

THE JOURNAL OF EDUCATIONAL PSYCHOLOGY

Vol. XII

MAY, 1921

No. 5

A SURVEY OF THE THREE FIRST GRADES OF THE HORACE MANN SCHOOL BY MEANS OF PSYCHOLOGICAL TESTS AND TEACHERS' ESTIMATES, AND A STATISTICAL EVALUATION OF THE METHODS EMPLOYED.

CLARA F. CHASSELL,

Psychologist of the Horace Mann School, Teachers College, Columbia University,

and

LAURA M. CHASSELL,

Instructor in Psychology, Ohio State University.

Part I of this article, published in the February issue of this magazine, contained a statement as to the causes which led to the undertaking of a survey of the three first grades of the Horace Mann School, and presented in full the data obtained in that survey. The measures employed included psychological tests, both group and individual, and rankings of the pupils by the teachers in maturity and ability in reading. In addition, it suggested a plan for utilizing the data thus secured, for the purpose of reclassifying the pupils into relatively homogeneous groups.

Part II continues the report of this survey, recording the correlations obtained between the various measures, evaluating these measures by comparing them with a composite of all the measures utilized, and giving a detailed account of the statistical methods employed in the conversion of these measures into mental ages.¹

¹For guidance in statistical method and evaluation of measures reported in Part II, the writers are indebted to Prof. T. L. Kelley, formerly of Teachers College, now of Leland Stanford University, and Prof. H. A. Ruger of Teachers College.

PART II. CORRELATIONS, EVALUATION OF MEASURES, AND STATISTICAL PROCEDURE

Presentation of correlations and evaluation of measures.—Table V presents the correlations obtained² between the various measures in the survey, including the composite of all. It should be noted that with the exception of the reliability coefficients reported³ these are based on mental age measures only.

TABLE V

Correlations between Mental Ages Secured by the Various Measures
Employed in the Survey

	Stanford Revision	Pressey Primer Scale	Meyer Tests	Com- bined Group Tests	Teachers' Esti- mates	Composite
Stanford Revision.....	(.93)	.45	.53	.64	.72	.89
Pressey Primer Scale.....	.45	.52 ⁴	.37			
Meyer Tests.....	.53	.37	.59 ⁴			
Combined Group Tests....	.64				.42	.87
Teachers' Estimates.....	.72			.42	(.61)	.85
Stanford Revision and Group Tests96
Stanford Revision and Teach- ers' Estimates.....						.91
Group Tests and Teachers' Estimates77					.98
Composite89			.87	.85	

²According to the product-moment method of Pearson

³The reliability coefficients actually taken into account in determining the weightings assigned to the various measures entering into the composite, referred to in Part I, were only approximate in the case of the Stanford Revision and the teachers' estimates. The coefficients reported in Table V were variously ascertained. The figure given for the Stanford Revision (.93) is the correlation found between earlier and later tests. (See Terman, L. M., *The Intelligence of School Children*, p. 142) The figures for the two group tests (.52 for the Pressey and .59 for the Meyer) were calculated by computing the correlation between the sum of the scores for odd tests and the sum of the scores for even tests. These coefficients are raised to .68 for the Pressey and .74 for the Meyer by the application of 'Brown's' formula (See Brown, William, *The Essentials of*

Mental Measurement, p. 102, fn. The formula as there stated is $r_{11} = \frac{1 + (n-1) r_1}{n}$

In this instance we are not interested in determining the number of applications of the test necessary to give any desired degree of reliability, however, but in determining the reliability of the entire test from the reliability of one-half of the test. Thus, in the case of the Pressey the correlation of .52 between the sum of the scores for odd tests and the sum of the scores for even tests is the reliability coefficient for only one-half of the test. In determining the reliability coefficient of the entire test, $n=2$ Applying the

formula, we have r_{11} , i. e., the reliability coefficient of the entire test, $= \frac{2(.52)}{1 + (2-1) .52} =$

$\frac{2(.52)}{1 + .52} = .68$. Cf also the formula given by Spearman, in *British Journal of Psychology*,

Vol. III, p. 281. The correlation given for teachers' estimates (.61) is that found to obtain between estimates of school work, made by two or more teachers of the same children in the Horace Mann School, at the time the criterion for the National Research Council Tests was being compiled

⁴Raised to .68 by the application of 'Brown's' formula

⁵Raised to .74 by the application of 'Brown's' formula

An examination of the correlations reported for the Stanford Revision shows that the highest correlation between this measure and any other single measure appears in the case of teachers' estimates, namely, .72 (60 cases being included).⁹ In this connection the findings of two other experimenters, while not directly comparable with our results,⁷ are sufficiently related to be of interest. Terman⁸ reports a correlation of .48 between intelligence quotients and teachers' estimates of intelligence, the rankings being on a scale of 5. Similarly, Dickson⁷ found a correlation of .79 between these same measures in the case of 149 first grade children, the teachers having been expressly cautioned to take the children's ages into account in making the ranking.

