AD-A066 779 AIR FORCE HUMAN RESOURCES LAB BROOKS AFB TEX F/G 5/10
RATER ACCURACY STUDY.(U)
FEB 79 C J MULLINS, K SEIDLING, J WILBOURN
UNCLASSIFIED AFHRL-TR-78-89 NL

| OF |
AD
A066779

END
DATE
FILMED
5-79
DDC

AFHRL-TR-78-89

AIR FORCE HUMAN RESOURCES LABORATORY

RATER ACCURACY STUDY

By

Cecil J. Mullins
Kevin Seidling
James Wilbourn
James A. Earles

PERSONNEL RESEARCH DIVISION
Brooks Air Force Base, Texas 78235

February 1979
Final Report for Period 9 March 1976 – 30 September 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, 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 and cleared for open publication and/or public release by the appropriate Office of Information (OI) in accordance with AFR 190-17 and DoDD 5230.9. There is no objection to unlimited distribution of this report to the public at large, or by DDC to the National Technical Information Service (NTIS).

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-78-89

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle): RATER ACCURACY STUDY •

5. TYPE OF REPORT & PERIOD COVERED: Final Rept. 9 March 1976 – 30 September 1978

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s): Cecil J. Mullins, James A. Earles, Kevin Seidling, James 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
2313T603
17 T6

11. CONTROLLING OFFICE NAME AND ADDRESS
HQ Air Force Human Resources Laboratory (AFSC)
Brooks Air Force Base, Texas 78235

12. REPORT DATE
February 1979

13. NUMBER OF PAGES
18

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
12 17 p.

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

SM Study Nrs. 6451, 6446, 6806

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

rater accuracy
ratings
peer ratings

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

Eight hundred eighty-two airmen were divided into more accurate and less accurate rating groups on the basis of their ability to estimate scores of their peers on a vocabulary test. To test whether the method actually did separate more accurate from less accurate raters, correlations were calculated within the more and less accurate groups between ratings of carefulness and scores on carefulness tests and between ratings of decisiveness and scores on decisiveness tests. The analysis consisted of counting the number of times the correlations between test scores and ratings in the more accurate group were larger than the analogous correlations in the less accurate group and computing the probability that this number of differences in the predicted direction might be expected by chance. It appears that this method of identifying accurate raters does work reasonably well. Several auxiliary questions

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

Unclassified

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

404415

79 04 02 112

Item 20 Continued:

concerning the best qualities to test in the estimating part of the study and the generalizability of rater accuracy to different personality characteristics were asked and answered.

PREFACE

This research was conducted under project 2313, Force Acquisition, Assignment, and Evaluation; task 2313T603, The Identification and Prediction of Rater Accuracy.

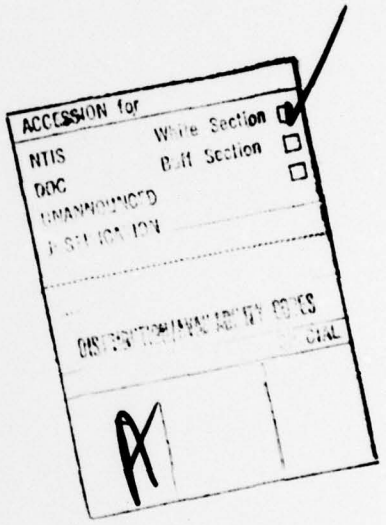

TABLE OF CONTENTS

LIST OF TABLES

RATER ACCURACY STUDY

I. INTRODUCTION

A previous study (Mullins & Force, 1962) indicated that there are measurable individual differences in rater accuracy. Briefly, the research design tested the hypothesis that raters who were more accurate in one rating situation would be more accurate in a second rating situation. Rater accuracy was identified by requiring all raters to estimate the scores of their peers on a vocabulary test and then summing for each rater the differences between the estimates and the actual scores. Ratings on carefulness and scores on five carefulness tests were similarly collected. It was found that the correlations between carefulness test scores and ratings were uniformly higher in the more accurate rating group than in the less accurate group.

