

AN OBJECTIVE PEER EVALUATION SCALE: CONSTRUCTION AND VALIDITY*

E. L. Hoffman

and

J. H. Rohrer

Urban Life Research Institute

Tulane University

AD No. 23628
ASTIA FILE COPY

The evaluation of Officer Candidates through the use of ratings made by the candidates' peers was demonstrated in military research of World War II, to be both a reliable and valid technique, but one that needed further refinement if it was to have general utility (cf., 1, 6, 8, and 10). Previous studies

*The data on which this report is based were collected by the Neuropsychiatry Branch, Bureau of Medicine and Surgery, Department of Navy, with the cooperation of the Basic School, Marine Corps School, Quantico, Virginia. The experimental design and data analyses were done by the writers. Recognition should be accorded to LCDR J. W. Bagby, Jr. and LCDR R. S. Herrmann of Navy Neuropsychiatry, and Col. J. G. Bouker and Lt. Col. H. B. Smith of the Marine Corps, for their collaboration and assistance in carrying out the study. This is one of a series of technical reports made under Contract NR 151-152 between Tulane University and the Office of Naval Research. The opinions expressed herein are those of the authors and do not necessarily reflect the opinion of the sponsoring agency, the U. S. Department of Navy, or the cooperating agency, the U. S. Marine Corps.

have employed nomination or candidate ranking by peers. In those studies any differences in the characteristics of the reference groups, (e.g., group size, general level of excellence, or group heterogeneity), seriously impaired

intergroup comparisons and thus limited the generality, and hence the usefulness, of the technique.

The present report presents a study of peer ratings obtained by having the raters evaluate their peers on objective criteria of officer excellence. These criteria, when standardized, do not have all of the technical limitations observed in previous studies.

PROCEDURE AND RESULTS

Subjects. The subjects in this study were enlisted Marines who were being screened for officer commissions in four-week screening courses conducted at Quantico, Virginia. Four different groups were involved in the study. Group I consisted of 518 candidates who were screened in 1951. Group II consisted of 172 candidates screened early in 1952, and Group III consisted of 142 candidates screened late in 1952. Group IV consisted of 145 candidates screened early in 1953.

The original group of 518 candidates was homogeneous in that all of the members had had experience serving as enlisted men in the Marine Corps. The average candidate had three and one-half years such service and, therefore, should have had sufficient experience to make valid judgments of desirable officer qualities. The modal service rank of the men was that of Sergeant. The average GCT score for the group placed them in the upper 7 per cent of the draft population. They had been carefully screened through recommendations before arriving at the final screening course. Those recommendations came first from the Field Commanders and then from the Commandant of the Marine Corps. The candidates were therefore, a highly select, homogeneous group of enlisted Marine men. Groups II, III and IV had gone through the same screening procedure but did not have as much previous experience in the Marine Corps as did Group I.

Throughout the screening course, the men were organized into platoons of about 45 men each. The platoons, in turn, were divided into three sections of approximately 15 candidates each. A complete description of the organization of the screening course and of the first candidate population used in this study is to be found elsewhere (7). Because of the nature of the situation the candidates' serious cooperation in the study was readily obtained.

Development of the Preliminary Scale. In the fourth week of the screening course, each candidate in group I was asked to rank, on the basis of excellence as an officer, every man in his platoon. Each individual included himself in the ranking. After they had placed the men in the section in rank order, they then wrote a descriptive paragraph about the five men they had ranked highest and the five men they had ranked lowest. They were instructed to state in the paragraph why they ranked the men as they did. Ten paragraphs from each of the 518 candidates provided a substantial body of information from which four groups of paragraphs were chosen. Those groups consisted of the paragraphs written about the candidates who had been ranked in the first position, in the fifth position, in the position fifth from last, and in the last position in the section. The last man had the rank of 13, 14, or 15, depending on the size of the section.

