The Reliability of the Binet Scale and of Pedagogical Scales

Arthur S. Otis & Herbert E. Knollin

To cite this article: Arthur S. Otis & Herbert E. Knollin (1921) The Reliability of the Binet Scale and of Pedagogical Scales, *The Journal of Educational Research*, 4:2, 121-142, DOI: [10.1080/00220671.1921.10879187](https://doi.org/10.1080/00220671.1921.10879187)

To link to this article: <http://dx.doi.org/10.1080/00220671.1921.10879187>

Published online: 15 Dec 2014.

Submit your article to this journal [↗](#)

View related articles [↗](#)

Citing articles: 1 View citing articles [↗](#)

THE RELIABILITY OF THE BINET SCALE AND OF PEDAGOGICAL SCALES¹

ARTHUR S. OTIS

Washington, D. C.,

AND

HERBERT E. KNOLLIN²

Leland Stanford Junior University

In connection with a study made by one of the writers (Knollin) on the intelligence of 180 adult males, it was suggested by Dr. Terman that it would be desirable to determine the reliability of the scale used, which was the 1915 edition of the Stanford Revision of the Binet Scale. The 180 subjects included 150 migrating unemployed and 30 business men. The usual precautions were taken in the administration of the tests.

As a measure of the reliability of the scale it was proposed to find the probable error of its score (the expression "probable error" being used in a restricted sense). This was found to be approximately six months in mental age. That is, in 50 percent of cases, mental ages of adults may be assumed to be correct within six months. It follows from this, theoretically, that in 90 percent of cases the score will probably be correct within 15 months, and in only one case in a hundred will the error probably be in excess of 23 months.³

¹ Involving a determination of the "probable error" of a mental age by the Binet Scale, an example of the use of a *difference* formula for correlation, and a discussion of the logical and mathematical aspects of the reliability of scales for measuring "mental" and "pedagogical" ability. This article was written in 1916. Publication was delayed on account of the war.

² Knollin is responsible for the testing and scoring involved in obtaining the material for this study; Otis is responsible for the method and proofs.

³ More recently Mr. Virgil E. Dickson of Stanford University, using the same method as that described herein, has found that the probable error of a mental age when the Stanford Revision is used with first-grade children, chiefly six to eight years of age, is approximately three months. Though less in absolute amount, this is about the same proportion of the mental age. For a child of seven years, an error of three months in mental age is one of 3.57 points in intelligence quotient. Taking 14 years as the median mental age of our miscellaneous adults, an error of six months in mental age is the same error of I. Q. This indicates that the probable error of a score varies with the amount of the score, and suggests that the probable error of an I. Q. is probably approximately constant, being about $3\frac{1}{2}$ points. From this it would follow theoretically that an I. Q. by the Stanford Revision is probably in error to the extent of about 6 points or more in a quarter of the cases, 10 points or more in one case in ten, and 14 points or more in one case in a hundred.

In finding the probable error of a score, a line of relation between the scores by the two halves of the scale was drawn by the "method of rank correspondence" described elsewhere.⁴ The determination of a coefficient of correlation between the two series of scores by a "Difference Method" is illustrated.⁵

The method deemed proper for expressing the degree of reliability of a scale or test is to give the probable error of a score, rather than to give only the coefficient of correlation between series of scores or between degrees of difficulty of the elements of a scale for different groups of individuals.

The maker of a scale, either pedagogical or intelligence, should give, therefore, in the interest of a proper interpretation of results, the probable error of a score by the scale, as a measure of its reliability.

THEORETICAL CONSIDERATIONS WITH REGARD TO METHOD

There are various ways in which we may conceive of the reliability of the Binet Scale. We may ask:

1. What is the probable deviation of a mental age by the Binet Scale from the average mental age that would be obtained by the same examiner testing the same individual many times, assuming no effect remained in any case from the previous testings? This deviation would result from fluctuation of attention, etc., on the part of the examinee, and from possible differences in method of giving the tests on the part of the examiner. (*Concept 1.*)

2. What is the probable deviation of a mental age by the Binet Scale from the average mental age that would be obtained by many different examiners testing the same individual with different scales made as nearly as possible like the present Binet Scale? This deviation would result not only from the above causes but also from the differences in personality of testers, and from the impossibility of making two scales exactly alike. It would therefore be greater than the first deviation. (*Concept 2.*)

3. What would be the probable deviation of a mental age by the Binet Scale from the true mental age of the individual tested,

⁴ See Reference 4 in the bibliography at the end of this article.

⁵ Since this formula was first presented by one of the authors (Reference 4) he has made diligent search in the literature in order to discover whether the same formula had been presented before. Such prior presentation has not been found and the formula in its present form is believed to be original with Otis.

as determined by a hypothetical scale which measured the intelligence perfectly, according to some definition. The average of a large number of measures of the intelligence of an individual by different testers using the same and different scales would not be free from the influence of environment, etc. Accordingly, this deviation would doubtless be even greater than the second. It would be the true measure of the probable error of a mental age by the Binet Scale. (*Concept 3.*)

We have no means of measuring this true probable error since we have no true measure of the intelligence of the individual. We have no way of measuring even the second mentioned deviation since we do not have different independent Binet scales. Moreover, a second testing by a different examiner would give results somewhat affected by the first testing, and for this reason even the first-mentioned deviation must be found in an indirect way.

Still another point to be noted here is that the probable error of a single score (in the sense of being the probable deviation of the score of any test from the average of a large number of scores of the same individual by the same scale) is less than the probable deviation of any one score from another. The former is, in fact, equal to $\sqrt{1/2}$ times, or 0.707 of, the latter theoretically. (See Appendix I for proof.) It is this probable error of a single score which we are seeking ultimately and to which we shall refer when speaking of the probable error of a mental age.

