AN EMPIRICAL STUDY OF THE VALIDITY OF THE SPEARMAN-BROWN FORMULA AS APPLIED TO THE PURDUE RATING SCALE

H. H. REMMERS, N W. SHOCK, AND E L. KELLY*  

Purdue University

HISTORICAL SUMMARY

When the same test is given at different times to the same persons, the correlation between the resulting scores gives a measure of the reliability of the test used. It is a well known fact that the reliability of a test is increased as its length is increased, but for some time there was no way of foretelling just how much the reliability would be increased in its length. There was need for a formula that would predict the reliability of a test \( n \) times as long as a given test. In 1911 William Brown published a formula which promised to meet this need. It was found to be only a special case of an earlier formula of Spearman's for finding the correlation between the sums or averages of scores. This revised formula is now in current use and is known as the Spearman-Brown prediction formula. If \( r_{12} \) is the coefficient of reliability of a test, then \( r_n \), the coefficient of reliability of a test \( n \) times as long, is given by the formula:

\[
r_n = \frac{nr_{12}}{1(n - 1)r_{12}}
\]

Since \( r \) is always less than 1, \( r_n \) will always be greater than \( r_{12} \), and thus the reliability of the longer test will increase as some logarithmic function in its length. In order for this formula to be valid, it must be assumed that the test material is homogeneous in character, or statistically speaking, that the intercorrelation between the units is equal and their standard deviations are equal. Were these assumptions strictly fulfilled, perfect prediction would result. In practice, however, such ideal conditions are never found, so our problem becomes that of determining just how accurate a prediction may be obtained for the material at hand.

It was only natural that the introduction of such a formula would be of much interest to all those interested in tests and measurements. Accordingly, studies soon began to determine the validity of the for-
mula for various classes of material. Holzinger, in studying the application of the formula to intelligence test material, found that the formula tended to predict the obtained reliability with fair accuracy for a pool of five tests or less, but for greater numbers the formula overestimated the empirical results obtained. His test material, however, violated to some extent the major assumption, i.e., homogeneity of test items. Later, Holzinger and Clayton applied the formula to spelling words and self-administered test material, and in each case found a close correspondence between the predicted and observed results. Kelley, using Gordon's data on lifted weights, also found a close agreement between the predicted and observed results on this material. Ruch and others applied the formula to spelling words, and they came to the conclusion that it gave a meaningful prediction for such test material. In two cases out of five, the tendency was to overpredict slightly the empirical scores. Wood applied the same formula to a set of scores on a hundred true-false questions given to language students, and he found that it tended very nearly to predict or else to underestimate the observed results. Furfey published results while the present study was in progress, showing that the formula slightly overestimated the reliability obtained by dividing the test into a number of components.

**Purpose of This Study**

It is the purpose of this study to determine the validity of the Spearman-Brown prediction formula when applied to the personal rating scale now in use at Purdue University. If it is found to be a valid means of predicting the reliability of the combined ratings of any number of judges, it will be quite useful to the personnel department in deciding on the number of judgments that should be obtained on an individual in order to obtain the desired reliability. At the present time, each engineering student is rated by five fellow students and by five instructors, all of his own choice. Thus the number of judgments has been arbitrarily placed at ten, without knowing whether it is too low to be dependable or too high to be economical. If it is shown that the Spearman-Brown formula will accurately predict the reliability for any given number of judgments with the scale, then the most efficient number of judgments, both for dependability and for economy, may be accurately determined.

In order that the reader may better understand the present study, a short description of the rating scale, which is now being used, is
Spearman-Brown Formula and Purdue Scale

given. It consists of a printed form containing a paragraph of instructions on the upper part of the sheet, which explains the purpose of the questionnaire and asks the reader to rate the individual whose name appears filled in the blank. The scale itself consists of ten traits on which the ratings are made. These traits are:

1. Address and Manner
2. Attitude
3. Character
4. Cooperative Ability
5. Disposition
6. Industry
7. Judgment
8. Initiative
9. Leadership
10. Mental Caliber

The ratings are made on a five point scale, five being considered perfect, one as very poor, and three as representing the average individual. Fractional judgments are permitted. It will thus be seen that fifty is the highest possible total score.

