

## WHAT IS READING ABILITY?

J. BENSON WYMAN AND MIRIAM WENDLE

Stanford University

Tests of reading ability have been devised—many of them. Their reliability coefficients have been determined and norms have been obtained; but do their reliability coefficients or their norms guarantee that they are good tests of reading ability or tests of reading ability at all? A reliability coefficient is only one criterion of a test. It measures the amount of agreement one would expect between an individual's score on one form of a test and his score on any other comparable form of the same test; but is there anything to show that the so-called reading tests do measure reading ability?

The present studies were undertaken to get at a method by which we could determine whether the so-called reading tests *do* measure reading ability. The questions arose: "What is reading ability? What criterion can be used for measuring it?" The first suggestion was to use teachers' estimates of the reading ability of their pupils as the criterion. But, knowing the fallibility of teachers' estimates, an alternative criterion was also used and the plan adopted to obtain it was the following: Two professors and three graduate students conversant with the tests used were asked to assign values to each test indicating the value of the test as a measure of silent reading ability (*i.e.*, they gave their judgments of the values of the tests as *tests of reading*, whereas for the first criterion the teachers gave their judgments of the reading ability of the pupils). These five values were made independently. The five values for each test were then averaged and this average value was taken as the scoring, or weight, for the test as a reading test. Then this alternative criterion for reading ability consisted of the combination of all the tests weighted according to the above average values. Having then these two criteria for reading ability the procedure was as follows:

*Reliability Coefficients.*—The correlation between a given set of scores in one test and another set similarly obtained on a similar form of the same test was determined. In cases where there were not two comparable forms of a test, the one form was divided into two halves measuring substantially the same thing. These halves were correlated; and then Brown's formula ( $\frac{2r}{1+r}$ )—where  $r$  was the obtained correlation—was applied; and this gave an estimate of the correlation

between two forms of the test. This correlation is the reliability coefficient of the test.

*Correlations.*—(1) Each test was correlated with every other test by means of Pearson's Product-Moment formula and the probable errors of these correlations were determined from

$$P.E_r = \frac{0.6745(1 - r^2)}{\sqrt{n}} \quad \text{where} \quad \begin{cases} r = \text{correlation} \\ n = \text{number of cases} \end{cases}$$

2. The average of the independent estimates of the reading ability of the pupils made by the two teachers was taken as the first criterion of reading ability. (Call this Reading Ability T.) The sum, or the average, of the pupils' scores on each test was taken as the test score. Then each test score was correlated with Reading Ability T.

3. The second criterion for reading ability (call this Reading Ability C) consisted of the combination of all the tests weighted according to their values as measures of reading ability—with the modification that, when a test was to be correlated with it, that test was taken out of the criterion. (Suppose, for example that the Thorndike Reading Alpha test were to be correlated with Reading Ability, then the criterion for reading ability would be a combination of all the weighted tests except Thorndike Alpha.) In order to determine this correlation, the following formula<sup>1</sup> was used:

$$* r_1(\Sigma w_x X) = \frac{\Sigma r_{1x} w_x \sigma_x}{\sqrt{\Sigma (w_x \sigma_x)^2 + 2 \Sigma r_{xy} w_x \sigma_x w_y \sigma_y}}$$

where

$\Sigma r_{1x} w_x \sigma_x$  = the sum of the correlations (each multiplied by its weight and standard deviation) of the type Thorndike Alpha and Monroe, Thorndike Alpha and Completion Beta.

$\Sigma r_{xy} w_x \sigma_x w_y \sigma_y$  = the sum of all the intercorrelations each one multiplied by the weights and standard deviations of both the tests correlated.

4. Spearman considered that if there were any errors in the original scores of the pupils due to chance mistakes, they would not compensate one another but would reduce the correlation. He devised the following formula by the appli-

<sup>1</sup> Kelley, T. L.: *Bulletin* 27, University of Texas, 1916.

cation of which the obtained correlations would be corrected for this, so that the corrected coefficient measures the extent to which the test would correlate with the criterion if the score of the individual were an accurate one both in the test and in the criterion:

$$\text{Corrected coefficient } R = \frac{r}{\sqrt{r_{13}} \sqrt{r_{24}}}$$

where

$r$  = correlation between test and criterion

$r_{13}$  = reliability coefficient of test

$r_{24}$  = reliability coefficient of criterion

$R$  then is the correlation between a true reading score and a true criterion of reading ability.

