DOCUMENT RESUME

ED 174 647 TM 009 495

AUTHOR Mullins, Cecil J.; And Others
TITLE Personnel Rating Effectiveness as a Function of Number of Rating Statements. Final Report for Period 24 February 1978--31 December 1978.
INSTITUTION Air Force Human Resources Lab., Brooks AFB, Texas.
REPORT NO AFHRL-TR-79-11
PUB DATE May 79
NOTE 21p.
AVAILABLE FROM National Technical Information Service, Springfield, Virginia 22161

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Cognitive Ability; Communication Skills; Evaluation Criteria; Job Skills; *Leadership Qualities; *Officer Personnel; Peer Evaluation; Personality Assessment; *Personnel Evaluation; *Predictive Validity; Profile Evaluation; *Rating Scales; Test Construction
IDENTIFIERS Air Force; *Test Length

ABSTRACT

Research on the comparative utility of varying numbers of rating statements per set--using an external criterion of recognition of rater profiles for evaluating "goodness" of the sets--was conducted on 132 noncommissioned Air Force officers. Groups of subjects rated groups of their peers on 20 factors, on a subset of 10 factors, and on a subset of five factors. Rating statements included: learning ability; leadership; quality of work; motivation; ability to follow instructions; bearing and behavior; accuracy; oral communication; problem analysis; initiative; quality of work; written communication; punctuality; adaptability; dependability; emotional stability; human relations; judgment; knowledge of duties; and honesty. Profiles, based on group ratings, were developed for each subject and given to the group members to identify. Profiles made from 20 rating statements were identified no better than profiles developed from 5 rating statements. Also, the use of two rating statements, learning ability and knowledge of duties, were found to produce information about a rate which could not be improved by the addition of many more rating factors. When the external criterion was used, results indicated that sets of statements larger than five did not provide better recognition of peers. (MH)

AFHRL-TR-79-11

AIR FORCE HUMAN RESOURCES LABORATORY

PERSONNEL RATING EFFECTIVENESS AS A FUNCTION OF NUMBER OF RATING STATEMENTS

By

Cecil J. Mullins
James A. Earles
James M. Wilbourn

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

May 1979
Final Report for Period 24 February 1978 – 31 December 1978

AIR FORCE SYSTEMS COMMAND
BROOKS AIR FORCE BASE, TEXAS 78235

NOTICE

When U.S. Government drawings, specifications, or other data are used for any purpose other than a definitely related Government procurement operation, the Government thereby incurs no responsibility nor any obligation whatsoever, and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data is not to be regarded by implication or otherwise, as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

This final report was submitted by Personnel Research Division, under project 2313, Force Acquisition, Assignment, and Evalution with HQ Air Force Human Resources Laboratory (AFSC), Brooks Air Force Base, Texas 78235. Dr. Cecil J. Mullins (PEP) was the Principal Investigator for the Laboratory.

This report has been reviewed by the Information Office (OI) and is releasable to the National Technical Information Service (NTIS). At NTIS, it will be available to the general public, including foreign nations.

This technical report has been reviewed and is approved for publication.

LELAND D. BROKAW, Technical Director
Personnel Research Division

RONALD W. TERRY, Colonel, USAF
Commander

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFHRL-TR-79-11	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) PERSONNEL RATING EFFECTIVENESS AS A FUNCTION OF NUMBER OF RATING STATEMENTS		5. TYPE OF REPORT & PERIOD COVERED Final 24 February 1978 – 31 December 1978
		6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Cecil J. Mullins James A. Earles James M. Wilbourn		8. CONTRACT OR GRANT NUMBER(s)
9. PERFORMING ORGANIZATION NAME AND ADDRESS Personnel Research Division Air Force Human Resources Laboratory Brooks Air Force Base, Texas 78235		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 61102F 2313T616
11. CONTROLLING OFFICE NAME AND ADDRESS HQ Air Force Human Resources Laboratory (AFSC) Brooks Air Force Base, Texas 78235		12. REPORT DATE May 1979
		13. NUMBER OF PAGES 18
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) Unclassified
		15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

personnel ratings
rating dimensions
rating factors
rating multiple factors
rating statements

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Previous work on sets of personnel rating statements leave unanswered the question of whether there is any advantage in using several "factor" rating statements over the use of a single statement. This is a study of the comparative utility of sets of rating statements varying in number of statements per set, using an external criterion.

A great deal of research effort has been expended in an effort to find "best" factors for collecting rating data. Most of this research has concentrated on internal psychometric characteristics of the rating data, such as means, standard deviations, and reliability coefficients. When internal psychometric considerations constitute the sole criterion, some small advantage is frequently found for one kind of rating statements over another. When external criteria for evaluating "goodness" of rating sets are applied, there are usually no differences found among sets.

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

4.

Item 20 Continued:

Indeed, there have been some indications that raters generally may not be able to evaluate ratees on more than one general quality of overall excellence and that collecting several "factor" ratings may be little more than time wasted.

