PROCEDURE FOR SCORING AN INTEREST TEST[1]

By EDWARD K. STRONG, JR.

Professor of Psychology, Stanford University

An interest blank consists of many items, to each of which the subject responds by indicating whether he likes or dislikes it. Freyd's blank called for two degrees of liking, two of disliking, and a measure of indifference. The interest blanks of Cowdery and of the author have provided for only three possible reactions, namely, liking, indifference, and disliking.

For example, to the first item on the Vocational Interest Blank, which is "*actor (not movie),*" the subject is asked to respond by circling one of the three letters, *L, I,* or *D,* meaning thereby that he would like, be indifferent to, or dislike that type of work. Of 575 engineers, 52 circled *L,* 172 circled *I,* and 351 circled *D.* We have records from 3,071 men, typical of 27 professions and occupations, as well as from 849 college seniors, making a total of 3,920 records. This group we refer to as our "*men in general*" group. There is no claim that it is a good sampling of the men in the United States, but it is the best approximation to such a sampling

TABLE 1. REACTIONS OF "MEN IN GENERAL" AND ENGINEERS TO FIRST FIVE ITEMS ON VOCATIONAL INTEREST BLANK

First Five Items on Vocational Interest Blank	Per cent of "Men in General" who like, are indifferent to, and dislike these items			Per cent of Engineers who like, are indifferent to and dislike these items			Differences in per cents between Engineers and "Men in General"			Scoring Weights for Engineering Interest		
	L	I	D	L	I	D	L	I	D	L	I	D
Actor (not movie)	25	34	41	9	30	61	—16	—4	20	—6	—1	4
Advertiser	32	39	29	13	38	49	—19	—1	20	—6	0	4
Architect	42	37	21	57	32	11	15	—5	—10	3	—1	—4
Army Officer	26	29	45	32	32	36	6	3	— 9	1	1	—2
Artist	33	37	30	29	38	33	— 4	1	3	—1	0	1

[1] Read at the annual meeting of the National Vocational Guidance Association, Atlantic City, Feb. 20-22, 1930.

that we have. Among these 3,920 men, 980 circled *L*, indicating thereby they would like the work of an actor, 1,333 circled *I*, and 1,607 circled *D*.

Having obtained similar totals for the three possible reactions to the 420 items on our interest blank, the next step has been to reduce such figures to percentages. Examples of such percentages are given in Table 1, based on the reactions of "*Men in General*" and engineers to the first five items on the Vocational Interest Test.

The differences between the reactions of engineers and "men in general" are similarly given in Table 1. These figures could be used for scoring the blank for engineering interest.

The scoring weights, however, are obtained by us in another manner through the use of the following formulæ, advocated by T. L. Kelly and found by K. M. Cowdery and the author to be extremely valuable in this connection.

$$\text{Weight} = 10\,\frac{\varphi}{(1-\varphi^2)\sigma} \quad \text{where } \varphi = \frac{ad-bc}{\sqrt{(a+c)(b+d)(a+b)(c+d)}}$$

$$\text{and } \sigma = \tfrac{1}{2}\sqrt{(a+c)(b+d)(a+b)(c+d)}$$

When all data are reduced to percentages a simplification of the formulæ is possible, reducing the above to:

$$\text{Weight} = 10\,\frac{\varphi}{(1-\varphi^2)\sigma} \quad \text{where } \varphi = \frac{a-c}{\sqrt{(a+c)(b+d)}}$$

$$\text{and } \sigma = \tfrac{1}{2}\sqrt{(a+c)(b+d)}$$

Furthermore it is possible to construct a table from which all weights may be read directly as soon as the two percentages represented by "a" and "c" are determined. For example, by looking on the chart for the weight opposite the two percentages of 25 and 9 (representing likes of "*men in general*" and engineers) the weight of 6 is obtained. This short cut saves an enormous amount of labor.[1]

[1] It may well be argued that reducing all data to percentages and thus permitting many short cuts is to be questioned since the factor of size of population is disregarded. Two defenses may be offered. Each occupational scale is based upon a population deemed large enough for our purposes. At first a population of 100 was considered ample. Later on it was found that this was not large enough, and since then all scales have been based upon at least 250 carefully selected cases. It is probable that scales and norms based on 500 cases would be a little more accurate, but the resulting slight

The data concerning any item on the blank is in the form of a six-fold table, thus, for the item actor:

Men in General		Men of Occupation in question	
Per cent liking	.25	Per cent liking	9
Per cent indifferent	.34	Per cent indifferent	.30
Per cent disliking	.41	Per cent disliking	.61

These must be reduced to a four-fold table in order to use the above formulæ. Cowdery accomplished this by dividing the indifferences among the likes and dislikes. Since in a few cases this did not give appropriate scores, the writer has followed the procedure of combining indifferences and dislikes together when calculating the weights for likes, combining likes and dislikes together when calculating the weights for indifferences, etc.

