AD NUMBER

ADB951172

LIMITATION CHANGES

TO:

Approved for public release; distribution is unlimited.

FROM:

**Distribution authorized to U.S. Gov't. agencies and their contractors;
Administrative/Operational Use; 07 OCT 1952.
Other requests shall be referred to Office of the Adjutant General (Army), Washington, DC 20310.**

AUTHORITY

AGO per DTIC form 55

(1)

RESEARCH NOTES

Number 52-68

(11) UNANNOUNCED

7 October 1952

(12) 23

EXPLORATORY ANALYSIS OF AGREEMENTS AMONG RATERS AT DIFFERENT LEVELS AND OF RELATIONSHIPS AMONG RATINGS, RANKINGS, AND OBJECTIVE MEASURES.

PJ-4903-03.

PERSONNEL RESEARCH SECTION

PR AND P BR, TAGO

EXPLORATORY ANALYSIS OF AGREEMENTS AMONG ...RS AT DIFFERENT LEVELS AND OF RELATIONSHIPS AMONG RATINGS, RANKINGS, AND OBJECTIVE MEASURES

CONTENTS

	Page
INTRODUCTION	1
The over-all program	
Purposes of this study	
RATER-LEVEL AGREEMENT	3
Method	
Results	
RELATIONSHIP AMONG RATINGS, RANKINGS, AND OBJECTIVE MEASURES	8
Method	
Results	
SUMMARY AND DISCUSSION	16
Summary of specific results	
Implications for further research	
APPENDIX	
Table A-1. Intercorrelations among objective measures and mean total ratings and rankings for NCO ratees at Fort Campbell.	19

TABLES

Table		Page
1.	Correlations between rater levels on ratings and rankings.	5
2.	Mean rater agreement coefficients for ranked and rated characteristics.	7
3.	Correlations between rater levels on certainty variables.	8
4.	Rater-ratee groups and number of ratees in each group.	9
5.	Correlations between objective measures and means of total ratings or rankings.	11
6.	Comparison of predictability coefficients for seven ratings and rankings conceptually consistent with different objective measures.	13
7.	Comparison of deviations of predictability coefficients for seven ratings and rankings conceptually consistent with different objective measures.	14

EXPLORATORY ANALYSIS OF AGREEMENTS AMONG RATERS AT DIFFERENT LEVELS AND OF RELATIONSHIPS AMONG RATINGS, RANKINGS, AND OBJECTIVE MEASURES

INTRODUCTION

THE OVER-ALL PROGRAM

The present study is an exploratory analysis of data available for a research program titled, "Improvement in Methods of Performance Evaluation with Particular Reference to the Infantry Soldier." This program had its origins in the need for criteria against which to validate the Infantry Military Occupational Specialty tests. This led to consideration of the various characteristics and abilities required for successful performance as an enlisted infantryman. As a result, a variety of rating scales were designed to obtain evaluations on these characteristics and abilities. These scales are embodied in DA AGO PRT 1800, "Ratings for Infantrymen."

The program has much wider ramifications, however, than the need to develop criteria for a particular set of infantry MOS tests. It became clear very early in the planning of the research program that it offered an opportunity to obtain answers to many basic questions related to rating methodology and procedures. Ratings are more widely used as criteria than are other types of measures. The ratings usually used are those already available as part of the administrative procedures being followed. However, such ratings are generally of little value as criteria. For this reason, an increasing emphasis has been placed on constructing rating scales designed to serve primarily as criteria for the validation of predictor instruments. Despite the fact that many studies utilizing criterion ratings have been made, little is really known concerning the effectiveness of the techniques available or of the best and most efficient methods of obtaining such criteria.

This program provided the opportunity to obtain data to throw light on these problems. The data gathered have proved to be quite formidable with respect to both quantity and complexity. Hence, the need for an exploratory analysis so that a point of reference can be established for future, more specific studies. Depending on the results obtained with this exploratory analysis, described in the present report, it will be possible to determine the feasibility of investigating various rating problems. Such problems are listed here so that an over-all view of the contemplated scope of this program can be obtained:

1. Can ratings be improved by selection of raters? Two phases of this question are:

   a. Comparison of ratings by self, peers, subordinates, and superiors.

   b. Comparison of results obtained from raters of different levels of intelligence.

2. Can ratings be improved by obtaining additional ratings from the same rater?

3. Can ratings be improved by increasing the number of raters?

4. Can ratings be improved by training raters in what to observe?

5. Can ratings be improved by special methods of dealing with rating scores?

6. What is the stability of criterion ratings over a period of time?

a. Who are the most consistent raters as judged by (1) rate rerate, and (2) by criterion relationships at test-retest?

7. Can ratings be improved by selection of the men to be rated?

8. Which of the above methods of improving ratings is best? Which combination of methods is best? Which method is most economical?

9. What factors are emphasized by raters in different relationships to the ratee?

10. What is the effect of ratings on rankings and vice-versa?

This listing of problems suggests an ambitious research program--with no guarantee that answers will be forthcoming for all the questions posed or that such answers as do emerge will have immediate applicability in Army personnel procedures. However, it is the lack of answers to such questions that has cast doubt on rating criteria in the past, with the result that the efficacy of predictors is questionable also. The program may therefore be construed as an attempt to obtain basic information that will contribute not only to the improvement of rating procedures in the Army, but also to the improvement of rating criteria used in the validation of various official predictor instruments.

