



A Group Scale of Intelligence for Use in the First Three Grades: Its Validity and Reliability

Luella Winifred Pressey

To cite this article: Luella Winifred Pressey (1920) A Group Scale of Intelligence for Use in the First Three Grades: Its Validity and Reliability, *The Journal of Educational Research*, 1:4, 285-294, DOI: [10.1080/00220671.1920.10879053](https://doi.org/10.1080/00220671.1920.10879053)

To link to this article: <http://dx.doi.org/10.1080/00220671.1920.10879053>



Published online: 15 Dec 2014.



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



View related articles [↗](#)

A GROUP SCALE OF INTELLIGENCE FOR USE IN THE FIRST THREE GRADES: ITS VALIDITY AND RELIABILITY

LUELLA WINIFRED PRESSEY
Indiana University

I. THE STATISTICAL PROBLEMS PRESENTED BY THE SCALE

In a recent number¹ of the *Journal of Educational Psychology*, the writer has presented a general descriptive account of the purpose, nature, and development of the "Primer Scale." At the time of writing that paper, little statistical evidence had been gathered as to the "validity" and "reliability" of the scale. By "validity" the writer means the extent to which the scale measures what it is supposed to measure—i.e., general intelligence; by "reliability" is meant the consistency or accuracy of the measures obtained. Thus, a scale might give very consistent measures that were not valid—that did not measure the qualities the scale was designed to measure—or, a scale might give reasonably valid measures of some quality, but the measures might be so inconsistent as to make the scale largely unsatisfactory. In other words, these two characteristics of a scale might exist in varying combinations with each other and should be kept separate in considering the value of the results. It is the purpose of the present paper to present the evidence that has been found, thus far, bearing on these two important points of (1) the validity and (2) the reliability of the Primer Scale.

II. THE VALIDITY OF THE PRIMER SCALE

1. *Sources of data.*—For a general description of the four tests composing the scale, the reader is referred to the above mentioned article. It may be briefly said that the first test is one of form discrimination in which dots are arranged in patterns; the second is a test of similarities and differences, using pictures of familiar objects as materials; the third is a form board test, arranged so that it can be presented on paper; and the fourth is a test of absurdities, again using pictures.

¹ Pressey, L. W. "A group scale of intelligence for use in the first three grades," *Journal of Educational Psychology*, 10:297-308, September, 1919.

The tests have been given to about twenty-five hundred public school children (June, 1919) in the first three grades;² also to 64 "children" with a mental age on the Stanford-Binet Scale of between six and eight in the state school for the feeble-minded in Minnesota. Many of the school children tested had been previously "Bineted."

2. *Methods*.—The writer has used three methods for arriving at a conclusion concerning the validity of the scale. (1) One hundred and twenty children who had been "Bineted" were divided into five groups on the basis of their Intelligence Quotients (I. Q.'s) as determined by their scores in the Binet Scale (Stanford Revision). The scores of these groups in the Primer Test were then examined for consistency with the groupings according to the Binet ratings. (2) Three correlations with the Binet ratings were calculated: first, the correlation between the Binet "mental age" and the Primer Test score for the 64 institutional feeble-minded cases; second, a similar correlation for 148 children in the primary grades of Council Bluffs, Iowa; and third, a correlation for 57 unselected six-year-old children in the first grade at Washington, Indiana. (3) The writer chose this last group for a more detailed analysis by the method of partial correlation. The separate tests were intercorrelated, each test was correlated with the Binet ratings, and the partial coefficients and regression equation were calculated.

3. *Results*.—The results of the first method are indicated in Table I which gives the distribution of scores on the Primer Scale of 120 unselected six-year-old children, for whom the writer had Binet ratings. The cases have been divided into five groups on the basis of the Binet I. Q.'s. Group I (the first column) includes the scores of those children with an I. Q. of 125 or above; Group II contains those children who obtain an I. Q. of 110-124; Group III those with an I. Q. from 90-109; Group IV those from 76-89; and Group V those with an I. Q. below 76. The five groups might be termed, very superior, superior, average, inferior, very inferior.

There is some over lapping from one group to another, but the only distribution for which the amount of over lapping is surprising is that of the "average" group (Group III) whose distribu-

² The writer wishes to express her obligations to the school superintendents of Washington, Bedford, and Bloomington, Indiana, and of Council Bluffs, Iowa; also, to Dr. F. Kuhlmann, Psychologist of the School for the Feeble-minded at Faribault, Minnesota for the materials here presented.

tion over laps that of every other group. The two extreme groups (I and V) are quite distinct from each other—that is, the lowest score in Group I is above the highest score in Group V. With the exception of one case, Groups I and II combined (i.e., the really bright children according to the Binet Scale) show no over lapping of test scores with the Groups IV and V combined (i.e., with the really dull children). The horizontal lines in Table I show the location of the 75-percentile, the median and the 25-percentile for the total distribution of these unselected children—i.e., the standard norms derived from the complete surveys made. As will be seen, the children in Group I make no scores below the 75-percentile for their age, while the children in Group V, with one exception, make no scores above the 25-percentile.