The four groups of paragraphs chosen were subjected to content analysis in which the content categories used were those described by White (9). The items in the descriptive paragraphs consisted of descriptive phrases or sentences which attributed to the men some personal characteristic that had been responsible for his being in the position ranked. These items were then assigned to White's value categories. Thirty-one different categories of values were identified.

The categories are listed in Table 1, and the percentage of statements occurring in each category that referred to the positive presence of the characteristic is also listed in that table for each group. A similar presentation is made in Table 2 for the occurrence of statements that refer to the absence of the characteristics.

Graphs were made showing the percentage of positive and negative mention of each characteristic for each of the four groups of candidates. From those graphs, 63 of the original statements were selected from the categories that showed, in going from the group ranked first in the section to the group ranked last, either a decrease in the number of times the presence of the characteristic was used to describe the candidate, or an increase in the number of times the absence of the characteristic was used to describe the candidates. These 63 items made up the preliminary scale. The instructions for using this scale requested each candidate to indicate whether he (a) strongly agreed, (b) agreed, (c) disagreed, or (d) strongly disagreed, with each statement as it pertained to the fellow candidate(s) he was evaluating.

This scale was administered to Group II composed of the 172 Officer Candidates in the screening course conducted at the Basic School in Quantico, Virginia, during the summer of 1952. The average time required by a candidate to evaluate approximately 14 fellow candidates using this preliminary scale was 125 minutes, with 180 minutes required for the slowest rater.

Refinement of the Preliminary Scale. From the preliminary scale a set of items was selected to form a one-hour test for predicting the final standing assigned to the candidates by their platoon officers. The platoon officers were line officers assigned to the platoons for the purpose of evaluating the candidates for recommendations for commissions at the end of the four-week screening course.

TABLE 1

Per cent of Description Mentioning the
Presence of Trait for Candidates in Ranks 1, 5, 11, and 15

Characteristics	<u>Rank</u>			
	<u>1</u>	<u>5</u>	<u>11</u>	<u>15</u>
Dominance	64.84	26.83	3.38	2.39
Determination	23.63	25.10	7.62	1.52
Intelligence	56.56	39.82	22.88	12.60
Works	13.93	10.82	8.05	2.82
Recognition	13.33	6.92	1.27	0.00
Appearance	46.06	20.77	7.20	1.95
Achievement	6.66	6.92	0.84	1.73
Self respect	3.43	2.59	7.62	1.52
Aggression	12.32	6.92	2.11	1.30
Emotional security	7.27	17.31	3.38	0.65
Knowledge	8.88	3.03	0.42	0.00
Obedience	7.27	3.03	0.00	1.08
Pleasant personality	12.52	8.22	7.62	1.08
Manners	6.66	3.89	0.84	0.00
Tolerance	13.73	9.09	2.02	0.65
Group unity	19.39	9.09	2.54	1.08
Strength	17.17	12.55	4.23	1.73
Adjustment	11.71	8.22	1.27	0.21
Interest	20.20	13.41	2.11	1.30
Practical knowledge	15.15	15.58	8.89	1.95
Friends	11.91	13.85	6.77	4.34
Value in general	0.80	4.32	1.27	0.00
Humor	6.86	4.32	0.00	0.00
Happiness	0.40	0.43	0.84	0.00
Morality	7.07	0.86	1.69	0.00
Truthfulness	5.05	2.16	1.69	0.00
Justice	0.60	4.32	1.27	0.00
Religion	2.10	0.00	0.00	0.00
Giving or generosity	1.61	3.03	0.42	0.00
Culture	0.60	0.00	0.00	0.00
Carefulness	0.80	2.59	0.42	0.00
Modesty	0.00	0.00	1.69	0.00
Creative	0.00	0.00	0.42	0.20

TABLE 2

Per cent of Description Mentioning the
Absence of Trait for Candidates in Ranks 1, 5, 11, and 15