The subjects for this study were 132 students at the NCO academy at Lackland A. B. Three treatment conditions were applied. Of the 132 subjects, 45 were required to rate their peers in their 15-man study groups on 20 rating factors. Another 44 subjects rated their peers on 10 factors, which were a subset of the 20 factors used by the first group. Still another 43 subjects rated their peers on a subset of only five rating factors. From the ratings, profiles were developed for each subject indicating how that individual had been rated by a peer group. These profiles, with no identifying information on them, were handed out to the group members, who were required to identify them. A record was kept of all correct identifications. Analyses of variance were done to see if there were any significant differences among the three groups. In addition, correlation coefficients were computed between the various sets of rating statements and a criterion of class standing upon graduation, to see whether 20 statements predicted this criterion better than 10 and whether 10 statements were more predictive than five.

The analysis of variance portion of the study produced no significant differences among the groups. The multiple linear regression analyses indicated that, when only two of the statements were used as predictors, addition of the other 18 to the predictor pool generated no useful additional prediction. The results of this study indicate that a very small set of rating factors (e.g., two in this study) produce information about a ratee which cannot be improved by the addition of many more factor statements.

TABLE OF CONTENTS

LIST OF TABLES

PERSONNEL RATING EFFECTIVENESS AS A FUNCTION OF NUMBER OF RATING STATEMENTS

I. INTRODUCTION

The literature contains a large number of studies issuing from the search for appropriate rating constructs to be used in the collection of rated data. The results have been rather disappointing, but still the search goes on. The pursuit of rating constructs (or "factors") is probably due to the enormous influence of Thurstone's work with the factor analysis of test data and to his conclusion that complex human characteristics can best be explained in terms of a few orthogonal factors–that is, factors which are not correlated with each other. American psychologists, in general, have accepted Thurstone's position. Those who have worked with rating data have started from the assumption that the concept of orthogonality is almost a natural law. If one accepts that assumption, it is reasonable that one of the primary goals of rating research has been to find that set of independent (orthogonal) constructs which best describes human behavior when rating data are used. It is, after all, merely an extension into rating data of a principle which has been accepted broadly as a fundamental concept in test data.

There are, however, at least four major difficulties that have beset researchers in their quest for simple structure in rating data. These difficulties are as follows:

1. *Orthogonality as a Concept.* Although the concept of orthogonality as a requisite for factors has been persuasive to American psychologists, not all prominent modern psychologists have succumbed to the attractiveness of Thurstone's arguments for the primacy of specific or orthogonal factors to describe human abilities (e.g., Horn, 1968; Humphreys, 1962; Jensen, 1966; McNemar, 1964, to name only a few). Indeed, McNemar (1964) has pointed out a serious weakness in the entire factor analytic process:

> In practically all areas of psychological research the demonstration of trivially small minutiae is doomed to failure because of random errors. Not so if your technique is factor analysis, despite its being based on the correlation coefficient that slipperiest of all statistical measures. By some magic, hypotheses are tested without significance tests. This happy situation permits me to announce a Principle of Psychological Regress: *Use statistical techniques that lack inferenttal power.* This will not inhibit your power of subjective inference.

In the same article (a discussion of the concept of intelligence), McNemar finds no advantage of fractionating general mental ability into differentially weighted independent separate factors, even in predicting meaningful criteria. The problem of finding separate rating "factors" is quite analogous. We have no convincing evidence that separate rating statements will provide data that are more useful than one global rating of all-around excellence. There may not *be* any set of rating "factors" in the simple structure sense.

2. *Theory Weakness.* If rating "factors" exist, it is not at all clear in what direction they may lie. There is no widely accepted theory which provides clues to the researcher to aid him in his search. Without such clues, the number of descriptive qualities, interacting with ways of expressing those qualities, is literally almost endless. This is one reason so much effort has been expended in the search for the best rating statements. In test theory, it is known that certain factors (e.g., verbal, numerical) are stable and replicable–although some have questioned the utility of large factor sets. In rating theory, we do not even know the best format for collecting data, much less which constructs are more likely to yield useful information.

It is not even clear how rating questions should be worded. For example, there has been considerable controversy over whether statements oriented around tasks performed ("adjusts the linkage on the clutch pedal") or statements oriented around personal characteristics of the ratee ("forceful and dominant in interpersonal relations") are more useful in describing ratees (Kavanagh, 1971; Massey, Mullins, & Earles, 1978). Generally, on this issue, task-oriented statements appear to be slightly better by internal psychometric standards (slightly less inflated means, slightly larger standard deviations, larger reliability coefficients), but no differences are usually observed when evaluation of the rating statements is made by applying an external criterion. This problem of theory weakness is much more severe in a search of rating data for rating factors than it is in a search of test score data for intellectual factors, because the universe of discourse is so much larger and harder to define.

3. *Differential Description.* Even if reasonably good factors could be deduced from some theory, and even if they really were present in a given rating situation, there is no assurance that they could be demonstrated from the rating data collected. All psychologists are familiar with the halo phenomenon in ratings, and the halo effect is possibly strong enough that the average rater simply cannot produce differentiation among ratee characteristics sharp enough and objective enough that the factors would show up in the data analysis. But the first burden of a set of rating factors–if they are really worthwhile–must be to describe differentially the members of a ratee group. If a set of rating statements does not paint a unique picture of each ratee with recognizable differences between his picture and that of each other member, it is difficult to see how that set of rating statements could produce useful validities against any reasonable outside criterion.