The purpose of this study was to replicate the Mullins and Force study, and to extend it in the following ways:

1. Rater accuracy was identified both by using differences between rater estimates of vocabulary scores and the scores actually made and also by the sum of differences between estimated and actual scores on a mathematics knowledge score and a composite of both mathematics and vocabulary difference scores.

2. The earlier work studied only flights (N about 50) as rating groups. After the earlier study was done, a hypothesis was developed that raters probably could more accurately estimate the performance of their seven squad members than they could the performance of those members of their flights who were not in their squads. This study used both flights and squads as rating groups. Since squads ranged in size from four to eight, the more-accurate and less-accurate groups were composed of the upper and lower 50% of squad members, rather than 25%. When flights were used as rating groups, the upper and lower 25% were identified as before.

3. Cronbach (1955) has proposed a correlational measure, called differential accuracy (DAr), for use as an indicator of rater accuracy. Cronbach pointed out that the absolute difference score in some situations may contain as many as seven unrelated sources of variance and that the DAr score may be more appropriate. The DAr score consists of a correlation of the ratings given by each rater for a characteristic with a true measure of that characteristic across all ratees. Later investigators (i.e., Borman, 1977; Borman, Hough, & Dunnette, 1976) have used the DAr score with r to z transformations. In addition to replicating the earlier study using raw difference scores as a measure of rater accuracy, the DAr score was also used in this study, but only on groups consisting of complete flights.

4. This study extended the earlier study by attempting to find predictors of rater accuracy.

II. METHOD

Sample and Procedure

The subjects used in this study were basic airmen reporting for experimental testing at Lackland AFB. During the experimental testing session, the airmen were required to indicate their flight on their answer sheets, and to which of the four squads they belonged, so that the proper rating groups could be set up later. During this session, a vocabulary test, a mathematics knowledge test, and four experimental tests selected in the hope that they might predict the rater accuracy scores, were administered to all subjects.

Approximately 3 weeks later, the airmen were required to estimate the scores that each of their flight members made on the vocabulary and mathematics knowledge tests, copies of which were given to the subjects to refresh their memories. They were told that on the vocabulary test the lowest score in the flight

was 6, the highest score was 29, and the average score was 17. The subjects were told that on the mathematics knowledge test the lowest score was 4, the highest score was 20, and the average was 10. These were arbitrary numbers and were given to the subjects only to furnish a standard frame of reference for making their estimates.

At the time the estimates were taken (i.e., during the second testing session), the airmen also rated the members of their flight on carefulness and on decisiveness, and two tests each of carefulness and decisiveness were administered (described in Appendix A).

When all the data had been collected and scored, each subject's estimate of the performance of each peer on the vocabulary test was compared with the actual performance of each peer, and an absolute difference score, ignoring sign, was assigned to each estimated-actual combination. These difference scores were then averaged to provide an average "miss" score for each subject.

On the basis of this average absolute difference score calculated for each rater, the upper half, upper quarter, lower half, and lower quarter of each rating group were identified. "Upper" and "lower," throughout this report, will refer to the most accurate (smaller average difference score) and less accurate (larger average difference score), respectively. This identification process was repeated using the difference between the mathematics knowledge test scores and the mathematics test score estimations and was repeated again using a composite of the two. The upper and lower quarters had rated their entire flights, and all their ratings were used.

In addition to this method of calculating rater accuracy scores, the DAr component, suggested by Cronbach and described above, provided another index of accuracy. The DAr score was used to identify the more accurate (upper) 25% and the less accurate (lower) 25% of raters in each flight. The DAr score was not used on squad-level groups, because the main purpose of using the DAr score was to compare it with the previously used absolute difference score, and the N's on the squad level were too small to make the comparisons on upper-lower 25% subsamples.

Only those airmen who remained in the same squad during the course of basic training were included in the sample. A number of airmen missed either the testing session or the rating session, so that the final N for the entire group was 882.