For the purpose of finding the probable difference between any two measures of the mental age of an individual by the Binet Scale, made by the same tester and assuming no lasting effect of the first testing, we are obliged, since we have but one scale, to divide it into two halves and find first the probable or median difference between the mental ages of single individuals by the two halves of the scale. Upon theoretical grounds it may be shown that the probable error of a mental age by the Binet Scale as a whole is equal to $\sqrt{1/2}$ times, or 0.707 of, the probable error of a mental age by one-half the scale. Therefore the probable error of a mental age by the whole scale is equal to $\sqrt{1/2} \times \sqrt{1/2}$ times, or $1/2$ of, the median difference between the mental ages by the two halves of the scale and hence can be very easily found from the latter by dividing it by 2. Proof of these propositions is given in the Appendix (I and III).

PROCEDURE

The scale was divided into two parts by placing the first half of the tests of each age group in one scale which we have called Scale A, and the second half of the tests of each age group in a second scale which we have called Scale B. The values of each test in months were then doubled so that the mental age by each half would be comparable to that by the whole. A point was then plotted as a small circle in Figure 1 for each individual, having an abscissa equal to his score by Scale A and an ordinate equal to his score by Scale B. A relation line was then drawn through these points as shown in the figure, according to a method which is called "the method of rank correspondence," and which may be called the *single relation line* to distinguish it from a regression line of which there are two for any pair of variables. This method⁶ is based on the assumption that, considering the values individually, the median of one distribution most probably corresponds to the median of the other, that the upper and lower quartile values of one distribution most probably correspond to the upper and lower quartile values respectively of the other, and similarly that the values having each of the other ranks in one distribution most probably correspond to the values having the same rank in the other distribution. For the purpose, therefore, of finding the position of the line of relation between pairs of scores, certain pairs of scores having the same ranks were selected. These were the middle values of each consecutive five in each distribution. Thus, beginning with the upper end of the distribution of Scale B values, the third, eighth, thirteenth, eighteenth, etc. values were indicated by blacking the centers of the circles; and the values having those ranks in Scale A were similarly indicated. By means of crosses points were then plotted in the quadrant whose respective abscissas and ordinates were equal to the values having the third, eighth, thirteenth, eighteenth, etc. ranks in each distribution. In this way 36 points were plotted, the abscissa and ordinate of each having the same rank in the two distributions. According to our assumption, then, these points best indicate the trend of the line of relation between the scores by the two scales. Inasmuch as there seemed to be no marked indication of a curvature in the line of relation, it was assumed to be rectilinear. Therefore a straight line was drawn as nearly as possible through the crosses, and this

⁶ See also Reference 4.

was assumed to be the line of relation between the scores by the two scales.⁷ If this line may be assumed to be correctly placed, then the ordinate of each point on the line represents the score by Scale B which corresponds to the score by Scale A represented by its abscissa. By means of the relation line, then, every score by Scale A may be transmuted into terms of Scale B in order to com-

FIGURE 1. THE CORRESPONDENCE BETWEEN THE MENTAL AGES OF THE 180 INDIVIDUALS BY THE TWO HALVES OF THE BINET SCALE (STANFORD REVISION)

⁷ This line is then presumably approximately as would be the line, $y = \frac{\sigma_y}{\sigma_x} x$, if the axes passed through the means of the two distributions. To draw this latter line in the first place, however, would have been to assume that the line of relation was a straight line without first determining whether it was or not. The above method makes no such *a priori* assumption.

compensate for differences in difficulty between the two scales. The difference between any actual score by Scale B and the score of the same individual by Scale A, after this has been transmuted into terms of Scale B, is one of the differences of which we are seeking the median. The value of this difference will be represented by the distance of the point for this individual above or below the line. The median of these distances, therefore, measures the probable difference, in terms of Scale B, between the scores of the several individuals in the two halves of the Binet Scale, when the difference in difficulty between them has been compensated for.

The distribution of the distances of the points above and below the line is shown in Figure 1 at the right. In order to find the median of these differences graphically with a reasonable degree of accuracy, it was necessary to construct Figure 2, in which the distribution of the differences has been plotted again at the left using a larger scale and with both plus and minus differences measured upward. These have been plotted once more at the right with each difference measured on a separate ordinate, the ordinates increasing in magnitude to the right. A smooth curve, being one-half an approximate ogive, was then drawn through these points as shown. The midpoint of this half ogive, measuring horizontally, was then found by erecting an ordinate between the ordinates of the ninetieth and ninety-first differences. This point may be seen to represent a difference of approximately 10.8 months in mental age. As a check upon this method, the average of the differences was calculated and found to be 12.62 months. Multiplying this by 0.8453⁸ gave 10.67 months, which is very nearly 10.8, as the theoretical median difference. These measures, of course, are in terms of Scale B. To find the interval in terms of Scale A corresponding to 10.8 months on Scale B, it is necessary to multiply 10.8 by the cotangent of the angle of the line of relation with the horizontal axis, since this is the ratio of the projections of any section of the line of relation upon the two scales. In this case the cotangent equals 0.96. Multiplying 10.8 by 0.96 gives 10.4 months as the median difference between scores in terms of Scale A. We may now take

⁸ The median deviation of a normal distribution equals 0.8453 times the average deviation.

FIGURE 2. THE DERIVATION OF THE MEDIAN DEVIATION OF A DISTRIBUTION OF VALUES (DIFFERENCES SHOWN AT THE RIGHT IN FIGURE 1) BY A GRAPHIC METHOD

the average of these values, 10.8 and 10.4, which is 10.6 months, as the median difference between scores in terms of the Binet Scale as a whole. As was stated above, the probable error of the Binet Scale, according to Concept 1, is equal to one-half of this value and therefore approximately 5.3 months.