4. *Criterion Problems.* Ratings are usually collected to serve as a criterion, rather than predictor, variable. One of the reasons rating data are used is that the investigator can find no other way of measuring the variable of interest. Therefore, the rating is the most "ultimate" score one can collect. There is no available metric closer to the true score than the ratings themselves. If one accepts the position that the rating score is the ultimate criterion, then of course one cannot question its validity. It is by definition perfectly valid. In such a case, one can only investigate certain internal psychometric characteristics, such as its reliability (Remmers, 1934, p. 621). In some situations, particularly in the operational use of ratings, this can sometimes be a reasonable position.

In doing research on rating methodology, however, it seems essential to have some other criterion available which allows one to compare the "goodness" of one rating or set of ratings against another. It is of little use to compare reliabilities, means, standard deviations, and other internal psychometric characteristics if one is trying to determine which set of ratings is better at measuring a particular condition.

Previous studies in this series of investigations (Curton, Ratliff, & Mullins, 1979; Massey, Mullins, & Earles, 1978) have attempted to discover qualities of "good" rating statements compared with "poor" rating statements. The methodology has been what one might expect from the difficulties labeled 3 and 4, above, discussing differential description and criterion problems, respectively. Different sets of rating statements were compared by observing their relative merits in differentially describing ratees, and in their prediction of external criteria. None of the sets of rating statements investigated so far have shown any superiority over any of the others.

Judging from the results available so far in the literature, it may well be that perhaps all that the average rater can do effectively is rate on some general overall idea of excellence and that requiring the rater to rate separate characteristics independently is beyond a person's capability. If this is so, it is another way of saying that halo error overwhelms the variance in sets of rating statements and that sets of rating statements are not more efficient than a single rating. This is a study of the relative effectiveness of requiring raters to rate varying numbers of statements, with effectiveness defined by criteria external to the ratings.

II. METHOD

Rating Rationale

The two studies cited previously were based on the premise that if rating statements are actually meaningful, raters should be able to identify unlabelled profiles made by their peers from those rating statements. The two previous studies indicated that correct identifications (hits) are too few to provide a reasonable degree of sensitivity. The average number of hits per rater, rating 14 peers, is only about 2.5. Therefore, a refined method of hits was also used, called the rank order (RO) method.

The RO method provides a way to credit near misses. The hits approach is all or nothing. The rating subject either guesses the profile correctly or does not. The rater may be sure that the profile being studied is either, say, peer B or peer F, so commits to B. If the profile really belongs to peer F, the rater not only gets no credit for being close on the identification of peer F's profile, but also misses peer B's profile as well. The RO method, though a little difficult to understand, is an approach designed to make the hits method more sensitive.

If, in addition to being asked for absolute hits, the rater is asked to rank the unidentified profiles in terms of some standard of excellence (e.g., how well the people with these profiles will do in a course of instruction), and if the rater is also given a list of the names of peers and is asked to rank these peers on the same standard of excellence, and then the rank differences are analyzed, this is in a real sense a measure of hits which gives credit for near misses. For example, if the rater believes correctly that the three profiles which appear to be the "best" in terms of most likely to succeed belong to peers B, C, and F, but is not sure which is which, the ranking approach will provide credit for placing these peers in the proper end of the ranking, whereas absolute hits might give the rater no credit at all.

Another external criterion for judging the relative efficacy of various sets of rating statements can be some typical success criterion, such as the final grade upon graduation from school. This criterion is not quite as direct as peer identification for judging the quality of peer ratings, because an additional element, validity of chosen statements, becomes a consideration. Using this criterion, not only must each rating statement contribute to a sound differential description of the ratee, but also it must be a statement which happens to be valid for that success criterion in order for differences among rating statements to appear.

For example, one might calculate the correlation coefficient between the criterion and a set of five rating statements and then calculate another correlation coefficient between the criterion and 10 rating statements, of which five were the same statements used in calculating the first correlation coefficient. By applying the proper statistics, one can then determine whether or not the set of 10 statements predicted better than the subset of five statements. Whether they do is determined not only by the quality of the ratings as descriptors of the ratees but also by the validity of the quality described in the rating statement for the chosen criterion. One should be able to rate one's peers rather accurately on, say, height, but that would probably not be a valid predictor of academic ability. Therefore, it is believed that the peer identification process is probably the most direct method of judging the relative accuracy of two sets of rating statements. Both approaches were used in this study.

Subjects

Nine seminar groups of Air Force non-commissioned officers (NCOs) (technical and master sergeants) assigned to the Air Training Command (ATC) NCO Academy at Lackland AFB Annex served as subjects for this study. Seven of the seminar groups were composed of 15 subjects each, one of 14 subjects, and one of 13 subjects, yielding a total N of 132. Length of military service for these subjects was 10 to 17 years.