The data for the entire program were obtained in two phases. In Phase 1, carried out in November 1949, rater groups at Fort Campbell and Fort Bragg completed DA AGO PRT 1800 on selected ratee groups. Other types of data, such as test scores and personal information, were also obtained for both raters and ratees. After the ratings were completed, discussion sessions were held with certain rater groups at Fort Campbell. One session covered the duties of a light and heavy weapons infantryman and the abilities and characteristics required to perform such duties. The contributions during this discussion came from the subjects rather than the AGO representative. The latter's function was to organize this information on the blackboard and to point out to the men that they were describing the duties and abilities on which they were supposed to rate when evaluating performance as an infantryman. Another discussion session was devoted to a presentation of the principles of rating. DA AGO PRT 1631, "Rater's Guide," was used as a text. At the conclusion of the

second discussion session, the men were instructed to observe their ratees carefully during the following month, and they were also advised that they would have another opportunity to make ratings at that time.

Discussion sessions were not held at Fort Bragg, and the raters there were not informed that there would be a rerating session a month later.

One month later, in December, 1949, rerating sessions to obtain Phase II data were conducted at both installations.

The present report is concerned only with Phase I data.

PURPOSES OF THIS STUDY

Two general problems are covered in the present report. The first is to determine degrees of agreement among different rater levels for various ratings and rankings. The second is to determine what relationships exist among the various ratings and rankings and objective rating measures utilized in this study. The findings are presented in two main sections, (1) Rater-Level Agreement and (2) Relationships Among Ratings, Rankings, and Objective Measures.

RATER-LEVEL AGREEMENT

METHOD

Populations. All subjects at Fort Bragg were members of the 82d Airborne Infantry Division, and all subjects at Fort Campbell were members of the 11th Airborne Division. Three types of persons served either as raters, ratees, or both: CO - Company commanders, executive officers, platoon leaders (officers), and company first sergeants; NCO - Men occupying the duty positions of squad leader, section leader, and platoon sergeant in light and heavy weapons:1/ EM - Men occupying the positions of assistant squad leader and below in light and heavy weapons platoons.

Three ratee populations were utilized to study rater-level relationships. These three ratee populations and the raters who evaluated them were as follows:

206 NCO's at Fort Bragg - rated by CO's, NCO's and EM's
737 EM's at Fort Bragg - rated by NCO's and EM's
400 NCO's at Fort Campbell - rated by CO's and NCO's

1/ At Fort Campbell, many of these men were persons who normally would not have occupied such duty positions. This was due to the fact that the 11th Airborne Division was under strength at the time this study was conducted.

At this writing it is not known how many raters comprised the rater populations, so that the average number of raters per ratee and ratees per rater is not known. It may be assumed that the number of EM raters for any particular ratee population was larger than the number of NCO raters, and the latter in turn were more numerous for any particular ratee group than the CO raters.

Variables. For this study, 15 ratings and rankings from PRT 1800 were used. (See Tables 1 and). For each of these, the following rater level variables were computed:

For Fort Bragg NCO ratees:

Mean CO rating
Mean NCO rating
Mean EM rating
Mean total rating[2/]
Sum of CO, NCO and EM means[2/]

For Fort Bragg EM ratees:

Mean NCO rating
Mean EM rating
Mean total rating[2/]
Sum of NCO and EM means[2/]

For Fort Campbell NCO ratees:

Mean CO rating
Mean NCO rating
Mean total rating[2/]
Sum of CO and NCO means[2/]

Statistical Processing. Three sets of intercorrelation matrixes were obtained, one set for each of the three ratee populations. Each set of matrixes contained a matrix for each of the 15 ratings and rankings. There were thus 15 matrixes for each of the three ratee populations, except for the Fort Bragg EM ratee group, for which there were 14 matrixes since this group was not rated on Teaching Ability. Altogether, therefore, there was a total of 44 matrixes. The variables for the individual matrixes were the rater-level variables described above. The basic datum used throughout was mean rating or mean ranking received by a ratee. All rankings were normalized before being subjected to statistical treatment.

2/ The purpose of obtaining mean total rating and the sums of the means was to study differences among rater levels in terms of weighted and unweighted means. Interpretations of the relevant data are not presented in this report and are deferred until information with respect to the number of raters per ratee is available.

RESULTS

Rater Level Agreements for Different Ratee Groups. Table 1 shows the correlations between various rater levels on each of the rating and ranking variables. (The certainty ratings and rankings are not included in this table since they are control variables rather than evaluations of characteristics or abilities; they are discussed later in this report.)

Table 1. Correlations between rater levels on ratings and rankings.