Plice in a recent study of the scale came to the following conclusions:

1. "Students rate each other higher than teachers rate them.
2. "Teachers agree better on the rating of students than do the students themselves.
3. "Teachers agreed fairly well with the students except on traits 7 and 10. (Trait 7 is Judgment and Trait 10 is Mental Caliber.)
4. "While the general agreement of teachers and students in rating is low, it is somewhat higher than chance agreement."

**Experimental Procedure**

If the Spearman-Brown prediction formula is to be used in predicting the reliability of a certain number of judgments, certain assumptions must be made. In the first place it is assumed that one judge is just as competent to judge traits of character and personality in an individual as another. This follows from the basic assumption that the test material is homogeneous, as is made in the original formula, since here instead of having test elements we have judges. Another assumption that follows directly from this is that an individual judge can be treated as a constant factor, comparable to a test element.
It was felt that we should have a group as large as possible with not less than twenty-five who would judge each other conscientiously and who also knew one another well enough to be fairly accurate in their judgments. Hence, a fraternal group was chosen. Here several factors entered which are possibly not the same as those encountered in the actual field of personnel rating. In the first place all of these men knew each other more intimately than would ordinarily be true of some ratings obtained in practice, since they had lived in the same house with one another for almost an entire semester. This fact would tend to raise the reliability of the ratings. On the other hand, the fact that each man was rated by every other man in the house, let petty jealousy and feeling creep into the ratings in some instances. This would tend to lower the reliability of the scale. This difficulty is obviated in practice since the individual rated chooses his own judges, thus possibly eliminating those who would rate him low merely because of personal feeling.

The first step was to get the men to rate each other. In order to facilitate the tabulation of the results each man in the fraternity was given a number as a judge and a letter as a subject to be judged. It was first thought best to have each man judge six others at a time, selecting each group by chance, in order to eliminate the factor of fatigue on the part of the judge. This was done, but with the first and second groups of six a decided change in standards on the part of the judge was noted. That is, the judges tended to compare and rank the men within each group of six, instead of judging the men within the fraternity as a whole. In order to avoid this difficulty, all of the remaining rating blanks were given out at the same time, thus minimizing the possibility of errors due to the shifting of the standards used by the individual in making his judgments.

The instructions given to the men were: "In order to carry out an investigation of the reliability of personal judgments, we are asking every man in the fraternity to rate every other man, using the Purdue Personnel Rating Scale. The thing that we want is a conscientious rating of the character and personality of each man, using the average man within the fraternity as a standard. The rest of the instructions will be found on the rating scale. You may be perfectly frank in your judgments, since all ratings will be anonymous." It was thus made clear to the men that each was to use his own conception of the average man within the fraternity as a standard by which he was to judge the others.
A fine spirit and keen interest was displayed by everyone. This was due to the fact that all were much interested in finding out what the rest of their fraternity brothers thought of them. This was possible only if each gave a good rating of the rest. As a result, the ratings came in promptly and were rather carefully filled out. It is reasonably certain that each rating was as carefully analyzed as is usually true in ratings made in actual practice.

After all the rating scales were received properly filled out, the individual's score was computed by taking the sum of the rating on each of the ten characteristics. After these data had been computed they were arranged in tabular form, subjects judged on the \( x \) axis, and judges on the \( y \) axis. The arithmetic mean of all the judgments on each subject were computed. The standard deviations (\( \sigma \)) of the distribution as well as probable errors (PE) of the means then computed and arranged in tabular form as found in Table I. All tables not shown in this article because of limitations of space are on file in the Department of Education, Purdue University.

**Table I**

Letters Indicate Subject. *Average* is Average of All Judgments on That Subject. \( \sigma \) is the Standard Deviation of the Distribution. PE\(_{av}\) is the Probable Error of the Average