It should be noted in this connection, however, that in the cases of the highest scores—those in the neighborhood of 18 or 19 years—the differences between the scores by the two halves of the scale could not be very great since very few tests were failed. That is, these differences are less than they would have been if there had been a larger number of the more difficult tests. This, no doubt, renders our obtained value of the probable error somewhat lower than its true value. It is thought likely, therefore, that one-half year would be more nearly the true value of the probable error of an adult's score by the Stanford Revision of the Binet Scale according to Concept 1.⁹

THE IMPROPER USE OF THE COEFFICIENT OF CORRELATION AS A MEASURE OF RELIABILITY

In this connection it might be well to make reference to a common improper use of a coefficient of correlation as an expression of the degree of reliability of a measure or scale. The manner deemed proper for expressing the degree of reliability of a score is, as in this study, to give the probable error of a score in the units in which the score is expressed. The difference between the two methods may be illustrated as follows. The correlation between the scores of the 180 individuals by the two halves of the Binet Scale has been calculated by the Pearson formula and found to be 0.850. Three other cases were considered. Calling the case of the 180 individuals Case 1, than as Case 2 were considered only the individuals whose mental ages fell between 13 and 16-11, in which case the correlation was found to be only 0.44. As Case 3 were

⁹ Since writing the above, one of the writers (Knollin) has tested a number of the convicts in a California state prison, using the latest Stanford Revision of the Binet Scale (1917) and from 133 of these testings the reliability of that revision has been determined by the same method. The median difference between mental ages by half the scale was found to be 10.9 months (corresponding to 10.6 months in the case mentioned above). It may be said, therefore, that the probable error of a mental age by the later revision is practically the same for adults as that of the former and may be taken also as one-half year.

considered only those individuals whose mental ages fell between 13 and 14-11, in which case the correlation was -0.14 . And as Case 4 were considered only those individuals whose mental ages were between 13 and 13-11, in which case the correlation was -0.62 .

Thus it may be seen that differences in the heterogeneity of the group make very great differences in the values of the coefficients of correlation between the scores, so great, in fact, as to rob the coefficient (0.85) of much of its significance. Doubtless the correlation would have been considerably higher than 0.85 if a large number of children of ages down to 3 or 4 had been included in the group. The probable errors of the scores in the four cases were also determined and found to be respectively 5.3 months, 5.8 months, 5.9 months, and 6.0 months, showing that heterogeneity of the group, as such, probably had no serious effect upon the value of the probable error of the score. These last three values of the probable error of a score are believed to be more nearly correct than the value 5.3 for reasons stated above. The further fact that in as small a group of individuals as 31 the probable error has practically the same value as that in the cases of the larger groups speaks well for the validity of that value.

The point to be noted in this connection is that the value of a coefficient of correlation between two series of measures depends upon two variables, first, the amount of difference between the members of each pair of x - and y -values (when expressed in the same terms) and second, upon the degree of heterogeneity of the group of individuals with regard to the magnitude being measured. Now the reliability of a scale has nothing at all to do with the heterogeneity of the group, except as we wish to consider the probable error in relation to the heterogeneity. The reliability should therefore be measured by values which are independent of the heterogeneity. The probable error of a score fulfills this condition and expresses the reliability in very significant terms; it tells us what limit of error may be expected in 50 percent of cases of measurement; and from it upon theoretical grounds the limit of error that may be expected in any other percent of cases may be calculated.

It should be said that in the special case in which it is desired to find the *relative* degrees of reliability of a number of tests which have been given to the *same group* of individuals, a coefficient of

correlation between duplicate sets of scores in the same test serves as a convenient measure of the degree of reliability of that test, with which to compare a similarly derived measure of the reliability of another test. This method of comparison is valid since the degree of heterogeneity of the group is the same in all cases. The coefficients derived from two different groups of individuals, however, would not be comparable unless the degrees of heterogeneity were the same for both groups.

THE USE OF A DIFFERENCE FORMULA FOR CORRELATION

It may be valuable in many cases to know the relation between the theoretical variability of the several scores of a single individual and the variability of the scores of the several individuals composing the group. A method for finding this relation has been described by one of the writers (Otis).¹⁰ By this method a value may be obtained from a measure of the variability of differences between pairs of scores by the same individuals and a corresponding measure of the variability of the scores of the several individuals, which corresponds exactly to the Pearson coefficient of correlation between the first scores of the group of individuals and their second scores. (The second scores may be by the same scale or a similar one.) The method may be illustrated as follows.

Let us suppose, as the simplest case, that the first scores (A scores) are directly comparable to the second scores (B scores). By this is meant that differences between the A and B scores, due to practice effect or to differences in difficulty between scales, etc., may be considered negligible. Then any difference, positive or negative, between the two scores of a single individual may be considered as only chance variation. Under these circumstances presumably the variability of the B scores will be practically the same as the variability of the A scores, so that these variabilities may be considered equal for the purpose of the illustration.

Pearson has given a formula for correlation:¹¹

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_v^2}{2\sigma_x\sigma_y} \text{ when } v = x - y$$

¹⁰ See Reference 4.

¹¹ See Reference 12.

Expressed in terms of our variables, A and B , this formula becomes

$$r_{AB} = \frac{\sigma_A^2 + \sigma_B^2 - \sigma_{(A-B)}^2}{2\sigma_A\sigma_B}$$

Now if we assume the simplest case where $\sigma_A = \sigma_B$, we have

$$r_{AB} = \frac{2\sigma_A^2 - \sigma_{(A-B)}^2}{2\sigma_A^2} \text{ or } \frac{2\sigma_B^2 - \sigma_{(A-B)}^2}{2\sigma_B^2}$$

or

$$r_{AB} = 1 - \frac{1}{2} \frac{\sigma_{(A-B)}^2}{\sigma_A^2} \text{ or } 1 - \frac{1}{2} \frac{\sigma_{(A-B)}^2}{\sigma_B^2}$$

The coefficient of correlation between A and B is seen, therefore, to be the expression of a certain relation between the variability of the differences between A and B and the variability of the values of A or B themselves. That is, the coefficient of correlation increases with decrease in the ratio of these measures of variability.

