

THE RELIABILITY OF JUDGMENT OF PERSONAL TRAITS

By JOHN SLAWSON, Division of Mental Defect and Delinquency, New York State Board of Charities

Two outstanding features in the science of mental measurement clearly indicate the necessity for the standardization of judgment evaluation of personal traits (1), an unfortunate lack of objective measures for human traits other than intelligence, or educational attainment, so that evaluation of an individual in the shop, factory or school is still largely a matter of comparative judgment, such as ranking in order of merit, no contention is made that the judgment method is as desirable as the standardized scale method employed in evaluating intelligence, but since the former is used of necessity, it must be raised to as objective a level as possible, (2), in the derivation of a great many of our objective tests for intelligence and specific aptitude, we incorporate judgment evaluation in the composite criterion which determines the authenticity of these tests

In this paper the writer intends to give a brief report of the results of a study¹ made in cooperation with the Bureau of Research of the Department of Education of the City of New York, the purpose of which was to determine the possibilities of the order of merit method in evaluating personal traits of individuals engaged in a specific profession or work. Eleven personal traits were chosen as defined below, and independent ratings were made by 31 judges distributed among six schools (5+5+4+7+6+4=31)

In each of the six schools, twenty-five teachers known to all of the judges chosen in that school were rated in the eleven traits, which were selected because of (a), supposed importance in the teaching profession, (b), competency of at least five raters to make judgments in them, and (c), distinctness or exclusiveness². The judges were asked to rate the teachers

¹ The original report bearing the same title as this paper is on file at the Psychological Library of Columbia University, in manuscript form

² Thorndike E. L., "A Constant Error in Psychological Rating," *Jour of App Psychol*, 1920, pp 25-29

in order of merit, i. e., placing the one who possessed most of the quality in question first and the one who possessed least of the quality last, by the use of the Stenquist "Card-list" Combination Ranking Card,³ which makes the task much less laborious to the rater and helps maintain uniformity in materials. Each judge, of course, made all ratings independently of the other judges.

- | | Trait | Definition |
|--------|---|---|
| No. 1 | <i>APPEARANCE</i> | (Personal neatness in dress, cleanliness, etc.) |
| No. 2 | <i>TACT</i> | (Ability to deal with others without giving offense) |
| No. 3 | <i>PUNCTUALITY</i> | (Habit of being on time) |
| No. 4 | <i>EFFORT</i> | (How hard does this person try?) |
| No. 5 | <i>JUDICIAL SENSE</i> | (Fairmindedness, impartiality) |
| No. 6 | <i>LEADERSHIP</i> | (Ability to lead, guide, direct, influence) |
| No. 7 | <i>COOPERATIVENESS</i> | (Willingness to work effectively with others) |
| No. 8 | <i>PROFESSIONAL INTEREST AND GROWTH</i> | (Interest in becoming a better teacher. Professional magazines, continued courses, lectures, associations, etc.) |
| No. 9 | <i>UNDERSTANDING OF CHILDREN</i> | (Insight into child nature. Success in handling children) |
| No. 10 | <i>COUNTERACTING FACTORS</i> | (This calls not for a judgment of the persons being ranked, but for a judgment of the <i>environment</i> in which each is working. Rank the person who is working under the greatest handicaps, No. 1, or first, the one working under the next greatest handicaps, 2d, etc.) Counteracting factors include such things as especially difficult classes, poor equipment, depressing relations with other teachers or with supervising staff, bad hygienic conditions, poor health, etc.— <i>ALL AS COMPARED WITH OTHERS OF THE GROUP BEING JUDGED</i> |
| No. 11 | <i>ALL-ROUND VALUE TO SERVICE</i> | (As compared with all others. Not a <i>total</i> of the other items. A single estimate in toto disregarding specific items above) |

³ Stenquist, J. L., "An Improved Form of Rating by the Order of Merit Method" *Jour. Ed. Psychol.*, Dec. 1920, p. 526

After an interval of about two weeks, the ratings were repeated by all of the judges independently of the first ratings. Sometime after the second set of ratings were made, the judges were asked to rate the same teachers using the same method as before, in acquaintance, i e., placing the person best known to the rator first, and the one least known last. The judges were also requested to answer a questionnaire, for the purpose of ascertaining the qualitative aspects of the data, such as the kind of criteria used in making ratings, etc. Statistical treatment of the first set of ratings the second set, the ratings in acquaintance and the answers to the questionnaire enabled the writer to answer either provisionally or definitely the following questions.