5. Then that test is more uniquely a reading test and less a test of any other function which shows the highest  $R$ .

*Detailed Procedure.*—Two studies were made. One was with grade VIII B. pupils where reading tests were the main tests, and the criteria were Reading Ability T and Reading Ability C. (We shall refer to this study as "VIII grade reading study.") The other study was with High School pupils, English tests were used and the criterion of English Ability—call it English Ability E—was the Teachers' grades. (This study we shall refer to as "High School English study.")

*Tests Given.*—(1) The following tests were given to 36 pupils in grade VIII just before promotion. Two teachers independently ranked the pupils in the order of their ability to read:

Thorndike's Reading Scale Alpha 2.

Monroe's Silent Reading Test II (forms 2 and 3).

Thorndike's Reading Test B—Visual Vocabulary (series x and y).

Kelley-Trabue Completion Exercise Alpha.

Terman Group Test of Mental Ability (form A).

Seven S Spelling Test (list 13).

Woody-McCall Arithmetic (forms 1 and 2).

Compositions (1) ("What I should like to do next Saturday.")

(2) "The most exciting ride I ever had.")

2. The following tests were given to 94 pupils of the Senior class of a High School:

Briggs' English Form Test (Beta).

Compositions (2).

Thorndike-McCall Reading Scale (form 1).

Kelley-Trabue Completion Exercise (Beta).

Abbott-Trabue tests of Poetic Appreciation (series x and y).

In this study the teachers' grades in English (English Ability E) for the previous semester were then secured as an objective measure against which to gauge these tests as tests of ability in English. The Terman Group Test of Mental Ability had been given, so that the scores on a reliable intelligence test were at hand for comparison. Since it was thought the correlations of the English tests with an arithmetic test might also bring out interesting relations, the scores on the arithmetic exercise in the Terman Group Test were used.

*Reliability Coefficients.*—For Thorndike's Visual Vocabulary, Woody-McCall Arithmetic, Monroe Silent Reading, Compositions and Abbott-Trabue tests one form was correlated with the second form.

For Spelling, Completion, Terman Group, Opposites Beta, Terman Arithmetic, Thorndike-McCall and Briggs two halves were obtained and correlated and then Brown's formula was applied.

For teachers' estimates (in the first study) one teacher's marks were correlated with the other teacher's and then Brown's formula was applied. But, since the Teachers' grades (in the High School English study) represented the marks of only one teacher per individual, it was necessary to estimate their reliability. In a study of teachers' estimates, T. L. Kelley ("Educational Guidance," sec. 4, page 15) found consistently low reliability coefficients. Teachers' gradings are probably somewhat more reliable. So the reliability of Teachers' Grades was estimated to be about 0.45.

Thorndike's Alpha 2 Test consists of passages that are to be read and questions on the passages are to be answered. In "Difficulty 7" there are two paragraphs in the passage with four questions to be answered on the first paragraph and three on the second. In "Difficulty 8" there are two paragraphs with four questions on each. In "Difficulty 8 $\frac{2}{3}$ " there is one paragraph with four questions on it; and in "Difficulty 9" there is one paragraph with five questions. It will be seen then that there are two ways in which the test can be divided into two comparable halves. The one way is to divide the test so that unbroken paragraphs and twelve questions are in either half. The other way is to divide the test so that there is the same number of questions on either side but neither side has unbroken paragraphs. The test was divided into two parts, according to the former method, by splitting it into its paragraphs so that the errors in

the first four questions in Difficulty 7, in the second four in Difficulty 8 and in all four in Difficulty  $8\frac{2}{3}$  formed one part while the rest of the errors formed the other part. These two parts were then correlated; and the reliability coefficient was obtained by applying Brown's formula to this correlation. The test could be divided into two parts, according to the second method, by taking the first half of the sum of the errors in Difficulty 7, and the second half in each of the Difficulties 8,  $8\frac{2}{3}$  and 9 as one part and the remainder of the errors as the other part, and correlating them.