Characteristics	<u>Rank</u>			
	<u>1</u>	<u>5</u>	<u>11</u>	<u>15</u>
Dominance	0.20	0.80	16.52	30.21
Determination	0.20	0.86	8.47	8.69
Intelligence	0.00	0.00	5.08	11.30
Works	0.00	0.43	6.77	7.17
Recognition	0.00	0.86	1.69	11.30
Appearance	0.00	2.59	7.52	15.30
Achievement	0.00	0.43	3.38	2.60
Self respect	0.20	0.43	0.42	0.86
Aggression	0.60	2.16	8.89	11.08
Emotional security	0.60	6.49	25.84	29.55
Knowledge	0.20	0.43	0.42	1.30
Obedience	0.00	0.43	1.27	3.04
Pleasant personality	0.00	0.43	1.27	5.86
Manners	0.40	1.29	2.96	2.39
Tolerance	0.60	2.02	12.00	10.43
Group unity	0.20	0.00	1.27	7.17
Strength	0.20	2.16	3.81	9.78
Adjustment	0.00	3.46	8.47	19.34
Interest	0.00	0.43	13.55	14.78
Practical knowledge	0.00	12.12	8.47	5.21
Friends	0.40	0.86	5.08	6.95
Value in general	0.00	0.00	0.00	0.20
Humor	0.00	0.00	1.27	0.40
Happiness	0.00	0.86	0.42	0.00
Morality	0.00	0.00	0.00	0.00
Truthfulness	0.00	0.43	0.84	0.00
Justice	0.00	0.00	0.42	0.20
Religion	0.00	0.00	0.00	0.00
Giving or generosity	0.00	0.00	1.27	1.41
Culture	0.00	0.00	0.00	0.20
Carefulness	0.00	1.29	6.35	3.63
Modesty	0.00	0.00	0.00	5.85
Creative	0.00	0.00	0.00	1.41

The first step in the selection of the set of items was to key the items for maximum validity for the criterion of final platoon standing as determined by the platoon officers. The keying of each item was done on the basis of the number of raters choosing each alternative and the mean criterion score of the candidates being rated. This method was developed by French (3) and involves keying the alternative having the largest value of $N_j(\bar{Y}_j - \bar{Y})$, where N_j is the number of testees choosing a particular alternative of item j ; $\bar{Y}_j$ is the mean criterion score for testees choosing a particular alternate of item j , and $\bar{Y}$ is the mean criterion score for all testees. The values of $N_j(\bar{Y}_j - \bar{Y})$ were converted to standard scores with a mean of zero and a standard deviation of 10. This criterion index is listed in Table 3 for each item.

On the basis of this analysis, weights were assigned to some alternatives of each item. The "strongly agree" alternative had the largest value of $N_j(\bar{Y}_j - \bar{Y})$ and was given a weight of two in every case but one. (Item 35 was eliminated from the analysis at this point because none of its alternatives were consistently related to this criterion). The "agree" alternative was related to the criterion for some items and was given a weight of one; these items are identified by an asterisk in Table 3.

A random sample of 35 individuals was then selected from the 172 men in group II. Each individual in this sample had had peer evaluation scales submitted on him by approximately 14 of his fellow candidates. For computational convenience, a random sample of 10 scales from these 14 was selected for each of the 35 candidates. The resulting group of 350 peer evaluation scales was scored according to the key which had been developed, and an average peer evaluation score was obtained for each individual.