Procedures

The nine seminar groups were randomly assigned to three treatment conditions of three seminar groups each. The subjects in Treatment Condition 1 were asked to rate their peers on five rating statements. Those in Treatment Condition 2 were asked to rate their peers on 10 rating statements, five of which were used by Treatment Condition 1. In Treatment Condition 3, the subjects rated their peers on 20 rating statements, including the 10 used by Treatment Condition 2. All 20 of the rating statements are given in the appendix. It should be noted that some of the statements are very general and person oriented in nature (e.g., 1, 2, 10) while others are more specific and job related (e.g., 6, 15, 19). This design provided 45 subjects on whom 20 rated statements were available, 89 on whom there were 10 rated statements, and 132 who had all rated the same five statements. Summary statistics for all nine seminar groups and three treatment conditions are available in Table 1.

Table 1. Number of Profile Identifications (Hits) by Treatment and by Seminar Group

	Seminar Group								
	Treatment 1 5 Statements			Treatment 2 10 Statements			Treatment 3 20 Statements		
Results	E	F	I	B	C	H	A	D	G
Group									
N	15	15	13	14	15	15	15	15	15
Total Hits	31	36	35	37	45	40	36	40	34
Mean Hits	2.07	2.40	2.69	2.64	3.00	2.67	2.40	2.67	2.27
SD Hits	1.65	.95	1.49	1.55	1.90	1.24	1.54	1.92	1.33
Treatment									
Total N		43			44			45	
Total Hits		102			122			110	
Mean Hits		2.37			2.77			2.44	
SD Hits		1.42			1.07			1.63	
T-Ratios									
Treatment 1 Versus 2			t = .219[a]						
Treatment 1 Versus 3			t = .163[a]						
Treatment 2 Versus 3			t = .246[a]						

[a] Not significant.

After the ratings had all been collected, rating scores were averaged across raters, and profiles were constructed, one for each ratee (see Appendix A for examples). The ratee's name was left off the profile, but the profile itself was reproduced in sufficient copies that all members of the group could have all the profiles of all their seminar group peers, unidentified as to name. In a second visit to the seminar groups, the subjects were given three more tasks to perform, in the following order. The first task was to study each of the profiles and rank the profiles according to how "a person" with that profile should do in the class the subjects were taking. The second task was to match each of their peers with one of the profiles (that is, indicate to whom each profile belonged). Third, each subject was given a list of the people in the seminar group and was told to rank them according to how well they would do in the course.

The data were subjected to two analysis of variance treatments (Tables 2 and 3), and then they were reanalyzed using multiple linear regression analysis (Tables 4 to 7).

III. RESULTS AND DISCUSSION

Table 2 shows that there were no differences among treatment conditions in the number of "hits." Those subjects who were looking at profiles made from 20 rating statements could identify their peers no better than those looking at profiles made from five statements. Number of hits, as mentioned above, is a relatively insensitive measure of peer recognition, however (Table 1 shows that each group averaged about 2.5 hits).

Table 2. Analysis of Variance of Number of Correct Profile Identifications (Hits) by Treatment and by Seminar Group

Source	Sum of Squares	DF	Mean Square	F
Treatment	3.739	2	1.870	2.172[a]
Seminar Groups Within Treatment	5.168	6	.861	.340[a]
Error (Within Groups)	311.717	123	2.534	

[a]Not significant.

Table 3 shows the results of analysis of variance treatment of the squared differences between the ranking of the unidentified profiles and the ranking of peers. Again, there is no significant difference among groups, even using this more sensitive measure of peer identification.

Table 3. Analysis of Variance of Squared Deviations between Unidentified Profile Rankings and Peer Rankings by Treatment and by Seminar Group

Source	Sum of Squares	DF	Mean Square	F
Treatment	15080.951	2	7540.475	1.29[a]
Seminar Groups Within Treatments	349887.789	6	58314.631	2.074[a]
Error (Within Group)	3458429.590	123	28117.313	

[a]Not significant.

Intercorrelation matrices among the various groups of rating statements and final school grade appear in Tables 4 to 6. The most striking aspect of these correlations is their size. The reliabilities of the separate rating statements are unknown, but it appears that the intercorrelations of each rating statement with the others must approach the statement reliabilities. For example, in Table 4, 143 of the 190 intercorrelations among rating statements are .70 or higher, and 30% are .80 or higher. All this argues rather strongly for the likelihood that little is being rated except a general idea of excellence.

The one worrisome feature of these intercorrelation matrices is the fact that there are some sizable differences among the 20 rating statements in their validities against final school grade, with Statement 1 exhibiting the highest relationship with the criterion. In each of the three matrices, this finding seems to argue against the proposition that the rater can evaluate only in general terms; otherwise, how could one statement be more valid than another? However, there are two possible explanations for the higher validity of Statement 1.