The correlations with the two group tests¹⁰ are much lower, namely, .45 for the Pressey Primer Scale, and .53 for the Meyer Tests¹¹ (based on 37 and 45 cases, respectively). A correlation of only .37 was found between the two group tests themselves. A combination of these two tests into a single measure results in a correlation of .64 with the Stanford Revision.

The correlation of .42 between the combined group tests and the teachers' estimates is surprisingly low, since a group examination, to a far greater extent than an individual, would seem to require a response similar to that met by the teachers in the usual classroom situation.

From the practical standpoint, the correlation of greatest interest is probably the one between the Stanford Revision and group tests combined with teachers' estimates. The correlation of .77 between these measures, based on only 30 cases, is sufficiently high to raise the question as to whether an individual psychological examination is really necessary for determining classification and promotion. If the possibility of partially explaining this relatively high correla-

⁹It should be borne in mind, however, that the teachers were already familiar with the intelligence quotients of a large number of their children. While it is probable that no direct comparison with these quotients was made, this previous knowledge doubtless had some influence upon the rankings assigned.

The rankings made by the teachers in our survey were of maturity and ability in reading, and in general extended over the entire range of the class.

⁷The Measurement of Intelligence, p. 75

⁸See Terman, L. M., *The Intelligence of School Children*, pp. 51-52.

¹⁰All correlations reported with the combined group tests are based on 30 cases only. This small number of cases in which direct comparison is possible is due to the prolonged absence of an unusually large number of children.

¹¹As explained in Part I, the Meyer Tests used in this survey are those constructed by Miss Helen Meyer, described in an unpublished thesis, "Group Tests for Grades I and II," on file in the Columbia University Library.

tion by the fact that the teachers were familiar with the intelligence quotients of a large number of their children be entirely overlooked, it should still be emphasized that a correlation of .77 means a reduction in the ratio of the variability around the regression line to the variability around the average of only .64." Even if this reduction in variability were much greater, in view of the great value of the Stanford Revision for the understanding of individual pupils, we should still hesitate to say that it was not an essential instrument for satisfactory classification and promotion. Certainly, however high this correlation may be found by other investigators, the Stanford Revision or its equivalent will continue to be indispensable at least for determining the placing of the problematical cases of any kind.

Aside from such information as may be afforded by the correlations with the Stanford Revision, already reported, for an evaluation of the various measures used in the survey we are dependent upon a comparison of these measures with the composite of all." As explained in Part I, this composite is the average mental age found by adding the mental ages obtained from the Stanford Revision, the two group tests, and the teachers' estimates, the Terman mental age being doubled, and dividing this total mental age by the number of measures available for each child, the Terman mental age being counted as two measures.

As judged from the correlations with this composite, presented in Table V, any one of the three types of measures which were used in the survey, that is, an individual examination or two group tests or two teachers' estimates, would approximate the same results for classification and promotion as the composite itself. The correlations for these three measures, based on 60, 30, and 60 cases, respectively, are .89, .87, and .85. In interpreting these correlations, as well as the other correlations to be reported with the composite, it must be borne in mind that the measures compared are not independent, but are already contained within the composite itself.

¹²Obtained by the formula, $\sqrt{1-r^2}$.

¹³It is impossible to evaluate this composite by comparing it with an altogether independent criterion of the value of the various measures employed for purposes of classification and promotion. Such an independent measure would be available subsequently if the results of instruction in groups classified on the basis of the composite could be compared with results obtained in control groups not so classified. One evidence that the composite employed is a fairly satisfactory criterion by means of which to evaluate the measures used in the survey, however, is the general satisfaction felt by the teachers with the results of the survey as summarized in this composite.

Hence they are far higher than would otherwise be the case." Even so, taking these correlations at their face value, since the ratio of the variability around the regression line to the variability around the average would be reduced approximately only one-half¹⁴ by the use of any one of the three types of measures, taken alone, they can be claimed, if used singly, to possess only a limited value for purposes of classification and promotion as compared with the composite.