If, however, the A and B scores are not directly comparable, either because of practice effect or differences in difficulty between scales A and B , then in order to find the probable error of a score by the scale, it becomes necessary, as has been explained above, to transmute the B values into terms of the A scale, or vice versa, by means of the line of relation between A and B .

In all cases of rectilinear relationship, the most probable position of the true line of relation between two variables, x and y , is the line which passes through the point in the plot representing the mean of the x values and the mean of the y values (assumed to be the origin of the plot) and which has a slope such that the tangent of the angle it makes with the horizontal axis is equal to the quotient obtained by dividing the standard deviation of the y values by the standard deviation of the x values. That is, the equation of the line which most probably expresses the true relationship between x and y is $y = \frac{\sigma_y}{\sigma_x}x$.¹² Now, if we measure the vertical deviations of the points in the plot from this line, $y = \frac{\sigma_y}{\sigma_x}x$, and call any one of these deviations d , then each value of d is the difference between a value of y and the corresponding

¹² The proof of this statement is quite involved. It is given by Otis in an article which is not yet published.

value of x after this has been transmuted into terms of y . And it may be shown that the coefficient of correlation is given by the formula,

$$r_{xy} = 1 - \frac{1}{2} \frac{\sigma_d^2}{\sigma_y^2}$$

In the previous description of this formula¹³ it was called the "Deviation Formula." It has since been considered preferable, however, to call it the "Difference Formula." This formula is identical with the Pearson product-moment formula, as is demonstrated in Appendix II.

In order to express this formula in terms of our values, A and B , we must give a notation to the B scores when rendered into terms of the A scale by means of the line of relation whose equation is $A = \frac{\sigma_A}{\sigma_B} B$. Let us call these transmuted B values, B_A values.

The difference formula then becomes

$$r_{AB} = 1 - \frac{1}{2} \frac{\sigma^2(A - B_A)}{\sigma_A^2} \dots \dots \dots (a)$$

This formula, then, expresses a certain relation between the variability of the difference between A and B values (when B values have been transmuted into terms of the A scale) and the variability of the A values. Conversely, of course, if the A values are converted into terms of the B scale, the difference formula takes the form:

$$r_{AB} = 1 - \frac{1}{2} \frac{\sigma^2(B - AB)}{\sigma_B^2} \dots \dots \dots (b)$$

Two modifications of the above difference formula are

$$r_{AB} = 1 - \frac{1}{2} \left(\frac{M. D. (A - B_A)}{M. D. A} \right)^2 \dots \dots \dots (c)$$

and

$$r_{AB} = 1 - \frac{1}{2} \left(\frac{A. D. (A - B_A)}{A. D. A} \right)^2 \dots \dots \dots (d)$$

in which $M. D. (A - B_A)$ and $M. D. A$ are the median deviations of the distribution of $(A - B_A)$ and A values, respectively, in terms of the A scale; and $A. D. (A - B_A)$ and $A. D. A$ are respectively the average deviations of these distributions. As has been shown, the probable error of a score (in terms of the A scale) is expressed by the equation:

$$P. E. A = \frac{M. D. (A - B_A)}{\sqrt{2}}$$

¹³ See Reference 4.

Hence if we wish to use the value of the probable error of a score directly in a correlation formula, we may take

$$r_{AB} = 1 - \left(\frac{P. E._A}{M. D._A} \right)^2 \dots \dots \dots (e)$$

or

$$r_{AB} = 1 - \left(\frac{P. E._B}{M. D._B} \right)^2 \dots \dots \dots (f)$$

Stated in words, the reliability coefficient of correlation between the duplicate scores of the individuals of a group is equal to 1 minus the square of the ratio of the probable error of a single score to the median variability of the scores, when these values are in the same terms. The advantage of enabling a correlation to be obtained directly from a value of the probable error is peculiar to the difference formula for correlation described herein. Any other corresponding measures of variability may be used, such as interquartile ranges.

To illustrate the use of form (c) of the difference formula in the present instance, the median deviation of the distribution of B values ($M. D._B$) was determined in the same way as the median deviation of the distribution of values ($A_B - B$), as shown in Figure 2. This value, $M. D._B$, was found to be approximately 20.1 months. Our value of $M. D._{(AB-B)}$ was 10.8. Therefore, solving the formula,

$$r_{AB} = 1 - \frac{1}{2} \left(\frac{M. D._{(AB-B)}}{M. D._B} \right)^2$$

we have

$$r_{AB} = 1 - \frac{1}{2} \left(\frac{10.8}{20.1} \right)^2$$

$$r_{AB} = 0.855$$

This is the coefficient of correlation between scores by the two halves of the scale. The corresponding value of r , found by the Pearson product-moment method, was 0.850.

The error of a coefficient of correlation found by this modification of the difference formula will depend, of course, among other things, upon the care with which the medians involved are determined, as shown in Figure 2. By the use of methods of approximation it has been found possible to obtain coefficients by the difference method about as quickly as by the method of unlike

signs, and the coefficients were believed to be much more reliable. By the use of formula (a) and (b), the reliability of coefficients is equal to that with the Pearson formula in cases of rectilinear relationship. When the relation is curvilinear the difference formula gives in certain instances a more accurate coefficient than the Pearson formula, since it corrects the coefficient, in certain instances, for skewness of the distributions in somewhat the same way as does the correlation ratio.¹⁴ Spearman's criticism of the Pearson formula, that too great weight is given to extreme values is obviated by the use of formulas (c) or (d). These facts will partly account for slight differences between results by the two methods.