The calculation of the weight on the engineering interest scale for *liking* the item *"actor (not movie)"* is as follows:

$$\frac{a\ .25 : .75b}{c\ .09 : .91b} \varphi = \frac{.25 - .09}{\sqrt{(.25+.09)(.75+.91)}} = \frac{.16}{.7512} = .213$$

$$\text{Scoring weight} = 10 \frac{.213}{(1-.213^2).3756} = 6$$

The weight is multiplied by 10 and the nearest whole number is taken in order to eliminate the use of decimals. A *plus* or *minus* sign is assigned to the weight depending upon whether the per cent of the occupational group in question is greater or smaller than the corresponding per cent for men in general. In this case a *minus* sign is added since engineers like *"actor"* less than men in general.

The weights are a mathematical expression of two factors: First, the extent to which the data in the four-fold table differentiate

---

gain is hardly to be justified in terms of the increased cost and the limits of our financial resources. The second defense has reference to the relationship between sampling and reliability. If our *"men in general"* group were a sampling of men in the United States, then undoubtedly, as the size of this group increased, the reliability of averages would increase; but our *"men in general"* group is merely a sum of all our cases, and the occupational groups that constitute the whole have been selected because of ease of securing the data, ease of setting up an objective criterion as to who belong to the occupation, and interest of college men in the occupation. Under such conditions there is no way of proving that 10,000 cases is a better sampling of men in general than 4,000 cases, to say nothing of the possible chance that it is a worse sampling.

between engineers and *"men in general"* (represented by the numerator of the equation); and secondly, by the extent to which the data might be the resultant of chance (represented by the denominator of the equation). The more, then, the data differentiate the two groups and the less likely the data are due to mere chance, the larger the weighting, and vice versa.

Furthermore, the whole procedure sets up standards for vocational selection and guidance on the basis of the *differences* between men in general and men in a particular occupation. The usual procedure has been to measure traits in terms of their *maximum*. It would seem that absolute amounts of this and that trait are not so important in this connection as the fact that a person is superior in those traits, which the members of the occupation he wishes to enter are superior in, and is inferior in those other traits in which the occupational group are inferior to the general run of mankind. It seems possible that if this principle were applied, some of the attempts to develop vocational aptitude tests, which have not turned out successfully, might be found to be more useful. Attempts to handle our data in the customary manner have given us very indifferent results, but when the same data are handled in terms of differences between men in general and the occupational group in question, satisfactory differentiation between groups has been obtained.

A Second Scoring Method

M. Freyd and several other investigators have used another method. They calculated the significance of the difference by dividing the difference by the standard error of the differences. The formulæ are:

$$\text{Significance of difference} = \frac{\text{Difference}}{\epsilon_{12}} \quad \text{where} \quad \epsilon_{12}^2 = \frac{p_1 q_1}{n_1} + \frac{p_2 q_2}{n_2}$$

Applied to the example used above we have:

$$\epsilon_{12}^2 = \frac{.25 \times .75}{3920} + \frac{.09 \times .91}{575} = .00019 \qquad \epsilon_{12} = .0138$$

$$\text{Significance of difference} = \frac{.16}{.0138} = 11.6$$

Instead of using such weights, Freyd adopted the rule that if the weight exceeded 2, the item would be given a weight of 1; if it

did not exceed this amount the item was weighted 0. Whether the weight of 1 was to be called +1 or —1 was determined in the manner used by us and described above.

As far as is known, all who have calculated weights in terms of the standard error of the difference, have followed Freyd's procedure of reducing all weights to +1, 0, or —1 and no one has made use of the weights themselves as obtained. How valuable they may be in this connection remains to be determined. In all probability, the reason that this procedure of Freyd has been followed, instead of using the weights as calculated, has been due to the small populations upon which the weights were based. If instead of populations of 3,920 and 575, populations of 100 and 100 were used in the example above, then the weight turns out to be 0.98 instead of 11.6, which, being smaller than 2, would be rated as 0. Unless relatively large population groups are used, very few weights above 2 will be obtained by this method.