TABLE 3

Item Analysis Data

<u>Item</u>	<u>Criterion Agreement</u> <u>Index "C"</u>	<u>Total Test Agreement</u> <u>Index "T"</u>	<u>Intra-item variability</u> <u>Index "V"</u>
1	+ 5	+ 9	+ 4
2	+25	+ 7	+ 3
3	+ 4	+ 9	-13
4*	+ 9	+ 2	+15
5	- 4	+ 1	+ 2
6*	+17	- 1	0
7*	+11	- 4	+21
8*	+ 3	- 6	+20
9*	- 3	-11	- 1
10*	+16	- 2	+ 7
11*	+ 5	- 9	- 8
12	-10	+ 3	- 6
13	- 4	+12	- 6
14	-13	+12	- 8
15	- 9	+ 7	-15
16	-11	+10	- 9
17	+ 7	+18	- 7
18*	+ 6	- 2	+16
19*	+ 2	-13	+ 5
20*	+12	- 9	-12
21	- 6	+14	+17
22	-11	+ 9	- 2
23	+ 3	+12	+ 6
24	+ 8	+12	- 4
25*	+ 9	- 9	- 8
26*	+ 7	- 6	-11
27*	+ 2	- 5	- 3
28*	+ 1	- 2	+19
29	-22	+ 2	+ 9
30	-10	-20	-17
31*	- 4	-11	- 8
32	-11	+ 5	- 2
33	-13	- 3	-16
34*	+ 3	-17	+ 1
36	-10	+ 8	+ 2
37*	+15	- 1	+14
38*	+16	- 6	+ 3
39	-18	+ 2	+11
40*	-11	-15	- 2
41	- 2	+ 8	- 7

Table 3 (Continued)

<u>Item</u>	<u>Criterion Agreement</u> <u>Index "C"</u>	<u>Total Test Agreement</u> <u>Index "T"</u>	<u>Intra-item variability</u> <u>Index "V"</u>
42	+ 5	+ 2	0
43*	- 1	- 7	+ 2
44	-10	+ 9	- 2
45*	+ 8	- 7	+ 1
46*	- 7	-16	-12
47	- 3	+ 3	+ 9
48	- 9	+ 4	+ 9
49	- 2	+16	+ 1
50	- 6	+13	-15
51	-17	- 2	+ 1
52*	+ 9	- 3	+ 9
53	- 4	+16	-15
54*	- 8	-15	- 9
55*	+ 8	-16	- 5
56	+ 1	-16	- 9
57*	- 3	-13	0
58*	+17	- 2	+20
59*	+ 2	-13	- 4
60	+ 5	+13	+10
61	- 2	+11	- 1
62	-16	+ 5	-18
63	+10	+11	+ 1

A relative measure of the agreement of each item with this total test score was then obtained for each item by calculating $N_j(\bar{X}_j - \bar{X})$ where $\bar{X}_j$ is the mean test score for candidates about whom the keyed alternative had been indicated; and where $\bar{X}$ is the mean test score for all candidates. These values of $N_j(\bar{X}_j - \bar{X})$ were converted to standard scores with a mean of zero and a standard deviation of 10 and are also listed in Table 3.

A measure of the variability of ratings on a given individual for each given item was also calculated from the data obtained from this sample of 350 peer evaluation scales. This measure consisted of the sum of the squared deviations from the mean rating of the individual on a given item. The average over individuals for these sums of squared deviations was used as an index for the precision or reliability of the item. These indexes were converted to standard scores with a mean of zero and a standard deviation of 10, and also are in Table 3.

The problem of maximizing the correlation between a criterion and the score on a subset of items of a specified size has been discussed by Gulliksen (4), and Horst (5), for the situation where the correlation of items with the criterion and the intercorrelations of the test items are considered. Both of these writers have observed that no practical analytical solution has been devised for the mathematical problems arising when rigorous solutions are attempted; they have also demonstrated that approximate graphic solutions can produce satisfactory empirical results. In the present selection problem a third index, item precision, was available for the items as a result of the same items being answered more than once for any given candidate.

Let us now consider how this index was used in conjunction with the more conventional indexes in an attempt to select the best subset of items for predicting the criterion. The following three indexes were available: C, a measure

of the relative agreement of the item with the criterion; T , a measure of the relative agreement of the item with the total test score; and V , a measure of the relative intra-item variability.

As a result of the criterion keying of the items, all items had some positive agreement with the criterion. The criterion index, indicated by C , is merely a measure of the relative agreement of the item with the criterion. The items with C indexes larger than zero are above the average in their agreement with the criterion, while those with indexes less than zero are below the average. Likewise all items had a positive relationship with the total test and items with T indexes larger than zero are above the average in their agreement with the total test. Furthermore, the intra-item variability indexes, V , that are larger than zero are above the average in variability.