THE COMMON FAILURE TO GIVE MEASURES OF RELIABILITY

Improper methods of expressing reliability constitute perhaps a less prevalent fault on the part of most scale makers than the failure to give any measure whatever of the reliability of single scores. To the knowledge of the writers no measure of the probable error of a mental age by any of the several varieties of the Binet Scale has been determined before. The same is true with regard to nearly all of the pedagogical scales.

Ayres says with regard to the reliability of his spelling scale:¹⁵ "By means of these standards children of the different grades of any locality may be tested as to their spelling attainments and the results compared with those which are found in the corresponding grades in city systems in general. . . . With less reliability the attainment of a small number of grades or of one grade may be tested. With still less reliability the attainment of a single child may be compared with these results." No measure of the P. E. of a score is given.

Thorndike says with regard to his "Scale for Measuring Ability in Reading":¹⁶ " . . . for exact measures of individual and so for small differences amongst them, a scale with more paragraphs and questions of each degree of difficulty is required. Until such scales, Beta, Gamma, Delta, and so on, are constructed, however, the present scale is the best to use." No measure of the reliability of a score is mentioned.

¹⁴ See Reference 11, p. 204.

¹⁵ Reference 2, p. 40.

¹⁶ Reference 9, 1915, p. 458.

No mention could be found of the probable error of a score in Starch's account of his grammar and arithmetic scales. He says merely, with regard to the grammar test:¹⁷ "In spite of these limitations, which fundamentally are not of a very serious character, these tests provide measures of grammatical knowledge which are quite accurate and far more accurate than ordinary methods of testing and marking."¹⁸

Starch also says with regard to his reading test:¹⁹ "It is recommended that the vocabulary test on page 37 be given in conjunction with the test for speed and comprehension. These three tests together will serve as a very adequate measure of a pupil's reading ability."

Perhaps no better comment can be made on the general attitude of makers and users of scales with regard to the reliability

¹⁷ Reference 6, p. 626.

¹⁸ It should be noted that no matter whether the general difficulties of the grammar tests for ten thousand children were found, the scale might still be far from reliable, due to the ambiguity of the questions and to faulty mathematics of standardization. Grammatical Scale A was submitted to three college professors of English who disagreed on the marking of sixteen different tests. Criticism could be made with regard to the theory of the construction of the scale. Such considerations, however, prove little. The proof of the pudding is in the eating and the proof of the reliability of a scale is in the P. E. of a score—at least to the extent of our saying that if the P. E. is large, the scale cannot be considered reliable. It may be invalid, as a measure of that which it purports to measure, even if the P. E. is small.

To get a rough idea of the amount of the probable error of the Starch grammar tests, Scales A and B were submitted to 25 children of grades IV to VIII of a small school. The papers were graded as accurately as possible with "keys" furnished by Starch. By comparing Scales A and B, the probable error of a score was found to be theoretically 1.2⁺ steps, which, if accurate, would mean that in half the cases, a score is in error by an amount slightly greater than the difference between the abilities standard for the seventh grade and for the junior class in high school, according to standards given by Starch. To be sure, no great reliance can be placed on our figures, derived as they were from so few individuals; and it is greatly to be hoped that Starch will be able to show that they are too high. Until such time, however, the presumption is that the tests are far from being "quite accurate."

The Pearson coefficient of correlation between the scores by the two scales was found to be only 0.13, while the coefficient of correlation between the numbers of individual tests passed was 0.47. (These values are rendered comparable since the individuals were the same in both cases.) Here is evidence for the further presumption that the very simple method of counting the individual tests passed gives a more reliable score than the method advocated by Starch. No doubt the reliability of the scales could be made greater by the adoption of a better method of scoring.

¹⁹ Reference 7, p. 21.

of the scores obtained by them than the first paragraph of chapter II of Starch's book. We will quote the paragraph with the single change of substituting the word "scales" for the word "marks." We believe it will depict with startling significance a situation which we are fast approaching. The paragraph as thus altered reads as follows:

"Scales are the universal measures of school work. Numerous and momentous problems in the operation of a school depend upon them, such as promotion, retardation, elimination, honors, eligibility for contests and societies, graduation, admission to higher institutions, recommendation for future positions and the like. Until a decade ago, no one questioned either the validity or the fairness of these measurements. It was tacitly assumed that scales were almost absolutely correct, or very nearly so, a fact attested by the surprisingly common practice of marking to a fractional part of a point, even on a hundred percentage basis."

The necessity for a measure of the reliability of the score of any test was discussed in 1912 by Otis and Davidson.²⁰ Obviously, no child's score in any pedagogical test can be safely used as a basis for promotion or grading, nor a test of intelligence for classification or commitment to an institution, when there are no means of taking into account the amount by which the score may be in error. Every maker of a scale for mental testing, when giving an account of the standardization of such scale, should therefore give the P. E. or some similar measure of the reliability of a score by the scale. Those who are using scales and drawing conclusions from the results are cautioned to bear in mind that every score is but an approximation, often only a very rough approximation, of the true measure of the ability in question. A mental age of fourteen years of the Binet Scale, when used with adults, is in reality a mental age of fourteen years plus or minus six months or more in half the cases. These measures of reliability are themselves subject to refinement, and at best they show only the amount of imperfection of a scale as a measure of that ability which it really measures, not necessarily the deficiency of the scale as a measure of the ability which it purports to measure. The latter deficiency may be greater.

²⁰ Reference 5.

APPENDIX

I. Proof of the Formula:

P. E. (Single measure) = $\sqrt{1/2} \times$ (Median difference between measures).

Let us suppose that we have n measures made upon a given magnitude. Let us denote these measures by m_1, m_2, m_3 , etc., and assume them to be measured from the mean, M , of all the measures.