A Third Scoring Method

The scoring of the interest blank with our procedure is a laborious task, since it must be scored separately for each occupational rating. We have spent a great deal of time and money endeavoring to discover worth while short cuts. Thanks to Mr. P. J. Rulon, a method has been developed of utilizing the Hollerith machine whereby the blank may be scored for twenty occupations in considerably less time than an hour. This represents a real contribution to the scoring of the blank, but it is not representing a change in the scoring method itself.

Our experiences with a third method of determining weights and of scoring the blank is worth reporting, since it throws further light on the method employed by Freyd and, in addition, focuses attention upon certain statistical difficulties that may not have been appreciated by many.

In order to obviate the labor of handling weights ranging from +30 to —30, we tried out the plan of considering all weights of +2 or more as +1, all weights of +1, 0 and —1 as 0, and all weights of —2 or less as —1, thus following Freyd to this extent. Several occupational scales were constructed on this basis and many blanks scored thereon. Correlations were then obtained between scores on the new scales referred to here as "*unit*" scales and the regular scales, referred to here as "*weighted*" scales. (See Table 2.)

TABLE 2. CORRELATION COEFFICIENTS BETWEEN "UNIT" AND "WEIGHTED" SCALES

	Based on data from 83 Lawyers	45 C.P.A.	50 Architects
Law unit vs. law weighted scales.....	.985	.971	.982
CPA unit vs. CPA weighted scales.....	.87	.962	.936
Architecture unit vs. Arch. weighted scales...	.892	.837	.905

The average of these correlations, in Table 2, is .927. From such figures one would naturally conclude that unit scales can be safely substituted for weighted scales.

To make sure that the selection of groups of about 50 blanks from these three occupations were truly typical of all our blanks in these groups, approximately an equal number of additional blanks were scored on both scales, with these results:

Law unit vs. law weighted scales, 83 lawyer blanks.....	r = .985
Law unit vs. law weighted scales, 190 lawyer blanks	.980
CPA unit vs. CPA weighted scales, 45 CPA blanks.....	.962
CPA unit vs. CPA weighted scales, 90 CPA blanks.....	.960
Architect unit vs. Arch. weighted scales, 50 architect blanks.....	.905
Architect unit vs. Arch. weighted scales, 103 architect blanks.....	.912

That there is very close agreement between these two scales is further shown when only the odd-numbered items are utilized with the unit scale, and scores obtained in this way are correlated against scores from the weighted scales using all the items on the blank. The correlations are given in Table 3.

TABLE 3. CORRELATIONS BETWEEN WEIGHTED SCALES AND ODD ITEMS OF UNIT SCALES

	Based on data from 83 Lawyers	45 C.P.A.	50 Architects
Law weighted vs. law unit (odd) scales.....	.96	.922	.944
CPA Weighted vs. CPA unit (odd) scales....	.803	.882	.824
Arch. weighted vs. Arch. unit (odd) scales...	.819	.702	.784

Although the average of these nine correlations is .849 and somewhat less than .927 from Table 2, where all the items are used with both scales, nevertheless the correlations are high enough to place genuine confidence in the unit scales as substitutes for the weighted scales.

When reliability of the two scales is considered, we find that the unit scales give slightly higher reliability based on the odd-even technique, an average of .839 in contrast with .799 for the weighted scales. The details are given in Tables 4 and 5.

TABLE 4. RELIABILITY OF WEIGHTED SCALES BASED ON CORRELATIONS BETWEEN ODD AND EVEN ITEMS (BROWN'S FORMULA)

	Based on data from		
	83 Lawyers	45 C.P.A.	50 Architects
Law (odd) vs. Law (even)	.907	.894	.897
CPA (odd) vs. CPA (even)	.635	.807	.638
Arch. (odd) vs. Arch. (even)	.827	.798	.792

Average of nine correlations is .80.

TABLE 5. RELIABILITY OF UNIT SCALES BASED ON CORRELATIONS BETWEEN ODD AND EVEN ITEMS (BROWN'S FORMULA)

	Based on data from		
	83 Lawyers	45 C.P.A.	50 Architects
Law unit (odd) vs. Law unit (even)	.936	.887	.912
CPA unit (odd) vs. CPA unit (even)	.755	.871	.846
Arch. unit (odd) vs. Arch. unit (even)	.818	.765	.760

Average of nine correlations is .84.

One would naturally assume that the unit scales could be substituted for the weighted scales, since the two correlate .93 and the unit scales have slightly higher reliability, but such is not the case. The unit scales do not differentiate occupational groups from one another as well as do the weighted scales. The data are set forth in Table 6.