The correlations with the composite are naturally very much higher still when any two types of measures are combined for purposes of comparison with it, on account of the fact that the four elements included in the two types of measures (e.g., the Terman mental age taken twice plus the mental ages obtained from the two group tests) are identical with four out of the six elements in the composite. Thus the Stanford Revision combined with the group tests gives a correlation of .96 with the composite, the Stanford Revision combined with the teachers' estimates, of .91, and the group tests combined with the teachers' estimates, of .98,¹⁵ (30 cases being included in each instance). The ratio of the variability around the regression line to the variability around the average for these correlations is, respectively, .28, .41, and .20. In interpreting these results, it should be noted that the presence of common elements in the measures correlated, which served in the first place to increase the correlations with the composite, results now in correspondingly lower regression values. Should the special conditions under which these correlations were obtained remain, if only two types of measures are to be used for purposes of classification and promotion, it would seem from these results to make little difference whether the Stanford Revision combined with group tests or group tests combined with teachers' estimates were selected.

¹⁴The reader who is interested in following up the implications of this statement statistically, may be referred to Yule, G. U., *An Introduction to the Theory of Statistics*, ch. XI. (See especially Exercises 6 and 7, p. 227.)

¹⁵Exactly one-half (i. e., 50) if the correlations were all .866. The exact regression values obtained by the formula, $\sqrt{1 - r^2}$ for these three correlations are .46, .49 and .53, respectively. It is apparent from this formula, however, that since these regression values involve the correlations with the composite, they are too low for the same reason that the correlations upon which they are based are too high.

¹⁶This last-mentioned correlation may have been artificially increased by still another factor, since as already stated, the teachers' judgments probably were influenced to a certain extent by previous knowledge of the intelligence quotients of the children. The extent of such influence can not be determined. Whatever it may have been, however, it did not succeed in raising the correlation between the Stanford Revision and the teachers' estimates above .72.

The statistical procedure utilized in the conversion of the measures into mental ages.—Reference has already been made in Part I¹⁷ to the fact that the incorporation of such varied data as were secured in the survey necessitated the reduction of all measures to a common basis. The one selected for this purpose was that of mental age.

Before concluding the report of the survey it is thus necessary to present in detail the statistical procedure involved in converting the various measures used into mental ages. The methods used are given in order for the measures employed.

1. The Stanford Revision.

In the case of the Stanford Revision the only computation necessary was that required on account of the fact that the giving of the individual examinations had extended over a long period. It was thus necessary to determine the mental ages of all the children at some specified time. Since the group tests had been given during that same month, the date selected was January 1. After the chronological ages of the children on that date had been computed, the corresponding mental ages were readily obtained by using the intelligence quotients already determined as multipliers.¹⁸ This process was facilitated by the use of an I. Q. slide rule.¹⁹

2. The Pressey Primer Scale and the Meyer Tests.

The method used in converting the scores made in both the Pressey Primer Scale and the Meyer Tests into mental ages, consisted of replacing the score made by a given child in one of these tests by the corresponding mental age, as shown in a table constructed with the median scores in the test for the various ages as a basis.

Since only year norms were available for the Pressey tests and no norms whatever for the Meyer, before the conversion into mental ages was possible it was necessary to build up a table of norms by months. In the case of the Pressey the procedure was as follows: First, since the various age norms provided for the Primer Scale actually represent the typical performances of children six months in advance of the ages specified, the norm for any given year being

¹⁷See *Journal of Educational Psychology*, Vol. XII, No. 2 (Feb., 1921), p. 74.

¹⁸This method is based on the assumption of the constancy of the I. Q., which, although not thoroughly established, seemed to be sufficiently so for use in the present instance.

¹⁹Published by the Reed College Co-Operative Store, Portland, Oregon.

based upon the scores made by all the children who have passed the birthday indicated and have not yet reached the succeeding birthday, the median score given as the norm for six years of age, for example, was taken to have a value in terms of mental age of six and one-half years. The amount of the interval between each of the median scores for the succeeding ages was then determined, and this amount divided by twelve in order to obtain the increment in terms of score which might be taken to correspond to a month of mental age. Each increment was then successively added to the appropriate year norm. The values thus secured were subsequently entered in a table opposite the equivalent year and month of mental age.

The procedure followed in the case of the Meyer Tests was naturally more complicated. No norms being available, it was necessary to estimate norms by means of a comparison of scores made by children in these tests with scores made by the same children in some other test. For this purpose the records from the Stanford Revision were utilized.