Now, whether an individual can answer Question 2 depends to a certain extent on whether he can answer Question 1, whether he can answer Question 3 depends to a certain extent on whether he can answer Question 2, and so on. This is the same for each paragraph. Let us call this correlation between questions based on the same paragraph  $\rho$  and the correlation between questions based on different paragraphs  $\eta$ . If we then call the value obtained by correlating the two parts of Alpha 2 according to the first method of division  $r_1$  and that obtained by the second method  $r_2$ , we can determine the value of  $\rho$  and  $\eta$  from the following equations:

$$r_1 = \frac{n^2\eta}{n + Mm(m-1)\rho + n(n-m)\eta}$$

where  $m$  = number of terms in a group  
 $M$  = number of groups  
 $n = Mm$

$$r_2 = \frac{n(\rho + \eta)}{1 + (n-1)\rho + n\eta}$$

where  $n$  = number of terms in either half.

By solving these equations we find

$$\rho = 0.179$$

$$\eta = 0.060$$

Then  $\rho$  being greater than  $\eta$  proves that the correlation between answers on a single paragraph is greater than that between answers on different paragraphs. Correlation  $\eta$  is due to a certain intelligence level (a child) acting upon certain independent tasks whereas  $\rho$  is due to this plus a factor (operating much as a chance factor) which aids in answering subsequent questions in a set if the first is answered correctly and which hinders if the first is answered incorrectly. Therefore  $\rho$  is spuriously high, due to correlation between errors, as a measure

of the correlation between independent questions. Since this is so,  $r_2$  is spuriously high as a reliability coefficient, because the two halves correlated in obtaining  $r_2$  are not composed of independent exercises. Accordingly  $r_1$  is the correct value for the reliability coefficient for Alpha 2.

1. The following are the reliability coefficients for each test (Grade VIII Reading study):

Terman Group Test	0.85 ± 0.03
Seven "S" Spelling	0.84 ± 0.03
Teachers' Estimates	0.84 ± 0.03
Thorndike's Visual Vocabulary	0.79 ± 0.04
Monroe Comprehension	0.75 ± 0.05
Woody-McCall Arithmetic	0.70 ± 0.06
Monroe Rate . . .	0.67 ± 0.06
Thorndike Alpha 2 . . .	0.53 ± 0.08
Kelley-Traube Completion	0.50 ± 0.08
Composition	0.25 ± 0.10

These reliability coefficients are measures of reliability based on the same pupils. So any difference in them cannot be charged to differences in range of talent. Hence, as regards reliability alone, we can place the above tests in the order indicated by the coefficients.

Terman Group, Teachers' Estimates and Spelling are the most reliable. Thorndike's Visual Vocabulary, Monroe Comprehension, Woody-McCall Arithmetic and Monroe Rate are satisfactory; but neither Alpha 2 nor Completion can be regarded as altogether satisfactory. The reliability coefficient for Composition, based on only two compositions, is very low. If compositions are to be used, by applying Brown's formula it can be seen that in order to have a reliability coefficient comparable to the Terman (0.85) it would be necessary to give 17 compositions:

$$\begin{aligned}
 \text{Reliability coefficient} &= \frac{nr}{1 + (n-1)r} \\
 &= 0.85 = \frac{n(0.25)}{1 + (n-1)0.25} \\
 &= n = 17
 \end{aligned}$$

where

$n$  = number of compositions

$r$  = reliability coefficient for 2 compositions = 0.25

2. The following are the reliability coefficients for each test (High School English study):

Terman Group Test	0.94 ± 0.01
Opposites (Beta)...	0.80 ± 0.025
Briggs' Four Test	0.78 ± 0.03
Completion (Beta) . . .	0.78 ± 0.03
Terman Arithmetic . . .	0.74 ± 0.03
Thorndike-McCall Reading	0.63 ± 0.04
Teachers' Grades. . . . .	0.45 (estimated)
Compositions.....	0.43 ± 0.06
Abbott-Trabue.. . . .	0.37 ± 0.06

The low reliability of the Abbott-Trabue Poetic Appreciation test accords with the results described by Abbott and Trabue in the article entitled "A Measure of Ability to Judge Poetry" (*Teachers' College Record, March, 1921*). Therefore it is not possible to measure poetic appreciation by means of this test. The high reliability coefficient of the Terman test agrees with the findings in the first study. The reliability of Teachers' grades, Composition and Trabue is too low to make them satisfactory measures. (See table, page 526.)