Then

$$(m_1 - m_2)^2 = m_1^2 - 2m_1m_2 + m_2^2$$

$$(m_1 - m_3)^2 = m_1^2 - 2m_1m_3 + m_3^2$$

$$(m_2 - m_3)^2 = m_2^2 - 2m_2m_3 + m_3^2$$

etc.

etc.

Let us call the summation of the squares of all the differences between m 's, $\Sigma(m-m)^2$. Then

$$\begin{aligned} \Sigma(m-m)^2 = (n-1) (m_1^2 + m_2^2 + \dots + m_n^2) &- m_1(m_2 + m_3 + \dots + m_n) \\ &- m_2(m_1 + m_3 + \dots + m_n) \\ &- \text{etc.} \end{aligned}$$

Here the first term of the right member is the sum of the $n-1$ differences involving m_1 , and the like number involving $m_2, m_3, \dots, m_n$. The remaining terms constitute the summation of the middle terms, $-2m_1m_2$, etc. Now to assist in simplifying this equation we will add $(m_1^2 + m_2^2 + \dots + m_n^2)$ to the first term and then immediately subtract these values by making the second term $-m_1(m_1 + m_2 + m_3 + \dots + m_n)$, etc., as shown below. Then

$$\begin{aligned} \Sigma(m-m)^2 = n(m_1^2 + m_2^2 + m_3^2 + \dots + m_n^2) &- m_1(m_1 + m_2 + m_3 + \dots + m_n) \\ &- m_2(m_1 + m_2 + m_3 + \dots + m_n) \\ &- \text{etc.} \end{aligned}$$

But since $m_1, m_2, \dots, m_n$ are measured from their mean, their sum is zero. Hence, in the second member all terms after the first vanish. By the definition of standard deviation,

$$\sqrt{\frac{m_1^2 + m_2^2 + m_3^2 + \dots + m_n^2}{n}} = \sigma_m$$

or

whence

$$\frac{m_1^2 + m_2^2 + m_3^2 + \dots + m_n^2}{n} = n\sigma_m^2$$

$$\Sigma(m-m)^2 = n(n\sigma_m^2)$$

The first member of this equation expresses the sum of all the differences between measures. Since there are n measures, there are as many differences as there are combinations of n things taken two at a time, or $\frac{n(n-1)}{2}$ differences. Dividing by the number of differences,

$$\frac{\Sigma(m-m)^2}{n(n-1)/2} = 2 \left(\frac{n}{n-1} \right) \sigma_m^2$$

Now the first member of this equation is the sum of the squares of a series of quantities, called $(m-m)$, divided by the number of such quantities. By definition, the square root of this quotient is the standard deviation of the quantities, which would be designated: $\sigma_{(m-m)}$. That is,

$$\frac{\Sigma(m-m)^2}{n(n-1)/2} = \sigma_{(m-m)}^2$$

Putting this value in place of the left member in the equation above, we have

$$\begin{aligned}\sigma^2_{(m-m)} &= 2 \left(\frac{n}{n-1} \right) \sigma^2_m \text{ from which} \\ \sigma^2_m &= \frac{1}{2} \left(\frac{n-1}{n} \right) \sigma^2_{(m-m)} \text{ and} \\ \sigma_m &= \sqrt{\frac{1}{2}} \sqrt{\frac{n-1}{n}} \sigma_{(m-m)}\end{aligned}$$

Now since we have considered the measures as all having been made from the mean, we may term the values of m , observed errors, and give σ_m the notation, $\sigma_{e, obs.}$. Then, in this notation,

$$\sigma_{e, obs.} = \sqrt{\frac{1}{2}} \sqrt{\frac{n-1}{n}} \sigma_{(m-m)}$$

Now if we had an infinite number of cases, the value of $\sqrt{\frac{n-1}{n}}$ would be practically unity, and denoting the expression, "when n equals infinity" by the subscript $n = \infty$, $\sigma_{e|n=\infty} = \sigma_{e, true}$, the true value of the standard error, so that $\sigma_{e, true} = \sqrt{\frac{1}{2}} \sigma_{(m-m)} [n=\infty]$. Now the obtained value of $\sigma_{(m-m)}$ is, of course, not always equal to exactly $\sigma_{(m-m)} [n=\infty]$. It has a probable error of $0.6745 \frac{\sigma_{(m-m)}}{\sqrt{2n}}$. But it is the best measure we have of the true standard deviation of the differences between measures, so we may say that

$$\sigma_{e, true} = \sqrt{\frac{1}{2}} \sigma_{(m-m)}$$

Then since the distribution of errors and of differences between scores are approximately normal, we may assume that the median deviation of the distribution of errors, which we may call the *probable error*, is equal to 0.6745 times the standard deviation of the distribution of errors, and similarly that the median deviation of the distribution of differences ($m-m$) is equal to 0.6745 times the standard deviation of the distribution of these differences. That is,

$$\begin{aligned}\text{P. E. (Single measure)} &= M. D. e, true = 0.6745 \sigma_{e, true} \\ \text{and} \quad M. D. (m-m) &= 0.6745 \sigma_{(m-m)}\end{aligned}$$

Hence we may assume that since

$$\sigma_{e, true} = \sqrt{\frac{1}{2}} \sigma_{(m-m)}$$

$$\text{Therefore, P. E. (Single measure)} = \sqrt{\frac{1}{2}} M. D. (m-m)$$

That is, the probable error of a single measure equals the-root-of-one-half times the median of the differences between the pairs of scores of single individuals. In the case of this study the single measures are scores (mental ages) by one or the other half of the Binet Scale and the differences between scores are found after inequalities of difficulty between the two halves of the scale have been compensated for.

It may be of interest to note that, following from the above proof as a corollary,

$$\begin{aligned}\sigma_{e, observed} &= \sqrt{\frac{n-1}{n}} \sigma_{e, true}, \text{ or conversely,} \\ \sigma_{e, true} &= \sqrt{\frac{n}{n-1}} \sigma_{e, observed} \text{ by which formula the most}\end{aligned}$$

probable value of the true standard error may be derived from an observed standard error taken from any given number of measures.