On the basis of the average absolute difference score, seven pairs of contrasting subgroups of raters were formed:

1. Upper and lower 25% of flight when estimating vocabulary scores.
2. Upper and lower 25% of flight when estimating mathematics knowledge scores.
3. Upper and lower 25% of flight on a composite of vocabulary estimates and mathematics knowledge estimates.
4. Upper and lower 50% of raters when they estimate vocabulary scores of members of their squad only.
5. Upper and lower 50% of raters when they estimate mathematics knowledge scores of members of their squad only.
6. Upper and lower 50% of raters when they estimate vocabulary scores of members of their flight not in their squad.
7. Upper and lower 50% of raters when they estimate mathematics knowledge scores of members of their flight not in their squad.

Analysis

Correlations were computed between the carefulness tests and carefulness ratings within each of the upper and lower rating groups as defined above. These correlations are given in Tables 1 through 4. If the method for identifying more and less accurate raters is efficient, and if the ability to estimate scores on vocabulary and mathematics knowledge is correlated with ability to rate carefulness, and if the carefulness

Table 1. **Correlations of Carefulness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 25% of Flights**
(N = 882)

	Carefulness Ratings					
Carefulness Tests	Flight Upper 25% Vocabulary	Flight Lower 25% Vocabulary	Flight Upper 25 % Math Knowledge	Flight Lower 25% Math Knowledge	Flight Upper 25% Composite	Flight Lower 25% Composite
1. Letter Comparison	.06	.03	.05	.04	.06	.03
2. Score Checking	.28	.24	.28	.21	.30	.21
Mean	3.09	2.93	3.08	2.98	3.11	2.94
SD	.58	.58	.57	.58	.59	.58

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 2. **Correlations of Carefulness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 50% of Squad Members**
(N = 882)

Carefulness Tests	Squad Upper 50% Vocabulary	Squad Lower 50% Vocabulary	Squad Upper 50% Math Knowledge	Squad Lower 50% Math Knowledge
Letter Comparison	.07	.01	.06	.05
Score Checking	.21	.22	.24	.20
Mean	3.24	3.17	3.20	3.19
SD	.64	.68	.65	.66

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 3. **Correlations of Carefulness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 50% of Non-Squad Flight Members**
(N = 882)

Carefulness Tests	Non-Squad Upper 50% Vocabulary	Non-Squad Lower 50% Vocabulary	Non-Squad Upper 50% Math Knowledge	Non-Squad Lower 50% Math Knowledge
Letter Comparison	.07	.02	.04	.05
Score Checking	.29	.24	.29	.25
Mean	3.03	2.91	3.02	2.92
SD	.56	.55	.56	.55

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 4. **Correlations of Carefulness Tests with Ratings from More and Less Accurate Raters Defined by DAr Scores**

(N = 882)

Carefulness Tests	Carefulness Ratings					
	Flight Upper 25% Vocabulary	Flight Lower 25% Vocabulary	Flight Upper 25% Math Knowledge	Flight Lower 25% Math Knowledge	Flight Upper 25% Composite	Flight Lower 25% Composite
1. Letter Comparison	.04	.04	.05	.03	.06	.03
2. Score Checking	.30	.18	.29	.19	.31	.16
Mean	2.99	2.99	3.03	3.00	3.01	3.02
SD	.59	.56	.60	.55	.62	.55

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

tests do measure carefulness as it is defined in the carefulness ratings, and if the carefulness ratings are also measuring carefulness as defined, then higher correlations would be expected in the group of more accurate raters (upper) than in the group of less accurate raters (lower). These are a lot of "ifs," and it was decided to compare the upper with the lower rating groups by a simple count of the number of times the differences between correlations were in the predicted direction, regardless of the size of the correlations.

This is an unusual method of analyzing correlational results, but it appears more sensitive than the usual method, which would require that a test of significance of differences between two such correlations produce a t-ratio significantly greater than chance. That is a very crude approach for use in a trend analysis of this type, where the question is, "Do the correlations in this group as a whole tend to be greater than the correlations in that group?" The conventional method would require access to tests of the rated characteristics (carefulness and decisiveness) which were known to have validities uncommonly high for tests in this general domain; it would require the collection of uncommonly good ratings from these groups of naive subjects; and it would require a very powerful method of separating good from bad raters before there could be any chance for achieving significant differences for *all* comparisons. With the conventional approach, if there should be several significant differences in the predicted direction, as is the case in this study (e.g., one of the three Score Checking comparisons in Table 1 and all three Score Checking comparisons in Table 4), and also several differences not statistically significant, the investigator is left in an ambiguous position. He has both confirmed and failed to confirm his hypothesis.