(a) No correlations are exceptionally high. As might be expected the correlations of Terman Group with Completion and Terman Group with Opposites are highest, since both Completion and Opposites are used as Intelligence tests in lieu of using Terman Group. The high correlation between the Terman Group test and the Terman Arithmetic is partly due to its being a correlation between Terman and part of itself.

(b) The highest correlations are those between Terman Group and certain English tests rather than between the English tests themselves. It may be that different aspects of English ability may not exist ordinarily in the same individual; but the implication may be that the Terman Group test, because of its greater reliability, is a better measure of English ability than anyone of the English tests.

(c) The correlation between Teachers' grades (English Ability E) and Thorndike-McCall Reading is the highest; Teachers' grades and Terman Group next; Teachers' grade and Completion next, and then

Teachers' grades and Opposites Beta. But none of these correlations is high, indicating

- (a) that these tests do not measure English ability as it is judged by the teachers, or
- (b) that the reliability of teachers' judgments is low. or
- (c) that other factors than English ability or General Intelligence affect English grades.

It seems that the Terman Group test is as indicative of English ability on the criterion of Teachers' grades as the more apparently English tests are.

2. Correlations, in the VIII Grade Reading study, between Reading Ability T and the tests. (The criterion for Reading Ability T was the average of the independent estimates of the pupils' reading ability as made by the two teachers):

Reading Ability T and Terman Group.	+0.77 ± 0.05
Arithmetic.....	+0.68 ± 0.06
Visual Vocabulary.....	+0.62 ± 0.07
Seven S Spelling ..	+0.59 ± 0.07
Completion.....	+0.58 ± 0.07
Composition.....	+0.54 ± 0.08
Alpha 2.....	+0.51 ± 0.08
Monroe Comprehension..	+0.49 ± 0.085
Monroe Rate.....	+0.10 ± 0.11
Age.....	-0.63 ± 0.07

These correlations show that what teachers call "reading ability" correlates more highly with what the Terman test measures than with what the so-called reading tests measure. The rate of reading, as measured by the Monroe tests, shows very little correlation with the teachers' estimates of reading ability. Age within a grade has, as would be expected, a negative correlation with reading ability.



Correlations.—(1) Correlations in the High School English study are:

	English Ability E	Completion	Abbott- Tarbue	Opposites Beta	Composition (Needelson)	Composition (Hillegas)	Briggs' Form	Thorndike- McCall	Terman Arithmetic
Completion . . . . .	0.39 ± 0.07								
Abbott-Tarbue . . . . .	0.31 ± 0.06	0.48 ± 0.06							
Opposites Beta . . . . .	0.35 ± 0.06	0.38 ± 0.06	0.58 ± 0.05						
Composition (Hudelson) . . . . .	0.03 ± 0.07	0.28 ± 0.06	0.42 ± 0.07	0.39 ± 0.06					
Composition (Hillegas) . . . . .	0.11 ± 0.07	0.32 ± 0.06	0.28 ± 0.06	0.29 ± 0.06	0.43 ± 0.06				
Briggs' Form . . . . .	0.17 ± 0.07	0.42 ± 0.06	0.19 ± 0.07	0.33 ± 0.06	0.30 ± 0.06	0.14 ± 0.07			
Thorndike-McCall . . . . .	0.49 ± 0.05	0.51 ± 0.07	0.54 ± 0.05	0.56 ± 0.05	0.52 ± 0.05	0.29 ± 0.06	0.32 ± 0.06		
Terman Arithmetic . . . . .	0.16 ± 0.07	0.32 ± 0.07	0.15 ± 0.07	0.36 ± 0.06	0.27 ± 0.07	0.14 ± 0.07	0.04 ± 0.07	0.35 ± 0.06	
Terman Group. . . . .	0.42 ± 0.06	0.61 ± 0.05	0.49 ± 0.06	0.64 ± 0.04	0.38 ± 0.06	0.36 ± 0.06	0.38 ± 0.06	0.56 ± 0.05	0.61 ± 0.06

Certain tendencies seem evident here.