Variable	NCO Ratees, Fort Bragg			EM Ratees, Fort Bragg	NCO Ratees, Fort Campbell	
	CO vs NCO	CO vs EM	NCO vs EM	NCO vs EM	CO vs NCO	Mean
1st Over-all Rate	.74	.49	.71	.67	.64	.65
1st Over-all Rank	.80	.60	.80	.76	.72	.74
Leadership Rank	.82	.46	.60	.73	.74	.67
Leadership Rate	.72	.51	.68	.69	.66	.65
Willingness to Work Rate	.68	.37	.64	.67	.60	.59
Learning Ability Rate	.62	.50	.70	.67	.56	.61
Teaching Ability Rate	.72	.65	.83	--	.70	.72
M-1 Rifle Rate	.45	.59	.62	.51	.49	.53
Care-of-Equipment Rank	.62	.60	.74	.71	.65	.66
Care-of-Equipment Rate	.52	.52	.64	.68	.55	.58
2nd Over-all Rank	.84	.67	.82	.76	.77	.77
2nd Over-all Rate	.73	.60	.76	.68	.64	.68
Mean	.69	.55	.71	.68	.64	.66

For NCO ratees at Fort Bragg, CO and EM raters show less agreement than do either CO and NCO or NCO and EM raters. This is true for all the rating and ranking variables. Stated another way, seniors and subordinates show greater disagreement than do either seniors and peers or peers and subordinates.

In rating Fort Bragg NCO's, CO's and NCO's show greater agreement than do NCO's and EM's on six of the scales; NCO's and EM's show greater agreement on eight of the scales; and there was equal agreement on one scale. There are thus no clear-cut indications that NCO's tend to agree better with CO's than with EM's. However, it is interesting that on Leadership Rank NCO's agree with CO's decidedly better than they do with EM's, i.e., peer raters agree with senior raters in ranking for leadership

much better than they agree with subordinates. (The same trend is manifest, although not so markedly, on Leadership Rate.) The greatest difference between CO and NCO agreement on the one hand and EM and NCO agreement on the other appears in Leadership Rank, Teaching Ability Rate, M-1 Rifle Rate, Care-of-Equipment Rank and Rate. Agreement is greater between NCO's and EM's on all of these variables except Leadership Rank.

A comparison of NCO vs EM agreement, when rating Fort Bragg NCO's and EM's indicates what differences exist when rater-level comparisons are similar but ratees differ. There are no marked differences except for Leadership Rank and M-1 Rifle Rate, with agreement higher for NCO ratees than for EM ratees only on Leadership Rank. In brief, NCO's and EM's agree better in ranking NCO's for leadership than in ranking EM's for leadership.

A comparison of CO vs NCO agreement in rating NCO's at different installations indicates what differences exist when both rater level and ratee level are similar, but installations differ. Agreement is higher between CO and NCO raters at Bragg than at Campbell for all but three scales. However, differences between the agreement are small, averaging .045 with a range of .02 to .10 for all 12 scales. With respect to rater-level agreement, therefore, installation differences appear to be of minor importance, if any.

If one may generalize at all concerning rater-level agreement, average correlations based on 12 rating and ranking scales show the following hierarchy for the rater groups studied (see means of columns in Table 1):

NCO vs EM, NCO ratees, Fort Bragg
CO vs NCO, NCO ratees, Fort Bragg
NCO vs EM, EM ratees, Fort Bragg
CO vs NCO, NCO ratees, Fort Campbell
CO vs EM, NCO ratees, Fort Bragg

This hierarchy suggests a hypothesis that future research might submit to experimental verification: Peer and subordinate raters show higher agreement than do peer and superior raters, and the latter show higher agreement than do senior and subordinate raters.

Rater-Level Agreements for Different Scales. If the 12 entries in means column of Table 1 are rearranged in order of magnitude, the rating and ranking variables would list in the following order:

Second Over-all Rank	.77
First Over-all Rank	.74
Teaching Ability Rate	.72
Second Over-all Rate	.68
Leadership Rank	.67

Care-of-Equipment Rank	.66
Leadership Rate	.65
First Over-all Rate	.65
Learning Ability Rate	.61
Willingness to Work Rate	.59
Care-of-Equipment Rate	.58
M-1 Rifle Rate	.53

It is noted that two so-called "specific" variables, M-1 Rifle Rate and Care-of-Equipment Rate, have the two lowest mean coefficients, and two general variables, Second Over-all Rank and First Over-all Rank, have the two highest mean coefficients, with Second Over-all Rate having the fourth highest. This suggests that there is less rater-level agreement on specific than on general variables, a conclusion which is in agreement with previous observations.

It appears that there is higher rater-level agreement on ranking than on ratings. Four of the rating scales listed in Table 1 have ranking counterparts. A comparison on these eight variables is shown in Table 2. It is to be noted that each of the four rankings has a higher mean rater-level-agreement coefficient than its rating counterpart. If agreement among judges at different levels is a criterion, then rankings are "better" than ratings.

Table 2. Mean rater agreement coefficients for ranked and rated characteristics.