On the other hand, the method used in this analysis is at least as accurate a method for subjecting the hypotheses in this study to tests of significance, and it is much more sensitive. It is a method which can be applied simultaneously to all comparisons, whether either of the correlations (or both) is significant and without having to be concerned with the distribution of the *r*'s involved.

If the data being studied are random or if the subjects are randomly divided into two groups, then correlation coefficients computed from those data should be greater in one of the groups equally as often as in the other. There should not be a significantly greater number of correlations in the predicted direction than in the other, within the limits set by the binomial distribution. Note that this is true whether the comparison is between correlation coefficients of "significant" size or not. Significance level of an individual correlation coefficient is an important consideration in many circumstances, but not in the one addressed by this study, where an overall trend is the phenomenon of interest.

In this analysis, the observations are the comparisons between analogous correlations in the two groups. Probabilities concerning the number of times larger correlations may be expected to appear by chance in the predicted direction can be computed from the basic binomial formula available in most mathematics and statistics textbooks, and tables are available (e.g., Bartz, 1976, p. 386; Freund, 1967, p. 392) from which the probabilities for this sort of occurrence can be read directly.

III. RESULTS AND DISCUSSION

It is clear from Table 1 that all six of the comparisons of interest are in the predicted direction for flights (p = .016), when absolute difference scores are used as the measure of accuracy, although the correlations between carefulness ratings and Letter Comparison are not significant beyond the .05 level. The rating groups in Table 1 were formed by using only the upper and lower 25% of flights, leaving out the middle 50%. This procedure forced more separation between the more accurate and the less accurate raters if it is granted that estimation of vocabulary and mathematics knowledge scores is a workable method of identifying rater accuracy.

Tables 2 and 3 show correlations involving smaller groups, comprised of the upper and lower 50% on rater accuracy, defined in terms of absolute difference scores. In both these tables, three of the four differences are in the predicted direction. The probability that three or more of these four comparisons will be in the predicted direction is .313, clearly not very impressive. One surprise in these two tables is that raters estimate the performance of members of their own squads no better than they do those members of their flights who are not in their squads. It was hypothesized that raters would know members of their own squads better than they would know members of other squads, but that does not appear to be the case. It was discovered in pursuing this point that squad members do not necessarily live closer together in the sleeping bays, nor do their daily activities bring them closer together, than non-squad members. A significant portion of a basic airman's time is spent sitting or lying on his bunk studying, repairing his clothes and equipment, and doing similar tasks. All the people facing him across a narrow aisle are members of another squad. His own squad members are arranged to his left and right, in a less convenient position for conversation. Also, recruits apparently seek out other people, from their own geographic area, whom they have previously met and came to know on the trip to Lackland AFB, and with whom they have interacted to a considerable extent during processing and before they have been assigned to squads. So, after the fact, this finding is not so surprising.

Table 4 is similar to Table 1, except that the DAr score was the basis for separating the raters into more and less accurate rating groups, rather than the average absolute difference score. The DAr score approach provided only five of the six important comparisons in the predicted direction (p = .109), although it yielded somewhat greater separation between more accurate and less accurate groups when scores on *Score Checking* (which gave the better results of the two carefulness tests) are correlated with carefulness ratings.

Tables 5 through 8 contain results of a similar analysis, using correlations between decisiveness ratings and decisiveness test scores the same way carefulness was used in Tables 1 through 4. Table 5 shows that

Table 5. **Correlations of Decisiveness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 25% of Flights**
(N = 882)

	Decisiveness Ratings					
Decisiveness Tests	Upper Flight Vocabulary	Lower Flight Vocabulary	Upper Flight Math Knowledge	Lower Flight Math Knowledge	Upper Flight Composite	Lower Flight Composite
1. Preference Scale	.05	.06	.06	.03	.08	.06
2. Dot Estimation	.10	.04	.06	.07	.12	.06
Mean	3.48	3.27	3.45	3.32	3.50	3.27
SD	.64	.66	.65	.63	.65	.66

