Table of Contents

Industrial Painting Helps Efficiency, by M. Rea Paul........................................ 681.
A University Tackles Accident Prevention, by C. W. Price........................................ 688
Salaries vs. Commissions for Salesmen, by Gorton James........................................ 692
What It Costs to Lose Sleep, by Donald A. Laird and William Wheeler, Jr.................. 694
The Object of Ventilation, by C. P. Yaglou................................................................. 697
Workers’ Stumbling Blocks, by Glenn L. Gardiner....................................................... 699
Progress in France, by Lorine Pruette.... ................................................................... 706
Clerical Test Agrees with Employers’ Ratings, by C. O. Weber and Miss
Mary Leslie ....................................................................................................................... 708
Just Naturally Bad, by Howard Signor ............................................................................ 712
Reliability of Ratings at Purdue University, by Herman Remmers and
M. J. Price ......................................................................................................................... 717
Editorial Comment ........................................................................................................... 722
Notes and Items .............................................................................................................. 724
New Books to Read ......................................................................................................... 729
Sante Naccarati .............................................................................................................. 680
To Cut Down Industrial Accidents.................................................................................... 693
Safety Dependent on Workers’ Eyes............................................................................... 705
Sex Differences in Handwriting ...................................................................................... 705
Suggested Tests for General Machinist ........................................................................... 716
Suggested Tests for Automobile Mechanics ................................................................... 716
Psychology of Compensation Neurosis .......................................................................... 716
Tests Become Futile ......................................................................................................... 722
International Congress of Psychology ............................................................................ 722
Results of Eight-Hour Day ............................................................................................. 725
Three Plans for Better Industrial Relations .................................................................... 725
Accidents Among Woman Workers................................................................................ 725
Safety Drive ..................................................................................................................... 725
Orlando Edgar Miller ..................................................................................................... 725
Being a Rosierucian ........................................................................................................ 726
The Shorter Workday ....................................................................................................... 726
Promoting Industrial Peace ............................................................................................ 727
Large-Scale Production .................................................................................................. 727

Cover illustration courtesy The Gorham Co.

Publication office 100 Liberty sent three weeks before the issue to
Street, Utica, N. Y. General office
Hamilton, N. Y.

Copyright 1926 by Donald A. Laird. Subscription rates $5.00 per year in
Entered as second class matter the United States, $5.50 in Canada,
June 1, 1926, at the Post Office at $6.00 in other countries of the Postal
Utica, New York, under act of March
Union. Changes of address should be
3, 1879.
Reliability of Ratings at Purdue University

H. H. Remmers and M. J. Plice

INAUGURATED in 1920 by Dean A. A. Potter of the engineering schools, a system of personal records is maintained for the students at Purdue University. Among the data collected are scores on the Scott Mental Alertness Test, and ratings on various character traits and scholastic grades. The purpose of the system is to make possible the following benefits:

1. It renders service to prospective employers by furnishing them with definite information concerning the man they wish to employ.

2. A student may refer to this department at any time after graduation for a recommendation regarding his character and general ability.

3. Records are immediately available, so that information concerning any student may be had on short notice.

4. Personality rating, by confronting the student with the judgments of his teachers and fellow students, stimulates him to improve his scholastic achievement and his attitude toward his fellow men.

5. It tends to promote better acquaintance among students and teachers.

6. It helps to systematize and standardize grading of students.

Engineering students are given the Scott Mental Alertness Test at the beginning of their Freshman year. Somewhat later in their career the students designate five instructors and five students who rate them on a scale of 10 character traits. A replica of the rating scale and record card is shown in Fig. 1. The ratings are then averaged and entered on a permanent filing card.

Dean Potter having expressed a desire for a statistical analysis of the data available in the personnel bureau, the authors undertook to make a study of the instruments used.

A crucial question with regard to any instrument of measurement is its reliability. Would a second measurement or rating under similar conditions produce the same results? How consistent are the students and instructors in their ratings? Do they agree closely or vary widely in their estimates? To answer these questions, reliability coefficients of the mental test scores and of the various ratings were calculated.