First, the mental ages of the twenty-nine children given the six Meyer tests in the regular manner,²⁰ all of whom had also been given the Stanford Revision, were arranged according to the respective chronological ages. Then for each successive half year of chronological age the median of the chronological ages and of the mental ages was found. Each median mental age was then divided by the median chronological age corresponding, and the median intelligence quotient for the children of each half year of chronological age found. The scores made by these same children in the Meyer Tests were similarly tabulated according to the chronological ages of the children, and the median score for each half year ascertained. The age medians thus secured were next divided by the intelligence quotients obtained as described above, in order to secure medians which would be typical of children in general, and which thus might serve as norms for the corresponding chronological ages. The values resulting were then employed in the same way in which the Pressey age norms had been utilized, and a table of scores with their

²⁰Sometime before plans had been made for the survey, two out of the six tests in the Meyer series had been given to the children in one of the rooms. Although, after the remaining four tests had been given to these children, their scores were adapted to make them comparable with those made by the other children, these records were not utilized in estimating the norms.

accompanying equivalents in terms of years and months of mental age built up.

3. The teachers' estimates.

The conversion of the teachers' estimates into mental ages was made by adapting two different methods. Only the general outlines of the methods will be indicated here; the adaptations employed were made in the course of the practical application of the procedures, and would not be of general interest. Both methods involved the transmutation of a given teacher's rankings in each trait, into an equivalent mental age, determined on the basis of the Terman mental ages on January 1st, already calculated, on the principle that for any group of children the distribution of mental age indicated by any measure would be similar to that already found by the application of the Stanford Revision. The first method, which is based on the assumption that the form of distribution concerned is that of the normal probability surface, was employed in the case of the two classes for which all or practically all of the Stanford Revision records were available; the second method, which makes no such assumption, was employed in the case of the third class because the data were relatively incomplete on account of the prolonged absence of a number of the children. For purposes of comparison a third method, used in the survey made in the spring throughout the elementary school, in which the ranks assigned were treated as gross scores, is included.

Method A, in which the distribution was assumed to follow the normal curve. The average and the standard deviation from the average of the Terman mental ages for the class concerned were first computed. Then the number of children assigned each of the teacher's rankings²¹ was determined, and the percentage that this number represented of the total number of the children in the group computed. The average distance of each percentage from the central tendency of the total group in terms of multiples of the standard deviation was then ascertained by reference to Table 54 in Thorndike's *Mental and Social Measurements*.²² The multiples thus obtained were multiplied by the value of the standard deviation from

²¹In a number of instances more than one child had been assigned the same rank.

²²See p. 221 ff.

the average of the Terman mental ages already obtained. The resulting quantities were then added algebraically to the average Terman mental age for the class, and the resulting figures replaced in each case by the nearest whole number. This figure, which represented the number of months of mental age desired, was then converted into years and months.

Method B, in which the distribution was treated as rectangular. In the first place, the highest rank was replaced by the highest mental age found for the children concerned, as its equivalent; and, similarly, the lowest rank was replaced by the lowest mental age. The range in months between the highest and the lowest age was then divided by $(n-1)$ the number of ranks, in order to determine the number of months of mental age which should be taken as equivalent to the interval between two successive ranks. The amount thus secured was subtracted from the highest mental age to secure the value of the second highest rank; then subtracted from this value to secure the value of the third highest rank, and so on, until a value had been found for each rank. The quantities resulting were in each case replaced by the nearest whole numbers, and these numbers, which represented the number of months of mental age appropriate to each rank, were then reduced to years and months.

Method C, in which ranks assigned were treated as gross scores. In the first place, since records for identical children in both measures, only, were usable, records for any children who had not been included in the teachers' rankings and examined by the Stanford Revision as well, were eliminated from the calculations. Next, the remaining ranks being treated as gross scores, the average and the standard deviation from the average of these ranks were computed. Similarly, the average and the standard deviation from the average of the Terman mental ages for the same children were found as in the first method described above. Then, the standard deviation of the ranks being considered as equivalent to the standard deviation of the mental ages, the number of months of mental age equivalent to the interval between two succeeding ranks was found. Thus, if the standard deviation of the ranks of a given class was 3, and the standard deviation of the mental ages for the same children was 6 months, the interval between two succeeding ranks would be considered equivalent to two months of mental age. Finally, with the

average mental age as the starting point, the appropriate number of months was added or subtracted for each following or preceding rank; the resulting figures, which represented mental ages, were replaced by the nearest whole numbers, and the latter reduced to years and months.