3. Intercorrelations between tests (in the Grade VIII Reading study):

	Terman Group	Visual vocab- ulary	Monroe compre- hension	Com- pletion	Thorn- dike Alpha 2	Compo- sition	Seven S spell- ing	Woody- McCall Arith- metic	
Visual Vocabulary	0 69								
Monroe Compre- hension . . . .	0 65	0 44							
Kelley-Trabue com- pletion . . . .	0 64	0 62	0 30						
Thorndike Alpha 2	0 58	0 56	0 20	0 54					
Composition	0 55	0 45	0 20	0 32	0 57				
Seven S Spelling	0 53	0 55	0 54	0 21	0 17	0 38			
Woody-McCall Arithmetic	0 53	0 39	0 45	0 50	0 24	0 40	0 42		
Monroe Rate	0 26	0 23	0 62	-0.10	0 07	0 01	0 80	-0 15	

4. Correlations, in the VIII Grade Reading study, between Reading Ability C and the tests:

*Weighting of Tests.*—The following are the average values or weights determined as described previously:

Thorndike Alpha 2 . . . .	10
Kelley-Trabue Completion	8
Terman Group Test	7
Visual Vocabulary . . . .	6
Woody-McCall Arithmetic	5
Monroe Comprehension	5
Composition . . . . .	4
Seven S Spelling . . . . .	3
Monroe Rate . . . . .	2

#### *Correlations*

Reading Ability C and Terman Group . . . .	0.85
Visual Vocabulary	0.76
Completion	0 64
Alpha 2 . . . . .	0 58
Composition . . . . .	0.57
Monroe Comprehension	0.53
Arithmetic . . . . .	0.52
Spelling . . . . .	0.49
Monroe Rate . . . . .	0.16

Here the highest correlation is with the Terman Group test; and, as the correlation is a high one, we must conclude they measure very much the same thing. The question arises, is our criterion for reading

ability really a measure of general intelligence, or does the Terman Group test measure reading ability?

Comparing the correlations 2 and these correlations, we see that, according to either criterion, the Terman Group test ranks highest as a measure of reading ability, just as the Monroe Rate of Reading test shows least relationship. The values for the other tests vary in the two lists. It would seem that the second of the two criteria (Reading Ability C) was the more reliable measure of reading ability, for the teachers' estimates of the reading ability of their pupils are more likely to be tempered by their knowledge of the general intelligence of the individuals.

The best test then for reading ability, as far as our data are concerned, is the Terman. Visual Vocabulary and then Completion and Thorndike Alpha 2 are the next. Rate of Reading cannot be considered a test of reading at all in so far as our criteria measure reading ability.

#### 5. Corrected Coefficients of Correlation:

(a) English Ability E and Thorndike McCall Reading....		0.92
Abbott-Trabue Poetic Appreciation .		0.76
Kelley-Trabue Completion Beta....		0.67
Terman Group Test of Mental Ability		0.65
Opposites Beta . . . . .		0.59
Briggs' Form . . . . .		0.29
Terman Arithmetic . . . . .		0.28
Composition (2)		0.26
Composition (1) . . . . .		0.07
(b) Reading Ability T and Composition..		1.29 ± 0.25
Terman Group . . . . .		0.98 ± 0.05
Completion . . . . .		0.98 ± 0.11
Arithmetic . . . . .		0.96 ± 0.07
Visual Vocabulary . . . . .		0.83 ± 0.08
Alpha 2... . . . .		0.83 ± 0.12
Spelling . . . . .		0.77 ± 0.09
Comprehension... . . . .		0.67 ± 0.11
Rate . . . . .		0.15 ± 0.17
(c) Reading Ability C and Composition . . . . .		1.14 ± 0.21
Terman Group.. . . .		0.92 ± 0.03
Completion . . . . .		0.90 ± 0.08
Visual Vocabulary.. . . .		0.85 ± 0.05
Alpha 2... . . . .		0.80 ± 0.09
Arithmetic . . . . .		0.62 ± 0.09
Comprehension . . . . .		0.61 ± 0.09
Spelling... . . . .		0.53 ± 0.09
Rate... . . . .		0.20 ± 0.14

For this group of correlations (c) where the criterion was Reading Ability C, the probable errors given are the values when the reliability coefficient of Reading Ability is assumed to be 1. Were it less than 1, the corrected coefficients would be  $\frac{1}{\sqrt{r_{24}}}$  times greater and the probable errors would be greater.