<table>
<thead>
<tr>
<th>Subject rated</th>
<th>Average ( \sigma )</th>
<th>Distr.</th>
<th>PE(_{av})</th>
<th>Subject rated</th>
<th>Average ( \sigma )</th>
<th>Distr.</th>
<th>PE(_{av})</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>29 0</td>
<td>4 31</td>
<td>±.68</td>
<td>N</td>
<td>33 4</td>
<td>5.20</td>
<td>± 70</td>
</tr>
<tr>
<td>B</td>
<td>31 1</td>
<td>4 97</td>
<td>± 67</td>
<td>O</td>
<td>34 8</td>
<td>4.90</td>
<td>± 66</td>
</tr>
<tr>
<td>C</td>
<td>30 8</td>
<td>5 42</td>
<td>±.72</td>
<td>P</td>
<td>33 8</td>
<td>4.77</td>
<td>±.64</td>
</tr>
<tr>
<td>D</td>
<td>35 8</td>
<td>4 90</td>
<td>±.66</td>
<td>Q</td>
<td>31 8</td>
<td>5.28</td>
<td>± 71</td>
</tr>
<tr>
<td>E</td>
<td>37 9</td>
<td>4 84</td>
<td>±.65</td>
<td>R</td>
<td>35 3</td>
<td>4.36</td>
<td>±.59</td>
</tr>
<tr>
<td>F</td>
<td>38 1</td>
<td>4 55</td>
<td>±.61</td>
<td>S</td>
<td>30 9</td>
<td>3.97</td>
<td>± 52</td>
</tr>
<tr>
<td>G</td>
<td>36 3</td>
<td>4 06</td>
<td>± 55</td>
<td>T</td>
<td>38 3</td>
<td>4.60</td>
<td>± 62</td>
</tr>
<tr>
<td>H</td>
<td>33 1</td>
<td>3 86</td>
<td>± 52</td>
<td>U</td>
<td>33 8</td>
<td>4.97</td>
<td>± 67</td>
</tr>
<tr>
<td>I</td>
<td>31 0</td>
<td>3 75</td>
<td>± 51</td>
<td>V</td>
<td>36 8</td>
<td>5.09</td>
<td>± 69</td>
</tr>
<tr>
<td>J</td>
<td>36 9</td>
<td>4 74</td>
<td>± 64</td>
<td>W</td>
<td>35 9</td>
<td>4.68</td>
<td>± 63</td>
</tr>
<tr>
<td>K</td>
<td>37 0</td>
<td>5 10</td>
<td>±.69</td>
<td>X</td>
<td>34 1</td>
<td>5.24</td>
<td>± 71</td>
</tr>
<tr>
<td>L</td>
<td>31 0</td>
<td>4 35</td>
<td>± 59</td>
<td>Y</td>
<td>33 8</td>
<td>4.62</td>
<td>± 62</td>
</tr>
<tr>
<td>M</td>
<td>35 0</td>
<td>5 10</td>
<td>± 69</td>
<td>Z</td>
<td>31 9</td>
<td>4.03</td>
<td>± 55</td>
</tr>
</tbody>
</table>

Averages of judgments varying in number from one to thirteen and randomly selected, were then correlated. That is, correlations of one judgment vs. one judgment, two judgments vs. two judgments and so on to thirteen judgments vs. thirteen judgments were obtained.
Column 2 of Table II gives the number of chance combinations correlated for the different groupings. A total of 96 \( r \)'s was computed.

Table II is a summary of the results of the entire investigation, since it shows the predicted and experimental reliability of the scale when used with given numbers of judgments.