1 Which of the traits chosen lend themselves best to objective evaluation?

2 Do the traits tend to retain their positions for relative objectivity in different groups?

3 To what degree can we reduce the differences in criteria employed?

4 What is the effect of acquaintance upon variability of trustworthiness of judgment?

5 What is the relation between judicial capacity and judicial consistency?

6 What is the effect of more than one trial of a rator's judgment upon judicial agreement?

7 What is the relation between judicial capacity and official position?

8 Does the distribution of judicial capacity tend to be specific or general? These questions will be dealt with separately in brief form. For a full account of statistical treatment employed and unabridged tables reference should be made to the original manuscript ⁴

1 Relative Objectivity of Personal Traits

The degree of objectivity of a personal trait may be determined by ascertaining the degree of group agreement, that is, the greater agreement there is among competent judges in assigning positions to subjects (independently of each other), the more objective is the trait ⁵. Table I gives the average coefficients of group agreement of the eleven personal traits arranged in order of magnitude. Each average coefficient of agreement ($\text{Av } \bar{r}_{pq}$) is the average of the average co-

⁴ *Ibid*

⁵ Hollingworth, H. L., "Experimental Studies in Judgment" *Archives of Psychol*, No 29, pp 116, 118

efficients of correlation between any judge, p, and any other judge, q, in each of the six groups of judges. In other words, each of the \bar{r}_{pq} 's in Table I is the average of 68 inter-correlations. The P E's are simply the variability coefficients for the six separate groups. The \bar{r}_{pq} , the coefficient of agreement for each of the six groups of judges, not shown in Table I, was obtained by the use of the following formula devised by Professor T. L. Kelley, which makes possible the computation of a group correlational index directly without the need of computing the individual correlations between the judges

$$\bar{r}_{pq} = 1 - \frac{M(2+4n)}{(M-1)(n-1)} + \frac{12\Sigma(SX)^2}{M(M-1)n(n^2-1)}$$

\bar{r}_{pq} is the same as Pearson's r.

TABLE I

Final Position	Name of Trait	Av \bar{r}_{pq}	P E (t-o)
1	All-Round Value to Service	603	032
2	Cooperativeness	522	013
3	Leadership	503	031
4	Effort	491	026
5	Understanding of Children	472	036
6	Professional Interest & Growth	470	027
7	Appearance	460	016
8	Tact	453	033
9	Punctuality	408	040
10	Judicial Sense	335	027
11	Counteracting Factors	251	038

Final Position of Traits with their average coefficients of agreement and unreliabilities of these averages

Table I shows that "All-Round Value to Service" holds first place in objectivity, it has the highest coefficient of intra-group agreement. Judging from its P E, its position in the series of eleven is relatively secure. It is to be noted that this is a complex or composite trait, the others are elemental or specific traits. This result, i. e., the greater objectivity of a composite trait as compared with specific traits is in accord with Wells' findings in his study of literary merit.⁶ "Counteracting Factors," the condition that tends to lower the efficiency of a teacher, retains the last position with about as high a

⁶ Wells, F. L., "A Statistical Study of Literary Merit," *Archives of Psychology*, No. 7, p. 14

reliability as "All-Round Value" retains first place "Appearance" occupies a surprisingly low position. If we consider the fact that the correlation between two intelligence tests, one verbal and one non-verbal, is about .60, we can see the significance of the first three coefficients in Table I as regards objectivity.

2 Intergroup Agreement as to Relative Objectivity

If we examine the differences between the coefficients together with the P.E.'s in Table I, we can see at a glance that the relative positions of the traits are insecure. This is, of course, due to the lack of a high degree of agreement between the six schools as regards the relative objectivity of these personal traits. Table II illustrates this fact more clearly. Here we see that the relative positions of the traits obtained from the coefficient of agreement in each school are by no means alike in each of the schools. Traits numbered 5, 7,

TABLE II

Trait No	Trait	Sch I	Sch II	Sch III	Sch IV	Sch V	Sch VI			
1	Appearance	9	7	4	4	5	8	6	5	
2	Tact	6	3	7	10	10		3	5	
3	Punctuality	2	2	9	9	9		10		
4	Effort	5	6	8	1	4		8		
5	Judicial Sense	8	10	11	11	7		9		
6	Leadership	3	5	9	3	6	2	5		
7	Cooperativeness	3	5	5	1	4	5	5	3	5
8	Prof. Int. and Growth	10	8	2	7	3		6	5	
9	Understanding of Children	7	4	10	2	5	6	1		
10	Counteracting Factors	11	11	6	8	11		11		
11	All-Round Value	1	1	5	2	5	1	2		

Relative positions of traits obtained from the coefficients of agreements in each school (1 indicates highest \bar{r}_{pq} and 11 indicates lowest \bar{r}_{pq})

10 and 11 hold their positions with a fair degree of consistency in that they at least stay in either the upper or lower half of the series of eleven. The positions of the rest of the traits are quite inconsistent. Trait No. 3, "Punctuality," occupying 2d position in the first two schools and 9th and 10th positions in the last four schools, shows indications that a variable operating in Schools I and II, in a manner opposed to that in Schools III, IV, V and VI is tending to make this trait objective in the former case, and subjective in the latter.