Ten per cent of architects rate A on the lawyer scale, that is, have the interests of lawyers, when the weighted scale is used; whereas thirty per cent are so rated when the unit scale is employed. In this case there is three times as much overlapping when the unit scale is used as when the weighted scale is utilized. Judging from the average of these seven comparisons, overlapping is increased more than three times by the unit scales in contrast with the weighted scales, when measured by A ratings. Furthermore, twice as many men will be definitely rated as not belonging to a group to which they do not belong when the weighted scales are

TABLE 6. EXTENT TO WHICH MEMBERS OF OCCUPATIONAL GROUPS ARE RATED A, B, AND C IN INTEREST IN LAW, ARCHITECTURE, AND PUBLIC ACCOUNTING, DEPENDING UPON WHETHER THE WEIGHTED OR UNIT SCALES ARE USED.

Occupational Group	Rated as to Interest in	Weighted Scale			Unit Scale		
		A Exceed -1 Q	B Bet.-1 Q & -3.5 Q	C Below -3.5 Q	A Exceed -1 Q	B Bet.-1 Q & -3 Q	C Be ow -3.5 Q
50 Architects	Law	10	72	18	30	50	20
45 C.P.A.'s	Law	31	65	4	42	56	2
45 C.P.A.'s	Architecture	2	20	78	9	51	40
83 Lawyers	Architecture	2	25	73	10	52	38
50 Architects	Accounting	2	38	60	24	48	28
83 Lawyers	Accounting	7	65	28	32	64	4
58 Journalists	Accounting	0	47	53	29	55	16
Average		7.7	47.4	44.9	25.2	53.7	21.1

used as when the unit scales are used. Such results eliminate the unit scales from further consideration.

Many are urging today that in the absence of an adequate method of measuring validity one may accept a test that gives high reliability. Our experience, as outlined here and in other connections, shows that two systems of testing may correlate over .90 and have equally high reliability and yet one may have much higher validity than the other.

OTHER POSSIBLE CHANGES IN SCORING TECHNIQUE

There are other possible short cuts in scoring the interest blank. A preliminary sampling suggests the possibility of disregarding items that are marked as indifferent. When this is done the separation of occupational groups was increased in four cases and decreased in four other cases. We will know the answer to this possibility in another month. If the change can be made it will cut the time of scoring the blank about one-third.

Very few items are weighted over +13 or under —13 on our occupational scales and most of these are for interests intimately associated with the occupation in question. Thus the item "*liking to be a clergyman*" is weighted +30 on the minister scale, the item "*liking to be a C.P.A.*" is also weighted +30 on the C.P.A. scale,

etc. Whether reducing the weights on these items will increase or decrease differentiation of occupational groups is not yet known. If the loss here is slight it probably will be well arbitrarily to limit weights between these limits of +13 and —13, as a vocational guidance test should not penalize lack of knowledge too heavily. As it is, a young man who should go into public accounting, but has never heard of it would probably mark the item C.P.A. as indifferent instead of liking, with a resulting loss of 39 points. (Being indifferent is weighted —9). This seems to be too heavy a penalty. A procedure which is statistically sound and of value in separating adult members of occupational groups may not be desirable from the standpoint of a test designed to study young men who do not yet know what careers to enter, nor very much about the possible careers themselves.

Two graduate students at the University of Minnesota, Miss Isabelle Rosenstein and Mr. Edward Birnberg have experimented with weights ranging from 1 to 9 instead of our weights from +30 to —30. They claim very high correlations between their results and ours. Whether as good separation can be obtained in this way is another problem we hope to solve within the next few weeks.

Conclusions

What is needed first of all is a scoring method that gives high reliability. The method employed by the author gives reliability coefficients between .75 and .90 depending upon the occupational scales used, the occupational groups scored, and methods of measuring reliability, whether the odd-even technique or that of repeating the test upon the same individuals after intervals of time as long as a year and a half.

What is needed second is high validity. There are two different measures of validity in this connection. First, the degree of separation obtained between occupational groups. It is this measurement of our test which we consider of prime importance. So far we have found no variation from our present procedure that does not lower the test's validity.

A second measure of validity has reference to the degree to which the ratings given young men agree with their subsequent careers. There has not yet been time enough to establish the degree of this second measure of validity. In a follow-up of Stanford 1927 seniors two years after taking the test, we found that 50 per cent

of those who were sure they had settled their life work were engaged in the work on which they rated highest on the test, and 71 per cent were engaged in the work on which they rated first or second highest. There were, however, 12 per cent who had entered an occupation in which, according to the test, they would not be interested. There is no doubt that many of these young men will change their plans in the years to come. Whether they will make changes in the direction of the test scores or not remains for the future to disclose.