Procedure

To determine the reliability of the Scott Test, one hundred test papers of students entering in 1924-25 were taken at random. The test was then split into chance halves by scoring separately the even-numbered and odd-numbered items. A correlation coefficient of these scores was then calculated and stepped up by the Brown-Spearman prophecy formula. This formula expresses the relationship between the length of a test and its reliability. By this formula it is also possible to determine the length of a test necessary to yield any desired coefficient of reliability when the latter is known for a given length of the test.
<table>
<thead>
<tr>
<th></th>
<th>SOPHOMORE RATINGS</th>
<th></th>
<th>SENIOR RATINGS</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>See scale at bottom of card*</td>
<td></td>
<td>See scale at bottom of card*</td>
</tr>
<tr>
<td></td>
<td>Teachers</td>
<td>Students</td>
<td>Others</td>
</tr>
<tr>
<td></td>
<td>Grade No.</td>
<td>Grade No.</td>
<td>Grade No.</td>
</tr>
<tr>
<td>1. ADDRESS AND MANNER: (Does he leave a good impression? Is he a good mixer? Is he popular?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. CO-OPERATIVE ABILITY: (Can he work with others? Is he accommodating? Willing to learn? Tolerant? Tactful?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. DISPOSITION: (Is he cheerful? Courteous? Enthusiastic and not conceited?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. INDUSTRY: (Is he a hard worker and has he perseverance?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. JUDGMENT: (Has he common sense? Observing and reasoning power? Foresight? Resourcefulness?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. INITIATIVE: (Is he a self-starter? Does he recognize and develop opportunities to a successful conclusion? Is he practical?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. LEADERSHIP: (Does he understand men and can he command their respect? Has he executive ability?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. MENTAL CALIBER: (Is he accurate, systematic, alert? Has he knowledge of facts and data? Can he concentrate? Does he learn readily?)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>X. Results of Intelligence Test</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Y. Honors Awarded</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Z. Faculty or Committee Action</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

REMARKS: Indicate under this head any additional information which may prove of value. If he has poor health, a deformity, or any peculiarity this should be reported.

*Please use the scale 5 to 1. 5 is the highest grade obtainable and means perfection. If he is above average in any characteristic rate him 4. 3 means average, 2 means below average and 1 means poor. Fractions such as 2.5 or 2.3 may be used to show intermediate ratings.
No attempt is here made to define the terms "character" and "personality." For purposes of this study they may be taken to mean whatever is included in one's concept of the ten traits rated. Another problem is whether an individual has these traits in absolute amounts or whether they represent relationships between him and his social environment, which varies from moment to moment. Polemics on this problem are also irrelevant to the present study.

To determine the reliability of the character ratings the averages of chance halves of the ratings of fifty seniors were correlated and the reliabilities of the total ratings found by the Brown-Spearman formula. Three correlations were calculated: teachers against teachers, students against students, and teachers against students. Our data are shown below, Tables 1, 2, and 3.

The meaning of a coefficient of reliability is enhanced by calculating in each case the coefficient of alienation, $k$, which is a measure of the lack of agreement between two variables. It is derived by using the formula $k = \sqrt{1 - r^2}$ in which $r$ is the observed coefficient of correlation. The resulting value expresses in terms of percentage the amount of error still present when 100 percent equals a purely chance result.

**Discussion of Results**

The coefficient of reliability for the Scott Alertness Test was found to be $r = + .96 \pm .005$.

There is great disagreement among both teachers and students as to the amounts of the various traits each student possesses. The teachers show significantly higher agreement among themselves than do the students. It is interesting to note in this connection the similarity of reliability coefficients of ratings to reliabilities of examinations in academic subjects, which, as ordinarily given, range from .40 to .60.

Although we did not tabulate average ratings, inspection of the rating sheets reveals the presence of the now familiar halo effect in ratings. An individual rating another tends to give values to the various traits which vary but slightly from each other. In other words, raters do not use the extremes of the scale to the same extent for one in-
individual as when different individuals are being rated. An individual tends to be rated either good or bad as a whole. A bias, favorable or unfavorable, is a factor in qualitative rating scales of this sort.

A question of individual ability to judge traits arises here. Does the possession of a certain trait accompany ability to judge that trait correctly? In general, the more desirable the trait, the closer is the relation between possession of it and ability to judge it. If a trait is defined as a desirable one, the inference is that in the long run possession of the trait and ability to judge it accompany each other.

The relatively low coefficients of reliability obtained do not, in our judgment, invalidate the rating scale technique as employed at Purdue. No one will argue that rating scales thus far developed are highly exact scientific instruments. They should rather be thought of as practical administrative devices. Several suggestions for improving the results of their use may be made.

Suggestions for Improvement

To yield reliable results, rating scales must be adapted to circumstances.

A training program for raters would probably do much to clarify the concepts involved in the trait names. It should be noted that an attempt at such clarification is the definition of each trait on the Purdue Scale. The training program would include a carefully worded rater’s manual; analysis of each set of ratings and detailed comparison with previous ratings; and, finally, the use of several devices for checking their validity and reliability.