Table 4. Intercorrelations, 20 Rating Statements, and Final School Grade
(N = 45)

	Rating Statements																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	FSG
1	1.00	.76	.85	.57	.67	.63	.81	.67	.76	.75	.78	.82	.57	.62	.70	.70	.48	.67	.73	.63	.75
2		1.00	.85	.77	.73	.76	.75	.78	.87	.84	.80	.79	.63	.71	.76	.74	.63	.73	.69	.67	.50
3			1.00	.75	.82	.76	.88	.71	.81	.85	.85	.79	.72	.74	.80	.74	.65	.74	.81	.73	.59
4				1.00	.67	.84	.75	.64	.70	.84	.74	.59	.65	.73	.72	.65	.68	.69	.65	.69	.35
5					1.00	.74	.85	.67	.84	.75	.82	.67	.74	.80	.81	.79	.67	.75	.84	.82	.36
6						1.00	.75	.66	.77	.83	.72	.61	.65	.78	.76	.75	.71	.75	.75	.74	.43
7							1.00	.76	.83	.87	.91	.76	.64	.76	.80	.79	.64	.79	.84	.76	.51
8								1.00	.81	.80	.80	.81	.61	.62	.71	.66	.52	.72	.73	.53	.44
9									1.00	.82	.85	.77	.73	.82	.85	.81	.70	.81	.82	.77	.44
10										1.00	.88	.76	.67	.77	.83	.76	.62	.78	.81	.71	.47
11											1.00	.82	.70	.76	.85	.78	.66	.77	.82	.71	.50
12												1.00	.57	.60	.71	.68	.48	.65	.68	.59	.61
13													1.00	.76	.79	.68	.73	.74	.78	.67	.32
14														1.00	.85	.83	.78	.81	.86	.83	.33
15															1.00	.81	.71	.79	.87	.78	.44
16																1.00	.76	.86	.83	.79	.43
17																	1.00	.84	.76	.75	.21
18																		1.00	.88	.77	.41
19																			1.00	.80	.40
20																				1.00	.32
FSG																					1.00
Mean	3.2	3.1	3.2	3.5	3.4	3.5	3.3	3.2	3.3	3.4	3.2	3.3	3.6	3.3	3.5	3.4	3.5	3.3	3.5	3.7	322.6
SD	.56	.47	.43	.40	.42	.41	.41	.48	.40	.45	.37	.36	.34	.42	.35	.42	.46	.43	.46	.35	22.4

Table 5. Intercorrelations, 10 Rating Statements, and Final School Grade
(N = 89)

	Rating Statements										FSG
	1	2	3	4	5	6	7	8	9	10	11
1	1.00	.71	.85	.64	.65	.49	.81	.72	.82	.73	.71
2		1.00	.83	.81	.79	.75	.80	.82	.84	.87	.44
3			1.00	.82	.80	.69	.92	.80	.87	.87	.64
4				1.00	.73	.74	.82	.70	.75	.88	.41
5					1.00	.74	.84	.68	.78	.78	.44
6						1.00	.71	.58	.63	.78	.39
7							1.00	.81	.86	.87	.61
8								1.00	.87	.81	.52
9									1.00	.82	.54
10										1.00	.50
FSG											1.00
Mean	3.2	3.1	3.3	3.5	3.5	3.6	3.4	3.3	3.3	3.5	328.3
SD	.57	.54	.47	.43	.44	.44	.44	.56	.45	.46	22.8

Table 6. Intercorrelations, Five Rating Statements, and Final School Grades
(N = 132)

	Rating Statements					FSG
	1	2	3	4	5	6
1	1.00	.71	.86	.64	.66	.68
2		1.00	.82	.80	.79	.36
3			1.00	.80	.80	.61
4				1.00	.74	.38
5					1.00	.38
FSG						1.00
Mean	3.2	3.1	3.3	3.4	3.4	329.1
SD	.55	.52	.44	.43	.43	21.8

The most obvious explanation is that Statement 1 (Learning Ability—acquires knowledge accurately and quickly) describes the factor among the set of 20 which is most important for success in this school environment. However, the learning at the NCO academy does not appear to be of the kind which taxes learning ability, such as difficult academic subjects might. Looking in from the outside, it appears that several others should be at least as important for success in this particular school (e.g., leadership, quality of work, motivation, knowledge of duties). Furthermore, if the raters really are making a distinction between learning ability and the other statements, it is difficult to explain the high intercorrelations among the statements.

An alternative explanation is that the order of presentation of the statements explains the higher validity of learning ability for the success criterion. This argument implies that whatever statement was presented first would exhibit the highest validity, and learning ability just happened to be the first in the series. Assuming that the raters really cannot consider the separate

statements independently, it would be natural enough for them to rate the first factor in terms of a global perception of the ratee's general excellence, which should exhibit the highest validity of which the rater is capable. When the rater is then faced with the task of rating the second and succeeding statements, the rater perceives an implication that these other statements should be somehow different from the first. So an implicit requirement is generated by the mechanics of the situation to rate the second and succeeding statements different from the first, but there is not (by assumption) an ability to do so accurately. If this scenario is accurate, intercorrelation matrices similar to those displayed in Tables 4, 5, and 6 should result.