This disagreement with respect to objectivity among apparently homogeneous groups is of practical significance, for

should the intergroup disagreement be inherent in rating by the order of merit method, local standardizations for each group would become necessary, thus greatly diminishing the usefulness of rating. The trait that would lend itself best to objective evaluation in one school of a large school system, or in one department of a large industrial organization, would not do so in another school or in another department.

3 The Question of Criteria Employed.

Although personal interpretation by the judges cannot and possibly should not be entirely eliminated, qualitative analysis of the questionnaire which was sent to the judges showed that standardization of the environmental conditions under which judgments are made and of the elements to be included in the interpretations will greatly reduce the disagreement and inconsistencies referred to above. The following are some of the more important points noted in regard to the standardization of criteria.

A Consulting Records. in schools I and II, practically all of the judges consulted records when rating for "Punctuality," but in the other four schools only two out of 21 judges consulted records. This explains the inconsistency referred to, which tended to increase intergroup disagreement, i. e., the trait occupied a high position for objectivity where records were consulted and a low one where they were not. This factor can of course be easily either eliminated or standardized. With the trait "Professional Interest and Growth," a somewhat similar situation occurred—records of professional courses and examinations taken were consulted by some judges and not by others.

B There was an indication that conditions were more favorable in some schools than in others for making ratings in a given trait. It became evident upon reading the questionnaire answers that certain traits had greater judicial advantage in one school than in another, because one school placed greater emphasis upon the activities involving these traits than another school. In schools II and VI where the trait "Tact" held a high position with respect to judicial agreement, we find statements characteristic of these schools only, such as, "Most teachers in this school have some executive position which makes judgment in this trait fairly easy." In the trait "Effort," some schools utilized as criteria extra-school activities, and others intra-school activities. "Counter-acting Factors" suffered in practically all schools due to lack of personal acquaintance with the conditions under which the

teachers were working, and it may be for this reason that it occupies a consistently low position for group agreement. The data points quite conclusively to the fact that ratings made in a trait which is brought more vividly or prominently to the attention of raters during an entire year are subject to less intra-group disagreement than ratings made in a trait rarely brought to the attention of the raters (or playing a relatively unimportant rôle in the life of that particular school)

C It appears that simple definitions of traits should be supplemented by specific items wherever possible, in order to reduce variability in criteria. In "Appearance," some allowed for financial condition of the persons rated; others emphasized how the wearing apparel suited the subject

4 Effect of Acquaintance

If we utilize the ratings made for acquaintance as heretofore described, and let

r_{1g} = the correlation between the rankings in merit given by judge I and the sum of the rankings given by the rest of the judges of his group *excluding* his own,

r_{1a} = the correlation between the rankings in merit given by judge I and his rankings in acquaintance,

r_{1g} = the correlation between the rankings in acquaintance given by judge I and the sum of the rankings in merit given by the rest of the judges of his group, *excluding* his own,

we can partial out the factor of acquaintance and get

r_{1ga} = the correlation between the rankings in merit given by judge I and the sum of the rankings given by the rest of the judges *excluding* his own, when acquaintance is kept constant (or when the judge in question is made equally acquainted with all the subjects rated)

By this process we determined to what extent *relative* acquaintance with the subjects rated contributed to the judge's agreement or disagreement with the rest of the judges. Does he deviate more from the group arrangement in rating the subjects whom he knows least, than in rating those subjects whom he knows most? It is impossible, due to lack of space, to give here the gains or losses when acquaintance is made constant for each of the 31 judges in each of the 11 traits. A summary is given in Table III by recording for each trait the average r_{1g} 's, i. e., the average judicial agreement when the factor of acquaintance is untouched and the average

$r_{Iga's}$, i. e., the average judicial agreement when acquaintance is made constant. In the fourth column are given the differences as indicated, and in the fifth the $P E$'s of these differences. A positive difference indicates, of course, that acquaintance was a handicap or hindered judicial agreement, while a negative difference shows that relative acquaintance promoted judicial agreement, because when acquaintance was made constant (ruled out) judicial agreement was decreased.

TABLE III

Trait No	Av r_{Iga}	Av r_{Iga}	Diff ($r_{Iga}-r_{Ig}$)	$P E$ (diff)
1	602	636	+ 034	026
2	562	583	+ 021	029
3	558	541	- 017	034
4	662	648	- 014	031
5	494	445	- 049	034
6	619	583	- 036	031
7	642	627	- 015	024
8	581	574	- 007	031
9	592	630	+ 038	036
10	400	385	- 015	038
11	661	651	- 010	024

Showing the influence of relative acquaintance in hindering or promoting judicial agreement in a given trait

It will be noticed that although there are 8 negative differences, the unreliabilities are so large (exceeding the differences in 7 out of the 11 traits, and in the remaining 4 being slightly less than the differences), we can fairly safely conclude that the influence of the difference in relative degree of acquaintance on judicial agreement in situations such as were here encountered is negligible.