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 6. **Correlations of Decisiveness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 50% of Squad Members**
(N = 882)

Decisiveness Tests	Squad Upper 50% Vocabulary	Squad Lower 50% Vocabulary	Squad Upper 50% Math Knowledge	Squad Lower 50% Math Knowledge
Preference Scale	.05	.03	.03	.03
Dot Estimation	.06	.03	.02	.07
Mean	3.60	3.50	3.54	3.55
SD	.74	.77	.75	.76

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 7. **Correlations of Decisiveness Tests with Ratings from More and Less Accurate Raters Defined by Absolute Difference Scores, Upper and Lower 50% of Non-Squad Flight Members**
(N = 882)

Decisiveness Tests	Non-Squad Upper 50% Vocabulary	Non-Squad Lower 50% Vocabulary	Non-Squad Upper 50% Math Knowledge	Non-Squad Lower 50% Math Knowledge
Preference Scale	.08	.05	.09	.04
Dot Estimation	.10	.06	.10	.06
Mean	3.40	3.25	3.40	3.26
SD	.61	.64	.62	.63

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

Table 8. **Correlations of Decisiveness Tests with Ratings from More and Less Accurate Rater Defined by DAr Scores**
(N = 882)

	Decisiveness Ratings					
Decisiveness Tests	Flight Upper 25% Vocabulary	Flight Lower 25% Vocabulary	Flight Upper 25% Math Knowledge	Flight Lower 25% Math Knowledge	Flight Upper 25% Composite	Flight Lower 25% Composite
1. Preference Scale	.06	.01	.07	.03	.07	.02
2. Dot Estimation	.08	.04	.09	.07	.09	.06
Mean	3.35	3.34	3.39	3.37	3.36	3.36
SD	.66	.64	.65	.62	.68	.62

Note. — .05 Level = .07; .01 Level = .09; "Upper" means more accurate raters.

when the average absolute difference scores are used to identify the accurate and inaccurate 25% subsamples, four of the six appropriate comparisons of *r*'s are in the predicted direction ($p = .344$). Of course, this is also not very impressive. Table 6, showing the results when accurate and inaccurate groups are formed on the basis of their ability to estimate performance of their own squad members, half of the comparisons are in the predicted direction and half are not. Table 7 indicates that the subjects can better estimate the performance of peers who are not in their own squads, because the four comparisons are all in the predicted direction ($p = .063$). The DAr score is examined again in Table 8, this time using decisiveness rather than carefulness as the rated dimension. All six of the appropriate comparisons in this table are in the predicted direction ($p = .016$).

A part of the purpose of this study was to find predictors for the rater accuracy variables. As is evident from Table 9, the predictions of both the absolute difference score and the DAr score were very poor. The highest correlation between either of the rater accuracy scores and any of the 13 predictor scores was only .20 (between DAr and ASVAB-Mech). None of the nine experimental predictor scores (described in Appendix A) correlated with either rater accuracy variable higher than .14. Several multiple R's computed against the average difference score were equally disappointing. The rater scores were simply not predictable to any practical degree by the predictor variables used in this study.

IV. SUMMARY AND CONCLUSIONS

As stated at the beginning of this paper, this study attempted to replicate the findings of the Mullins and Force (1962) study and added to it in five ways. It seems appropriate now to take stock of the results as they bear on these issues. The replication is a rather small part of this study and occupies only the first two columns of Table 1. As can be seen, all six of the comparisons of the two carefulness tests are in the predicted direction ($p = .016$), as in the 1962 study.

Extension 1. Extending the basic technique of selecting better and poorer raters by using differences between mathematics test scores and estimates of those scores, and by using a combination of these differences with the vocabulary differences as used in the previous study, was, in general, rather successful. Although identification by mathematics difference scores was apparently not as good as identification by vocabulary scores, the best technique of all seems to be identification on the basis of the composite scores (vocabulary and mathematics). Each one of the eight comparisons involving the composite scores (Tables 1, 4, 5, and 8) was in the predicted direction ($p = .004$) for both the carefulness and decisiveness sections of the analysis.