There is no way within the limits of this study to determine which of these two explanations is correct, but another study in the same context varying the order of presentation of the statements might clarify the relationships and should be easy enough to accomplish.

The results of four regression analyses appear in Table 7. The first question to be addressed by this table is "When five rating statements are available on a set of ratees, is anything of value added by considering an additional 15 rating statements when one is predicting some meaningful criterion such as school success?"

Table 7. **Regression Analyses of Varying Numbers of Rating Statements**

Problem	Rating Statements	R^2	N	Difference	F
A	1–20	.693	45		
	1–5	.605	45		
				.088	.458[a]
B	1–10	.575	89		
	1–5	.549	89		
				.026	.872[a]
C	1, 19	.614	45		
	1 alone	.566	45		
				[illegible]	5.212*
D	1, 19, 9	.623	45		
	1, 19	.614	45		
				.009	.991[a]

[a]Not significant.

*Significant at the .05 level.

Only 45 of the subjects (Treatment Condition 3) were available to investigate this question, since only 45 subjects rated their peers on all 20 rating statements. Problem A in Table 4 shows clearly that there is no significant difference between the full model R^2 (using all 20 statements) and the restricted model R^2 (using only five statements).

The next question that arises is, "Are 10 statements better than five in predicting school success?" Data pertinent to this question were available from 89 subjects, and are shown in problem B. Again, there is clearly no significant advantage in using 10 statements, rather than five.

Finally, it seems important to ask, "What is the smallest subset of the 20 rating statements which carries the predictive burden of the entire set?" Problems C and D in Table 7 address this issue. The 45 subjects in Treatment Condition 3 were used, since these were the only subjects who rated the entire set of 20 statements.

When rating statement 1 (Learning Ability) is the only variable in the prediction system, the R^2 is .566. Adding rating statement 19 (Knowledge of Duties) to the prediction system increases the R^2 to .614, an increase which is significant at the .05 level. When both rating statements 1 and 19 are in the prediction system, the addition of any of the other 18 statements does not improve prediction significantly.

It is not unusual that the results of a study support a simply stated hypothesis only partially. This study began with the hypothesis that untrained raters can rate only on some general idea of excellence and that ratings cannot be made better by requiring the rater to rate several separate characteristics. This hypothesis was tested, using two different external criteria of goodness-of-rating scores.

When an external criterion of recognition of ratee profiles is used to evaluate the goodness of sets of ratings, the results indicate that sets of rating statements larger than five do not provide better recognition of peers. The analysis of variance design did not permit any conclusions concerning whether five rating statements produced significantly better recognition than one statement.

When the external criterion is class standing and the 20 rating statements are subjected to multiple linear regression analysis, the results indicate unequivocally that large sets of rating statements do not provide better measurement than small sets. Apparently, however, a single rating statement does not carry as much predictive power as two. One does not know, of course, whether a single rating statement deliberately designed to be as broad and global as possible might have provided all the prediction attainable from all combinations of "factor" statements–since the 20 statements studied were selected partially on the basis of their apparent independence of each other. But that does not change the fact that the beginning hypothesis had to be rejected in favor of an alternate one that states raters can effectively use only a very small subset of the 20 rating statements investigated in this study. Future studies will determine whether a single global rating statement will provide all the useful information available from any number of "factor" rating statements and will find out in several different contexts what is the most likely maximum number of useful "factor" rating statements.

REFERENCES

Curton, E.D., Ratliff, F.R., & Mullins, C.J. Content analysis of rating criteria. Chapter XIII.1 in *Criterion Development for job performance evaluation: Proceedings from symposium.* AFHRL-TR-78-85. AD-A000 000. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, February 1979, 116–122.

Horn, J.L. Organization of abilities and the development of intelligence. *Psychological Review*, 1968, **75**(3), 242–259.

Humphreys, L.C. The organization of human abilities. *American Psychologist* 1962, **17**, 475–483.

Jensen, A.R. Individual differences in concept learning. Chapter 7 in *Analyses of concept learning.* New York: Academic Press, 1966.

Kavanagh, M.J. The content issue in performance appraisal: A review. *Personnel Psychology*, 1971, **24**, 653–668.

Massey, R.H., Mullins, C.J., & Earles, J.A. *Performance appraisal ratings: The content issue.* AFHRL-TR-78-69, AD-A064 690. Brooks AFB, TX: Personnel Research Division, Air Force Human Resources Laboratory, December 1978.

McNemar, Q. Lost: Our intelligence? Why? *American Psychologist*, 1964, **19**, 871–882.