Characteristic	Coefficients: Rate	Rank
First Over-all	.65	.74
Leadership	.65	.67
Care of Equipment	.58	.66
Second Over-all	.68	.77

Rater-Level Agreement for Certainty Variables. In PRT 1800, the rater is required to record the degree of certainty he feels about his over-all evaluations of the ratees. The 1st Certainty Rate refers to the 1st Over-all Rate and Rank; the [illegible]ertainty Rank and 2nd Certainty Rate refer to the 2nd Over-all Rate [illegible]l Rank. Certainty rates and ranks were obtained for use as control va[illegible]bles throughout the program.

Table 3 shows the correla[illegible]ns between various rater levels on the three certainty variables.

Table 3. Correlations between rater levels on certainty variables

Variable	NCO Ratees, Fort Bragg			EM Ratees, Fort Bragg	NCO Ratees, Fort Campbell	Mean
	CO vs NCO	CO vs EM	NCO vs EM	NCO vs EM	CO vs NCO	
1st Certainty Rate	.49	.33	.40	.36	.54	.42
Certainty Rank	.76	.69	.80	.62	.71	.72
2nd Certainty Rate	.55	.52	.59	.43	.57	.53
Mean	.60	51	.60	.47	.61	.56

In rating NCO's at Fort Bragg, CO's and EM's apparently show less agreement on certainty of evaluations than do either CO's and NCO's or NCO's and EM's. This is in line with results obtained for the 12 evaluations of ratee characteristics above. There appears to be greater agreement on certainty between NCO's and EM's when NCO's, rather than EM's, are the ratees.

Agreement on certainty rank is decidedly higher than agreement on certainty rate. This is true for all rater-level comparisons. The mean agreement coefficient for certainty rank is .72 as compared with .42 and .53 for the two certainty rate scales.

RELATIONSHIP AMONG RATINGS, RANKINGS, AND OBJECTIVE MEASURES

METHOD

Populations. Eleven rater-ratee groups comprised the basic population for this aspect of the study. From these 11 groups, an additional 6 groups were created so that there was a total of 17 rater-ratee groups. These groups and the number of ratees in each group are shown in Table 4. It will be noted that there were no EM raters at Fort Campbell.

Variables. There was a total of 19 variables. Fifteen of these were the rating and ranking variables listed in Tables 1 and 3. In addition, there were 4 objective measure variables, as follows:

1. Aptitude Area I Score
2. Biographical Information Blank, Section I-II Score (DA AGO PRT 2009)
3. Biographical Information Blank, Section III Score (DA AGO PRT 2009)
4. Light Weapons Proficiency Test or Heavy Weapons Proficiency Test (DA AGO PRT 1761, 1765 1781, or 1785)

Table 4. Rater-ratee groups and number of ratees in each group.

Raters		Ratees		
Rel. to Ratee	Grade	Grade	Bragg	Campbell
Peers	NCO	NCO	251	214
	EM	EM	346	---
Seniors	CO	NCO	251	284
	NCO	EM	345	691
Subordinates	EM	NCO	129	---
Self	NCO	(self)	176	236
	EM	(self)	190	---
Mean of total ratings*		NCO	251	285
Mean of total ratings*		EM	346	---
Sum of mean ratings**		NCO	251	284
Sum of mean ratings**		EM	346	---

*Mean of ratings received from all raters, no matter what level.

**Sum of mean ratings received from different rater-level groups, e.g., sum of mean rating from CO raters, mean rating from NCO raters, and mean rating from EM raters.

Statistical Processing. Seventeen matrixes were computed, one for each 17 rater-ratee group in Table 4. The variables for the matrixes were the 15 rating and ranking variables and the 4 objective measures. All matrixes for Fort Bragg NCO ratees and Fort Campbell EM ratees lacked the two BIB variables. A typical matrix is presented as Table A-1 in the Appendix.

RESULTS

General. Visual inspection of the 17 matrixes seemed to indicate that the general pattern of the intercorrelations is quite similar for all matrixes and that the intercorrelations are of such magnitude that one general factor would account for practically all the variance in any one of the matrixes. There are no apparent significant installation differences. Intercorrelations among the rating and ranking variables are quite high, (mostly in the .70's, .80's, and .90's), except for the three certainty variables which are lower (mostly in the .50's, .60's, and .70's). Correlations between rating and ranking variables on the one hand the objective measures on the other are generally low, ranging from low negative or zero values to values in the low .40's. The intercorrelations in the self-rating matrixes are generally slightly lower than those in the other matrixes.

Reliability. Six of the ratings and rankings offer an opportunity to study reliability. These are the two over-all rates, and two over-all ranks and the two certainty rates. Reliability coefficients for over-all rate range from the middle .80's to the low .90's, except for the self-rating matrixes in which they range from the middle .60's to the middle .70's. Reliability coefficients for over-all rank range from the middle .80's to the high .90's, and in the self-rating matrixes they are also within this range. With respect to reliability, therefore rankings are slightly superior to ratings.

Reliability coefficients for the certainty rate range from the middle .70's to the low .90's, except for the self-rating matrixes where the range is from .50 to .61. The lower reliability coefficients for the certainty variables may be due to the true change in feeling of certainty during the interim between the first and second administration of the scale. The first certainty rate scale was the third scale administered in the rating session the second certainty scale was the last scale administered in the session.