Extension 2. The extension of generality of rating accuracy to both carefulness and decisiveness ratings was moderately successful. Of the 12 comparisons made between carefulness test scores and carefulness ratings (Tables 1 and 4), 11 were in the predicted direction ($p = .003$). Of the 12 analogous decisiveness comparisons (Tables 5 and 8), 10 were in the predicted direction ($p = .019$). Both of these are significantly different from chance expectation, although carefulness was slightly better. It appears, then, that ability to estimate performance of peers on vocabulary and on mathematics tests is associated with ability to rate both carefulness and decisiveness. It seems clear that rater accuracy is not specific to a single dimension or quality, but is rather generalizable.

Extension 3. In this study, basic airmen did not rate their squad peers more accurately than they did other members of their flight. This does not mean that proximity and familiarity with the ratee are unimportant in the accuracy of ratings, but that the particular mechanics of squad formation and duties appear to throw the basic trainee into contact with non-squad flight members at least as often as with squad members. Because of these administrative arrangements in squad formation, this extension could not be adequately tested.

Extension 4. The DAr score seems to be slightly better as a measure of rater accuracy than the absolute difference score. Of the 12 comparisons made between groups formed using the DAr score (Tables

Table 9. **Intercorrelations Among Difference Scores, DAr Scores, and Predictor Variables**
(N = 882)

	ASVAB				Experimental Predictors									Difference Scores			DAr Scores		
Rated Dimension	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**
1. ASVAB–MECH	1.00	.52	.33	.56	.12	.11	–.03	.24	.01	.00	.03	.01	.11	.00	–.13	–.06	.16	.16	.20
2. ASVAB–ADM		1.00	.07	.25	.14	.13	.04	.18	.08	–.03	–.01	.03	.09	–.03	–.12	–.08	.12	.11	.13
3. ASVAB–GEN			1.00	.57	.10	.10	–.29	.14	–.02	.05	.02	.04	–.04	.01	–.07	–.02	.05	.07	.05
4. ASVAB–ELEC				1.00	.11	.11	–.17	.17	–.04	.01	.03	–.02	.07	.03	–.09	–.02	.05	.07	.09
5. Unusual Facts I					1.00	.92	–.02	.16	.08	–.02	–.01	.09	.02	.01	–.02	–.01	.03	.05	.03
6. Unusual Facts II						1.00	–.02	.14	.08	.00	.01	.11	.04	.00	–.02	–.01	.02	.05	.02
7. Activities Pref							1.00	.10	.32	–.16	–.24	.18	–.09	–.03	–.03	–.04	.05	.02	.05
8. Work Mngs–C								1.00	.04	–.05	–.06	–.05	–.01	–.12	–.10	–.14	.09	.07	.11
9. FCSRI–A I									1.00	–.62	–.77	.62	–.45	.10	.02	.09	.00	.01	.03
10. FCSRI–A II										1.00	.75	–.18	.37	–.08	–.03	–.08	.03	–.02	–.02
11. FCSRI–A III											1.00	–.43	.53	–.09	–.01	–.07	.03	–.02	–.02
12. FCSRI–A IV												1.00	–.41	.02	–.01	.01	.00	.01	.01
13. FCSRI–A V													1.00	–.06	–.04	–.06	.05	.02	.04
14. Vocab Diff Score														1.00	.32	.91	–.36	–.06	–.24
15. Math Diff Score															1.00	.69	–.17	–.52	–.36
16. Comp Diff Score																1.00	–.35	–.28	–.34
17. Voc DAr Score																	1.00	.31	.76
18. Math DAr Score																		1.00	.71
19. Comp DAr Score																			1.00

Note. — .05 Level = .07; .01 Level = .09.

4 and 8), 11 were in the predicted direction (p = .003). Only 10 of the 12 comparisons using the average absolute difference score (Tables 1 and 5) were in the predicted direction (p = .019).

Extension 5. Obviously, it may be that future work may uncover variables which are useful in predicting rater accuracy scores. The predictor instruments used in this study, however, were of no practical usefulness for this purpose.