In the selection of items that were relatively high in their agreement with the criterion and, at the same time, relatively low in their agreement with the bulk of the selected subsets of items, it was possible to place each item on a scatter plot with C as the ordinate and T as the abscissa. Items in the upper left-hand portion of the scatter plot were considered the most desirable items to the extent that agreement with the total of the selected subset was approximated by using the total score of all of the original items.

In order to identify items that were relatively high in their agreement with the criterion and at the same time relatively low in intra-item variability, it was possible to locate the items on a scatter plot with C as the ordinate and V as the abscissa. Again, items in the upper left-hand portion of the scatter plot were the most desirable.

In order to plot the T and V indexes so that the most desirable items would be in the upper left-hand portion of the scatter plot, it was necessary to reverse

the signs of the T indexes and locate the items on the scatter plot with T as the ordinate and V as the abscissa. For two items on this scatter plot having the same agreement with the total test, the item with the lower variability was considered more desirable.

Three points of a triangle may be located if the three pairs of indexes for any item are plotted in the manner indicated above and the three grids superimposed to form one common grid. The ordinate-abscissa labels on this grid would change for each pair of indexes. The centroid of the triangle thus formed, can be taken as the single point that best represents the triangle. If the centroid for each item is located on a scatter plot, the points in the upper left-hand portion of the scatter plot will be the most desirable items. Fortunately, the actual plotting of the three pairs of points for each triangle is unnecessary since it is possible to readily calculate the coordinates of the centroid by a pair of formulae.² The equations for the coordinates of the centroid are $\bar{X} = \frac{2V+T}{3}$ and $\bar{Y} = \frac{2C-T}{3}$.

For each of the original 63 items except item 35, $\bar{X}$ and $\bar{Y}$ were calculated and the items were located on a scatter plot. A line with slope of plus one was shifted from the upper left-hand region of the scatter plot towards the lower right-hand region. The items above this line were considered more desirable than those below the line so that the selection of n items necessitated moving the line towards the lower right-hand region until n items were above the line.

2.

These equations have this simple form because the triangle formed by the three pairs of indexes for any item is, in every case, a right triangle with one vertical and one horizontal leg. This is due to the fact that each index is used to locate two points on the triangle.

Eighteen of the first 20 items segregated by this procedure were used in the final form. The two items excepted were duplicated in content very closely by items previously selected, and were replaced with the 21st and 22nd items. These 20 items and their scoring weights are presented in Table 4.

Validation Results. The refined Peer Evaluation Scale was scored and the rank-difference correlation between the average of this score for each individual and his final platoon standing was calculated for the 35 individuals from Group II that were used in the item analysis procedure. The Rho coefficient obtained was .84. It was recognized that this coefficient was contaminated and possibly inflated. However, the 35 individuals were selected from three different platoons and any platoon differences would lower the correlation between platoon standing and the peer score.

The Peer Evaluation Scale was then used on Group III and IV, two independent samples consisting of 142 and 145 individuals respectively. These groups each contained three new platoons. The rank-difference correlations of peer scores with final platoon standing were calculated for these platoons. The results are presented in Table 5. The six Rho coefficients obtained were homogeneous and their appropriate average is .85.

Norms were determined for each of the six platoons in Group III and IV. A Chi Square test of homogeneity of the norms for these groups was not significant ($p = .51$) indicating that the norms for the various groups may be regarded as chance variations from a common set of norms. The homogeneity of these norms lends support to any generalizations regarding the relationship of the Peer Evaluation scores with inter-platoon rankings. It would, therefore, seem that one of the main limitations to peer evaluation may be circumvented by using the Evaluation Scale since scores on this scale have meaningful use beyond ranking individuals in a single platoon.