It is noted too that the mean rating on over-all ability tends to decrease from first to second rating, whereas the mean rating on certainty tends to increase. The changes for both scales are very small and probably not statistically significant, but the trend is manifest. In brief certainty shows a tendency to improve with practice in rating and is accompanied by slightly lower ratings on over-all infantry ability.

Predictability and Correlational Conceptual Consistency. The predictability of a rating or ranking variable may be estimated in terms of the magnitude of the correlation between it and objective variables. Table 5 shows the correlations between objective variables and the mean of total ratings or rankings. Complete data were available only for two ratee groups, NCO ratees at Campbell and EM ratees at Bragg. The entries in columns headed "BIB" all represent correlations between the ratings and rankings and the sums of scores on BIB's I-II and III. Data for Aptitude Area I and Proficiency Test scores, but not the BIB, were available for NCO ratees at Bragg. Data for EM ratees at Campbell are not presented since they were rated only by NCO's.

The predictability coefficients shown in Table 5 are generally low, ranging from -.04 to .40 with a mean of .22. For NCO ratees, predictability is generally better at Bragg than at Campbell. Predictability for the EM ratees at Bragg is generally better than predictability for the NCO ratees at Campbell with respect to the Proficiency Tests and the BIB, but not with respect to Aptitude Area I.

Table 5. Correlations between objective measures and means of total ratings or rankings.

Variable	NCO Ratees, Fort Campbell			EM Ratees, Fort Bragg			NCO Ratees, Fort Bragg	
	AA-I	(Prof. Test)	BIB	Area I	(Prof. Test)	BIB	AA-I	Prof. Test
1st Over-all Rate	.27	.11	.20	.20	.36	.30	.28	.19
1st Over-all Rank	.26	.01	.22	.21	.36	.31	.28	.10
Leadership Rank	.23	-.04	.23	.23	.33	.30	.27	.10
Leadership Rate	.22	.02	.20	.26	.40	.31	.28	.12
Willingness to Work Rate	.21	.09	.18	.17	.30	.28	.22	.19
Learning Ability Rate	.40	.17	.28	.33	.38	.32	.38	.24
Teaching Ability Rate	.31	.11	.26	-	-	-	.39	.20
M-1 Rifle Rate	.15	.09	.19	.14	.33	.26	.35	.16
Care-of-Equipment Rank	.23	.06	.23	.16	.29	.31	.19	.12
Care-of-Equipment Rate	.16	.09	.13	.13	.27	.30	.11	.09
2d Over-all Rank	.25	.02	.24	.38	.26	.39	.28	.12
2d Over-all Rate	.24	.08	.22	.35	.25	.34	.26	.18
Mean	.24	.07	.22	.23	.32	.31	27	.15

The distinctly higher predictability against the Proficiency Test and BIB at Bragg as compared to Campbell may be due to the fact that the subjects at Bragg were in a training or school environment, whereas the subjects at Campbell were largely recent returnees from Japan. The assumption here is that men in training environment will perform better on an examination like the Proficiency Tests, and that the training environment is more like the situation in which the BIB was originally validated.

It should also be mentioned that the Proficiency Test and BIB were administered during the rating sessions, whereas Aptitude Area 1 scores were obtained months or years previously for most of the subjects. It should be noted that differences for Aptitude Area 1 are generally smaller than those for the Proficiency Tests and BIB.

Certain ratings and rankings are logically and theoretically expected to correlate higher with certain objective measures than with others. In the present study, ratings and rankings would theoretically pair off with objective measures as follows:

Leadership Rank.....	BIB
Leadership Rate.....	BIB
Learning Ability Rate.	AA-1, Prof Test
Teaching Ability Rate...	AA-1, Prof. Test
M-1 Rifle Rate.....	Prof. Test
Care-of-Equipment Rank	Prof. Test
Care-of-Equipment Rate...	Prof. Test

It has been suggested that the data of Table 5 support the hypothesis of conceptual consistency. To illustrate this, the predictability coefficients for Leadership Rate and Learning Ability Rate may be used as examples. If conceptual consistency holds, the sum of the coefficients for Learning Ability Rate vs. AA-1, Learning Ability Rate vs. Proficiency Tests and Leadership Rate vs. BIB would be greater than the sum of the coefficients for Leadership Rate vs. AA-1, Leadership Rate vs. Proficiency Tests, and Learning Ability Rate vs. BIB. For the NCO ratees at Fort Campbell, the theoretically larger sum is .77 while the theoretically smaller sum is .52; for the EM ratees at Fort Bragg, the theoretically larger sum is 1.02 while the theoretically smaller sum is .98. According to this schema conceptual consistency for Leadership Rate and Learning Ability Rate has been demonstrated for both installations.

If this type of schema is extended to include all seven scales listed above, a table such as Table 6 results. In Table 6 the asterisked coefficients should theoretically correlate higher than the remaining coefficients.

Table 6. Comparison of predictability coefficients for seven ratings and rankings conceptually consistent with different objective measures.