II. The Derivation of the Difference Formula and its Relation to the Pearson Product-Moment Formula

As defined in the text, for any values of x and y , $d = y - \frac{\sigma_y}{\sigma_x} x$. For convenience, let $m =$ the tangent of the angle of the line of relation with the horizontal axis.

Then

$$m = \frac{\sigma_y}{\sigma_x}$$

Therefore,

$$d = y - mx$$

Squaring,

$$d^2 = y^2 - 2ymx + m^2x^2$$

Summing,

$$\Sigma d^2 = \Sigma y^2 - 2m \Sigma xy + m^2 \Sigma x^2$$

Transposing,

$$2m \Sigma xy = \Sigma y^2 + m^2 \Sigma x^2 - \Sigma d^2$$

Dividing by n ,

$$2m \frac{\Sigma xy}{n} = \frac{\Sigma y^2}{n} + \frac{m^2 \Sigma x^2}{n} - \frac{\Sigma d^2}{n}$$

Or

$$2m \frac{\Sigma xy}{n} = \sigma_y^2 + m^2 \sigma_x^2 - \sigma_d^2$$

Dividing by

$$\frac{\Sigma xy}{n \sigma_x \sigma_y} = \frac{\sigma_y^2 + m^2 \sigma_x^2 - \sigma_d^2}{2m \sigma_x \sigma_y}$$

$$2m \sigma_x \sigma_y$$

Now since

$$m^2 \sigma_x^2 = \frac{\sigma_y^2}{\sigma_x^2} \sigma_x^2 = \sigma_y^2$$

and

$$m \sigma_x = \frac{\sigma_y}{\sigma_x} \sigma_x = \sigma_y$$

therefore

$$\frac{\Sigma xy}{n \sigma_x \sigma_y} = \frac{\sigma_y^2 + \sigma_y^2 - \sigma_d^2}{2\sigma_y \sigma_y}$$

or

$$\frac{\Sigma xy}{n \sigma_x \sigma_y} = \frac{2\sigma_y^2 - \sigma_d^2}{2\sigma_y^2}$$

or

$$\frac{\Sigma xy}{n \sigma_x \sigma_y} = r_{xy} = 1 - \frac{1}{2} \frac{\sigma_d^2}{\sigma_y^2}$$

The left member is the Pearson product-moment formula and the right member is the difference formula.

In a normal distribution,

The median deviation ($M. D.$) of the distribution = 0.6745 times the standard deviation (σ) of the distribution.

The average deviation ($A. D.$) of the distribution = 0.8453 times the standard deviation of the distribution.

The interquartile range ($I. Q. R.$) of the distribution = 1.349 times the standard deviation of the distribution.

And in cases of all distributions which are approximately normal, the same equalities are approximately true. Therefore,

$$\begin{aligned} 1 - \frac{1}{2} \frac{(M.D.)^2_d}{(M.D.)^2_y} &= 1 - \frac{1}{2} \left(\frac{\text{approximately } 0.6745\sigma_d^2}{\text{approximately } 0.6745\sigma_y^2} \right) \\ &= \text{approximately } 1 - \frac{1}{2} \frac{\sigma_d^2}{\sigma_y^2} \end{aligned}$$

Similarly,

$$1 - \frac{1}{2} \frac{(A.D.)^2_d}{(A.D.)^2_y} = \text{approximately } 1 - \frac{1}{2} \frac{\sigma^2_d}{\sigma^2_y}, \text{ and}$$

$$1 - \frac{1}{2} \frac{(I.Q.R.)^2_d}{(I.Q.R.)^2_y} = \text{approximately } 1 - \frac{1}{2} \frac{\sigma^2_d}{\sigma^2_y}$$

It may be interesting to note the relation between the difference formula and the Pearson formula:

$$r_{xy} = \frac{\sigma^2_x + \sigma^2_y - \sigma^2_v}{2 \sigma_x \sigma_y} \text{ in which } v = y - x.$$

If instead of measuring the deviations (d) of the points of the plot from the line of relation, $y = \frac{\sigma_y}{\sigma_x}x$, these deviations are measured from the line which passes through the origin making an angle of 45° with the horizontal axis, the equation of which is $y = x$, then m , the tangent of the angle of the line, $= 1$, and $d = y - x$. Beginning with this equation, the derivation of the Pearson formula is identical with the first eight steps of the derivation of the difference formula given above when all m 's are omitted because in this case $m = 1$.

In a sense, therefore, this Pearson formula may be considered as a special case of the more general difference formula given above.

III. Proof of the Theorem that the Probable Error of a Score by the Whole Scale $= \sqrt{1/2}$ times the Probable Error of a Score by Half the Scale

We have seen from Appendix II that the reliability coefficient of correlation between the duplicate scores of single individuals is

$$r = 1 - \frac{1}{2} \left(\frac{\sigma(x-y)}{\sigma_x} \right)^2$$

(assuming x and y to be in the same terms).

We have seen also, from Appendix I, that

$$\sigma_e = \frac{\sigma(x-y)}{\sqrt{2}}$$

or

$$\sigma(x-y) = \sqrt{2} \sigma_e$$

in which σ_e is the standard deviation of true errors, that is, of differences between the several measures and the corresponding true measures.