### TABLE I

<table>
<thead>
<tr>
<th>Trait</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>.38</td>
<td>.64</td>
<td>.43</td>
<td>.36</td>
<td>.46</td>
<td>.39</td>
<td>.49</td>
<td>.40</td>
<td>.15</td>
<td>.27</td>
</tr>
<tr>
<td>P.E.</td>
<td>.082</td>
<td>.056</td>
<td>.078</td>
<td>.083</td>
<td>.075</td>
<td>.081</td>
<td>.072</td>
<td>.080</td>
<td>.090</td>
<td>.090</td>
</tr>
<tr>
<td>r2</td>
<td>.55</td>
<td>.17</td>
<td>.19</td>
<td>.13</td>
<td>.15</td>
<td>.19</td>
<td>.15</td>
<td>.13</td>
<td>.17</td>
<td>.55</td>
</tr>
<tr>
<td>k</td>
<td>.92</td>
<td>.77</td>
<td>.72</td>
<td>.92</td>
<td>.93</td>
<td>.93</td>
<td>.91</td>
<td>.66</td>
<td>.90</td>
<td>.98</td>
</tr>
</tbody>
</table>

\[ r_{1/2} \text{ I/II} = \text{Coefficient of correlation of first half of ratings with second half.} \]

\[ \text{P.E.}\pm = \text{Probable error of coefficients of correlation.} \]

\[ r_2 = \text{Coefficient of reliability of the total ratings.} \]

\[ k = \text{Per cent of error still present.} \]

### TABLE II

<table>
<thead>
<tr>
<th>Trait</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>.17</td>
<td>.31</td>
<td>.44</td>
<td>.26</td>
<td>.37</td>
<td>.20</td>
<td>.12</td>
<td>.07</td>
<td>.29</td>
<td>.44</td>
</tr>
<tr>
<td>P.E.</td>
<td>.026</td>
<td>.087</td>
<td>.087</td>
<td>.089</td>
<td>.083</td>
<td>.093</td>
<td>.094</td>
<td>.085</td>
<td>.087</td>
<td>.077</td>
</tr>
<tr>
<td>r2</td>
<td>.30</td>
<td>.47</td>
<td>.62</td>
<td>.41</td>
<td>.53</td>
<td>.33</td>
<td>.21</td>
<td>.014</td>
<td>.47</td>
<td>.62</td>
</tr>
<tr>
<td>k</td>
<td>.98</td>
<td>.95</td>
<td>.84</td>
<td>.96</td>
<td>.93</td>
<td>.97</td>
<td>.99</td>
<td>1.0</td>
<td>.95</td>
<td>.84</td>
</tr>
</tbody>
</table>

### TABLE III

<table>
<thead>
<tr>
<th>Trait</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>r</td>
<td>.40</td>
<td>.21</td>
<td>.31</td>
<td>.32</td>
<td>.36</td>
<td>.50</td>
<td>.10</td>
<td>.36</td>
<td>.52</td>
<td>.014</td>
</tr>
<tr>
<td>P. E.</td>
<td>.080</td>
<td>.090</td>
<td>.086</td>
<td>.086</td>
<td>.083</td>
<td>.071</td>
<td>.095</td>
<td>.083</td>
<td>.070</td>
<td>.095</td>
</tr>
<tr>
<td>k</td>
<td>.90</td>
<td>.97</td>
<td>.95</td>
<td>.95</td>
<td>.93</td>
<td>.86</td>
<td>.99</td>
<td>.93</td>
<td>.88</td>
<td>1.0</td>
</tr>
</tbody>
</table>
Cooperation among both students and instructors should be increased by all possible means, particularly by bringing to their attention the possible injustice of careless ratings. Adequate clerical assistance should of course be available in order to keep the machinery of the rating system operating promptly.

Increasing the number of judges per individual rated would measurably increase the reliability of the averaged ratings. To ask an instructor to rate a student who was one of a class of perhaps forty students a year ago is hardly conducive to accurate rating.

A careful study of the relationship between ratings and employers' estimates would serve to answer the question of validity of ratings. Such an analysis, although we desired very much to make it, was impossible because too few employers' ratings were available. This is explained by the fact that the present rating scale is a revision of an earlier form and has been in use only one year at the time of this study.

Conclusions

Several general conclusions may be drawn from this work:

1. Strictly objective measurements like the Scott Mental Alertness Test show a high degree of reliability as compared with ratings.

2. Students rate each other higher than teachers rate them.

3. Teachers agree better on rating of students than do the students themselves.

_before a rating scheme can be trusted it must be tested. The authors show how one can be tested._

4. Teachers' ratings agreed fairly well with students' ratings except in Traits 7 and 10 (Judgment and Mental Calibre).

5. While the general agreement of teachers and of students in rating is low, yet it is much higher than chance agreement.

6. Greater agreement in rating will result if more care and thought is given to the rating.

7. The value of university ratings will be increased when more employers' ratings are obtained.

4 Journal of Educational Psychology, May, 1925.
5 Wood, "Measurement in Higher Education."