TABLE 4

PEER EVALUATION SCALE AND SCORING WEIGHTS

Item number on final scale		Strongly Agree	Agree	Undecided	Disagree	Strongly Disagree
1.	This man shows confidence in himself.	2	0	0	0	0
2.	His appearance is good.	2	0	0	0	0
3.	He shows leadership in the field.	2	1	0	0	0
4.	He has the ability to stand up under pressure.	2	1	0	0	0
5.	He takes the initiative.	2	1	0	0	0
6.	He is a fine athlete and enjoys taking part in sports.	2	1	0	0	0
7.	He is well educated.	2	0	0	0	0
8.	This man has command presence.	2	1	0	0	0
9.	He has stamina and endurance.	2	1	0	0	0
10.	His actions show that he has a familiarity with many things and situations.	2	1	0	0	0
11.	He gives orders well.					
12.	He exhibits imagination in solving problems.	2	1	0	0	0
13.	He does things to help other people.					
14.	This man thinks quickly and well in a crisis.	2	1	0	0	0
15.	He is the type of man who will carry through in any situation.	2	1	0	0	0
16.	His attitude is neither overbearing nor subservient.					
17.	He follows orders well.					
18.	He exhibits poise in most situations.	2	1	0	0	0
19.	He considers the consequences before he acts or says something.					
20.	He has shown himself to be a gentleman of high character.					
21.	He has personal pride in himself and his work.	2	0	0	0	0
22.	His decisions show sound judgment.	2	1	0	0	0
23.	He does not lose his temper.					
24.	He performs well before the group.	2	1	0	0	0
25.	He learns quickly and remembers details.					
26.	He has experience in the military line which he utilizes to advantage.	2	1	0	0	0
27.	He has good training and knows the duties and responsibilities of an officer.	2	1	0	0	0
28.	He exhibits practical judgment.					
29.	He has proven himself to be honest and dependable.					
30.	Men will follow him gladly.	2	1	0	0	0

TABLE 5

Rank-Difference Coefficients for
Correlation of Peer Score with Final Platoon Standing

Platoon	N	Rho
A	47	.88
B	49	.78
C	46	.84
D	49	.84
E	49	.90
F	47	.85

Summary

1. A Peer Evaluation Scale was described in which statements describing the peer were selected on the basis of content analyses of paragraphs describing outstanding and inferior candidates for the billet of Marine Corps Officer. The statements were written by enlisted Marines who had had an average of three and one-half years of active service in the Marine Corps.

2. The items used in the Peer Evaluation Scale were selected from the above mentioned statements on the basis of their ability to discriminate between the men ranked first, fifth, tenth, and last, in a section of approximately 15 men.

3. This preliminary scale was then administered to a second independent group of candidates and three indexes for item selection purposes were calculated. These indexes were: a criterion index which was a measure of the relative agreement of the item with the criterion; a total test index, which was a measure of the relative agreement of the item with the total test score; and an item variability index, which was a measure of the relative intra-item variability. Using these three indexes, 20 items were segregated for use in the final scale.

4. The final peer evaluation scale was then administered to an independent sample of candidates. The rank-difference correlation between the average peer

score for a candidate and his rank position in the platoon as determined by four line officers was calculated. An average Rho coefficient of .85 was obtained.