This inability of acquaintance to definitely either raise or lower judicial agreement becomes evident upon considering the several ways in which this factor may operate. For although lack of acquaintance with one or several subjects may result in chance ratings, thus lowering the correlation between the unacquainted judge, and the rest of the group of judges, intimate acquaintance between a rator and subjects may also lower the correlation between the intimately acquainted judge and the rest of the judges, by the exercise of prejudice either due to friendship or to the discovery of peculiarities in the subject which are particularly abhorrent to the rater (and probably unknown to the less intimately acquainted judges). The positive and negative influences would then in the long run tend to balance each other, that is, acquaintance would have little or no effect. This finding is, of course, no argument in favor of phrenology, because zero acquaintance was

eliminated before we started by choosing as judges those who had at least some acquaintance with the subjects. The discussion refers only to relative degree of acquaintance and not to a total lack. This result should also not be confused with what we said about the advantage that the familiar trait has over the non-familiar one as regards group agreement or objectivity—"trait acquaintance" is an important factor, but relative degree of personal acquaintance with subjects is not.

5 Judicial Capacity and Judicial Consistency

Judicial capacity is determined by the degree of the judge's agreement with the competent group of judges of which he is a member.⁷ Judicial consistency is determined by the degree of agreement between two or more independent ratings made by a judge.⁷ No demonstrable relation was found between the former and the latter. This relation was investigated with the intention of ascertaining whether it was possible to determine capacity from consistency, since the latter can be determined so quickly and easily. The results were negative.

6 The Effect of More Than One Trial.

Contrary to expectations, we found that intercorrelating the average rank for two trials given by each judge with the average rank for two trials given by each of the other judges did not increase group agreement or objectivity. As a matter of fact, the latter was lowered in some of the traits. In rating by order of merit, it seems that increasing the sampling of measures (ratings) does not bring us nearer the truth, i. e., increase group agreement.

7 Judicial Capacity and Official Position

Table IV enables us to approximate roughly the relation

TABLE IV

Official Position	No	Position for Judicial Capacity						Total for 1st 3 pos		Total for Pos below the third	
		1		2		3					
		No	%	No	%	No.	%	No.	%	No	%
Principals	4	1	25	2	50	1	25	4	100	0	00
Assis. Principals	10	3	30	3	30	2	20	8	80	2	20
Teachers	17	2	12	1	6	3	18	6	36	11	64

Showing the relation between official position and judicial capacity

⁷ Hollingworth, H. L., op cit, p. 109, "Vocational Psychology," pp 158, 159, 167 (New York, Appleton, 1916)

between official position and judicial capacity. In interpreting the table the percentages of principals, assistant principals and teachers occupying the first three positions in judicial capacity should be compared with the percentages occupying positions below the third. The superiority of the principals and assistant principals as judges is clearly seen. The assistant principals are better judges than the teachers, and the principals are better than the assistant principals. It is, of course, unsafe to make a positive conclusion from the small number of representatives that we have in each official position, but the indications are that executives and supervisors make better judges than associates or, stated differently, there is a positive correlation between official position and judicial capacity.

8 Distribution of Judicial Capacity

The distribution of judicial capacity for each judge was investigated by plotting the number of times every judge occupied each designated position for judicial capacity in his group. In Fig. (a) below, we see that Judge A of School I, where there were 5 judges and, therefore, 5 possible positions for judicial capacity in each of the 11 traits, occupies the 5th or last position in 10 out of the 11 traits. He is, therefore, a generally poor judge, in Fig. (b) we see a generally good judge (but not so marked as the former)

Fig. (a) Position-Judge A-School I	Fig. (b) Position-Judge C-Sch II	Fig. (c) Position-Judge D-Sch IV
1st	1st *****	1st *
2nd	2nd *****	2nd ***
3d	3d *	3d *
4th *	4th	4th **
5th *****	5th	5th **
		6th *
		7th *

On the other hand, in Fig. (c) we see a judge occupying every possible position in his group (there were 7 judges in this school, and, therefore, 7 possible positions), from best to poorest. Figures (a) and (b) are illustrations of general judicial capacity, and (c) illustrates specific judicial capacity. If we plotted all the graphs here for each of the 31 judges, it would be seen that, roughly speaking, figures of type (a) and (b) are predominant, indicating that there is a greater tendency toward a general distribution of judicial capacity for each judge than toward a specific distribution. There is,

however, hardly strong enough evidence to conclude that if a person is a good judge in one trait he will be an equally good judge in another trait. There is an indication that he will. This relation is somewhat similar to the one found with intelligence—the preponderance of a general distribution of intelligence over a specific distribution