The lowest group is the most compact of the five. There is, in general, not as much over lapping of one group upon the other at the lower end of the distribution as at the upper end. This is understandable, as the brilliant children might not give their best performance and would thus make a lower score than they really should; while, on the other hand, it is not to be expected that adventitious circumstances would cause a dull child to be rated too high to the same degree. On the whole, the Primer Scale seems to have marked off degrees of ability as well as could be expected of a group test given to such young children. In this connection, it might be emphasized that some six-year-old children who make a low score on a group test receive a higher rating on the Binet examination because of the very different testing environment. This aids in explaining the dropping down of scores in the first three groups into the distributions of the lower groups. Some of the children making these scores lack, not intelligence, but independence. When asked to perform certain tasks in a group, they do not have the initiative to do their best; whereas they show a Binet rating more nearly commensurate with their true intelligence, because the examiner can constantly recall their attention and their efforts to the task in hand. The writer feels that the relationship shown in Table I is probably as high as can be expected between a group test and an individual test,³ especially with such young children.

³ Lowell, Frances. "A group intelligence scale for primary grades," *Journal of Applied Psychology*, 3:215-48, September, 1919. The scale described in this article is a group form of the younger years of the Binet; yet the correlation between group score and the rating on the Binet Scale when given to the same children individually is only +0.75.

TABLE I. THE DISTRIBUTION OF SCORES MADE BY 120 SIX-YEAR-OLD CHILDREN ON THE PRIMER SCALE, GROUPED ACCORDING TO I. Q. ON THE STANFORD-BINET SCALE

Primer Scale Scores	Group I (I. Q. above 125)	Group II (I. Q. 110-124)	Group III (I. Q. 90-109)	Group IV (I. Q. 76-89)	Group V (I. Q. below 76)
	1	2	3	4	5
80	1				
78					
76					
74					
72	2		1		
70	1				
68					
66	1	2			
64	1	1	1		
62	4	4	2		
60	3	2	1		
58	2	2	3		
56	4	1	4		
54		1	4		
52			5		
50		1	1		
48	2	2	1		
<i>75-percentile</i>					
46		3	3	1	
44			5		
42		2	4		
40		1	2		
38		1	3		
<i>Median</i>					
36		2	1		
34			1		
32			1		
30			2		
28			4		1
26					
24			2		
22					
20			1		
18					
<i>25-percentile</i>					
16			1		1
14			2	2	
12					
10			1	2	
8			1	2	
6			1	1	1
4					1
2					1
0					3
No. cases . . .	21	25	58	8	8
Medians	61	55	45	11	4

The second method of studying the validity of the Primer Scale involved the calculation of the Pearson coefficient of correlation between scores derived from it and scores obtained by using the Binet Scale. The crude coefficients of correlation for each of the three groups of children are as follows:

64 children from a feeble-minded school	+0.75
148 children in primary grades	+0.66
57 unselected children 6 years of age	+0.62

In all cases, the correlation is between group test scores and "mental ages" on the Binet examination.⁴ The correlations are, the writer feels, as high as can be expected from the very different nature of the testing conditions for the two types of tests (group and individual).

The scores of the 57 unselected six-year-old children were subjected to further analysis by means of partial correlation. The crude coefficients of the scores of each of the four tests which make up the Primer Scale with each other and with the Binet ratings are given in Table II.

TABLE II. CRUDE COEFFICIENTS OF CORRELATION BETWEEN TESTS OF THE PRIMER SCALE AND THE BINET SCALE, TAKEN PAIR BY PAIR

Tests	1	2	3	4	LEGEND
1					1 = Test 1, Primer Scale
2	0.59				2 = Test 2, " "
3	0.48	0.46			3 = Test 3, " "
4	0.57	0.55	0.33		4 = Test 4, " "
5	0.53	0.45	0.42	0.47	5 = Binet Scale

This table should be read: the correlation between Tests 1 and 2 of the Primer Scale was found to be 0.59; between Tests 1 and 3, 0.48; etc.