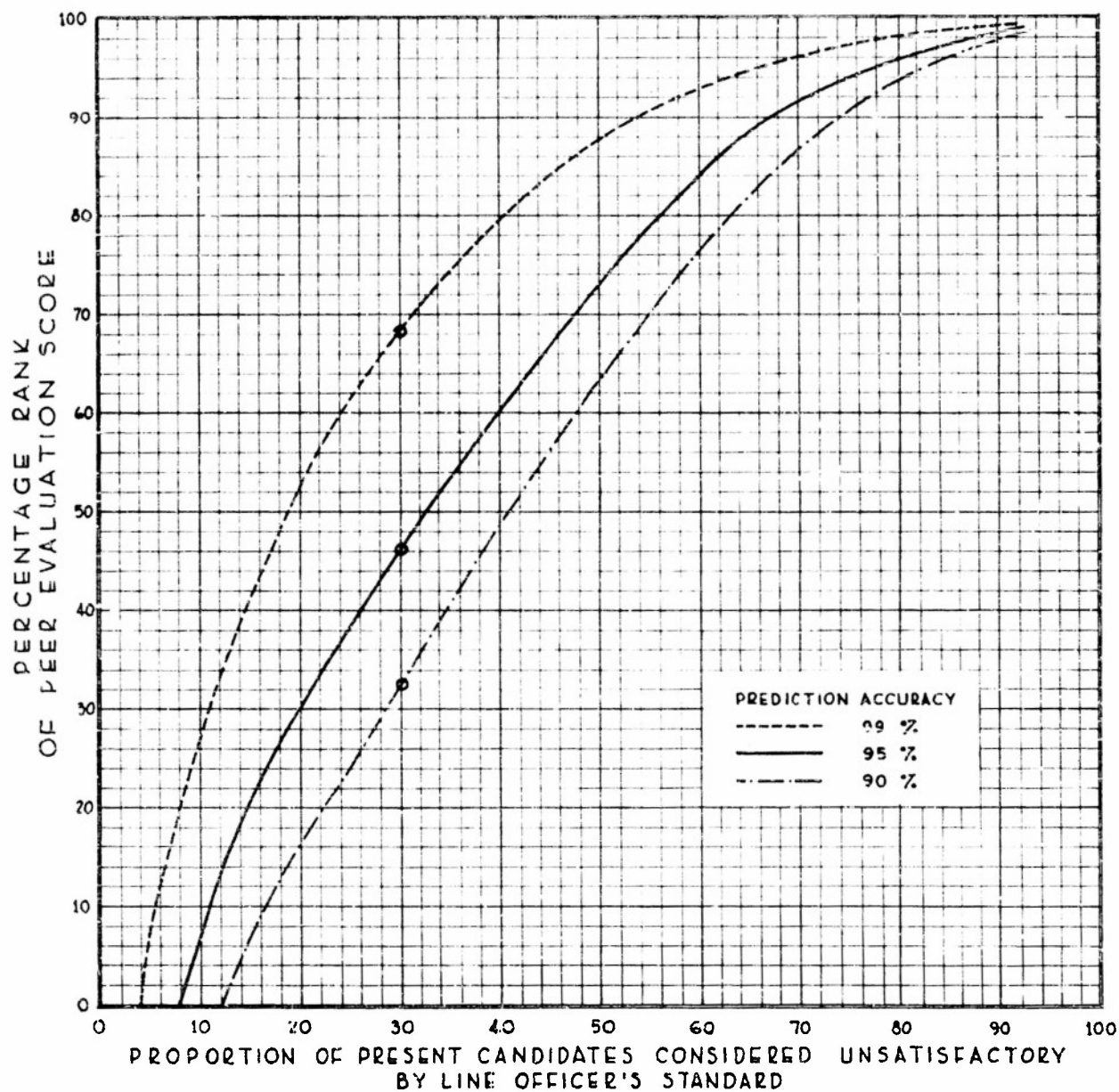
5. Norms for the three platoons were shown to be homogeneous on the Peer Scale. The finding that the norms were homogeneous is of importance in that it permits a more general interpretation of a score on the Peer Evaluation Scale. Thus, in addition to intra-group validity of considerable magnitude it is possible to generalize these findings to an inter-group situation in which the Peer Evaluation Scale scores have meaning beyond the reference group from which the score was obtained. The average rating by a candidate's peers might thus serve as a criterion for the evaluation of ratings obtained from less experienced line officers serving on screening programs. However, averages of several raters are, in general, more satisfactory (cf., 2, p. 433) and should be continued .

Recommendations

The Peer Evaluation Scale has considerable agreement with the lineal ranking of line officers under the present procedures for screening Marine Officer Candidates in the Basic School. It could serve as a valuable adjunct to the screening program if it was administered at the end of the second or third week. The results then could be used to identify those candidates with a high probability (say 95 chances in 100) of being approved by the Line Officers. Candidates with chances for satisfactory completion who do not fall in this group could then be screened more carefully by the Line Officers during the remaining portion of the screening period.