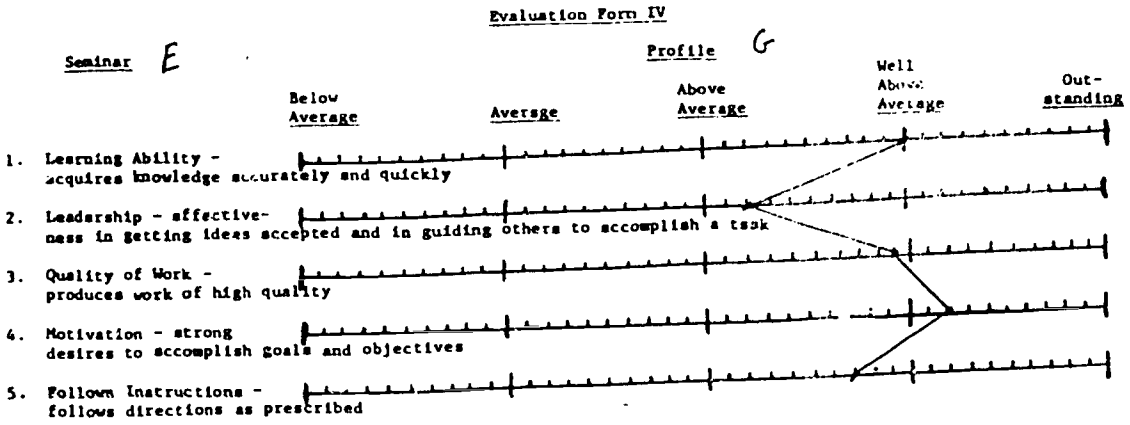
Remmers, H.H. Reliability and halo effects of high school and college students' judgment of their teachers. *Journal of Applied Psychology*, 1934, **18**, 619–630.

APPENDIX A: EVALUATION FORMS

EVALUATION FORM

	Below Average	Average	Above Average	Well Above Average	Outstanding
1. Learning Ability – acquires knowledge accurately and quickly	(A)	(B)	(C)	(D)	(E)
2. Leadership–effectiveness in getting ideas accepted and in guiding others to accomplish a task	(A)	(B)	(C)	(D)	(E)
3. Quality of Work – produces work of high quality	(A)	(B)	(C)	(D)	(E)
4. Motivation – strong desires to accomplish goals and objectives	(A)	(B)	(C)	(D)	(E)
5. Follows Instructions – follows directions as prescribed	(A)	(B)	(C)	(D)	(E)
6. Bearing and Behavior – maintains professional conduct and appearance	(A)	(B)	(C)	(D)	(E)
7. Accuracy – precision and carefulness in work performance	(A)	(B)	(C)	(D)	(E)
8. Oral Communication – expresses ideas clearly, logically, and grammatically in conversation	(A)	(B)	(C)	(D)	(E)
9. Problem Analysis – identifies and analyzes problems which require action	(A)	(B)	(C)	(D)	(E)
10. Initiative – self-starting, rarely needs a push to get going	(A)	(B)	(C)	(C)	(E)
11. Quantity of Work – accomplishes a large amount of work	(A)	(B)	(C)	(D)	(E)
12. Written Communication – expresses ideas clearly in writing with good grammatical form	(A)	(B)	(C)	(D)	(E)
13. Punctuality – prompt in keeping engagements	(A)	(B)	(C)	(D)	(E)
14. Adaptability – changes attitude and behavior to meet the demands of the situation	(A)	(B)	(C)	(D)	(E)
15. Dependability – does assigned tasks conscientiously without close supervision	(A)	(B)	(C)	(D)	(E)
16. Emotional Stability – stability and calmness under pressure and opposition	(A)	(B)	(C)	(D)	(E)

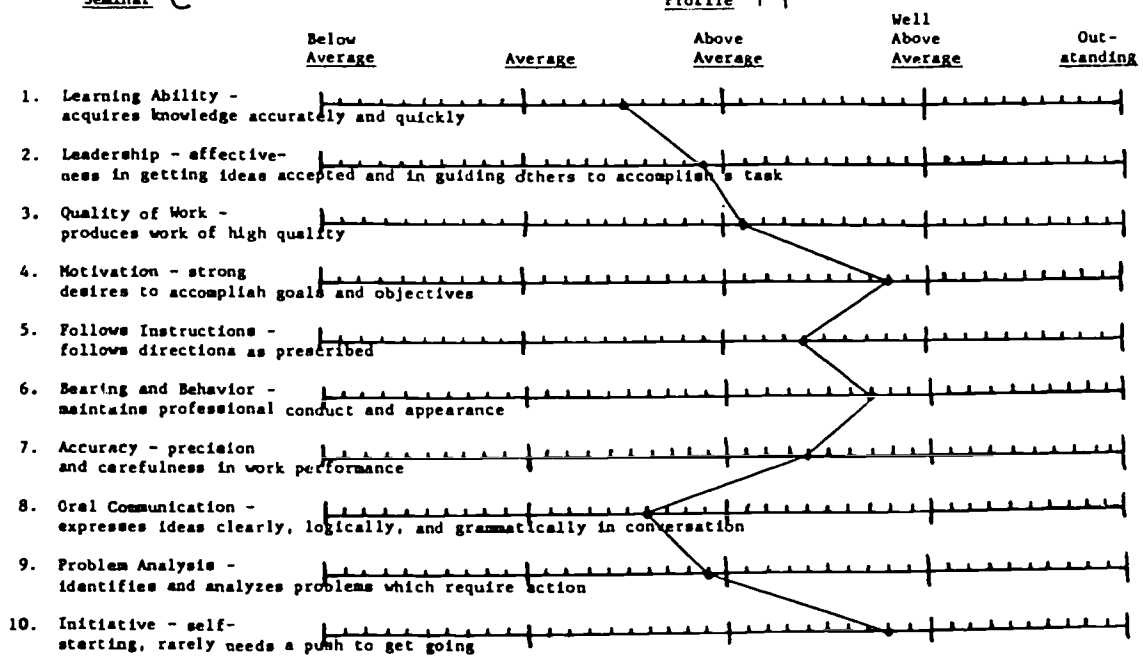
	Below Average	Average	Above Average	Well Above Average	Outstanding
17. Human Relations – gets along well with fellow workers and works effectively with them	(A)	(B)	(C)	(D)	(E)
18. Judgment – makes good decisions among competing alternatives	(A)	(B)	(C)	(D)	(E)
19. Knowledge of Duties – understands the requirements for effective work performance	(A)	(B)	(C)	(D)	(E)
20. Honesty – straight-forward and truthful in dealing with others	(A)	(B)	(C)	(D)	(E)