Variable	NCO Ratees, Fort Campbell				EM Ratees, Fort Bragg			
	AA-I	Prof. Test		BIB	AA-I	Prof. Test		BIB
Leadership Rank	.23	- 04		.23*	.23	.35		.30*
Leadership Rate	.22	02		.20*	.26	.40		.31*
Learning Ability Rate	.40*	.17*		.28	.33*	.38*		.32
Teaching Ability Rate	.31*	11*		.26	--	--		--
M-1 Rifle Rate	.15	.07*		.15	.14	.33*		.26
Care-of-Equipment Rank	.23	.06*		.23	.16	.29*		.31
Care-of-Equipment Rate	16	.09*		.13	.13	.25*		.27

*Coefficients theoretically larger if correlational conceptual consistency holds.

The average of the asterisked coefficients for Fort Campbell ratees is .18 and the average of the remaining coefficients is .17. For Fort Bragg ratees, the average of the asterisked coefficients is .31 and of the remaining coefficients, .26. It would appear therefore that conceptual consistency obtains when all seven variables are considered.

It may be noted that the Campbell coefficients are generally lower than the Bragg coefficients and those for the Proficiency Test at Campbell are especially low. This makes it difficult to demonstrate the pattern of conceptual consistency inherent in the data. The test variables serve as "criteria" or standards against which the conceptual effectiveness of the ratings and rankings are evaluated. However, the objective measures differ from each other in the range and general magnitude of the predictability coefficients listed for each. A simple way to lessen such differences is to substitute for each raw predictability coefficient its deviation from the mean predictability coefficient of the objective measures. Table 7 shows these deviations. The asterisked deviations are those which should theoretically be greater than the others and in the positive direction.

From another point of view, the discussion above gives a general picture which supports the hypothesis of conceptual consistency. With respect to specific rating and ranking variables, however, the picture is slightly different. Leadership Rank and Leadership Rate are conceptually consistent for NCO ratees (Campbell) i.e., their deviations are higher in the positive direction for the BIB than for AA-I or the Proficiency Tests. For EM ratees (Bragg) Leadership Rate and Leadership Rank do not show this consistency. Conceptual consistency with respect to leadership obtains therefore for men in leadership positions (squad leader, section leader, platoon sergeant) but not in followership position (assistant squad leader and below).

Table 7. Comparison of deviations of predictability coefficients for seven ratings and rankings conceptually consistent with different objective measures.

Scale	NCO Ratees, Campbell			EM Ratees, Brage		
	AA-I	Prof. Test	BIB	AA-I	Prof. Test	BIB
Leadership Rank	-.014	-.107	.015*	-.003	.009	-.011*
Leadership Rate	-.024	-.047	-.015*	.027	.079	-.001*
Learning Ability Rate	.156*	.103*	.065	.097*	.059*	-.009
Teaching Ability Rate	.006*	.043*	.045	---	---	---
M-1 Rifle Rate	-.034	.023*	-.025	-.093	.009*	-.051
Care-of-Equipment Rank	-.014	-.007*	.015	-.073	-.031*	-.001
Care-of-Equipment Rate	-.084	.013*	-.085	-.103	-.051*	-.011

*Deviations theoretically positive and greater if correlational conceptual consistency holds.

Learning Ability Rate is conceptually consistent at both installations. There is a larger positive deviation for Aptitude Area I score than for Proficiency Test scores, and the deviation for the latter is larger than that for BIB score.

Teaching Ability Rate is also conceptually consistent, although the trend is not so marked or clear-cut as it is for Learning Ability Rate.

Care-of-Equipment Rank is not conceptually consistent for either NCO or EM ratees. Care-of-Equipment Rate is conceptually consistent only for EM ratees. It may be that the Proficiency Test is not the proper standard content-wide for evaluations of care-of-equipment behavior.

In summary, therefore, there is conceptual consistency for all the rating and ranking variables, except those for care of equipment.

Predictability of Ranking Compared with Ratings. Eight of the variables provide an opportunity to compare the predictability of rankings and ratings:

First Over-all Rate and First Over-all Rank

Leadership Rate and Leadership Rank

Care-of-Equipment Rate and Care-of-Equipment Rank

Second Over-all Rate and Second Over-all Rank

With respect to Aptitude Area I, BIB I-II, and BIB III scores, the rankings are generally more predictable than the ratings. The differences among the coefficients, however, are extremely small, averaging only about .03 higher against Aptitude Area I, about .004 higher against BIB I-II and

about .01 higher against BIB III. Against the Proficiency Tests the ratings are generally more predictable than the rankings for the First Over-all, Leadership, and Second Over-all scales; and the ranking is generally more predictable for the Care-of-Equipment. However, it must be pointed out again that the differences in predictability between these ratings and rankings are, with few exceptions, very small and of questionable significance.

The practical import of the findings concerning the comparative predictability of rankings and ratings lies in their contribution to the construction of ranking and rating criteria. One might say, on the basis of the results obtained, that normalized rankings are better than ratings. Although the differences with respect to predictability are small and of questionable statistical significance they are nevertheless in favor of rankings and this, together with the higher reliabilities obtained for rankings, seems to indicate the use of rankings rather than ratings as criteria wherever possible.