Figure 1 presents a curve which will enable one to determine cutting points, on the distribution of Peer Evaluation Scores, which will correspond with the Line Officers' standards of selection.

FIGURE 1*



*These curves were developed from tables developed by H. C. Taylor and J. T. Russell, The relationship of validity coefficients to the practical effectiveness of tests in selection; discussion and tables. J. appl. Psychol., 1939, 23, 565-578.

An hypothetical example may aid in the interpretation of the table. The Scale could be administered during the third week of the course. Should it be deemed desirable by the Line Officers to eliminate the bottom 30 per cent of the candidates in a particular screening group, one could determine from the solid-line curve in Figure 1 those candidates scoring above the bottom 46 per cent of the screening group on the Peer Evaluation Scale. These candidates would have 95 chances in 100 of being above the lower 30 per cent on the rankings by the Line Officers at the end of the screening course. The candidates comprising the lowest 46 per cent could then be subjected to a more intensive and critical screening during the final week of the screening course. If one wanted a higher level of assurance of success regarding the men who were selected on the basis of the Peer Evaluation Test score, then one could use the broken-line curve in Figure 1. Individuals with Peer Scores above this curve have 99 chances in 100 of being screened in by the Line Officer criteria. Likewise if one wished to use a less stringent criterion, the dotted line in Figure 1 could be used. This is the line that represents the 10 per cent level of confidence.

Decile norms for the combined III and IV groups are presented in Table 6. The Peer Evaluation scores listed are those that occurred at the dividing points when the scores for the combined group were arranged in order of size and then divided into tenths. Thus, from Table 6 one may see that a score of 13 corresponds to decile 4 which indicates that 40 per cent of the group had a score of 13 or less.

It is suggested that the averages of several raters be continued in use rather than a score from a single rater. Baier has suggested elsewhere that ". . . much more is gained by combining ratings made by different raters than by improving the rating of a single rater through the use of special technique." (2,p.433). The average rating by a candidate's peers might further serve as the

criterion to evaluate ratings obtained from inexperienced Line officers serving on the screening program.

TABLE 6

Decile Norms for Peer Evaluation Scale

N = 288

Decile	Peer Score
10	30
9	25
8	21
7	19
6	17
5	15
4	13
3	11
2	8
1	6

References

1. Anderhalter, O. F., Wilkins, W. L., and Rigby, Marilyn K. Peer Ratings. Technical Report #2, prepared at St. Louis University under Office of Naval Research Contract N7 onr-40802 (NR 151-092) 1952.
2. Baier, D. E. Reply to Traver's "A critical review of the validity and rationale of the forced-choice technique." Psychol. Bull., 48, 1951, 421-434.
3. French, J. W. A technique for criterion-keying and selecting test items. Psychometrika, 1952, 17, 101-106.
4. Gulliksen, H. Item selection to maximize test validity. Proceedings of the 1948 Invitational Conference on Testing Problems. Princeton: The Educational Testing Service, 1949.
5. Horst, A. P. Item selection by means of a maximizing function. Psychometrika, 1936, 1, 229-244.
6. McClure, G. E., Tupes, E. C. and Dailey, J. T. Research on criteria of officer effectiveness. HRRC Research Bulletin, 51-8, San Antonio: Lackland AFB, 1951.
7. Rohrer, J. H., Bagby, J. W. Jr., Wilkins, W. L. The Potential Combat Officer. A technical report to the Neuropsychiatry Branch, Bureau of Medicine and Surgery, Department of Navy, Washington, 1952.
8. Wherry, R. J. and Fryer, D. H. Buddy ratings: popularity contest or leadership criteria? Personnel Psychol., 1949, 2, 147-159.
9. White, R., Value analysis: A quantitative method for describing qualitative data. J. soc. Psychol., 1944, 19, 351-358.
10. Williams, S. B. and Leavitt, H. J. Group opinion as a predictor of military leadership. J. Consult. Psychol., 1947, 11, 283-292.