Evaluation Form V

Seminar C

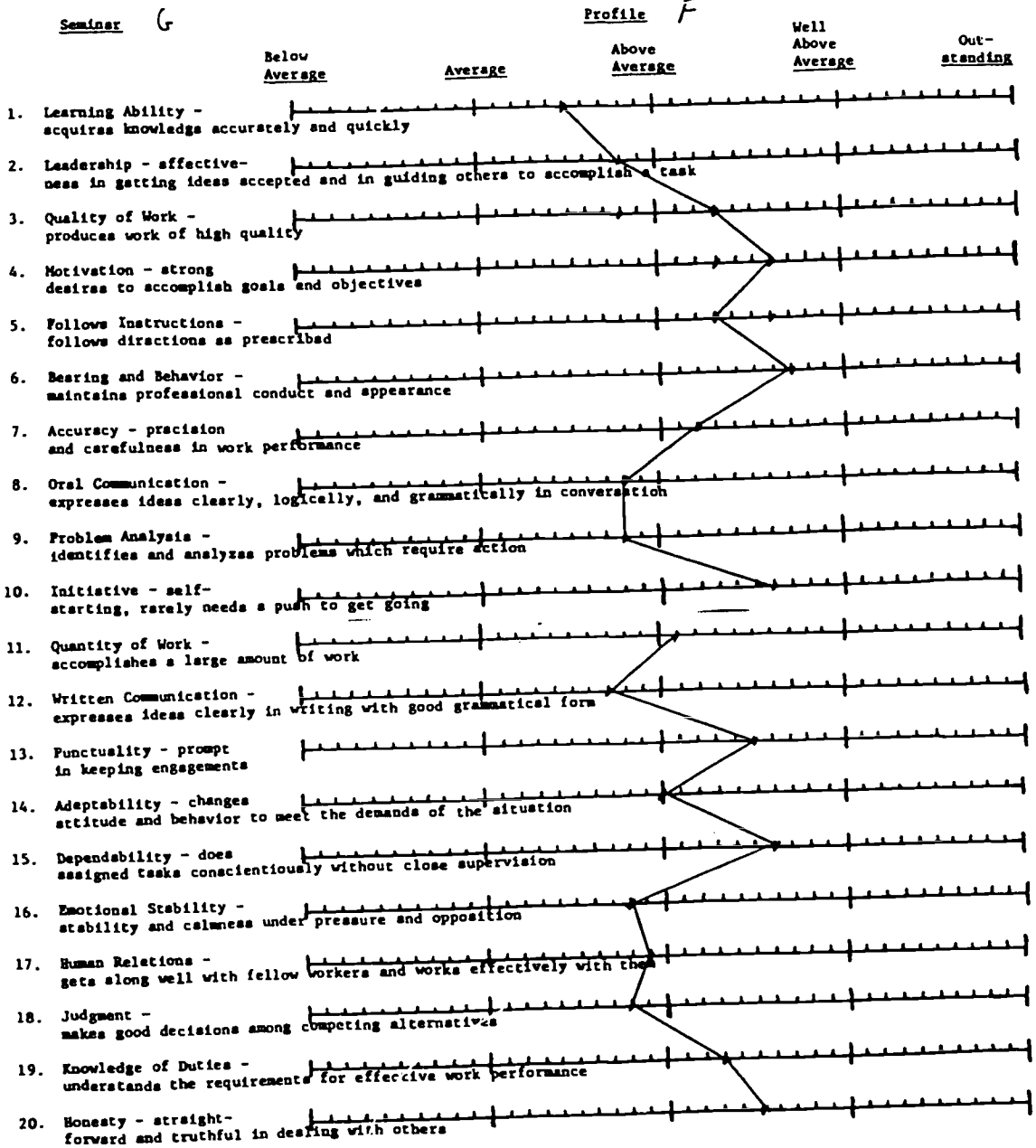
Profile M



Evaluation Form VI

Seminar G

Profile F



☆U.S. GOVERNMENT PRINTING OFFICE:1979- 671-056/49