**Table II.—Correlations Predicted and Observed**

<table>
<thead>
<tr>
<th>Number</th>
<th>Predicted ( r )</th>
<th>PE ( \pm r )</th>
<th>Observed ( r )</th>
<th>PE ( \pm r )</th>
<th>Difference observed</th>
<th>Predicted observed</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ones</td>
<td>25</td>
<td></td>
<td>251 ± 0.40</td>
<td>± 0.04</td>
<td>-0.13</td>
<td></td>
</tr>
<tr>
<td>Twos</td>
<td>23</td>
<td>401 ± 0.11</td>
<td>388 ± 0.20</td>
<td>± 0.09</td>
<td>+0.02</td>
<td></td>
</tr>
<tr>
<td>Threes</td>
<td>8</td>
<td>0.500 ± 0.099</td>
<td>0.592 ± 0.043</td>
<td>± 0.087</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>Fours</td>
<td>6</td>
<td>0.572 ± 0.087</td>
<td>0.538 ± 0.041</td>
<td>± 0.080</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>Fives</td>
<td>10</td>
<td>0.626 ± 0.080</td>
<td>0.631 ± 0.046</td>
<td>± 0.077</td>
<td>+0.05</td>
<td></td>
</tr>
<tr>
<td>Sixes</td>
<td>6</td>
<td>0.677 ± 0.070</td>
<td>0.661 ± 0.013</td>
<td>± 0.067</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>Sevens</td>
<td>5</td>
<td>0.700 ± 0.067</td>
<td>0.777 ± 0.003</td>
<td>± 0.062</td>
<td>+0.07</td>
<td></td>
</tr>
<tr>
<td>Eights</td>
<td>3</td>
<td>0.728 ± 0.062</td>
<td>0.667 ± 0.040</td>
<td>± 0.067</td>
<td>-0.01</td>
<td></td>
</tr>
<tr>
<td>Nines</td>
<td>2</td>
<td>0.751 ± 0.058</td>
<td>0.663 ± 0.027</td>
<td>± 0.051</td>
<td>-0.08</td>
<td></td>
</tr>
<tr>
<td>Tens</td>
<td>2</td>
<td>0.770 ± 0.051</td>
<td>0.731 ± 0.015</td>
<td>± 0.050</td>
<td>-0.03</td>
<td></td>
</tr>
<tr>
<td>Elevens</td>
<td>2</td>
<td>0.786 ± 0.050</td>
<td>0.834 ± 0.041</td>
<td>± 0.048</td>
<td>+0.04</td>
<td></td>
</tr>
<tr>
<td>Twelves</td>
<td>2</td>
<td>0.801 ± 0.048</td>
<td>0.865 ± 0.034</td>
<td>± 0.048</td>
<td>+0.06</td>
<td></td>
</tr>
<tr>
<td>Thirteens</td>
<td>2</td>
<td>0.813 ± 0.046</td>
<td>0.852 ± 0.044</td>
<td>± 0.048</td>
<td>+0.03</td>
<td></td>
</tr>
</tbody>
</table>

The first column of numerals is the number of correlations in the average.

Predicted \( r \) is reliability predicted by Spearman-Brown formula. Observed \( r \) is reliability obtained by experiment.

Inspection of this table reveals certain interesting trends in the data: (1) A relatively rapid increase in reliability in the first few groups of correlated judgments. (2) A very close approximation to the predicted value when the number of correlations is relatively large. (3) The last three values in the observed series are higher than those in the predicted series. This, however, may be merely a chance fluctuation due to limited sampling.

A graphical representation of these results is given in Fig. 1. It will be seen that the experimental value for the reliability falls outside of one probable error of the predicted value only three times, and that it is always within one standard deviation. One reason for the irregularity of the curve of the experimental reliabilities is the relatively
small number of correlations which were possible in the groups containing the higher numbers.

Figure 2 is made from the same data, but with the curve of predicted values represented by a straight line in order to show that the probable error of the correlations decreases as the coefficient of correlation increases. This fact is not evident from Fig. 1 since the PE is measured on the axes rather than on line perpendicular to the curve of predicted reliability. Holzinger and Clayton in a graph showing the results of a similar study to test the validity of this formula when applied to spelling words did not seem to take this into account, but seemed to measure the probable error on a line perpendicular to the curve at the given point. This did not give the true location of the PE curve since all the distances must be measured on the abscissa or ordinate.

SUMMARY

A review of previous experimental work designed to test the validity of the Spearman-Brown prediction formula shows it to give a meaningful prediction on such materials as mental test items, spelling words, lifted weights, true-false test items in language, and component units of rating scales.

The present investigation using judges as equivalent to test items in the sense required by the formula, concerned itself with correlations between groups of judgments varying from 1 to 13. These judgments were obtained by having each man in a fraternity group of 26 judges every other man on certain traits of character and personality as given in the Purdue Personnel Rating Scale.

CONCLUSIONS

1. The reliability of the average of a number of judgments with this scale is higher than that of one judgment.

2. The Spearman-Brown formula predicts within two probable errors or one standard deviation the empirical reliability obtained by experiment up to and including thirteen judges.

3. Although the increase in reliability diminishes as the number of judgments is increased, the decrease in per cent of error of the average judgment continues to be appreciable.

4. If the Spearman-Brown formula holds for numbers of judgments beyond thirteen, the number of judgments required for any desired reliability can be readily determined. The practical import of this
generalization is of considerable consequence in actual rating or judging situation.

REFERENCES

2. Ibid., page 205.
10. Plice, Max Jennings: A Study of Some of the Psychological Aspects of Personnel Selection with Special Reference to Purdue University, page 50. (To be published in an early number of the *Journal of Industrial Psychology*.)
12. Ibid., page 331.