Halo. As already indicated, the various rating and ranking variables generally correlate high with each other, and correlations between the scales and the test measures are generally low. Correlations among scores on the different objective measures were also found to be comparatively low, ranging from about .15 to about .40. As already indicated, correlations between certain rating and ranking variables and objective measures with which they should theoretically correlate high are not particularly higher than correlations between these variables and objective measures with which they should not theoretically correlate high. This general picture would lead one to conclude that halo is operating to a high degree - if it is assumed that these are adequate objective measures of such characteristics as learning ability (Aptitude Area 1) care of equipment and knowledge of weapons (Proficiency Tests), and leadership (BIB I-II and BIB III).

Mean of Total Compared with Sum of Means. For all ratee groups, the mean of total ratings is generally more predictable with respect to Aptitude Area 1 score than is the sum of means. The differences in predictability between mean of total and sum of means are, however, quite small, averaging only .02 for the EM ratee group at Bragg and the NCO ratee group at Campbell, and .06 for the NCO ratee group at Bragg.

The same tendency is true, only less so, with respect to the Proficiency Tests. The average difference in favor of the mean of total is .01 for the EM ratee group and .06 for the NCO ratee group at Bragg, and .02 for the NCO ratee group at Campbell.

With respect to BIB I-II score, the average difference is again in favor of the mean of the total, but the magnitude of the difference is practically zero.

In the case of the BIB III there is the first indication that the sum of the means is more predictable than the mean of the total. For the NCO ratee group at Campbell, the difference is in favor of the sum of the

means on every rating and ranking variable. Again, the differences are very small, averaging less than .04. For the EM ratee group at Bragg the average difference in predictability between sum of means and mean of total with respect to BIB III is practically zero.

SUMMARY AND DISCUSSION

SUMMARY OF SPECIFIC RESULTS

The purpose of this study was twofold: (1) To determine degree of agreement among different rater levels on various ratings and rankings; and (2) to determine what relationships exist among these ratings and rankings and various objective measures.

1. There is evidence to suggest the hypothesis that peer and subordinate raters show higher agreement than do peer and superior raters, and the latter in turn show higher agreement than do senior and subordinate raters.

2. The data suggest that there is higher rater-level agreement in evaluating on general or over-all behavior than on specific behavior.

3. There is higher rater-level agreement on rankings than on ratings.

4. With respect to certainty of evaluations, the correlations between peer and subordinate raters are generally higher than the correlations between peer and senior raters; and the latter correlation is higher than that between senior and subordinate raters. At all rater levels, agreement on certainty rank is higher than agreement on certainty rate.

5. Intercorrelations among ratings and rankings of such characteristics as over-all infantry ability, leadership, willingness to work, learning ability, teaching ability, ability to handle M-1 rifle, and care-of-equipment are quite high, i.e., in the .70's, .80's, and .90's for various rater-ratee groups. Correlations among ratings and rankings of certainty are lower, i.e., in the .50's, .60's, and .70's. Correlations between rating and ranking variables on the one hand and objective measures on the other hand are generally low, ranging from low negative or zero values to the low .40's. Intercorrelations in self-rating matrixes are generally slightly lower than those in other matrixes.

6. Rate-rerate reliability coefficients for over-all infantry ability range from the middle .80's to the low .90's, except for the self-rating matrixes, in which they range from the middle .60's to the middle .70's. Rank-rerank reliability coefficients range from the middle .80's to the high .90's, and in the self-rating matrixes they are also within this range. With respect to reliability therefore, rankings appear to be slightly superior to ratings.

7. Rate-rerate reliability coefficients for the certainty scales range from the middle .70's to the low .90's except for the self-rating matrixes where the range is from .38 to .61.

8. Rate-rerate reliability coefficients for certainty of evaluation scales range from the middle .70's to the low .90's, except for the self-rating matrixes where the range is from the high .30's to the low .60's.

9. Leadership Rank and Leadership Rate are conceptually consistent with objective measures of leadership (BIB) for ratees who are in leader positions, but not for ratees who are not in leader positions.

10. Learning Ability Rate and Teaching Ability Rate are conceptually consistent with Aptitude Area I score.

11. Care-of-Equipment Rank is not conceptually consistent, when the Proficiency Tests are the standard for such consistency. Care-of-Equipment Rate is conceptually consistent for FM, but not for NCO ratees.

12. Rankings are generally more predictable than ratings. However the differences are very small and of questionable statistical significance.

13. Pending evidence to the contrary, it appears that halo operated to a high degree for the ratings and rankings in this study.

IMPLICATIONS FOR FURTHER RESEARCH

In future studies utilizing these data, the number of rating variables can be reduced. Although none of the matrixes of rating and ranking variables described in this report was submitted to a factor analysis, it would appear from observation that one general factor accounted for most of the variance in the matrixes. It is believed selection of rating variables for future studies should be made primarily on the basis of conceptual consistency with respect to the purposes of such studies, and the number of rating variables selected should be kept to a minimum.