In determining the probable errors for these corrected coefficients the following formula<sup>1</sup> was used:

$$P.E. = 0.6745 \sqrt{\frac{R^2}{N} \left[ \frac{(1-r^2)^2}{r^2} + \frac{(1-r_{13})^2}{4r_{13}^2} + \frac{(1-r_{24})^2}{4r_{24}^2} + \frac{(1-r_{13})(2-2r^2+r_{13}-r_{13}^2)}{2r_{13}} - \frac{(1-r_{24})(2-2r^2+r_{24}-r_{24}^2)}{2r_{24}} + \frac{r^2(1-r_{13})(1-r_{24})}{r_{13}r_{24}} \right]}$$

Spearman found some corrected coefficients greater than unity. He says, "At most, the corrected coefficient is only the true coefficient plus the error due to testing a limited sample—the general magnitude of such an error is indicated by the so-called probable error; and though a true coefficient cannot exceed unity there is no reason why a coefficient plus an error should not do so. In such a case the coefficient must be taken as 1—this being its most probable value."

The only corrected coefficients that would support Spearman's contention that these coefficients are 1 would be those near 1.00 and having small probable errors.

Suppose a corrected correlation  $1 + a$ . If its probable error be equal to, or less than,  $\frac{a}{3}$  then the fallacy of Spearman's contention (that the most probable value of the coefficient is 1.00) is very evident. It proves that his hypotheses (lack of correlation between errors) are unsound. Suppose, for example, that we have a population of 3600, and the corrected correlation between Reading Ability and Composition is  $1.29 \pm 0.025$ . Spearman would say the true correlation was 1.00—without any regard to the probable error value. This correlation (1.00) is about as unreasonable a value as could be chosen, for the chances that the correlation 1.29 would ever be 1.00 are, in the light of its probable error, infinitely remote.

<sup>1</sup> This formula for the P.E. of a coefficient of correlation corrected for attenuation was derived by Dr. Truman L. Kelley, but has not hitherto appeared in print.

Had Spearman known the probable errors of his corrected coefficients, he would have seen the absurdity of claiming that coefficients well above 1.00 supported his argument that all corrected coefficients tended toward 1.00.

#### GENERAL CONCLUSIONS REGARDING THE TESTS

1. Certain of the English tests are too unreliable to be worth much in the classification of pupils (*e.g.*, Abbott-Trabue Poetry Appreciation test). Opposites Beta is more reliable, Briggs' Form test and Completion Beta are slightly less satisfactory and the Thorndike McCall still less. None of the English tests has as high reliability as the Terman Group test.

2. As far as raw correlations with English ability are concerned, using Teachers' grades in English as the criterion, none of the correlations for any of the tests is marked, the highest being 0.49.

3. The arithmetic test was included in the battery in the second study to see whether the criterion of English ability and the treatment would result in the Arithmetic test falling into a low position as an English test which would be expected on the *a priori* assumption that English and arithmetic are different functions. Since this corrected coefficient is so low (0.293) it affords objective evidence that arithmetic and English ability constitute two separate capacities.

4. From the point of view of the classification of pupils, it is probable better results can be obtained on the basis of these tests in the second study (except Terman Arithmetic, Briggs' Form and Abbott-Trabue Poetic Appreciation) than can be obtained from Teachers' Grades. Opposites Beta, a 15 minute-test, differentiates more correctly than Teachers' grades and could be used very profitably.

5. If compositions are to be used as measuring ability, the average score on from 15 to 20 compositions must be taken.

6. According to our criteria for reading ability, the Terman Group test of Mental Ability is a better measure of reading ability than any of the other tests used.

7. Of the so-called Reading tests used, the best of them as a test of reading ability is Thorndike's Visual Vocabulary, while Monroe's Rate of Silent Reading test shows almost no correlation with our criteria for reading ability.

8. From these studies can be seen the necessity for having other objective information about a test than its reliability coefficient before the function it measures can be stated.

To these conclusions might be added a warning with regard to the use of Spearman's formula—Care must be taken

- (a) that the halves of the tests used in determining the reliability coefficients be strictly independent and comparable samplings of ability. The tests must be given under similar conditions such as will not lead to spurious correlations.
- (b) that the probable errors be determined, and the results be interpreted in the light of the probable errors.