It will be noticed that the coefficients of correlation between the tests of the Primer Scale are rather high, except between the third and fourth. Because of this difference in the relationship

⁴ The I. Q., which is frequently used for such correlations, is a percent statement. No similar statement was readily possible with the Primer Scale. Therefore, it was necessary to use *total* scores on both scales. Even in correlating the scores of children of the same age, the error introduced in using I. Q. is considerable, as the basis for calculating the I. Q. may range (in the case above) from 6 years and 0 months to 6 years and 11 months—or nearly a whole year.

between the tests, it is especially desirable to find the partial correlations of each test with the Binet Scale, to determine just what the independent value of each test might be. The resulting partial coefficients of correlation are as follows:

$$r_{51.234} = 0.2436$$

$$r_{62.134} = 0.0951$$

$$r_{63.124} = 0.1834$$

$$r_{64.123} = 0.2006$$

These partial coefficients should be read: the correlation of 5 (Binet rating) with 1 (Test 1), the other three tests being constant, is 0.2436; of Binet rating and 2 (Test 2), the other three tests being constant, is 0.0951.

The partial coefficients are interesting in themselves aside from the part they play in calculating the regression equation.⁵ The relative standing of the four tests in their relation to the Binet Scale, has not changed materially from that indicated by the original correlations (Table II). The first test still has the highest correlation, and the order of decreasing coefficients is now the fourth, the third, and the second according to the partial correlations shown above. The second test seems to contribute almost nothing to the correlation of the whole scale with the Binet Scale, which is not already contributed by some other test. When the other tests are held constant, its correlation with the Binet Scale becomes practically zero (0.0951). Thus, the second test is, according to this criterion, the least valuable test of the scale⁶ and the first test is the most valuable.

III. THE RELIABILITY OF THE PRIMER SCALE

1. *Materials*.—The papers of all the children in the first and second grades, and of all the six- and seven-year-old children in

⁵ The regression equation of the Binet ratings on the four tests of the scale was found to be the following:

$$x_5 = 0.5719x_1 + 0.2142x_2 + 0.2844x_3 + 0.5380x_4,$$

where x is the deviation of a pupil's score in Test 1 from the average, x_2 his deviation from the average in Test 2, etc. x_5 is the most probable deviation from the average Binet score. The scores of the original 57 papers used in the derivation of the equation were weighted in accordance with the above equation. This weighting raised the correlation from 0.62 to 0.65.

⁶ Since writing the above, the writer has tried a new method of scoring the second test, which takes account of possible chance successes, and has obtained a correlation of 0.52 instead of 0.45 as given in Table II. This difference in correlation might make some difference in the relative standing of the second test as revealed by partial correlation.

the city most recently tested were chosen as the material from which to estimate the reliability of the scale.

2. *Methods.*—The degree of the reliability of the measures obtained by the scale and of its separate tests has been calculated in two ways. Coefficients of reliability have been found for each of the four groups just mentioned (grades I and II and ages six and seven). As there was only one form of the scale, the coefficients of reliability were calculated by dividing the scale into halves (using alternate items), correlating the scores of the half tests, and applying Brown's formula.⁷ This formula represents the extent to which the amalgamated results of two tests would correlate with a similar amalgamated series of two other applications of the same test. The formula is:

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

Both the total scale and each test have been treated in this way for the four age and grade groups. (2) The reliability has been calculated in terms of the probable error (P. E.), by the following method. Each pupil's score on one half of the scale was subtracted from his score on the other half. These differences were then averaged, giving an "average difference."

$$\begin{aligned} \text{The P. E. of a scale} &= \frac{0.8453}{\sqrt{2}} \times \text{Av. Diff.} \\ &= 0.5978 \times \text{Av. Diff.}^8 \end{aligned}$$

This formula applies whether the two tests from which the average difference is found are two separate equivalent tests or two halves of the same test. If two separate forms of the same test had been used, the above formula (P. E. = $0.5978 \times \text{Av. Diff.}$) would give the correct value, but in the case of the two halves of the same test, a further calculation must be made. The formula gives the P. E. of the half tests, but what is desired to obtain finally, is the P. E. of the *whole* test—not the P. E. of one of its halves. The P. E. of a test increases as the square root of its length; that is, in this case, the whole test is *twice* as long as its halves, and its P. E. consequently increases by $\sqrt{2}$. The P. E. of the whole test would, then, be equal to that of the

⁷ Brown, William. *Essentials of mental measurement*. New York: G. P. Putnam's Sons, 1911, p. 101.