REFERENCES

Bartz, A.E. *Basic statistical concepts in education and the behavioral sciences.* Minneapolis: Burgess Publishing Company, 1976.

Borman, W.C. Consistency of rating accuracy and rating errors in the judgment of human performance. *Organizational Behavior and Human Performance,* 1977, **20**, 238–252.

Borman, W.C., Hough, L.M., & Dunnette, M.D. *Performance ratings: An investigation of reliability, accuracy, and relationships between individual differences and rater error.* Minneapolis: Personnel Decisions, Inc., 1976.

Cronbach, L.J. Processes affecting scores on "understanding of others" and "assumed similarity." *Psychological Bulletin,* 1955, **52**(3), 177–193.

Freund, J.E. *Modern elementary statistics.* Englewood Cliffs: Prentice-Hall, 1967.

Mullins, C.J., & Force, R.C. Rater accuracy as a generalized ability. *Journal of Applied Psychology,* 1962, **46**(3), 191–193.

Norman, W.T. Development of self-report tests to measure personality factors identified from peer nominations. ASD-TN-61-44, AD-267 779. Lackland AFB, TX: Personnel Laboratory, Aeronautical Systems Division, May 1961.

APPENDIX A: DESCRIPTION OF TESTS

Letter Comparison. This test requires the subject to compare a group of randomly typed letters on the left side of the page with a similar group of letters on the right side of the page, and to mark whether the two groups are identical. Each pair of letter groups constitutes one item, and there is never more than one letter difference between the groups. Twenty items, 28-minute time limit. (Carefulness)

Score Checking. The subject is presented a sheet showing 10 three-digit scores for each of 43 persons. On the back side of the same page is, supposedly, a reproduction of the first page. The tabulation on the back of the page has been rotated 90 degrees, and the subject must check for transcription errors. There are 43 score discrepancies. Forty-three items, 28-minute time limit. (Carefulness)

Preference Scale. The subject is presented a list of 62 activities for which he is to indicate a degree of preference using a 5-point scale. He does not know that it will be timed and that he will be stopped in 5 minutes. Score is number of preferences indicated. (Decisiveness)

Dot Estimation. Each item of this test consists of a pair of 1-inch squares. In each square are 10 to 25 small dots. The subject must choose which square has the most dots, for 110 items, within a 4-minute time limit. Score is the number of items attempted, right or wrong. (Decisiveness)

Unusual Facts. This test was designed to measure the tendency for subjects to over-estimate or under-estimate. For example, one item is "the heaviest human being weighed," and four alternatives follow, ranging from 900 pounds to 3,400 pounds. There are no right or wrong answers. There are two scores, both reflecting the tendency for subjects to give responses toward the extreme ends of the range of the alternatives. There are 70 questions. The test was carried to 100% completion.

Activities Preference. This appears to be a forced-choice interest test; however, each pair of activities comprising an item was selected so that one of the pair involved working or playing alone. It is a test of gregariousness; for example, "I would rather (a) collect autographs, (b) collect stamps." There are 60 questions. The test was carried to 100% completion.

Word Meanings—C. This test is composed of ten items, each stating a category such as "Which of the following are carpentry terms?" Each item is followed by fifteen alternatives, at least one of which is obviously in the stated category (e.g., "hammer") and at least one of which is not (e.g., "tiger"). The other alternatives could be interpreted either way (e.g., "oak"). The subject is told to mark all the alternatives he thinks belong in the stated category. His score is the total number of alternatives marked. This test was designed as a measure of conceptual complexity. It was carried to 100% completion.

FCSRI—A. The initials stand for Forced–Choice Self Report Inventory, Form A. The FCSRI–A is one of a battery of tests prepared by Warren T. Norman (Norman, 1961) under contract. It consists of 192 forced-choice questions, such as "A. I have no fear of spiders," or "B. I am a modest person." Each question is a pair of such statements from which the subject picks one as more descriptive of himself. The test was designed to measure five factors (surgency, agreeableness, dependability, emotional stability, and culture), and it produces five scores.

☆U.S. GOVERNMENT PRINTING OFFICE: 1979– 671-056/18