With respect to combining Fort Bragg and Fort Campbell data, a great deal depends on the particular study contemplated and the variables that must be used. In most instances combining data from the two installations should be quite feasible; however, with respect to certain variables, caution should be observed before such a step is taken, and the matrixes described in the present report should be carefully examined. For example, if the Proficiency Tests were one of the variables under analysis, it would not be advisable to combine NCO ratees from both installations since the pattern of correlations between the Proficiency Tests and the rating and ranking variables were different at the two installations, and the magnitude of the coefficients was decidedly higher at one installation. With respect to combining various rater-ratee groups, the same precaution must be observed.

The question of what future specific studies should be undertaken with the data available must be answered by an evaluation of the extent to which such studies are likely to contribute to improving criterion ratings. It may be mentioned that two research programs have already been established. The first of these programs will attempt to determine the personal characteristics of raters who are reliable, valid, free from halo, and neither hard nor easy raters. The second program will attempt a similar analysis for ratees who are reliably and validly rated.

Probably the most important research program that can be imagined to utilize these data is one to study the effect of training on the reliability, validity, spread of ratings, and halo. Other, perhaps more specific, research questions to which the data of the present study can be applied are those already listed in the introduction to this report.

PERSONNEL IN CHARGE

Program Coordinator: Edward A. Rundquist
Richard H. Gaylord
Hubert E. Brogden
David J. Chesler
Project Director: David J. Chesler
Statistical Adviser: Neil J. Van Steenberg

COLLECTION OF DATA: November 1949

PREPARATION OF REPORT: 14 December 1951

APPENDIX

APPENDIX A

Table A-1. Intercorrelations among objective measures and mean total ratings and rankings for [illegible] N = 28)

Variable	Intercorrelations*													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1. 1st Over-all Rate														
2. 1st Over-all Rank	.84													
3. 1st Certainty Rate	.57	.67												
4. Leadership Rate	.88	.88	.67											
5. Willingness to Work Rate	.85	.77	.53	.80										
6. Leadership Rank	.81	.94	.66	.89	.74									
7. Teaching Ability Rate	.82	.80	.61	.86	.73	.81								
8. M-1 Rifle Rate	.72	.71	.60	.77	.65	.70	.71							
9. Care-of-Equipment Rate	.69	.60	.40	.65	.68	.58	.61	.66						
10. 2nd Over-all Rank	.84	.94	.67	.88	.77	.95	.82	.72	.63					
11. Learning Ability Rate	.85	.81	.57	.83	.78	.80	.85	.71	.62	.81				
12. Care-of-Equipment Rank	.73	.78	.54	.70	.71	.75	.67	.62	.79	.80	.70			
13. 2nd Over-all Rate	.91	.85	.60	.91	.86	.83	.85	.76	.75	.86	.86	.75		
14. Certainty Rank	.74	.86	.76	.81	.68	.87	.76	.68	.55	.89	.73	.75	.78	
15. 2nd Certainty Rate	.64	.68	.85	.70	.59	.67	.67	.65	.48	.69	.62	.57	.69	.79
16. AGCT	.27	.26	.10	.22	.21	.23	.31	.15	.16	.25	.40	.23	.24	.24
17. Proficiency Tests	.11	.01	-.05	.02	.09	-.04	.11	.09	.09	.02	.17	.06	.08	-.02
18. BIB I and II	.10	.14	.11	.11	.10	.12	.18	.15	.10	.13	.19	.06	.12	.08
19. BIB III	.22	.21	.18	.20	.18	.24	.24	.15	.11	.25	.26	.15	.22	.22

*The ratings and rankings (variables 1-15) have been reflected to reverse the scale of measurem of the correlation coefficients of the objective measures (variables 16-19) with the ratings a

APPENDIX A

measures and mean total ratings and rankings for NCO ratees at Fort Campbell.

Intercorrelations*																Mean	S.D.
4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19		
																5.82	1.59
																8.05	1.81
																5.24	1.59
																6.07	2.03
.80																4.56	1.53
.89	.74															8.06	1.79
.86	.73	.81														6.33	1.95
.77	.65	.70	.71													4.78	1.27
.65	.68	.58	.61	.66												5.10	1.34
.88	.77	.95	.82	.72	.63											8.08	1.81
.83	.78	.80	.85	.71	.62	.81										4.59	1.33
.70	.71	.75	.67	.62	.79	.80	.70									8.07	1.36
.91	.86	.83	.85	.76	.75	.86	.86	.75								5.72	1.66
.81	.68	.87	.76	.68	.55	.89	.73	.75	.78							8.11	1.46
.70	.59	.67	.67	.65	.48	.69	.62	.57	.69	.79						5.28	1.46
.22	.21	.23	.31	.15	.16	.25	.40	.23	.24	.24	.10					97.50	16.50
.02	.09	-.04	.11	.09	.09	.02	.17	.06	.08	-.02	-.04	.43				102.90	20.40
.11	.10	.12	.18	.15	.10	.13	.19	.06	.12	.08	.08	.37	.29			46.11	5.54
.20	.18	.24	.24	.15	.11	.25	.26	.15	.22	.22	.23	.19	.12	.30		35.32	6.46

e been reflected to reverse the scale of measurement. This changes the obtained signs tive measures (variables 16-19) with the ratings and rankings.