Therefore

$$r = 1 - \frac{1}{2} \left(\frac{\sqrt{2} \sigma_e}{\sigma_x} \right)^2$$

or

$$r = 1 - \left(\frac{\sigma_e}{\sigma_x} \right)^2 \dots \dots \dots (1)$$

Now by Spearman's theorem that

$$r_{x(p), x(q)} = \frac{pr_{x(q), x(q)}}{q + (p-q)r_{x(q), x(q)}}$$

when p is one number of measures and q another (in this case $p = 2$ and $q = 1$),

$$r_2 = \frac{2r_1}{1 + r_1}$$

in which r_2 = the reliability coefficient of the whole scale and r_1 = the reliability coefficient of half the scale.²¹

$$\begin{aligned}\text{From (1)} \quad \sigma^2_x r &= \sigma^2_x - \sigma^2_e \\ \sigma^2_e &= \sigma^2_x - \sigma^2_x r \\ \sigma^2_{e_1} &= \sigma^2_{x_1}(1 - r_1) \text{ half scale} \dots \dots \dots (3)\end{aligned}$$

$$\text{or} \quad \sigma^2_{e_2} = \sigma^2_{x_2}(1 - r_2) \text{ (whole scale)} \dots \dots \dots (4)$$

(Hereafter, subscripts 1 and 2 refer respectively to the half scale and whole scale.)

Now, since the obtained score by the half scale has been multiplied by 2 (see page 123), each score by the whole scale is the average of the scores by the two half scales; that is, $x_2 = \frac{x_1 + x'_1}{2}$ in which x_1 is the score by the first half of the scale and x'_1 the score by the second half.

$$\begin{aligned}\text{Whence} \quad x^2_2 &= \frac{x^2_1 + 2x_1x'_1 + (x'_1)^2}{4} \\ \Sigma x^2_2 &= \frac{\Sigma x^2_1 + 2 \Sigma x_1x'_1 + \Sigma (x'_1)^2}{4}\end{aligned}$$

Since $\frac{\Sigma x_1x'_1}{n \sigma_{x_1} \sigma_{x'_1}} = r_1$ and $\sigma_{x'_1} = \sigma_{x_1}$ because the variabilities of the two half scales are presumably the same, therefore

$$\begin{aligned}\sigma^2_{x_2} &= \frac{\sigma^2_{x_1} + 2r_1\sigma^2_{x_1} + \sigma^2_{x_1}}{4} \\ \sigma^2_{x_2} &= \frac{\sigma^2_{x_1}(1 + r_1)}{2} \dots \dots \dots (5)\end{aligned}$$

From (4), (5), and (2)

$$\sigma^2_{e_2} = \frac{\sigma^2_{x_1}(1 + r_1)}{2} \left(1 - \frac{2r_1}{1 + r_1} \right)$$

$$\text{Simplifying,} \quad 2\sigma^2_{e_2} = \sigma^2_{x_1}(1 - r_1) \dots \dots \dots (6)$$

$$\text{But from (3),} \quad \sigma^2_{e_1} = \sigma^2_{x_1}(1 - r_1)$$

$$\begin{aligned}\text{Therefore} \quad 2\sigma^2_{e_2} &= \sigma^2_{e_1} \\ \text{and} \quad \sigma_{e_2} &= \sqrt{\frac{1}{2}} \sigma_{e_1} \dots \dots \dots (7)\end{aligned}$$

Now since the distribution of errors is presumably normal, we may say that

$$\begin{aligned}\text{and} \quad \text{P. E. (whole scale)} &= 0.6745 \sigma_{e_2} \\ \text{P. E. (half scale)} &= 0.6745 \sigma_{e_1} \\ \text{whence} \quad \text{P. E. (whole scale)} &= \sqrt{\frac{1}{2}} \text{ P. E. (half scale)} \\ &\text{Q. E. D.}^{22}\end{aligned}$$

²¹ Spearman, C. "Correlation calculated from faulty data," *British Journal of Psychology*, 3: 271-95, October, 1910.

²² For suggestions concerning this proof, the writers are indebted to Dr. Truman L. Kelley.

BIBLIOGRAPHY

1. ABELSON, A. R. "The measurement of mental ability of 'backward' children," *British Journal of Psychology*, 4:268-314, December, 1911.
2. AYRES, LEONARD P. *Measurement of ability in spelling*. New York: Russell Sage Foundation. (Educational Monograph E139.)
3. BUCKINGHAM, B. R. *Spelling ability: Its measurement and distribution*. (Teachers College Contributions to Education No. 59.) New York: Teachers College, Columbia University, 1913. 116 pp.
4. OTIS, ARTHUR S. "Reliability of spelling scales, involving a 'deviation formula' for correlation," *School and Society*, 4: 676, 716, 750, 793, October 28-November 18, 1916.
5. OTIS, ARTHUR S. and DAVIDSON, P. E. "Reliability of standard scores in adding ability," *Elementary School Teacher*, 13:91-105, October, 1912.
6. STARCH, DANIEL. "The measurement of ability in English grammar," *Journal of Educational Psychology*, 6:615-626, December, 1915; and "A scale for measuring ability in arithmetic," *Ibid.*, 7:213-222, April, 1916.
7. STARCH, DANIEL. *Educational measurements*. New York: Macmillan Company, 1916. 202 pp.
8. THORNDIKE, EDWARD L. *An introduction to the theory of mental and social measurements*. (Second edition.) New York: Teachers College, Columbia University 1913. 277 pp.
9. THORNDIKE, EDWARD L. "An improved scale for measuring ability in reading," *Teachers College Record*, 16:445-467, 17:40-67, November, 1915 and January, 1916.
10. KRUEGER, F. und SPEARMAN, C. "Die Korrelation zwischen verschiedenen geistigen Leistungsfähigkeiten," *Zeitschrift für Psychologie*, 44:50-114, 1906.
11. YULE, G. UDNY. *An introduction to the theory of statistics*. London: C. Griffin and Company, 1919. 398 pp.
12. PEARSON, KARL. "On further methods of determining correlations," *Mathematical Contributions to the Theory of Evolution*, XVI, Drapers Co. Research Memoirs, Biometric Series, IV.