education, 1951. 819 pp.

_____, Statistical Analysis in Educational Research. Boston: Houghton Mifflin Company, 1940. 266 pp.

Thorndike, Edward L., E. O. Bregman, M. V. Cobb, Ella Woodyard, and Staff, The Measurement of Intelligence. New York: Bureau of Publications, Teachers College, Columbia University, 1925. 616 pp.

Thurstone, Louis L., The Reliability and Validity of Tests. Ann Arbor, Michigan: Edwards Brothers, Inc., 1939. 113 pp.

_____, The Vectors of Mind. Chicago, Illinois: University of Chicago Press, 1935. 266 pp.

Spearman, Charles E., The Abilities of Man. New York: Macmillan Co., 1927. 415 pp.

B. TEST MANUALS

Courtis, Stuart A., "Explanations Essential to Understanding." Detroit, Michigan: Stuart A. Courtis. 1951.

Courtis, Stuart A., "Instructions for Giving the General Development Tests." Detroit, Michigan : Stuart A. Courtis. 1951.

_____,"The Interpretation of Scores in the General Development Tests." Detroit, Michigan: Stuart A. Courtis. 1951.

Otis, Arthur S., "Manual, Otis Quick-Scoring Mental Ability Tests, Gamma Test: Forms Am and Bm." Yonkers-on-Hudson, New York: World Book Company, 1937. 6 pp.

Sullivan, Elizabeth T., Willis W. Clark, and Ernest W. Tiegs, "Manual, California Short-Form Test of Mental Maturity, Advanced, Grades 9 to Adult, 1950 S-Form." Los Angeles, California: California Test Bureau, 1950. 20 pp.

C. PERIODICAL ARTICLES

Courtis, Stuart A., "What is a Growth Cycle," Growth, I (May, 1937).

D. UNPUBLISHED MATERIALS

(Anonymous) "Tabulations for norms based on groups of children ALIKE IN SEX-AGE-GRADE." 3 pp.

Courtis, Stuart A., "A New Point of View in Psychological Measurement." Unpublished paper presented to the Michigan Academy of Science, Arts and Letters, Psychology Section, East Lansing, Michigan, March 23, 1951. 5 pp.

_____, "Differential Testing as a Method of Psychological Analysis." Address of retiring Vice-President, Section Q, American Association for the Advance of Science, Education, Boston, December 29, 1933.

_____, "The Inside Story of the New Deal in Educational Measurements." Ann Arbor, Michigan, 1934.

DeLong, Arthur R., "How Does a Constant Disturbance Factor Affect the Developmental Ratios on the Courtis General Development Test." Unpublished paper read before the meeting of the Michigan Academy of Science, Arts, and Letters, Psychology Section, Ann Arbor, Michigan: April 11, 1952. 7 pp.

Rusch, R., "Psychology Seminar." Unpublished paper, Naperville, Illinois, 1950. 6 pp.

E. VERBAL COMMUNICATION

Jacobs, James, Verbal report of findings in a study conducted at Boys Vocational School, Lansing, Michigan, May, 1952.

ROOM USE ONLY

NO 16 '54

NO 23 '54

[illegible] 14 '55

JI 26 '54

Feb 4 '55

Feb 26 '5[illegible]

Jul 7 [illegible]

Mar 24 '56

Jul 30 '5[illegible]

Aug 6 '56

Jul 2[illegible] '57

[illegible] 24 1959

DEC 12 1959

APR 20 1964

[illegible] 7 [illegible]

MICHIGAN STATE UNIVERSITY LIBRARIES
3 1293 03083 0545