⁸ Thorndike, E. L. *An introduction to the theory of mental and social measurements*. New York: Teachers College, Columbia University, 1916, pp. 190-93.

half test, multiplied by the $\sqrt{2}$. The resulting formula for deriving the P. E. of a test from that of its halves becomes:

$$P. E. = 0.5978 \times Av. Diff. \times \sqrt{2}.$$
⁹

3. *Results.*—The coefficients of reliability for grades I and II and for ages six and seven are given in Table III. For grade I and ages six, the first column represents the reliability of the scale, including all zero scores; the second column gives the reliability

TABLE III. THE COEFFICIENTS OF RELIABILITY BETWEEN THE TWO HALVES OF THE PRIMER SCALE FOR GRADES I AND II AND FOR AGES SIX AND SEVEN, AFTER APPLYING THE FORMULA:

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

(In this formula r_1 is the coefficient of correlation obtained by correlating the two halves of the scale and r_2 is the coefficient of reliability given in the table below.)

	GRADE I (142 CASES)		AGE SIX (90 CASES)		Grade II (120 Cases)	Age Seven (103 Cases)
	Zero Scores Included	Zero Scores Omitted	Zero Scores Included	Zero Scores Omitted		
	1	2	3	4	5	6
Total score.....	0.96	0.95	0.95	0.94	0.89	0.92
Test I.....	0.94	0.90	0.92	0.89	0.89	0.84
Test II.....	0.92	0.81	0.88	0.81	0.67	0.68
Test III.....	0.89	0.81	0.88	0.81	0.93	0.93
Test IV.....	0.82	0.78	0.84	0.79	0.54	0.75

omitting the zero scores. Neither method is satisfactory. If the test were to be given again, some but probably not all, of those children making zero scores would make a higher score. Since some of them would, the first column of coefficients, including all the zero scores, is too high: but, since not all of them would, the second column of coefficients, omitting all zero scores, is too low. Thus the true coefficient probably lies somewhere between the two.

The coefficients of reliability are uniformly high for the whole scale and for the first and third tests, but vary considerably for

⁹ The writer is indebted to Dr. Walter S. Monroe for this method of determining the P. E. of a scale from the P. E. of its halves; as well as for many other helpful suggestions and criticisms in the course of the statistical analysis.

the second and fourth tests in the different groups. The reliability, as measured by the correlation of one half of a scale to the other, is in part a measure of the consistency of the performance of the pupils, and in part a measure of the consistency of the test from item to item. It is the judgment of the writer that the reliability coefficient is mainly an index of the latter factor. If the halves of a test correlate highly, it is evident that the items composing the two halves demand the same, or approximately the same kind of mental effort. For Tests I and III this consistency is high; these tests are both "form" tests, they demand no information of any kind, and it apparently makes little difference which set of items is used. Tests II and IV call for information of a practical sort; the two halves of these tests correlate rather poorly—sometimes very poorly. Evidently, practical information is not at all the unitary thing that form discrimination is; and the kind of information demanded by one item is different from that demanded by another. If still another form of the test were to be used, it is probable that these two tests would still correlate poorly, since these new items would involve still different informational elements.

The P. E. of the scale was derived in the manner explained above. The P. E. of grade 1 (142 cases) was 2.39 "points," or units; that of age six (90 cases) was 2.20 "points," or units on the scale. Since the difference between the median of one age and the median of the next age is, on the average, twelve "points," or units, a P. E. of 2.39 points would equal 2.39 months, and a P. E. of 2.20 points would equal 2.20 months. Again, since there are about twelve points from one age to the next, a statement, for example, that a given child scores at the median for the age above his own, does not seem to be invalidated by the unreliability of the scale, since the differences between the successive medians are well beyond the P. E. of the scores.

The P. E. of the Stanford-Binet Scale has been estimated by Otis to be about three and a half months.¹⁰ The P. E. of the Primer Scale is even less. If the same reliability can be obtained with a group scale as with an individual scale, the group scale can,

¹⁰ Otis, A. S. "An absolute point scale for the group measurement of intelligence," *Journal of Educational Psychology*, 9:341, June, 1918.

to some extent, take the place of the individual scale, at least for the preliminary sorting of pupils.

On the whole, the scale seems very reliable—unusually so when the ages of the children are considered. Considerable confidence may, therefore, be placed in the accuracy of the scores obtained.

IV. SUMMARY

1. The writer presents data to show the reliability and the validity of the Primer Scale.

2. The validity is shown (1) by the distribution of scores made on the Primer Scale by children who have been classified according to I. Q. on the Binet examination; (2) by the correlations for various groups of children between the group scale and the Binet ratings; and (3) by partial coefficients of correlation.

3. The reliability is shown by (1) the coefficients of reliability for different ages and grades and (2) by the P. E. of the whole scale, as derived from the P. E. of the half scale.

4. It is concluded that the Primer Scale is reasonably valid and reliable.