RESEARCH ON THE DEVELOPMENT OF SHIPBOARD PERFORMANCE MEASURES AND PERFORMANCE JUDGMENTS

Final Report

HUMAN FACTORS RESEARCH, INCORPORATED
SANTA BARBARA • LOS ANGELES • SAN DIEGO

RESEARCH ON THE DEVELOPMENT OF SHIPBOARD PERFORMANCE MEASURES AND PERFORMANCE JUDGMENTS

Final Report

Prepared for

Personnel and Training Branch
Psychological Sciences Division
Office of Naval Research
Department of the Navy

Prepared by

Human Factors Research, Incorporated
Santa Barbara * Los Angeles * San Diego

January 1965
Contract No. Nonr 1241(00)
NR-153-165

Table of Contents

RESEARCH ON THE DEVELOPMENT OF SHIPBOARD PERFORMANCE MEASURES AND PERFORMANCE JUDGMENTS

This is the final report on Contract Nonr 1241(00), "Research on the Measurement of Shipboard Performance Aboard Submarines," between the Personnel and Training Branch, Psychological Sciences Division, Office of Naval Research and Human Factors Research, Incorporated.

This report contains a brief summary of the research conducted and a list of technical reports prepared and distributed under the contract.

Research on Shipboard Performance Measures

The initial work was concerned with the measurement of the performance of enlisted personnel serving aboard submarines. The nine technical reports of the research in this area are described below.

Wilson, C. L. and Mackie, R. R. Research on the development of shipboard performance measures, Part I. The use of practical performance tests in the measurement of shipboard performance of enlisted naval personnel, 1952.

This report describes the results of the administration of two performance test batteries, one designed for enginemen (ENs) and one for electrician mates (EMs) serving aboard submarines. It was concluded that performance tests can be designed which are reliable, objectively scorable, easily administered, and which measure performance factors not generally measured by rating methods or written tests. It was suggested that job sample performance tests could be used by the Navy as part of the advancement in rate examinations, as criteria for training programs, and as shipboard criteria for validating selection and training procedures.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part II. The use of a performance rating scale in the measurement of shipboard performance of enlisted naval personnel, 1954.

This report describes the development of a (10-trait) graphic man-to-man rating scale and the results of its use in evaluating 320 EMs and 487 ENs serving in the Atlantic and Pacific submarine fleets. The reliability of the ratings was found to be high: the rate-rerate reliability estimate was .88, and the interrater agreement index was .70.

Factor analyses of the trait intercorrelations yielded two major factors: technical competence and personal adjustment to submarine life.

The following were some of the conclusions of the study:

1. Reliable and discriminating ratings can be made by Navy officer and petty officer personnel. A man-to-man rating format appears to enhance discrimination and reliability.

2. Ratings should be made separately for different rates and pay grades.

3. Ratings should be supplemented by performance tests.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part III. The use of performance check lists in the measurement of shipboard performance of enlisted naval personnel, 1954.

This report contains a description of the development of performance check lists and the results of their subsequent use in assessing the performance of 313 EMs and 511 ENs serving aboard submarines of the Atlantic and Pacific fleets.

It was concluded from the results of the analyses that reliable and discriminating performance check lists can be constructed to measure the shipboard performance of Navy enlisted men. However,

this general conclusion had to be evaluated in the light of the following more specific conclusions, some of which point to shortcomings in the check-list approach.

1. Pay Grade. Since many administrative decisions in the Navy require distinguishing among the abilities of men within a particular pay grade, it is essential, if ratings are to be used as criteria of performance, to achieve satisfactory interrater agreement and discrimination within pay grade. Results of this study indicated that check lists can be constructed which will produce the desired discrimination and interrater agreement within pay grade. Agreement may be greater within the higher pay grades and more substantial for some rates than others, depending generally on how discrete and observable the job-tasks are and how much opportunity the raters have had to observe performance.

2. Selection of Tasks. To insure that performance check lists satisfactorily fulfill their purpose, the task-items included should be

   a) Representative of the job

   b) Important to the job

   c) Of appropriate difficulty level

   d) Performed frequently enough so that raters will have had an opportunity to observe their men performing them

   e) Of such a discrete nature that it is reasonably clear to the raters whether or not their men can perform the task independent of the cooperation and suggestions of co-workers.

   The last two conditions will restrict the selection of tasks rather severely for some Navy rates and pay grades. For example, it was easier to select observable discrete tasks for EMs than for ENs. However, in neither rate was there a satisfactory number of tasks that most raters had had an opportunity to observe their men performing.

3. Officer vs. Enlisted Raters. The results of this study revealed no evidence of systematic differences in the reliability or leniency-stringency characteristics of the ratings assigned by officers and

leading petty officers. However, there was a real difference in the number of items checked by the two groups of raters. On the average, officers were able to check far fewer items than petty officers, and they readily admitted they lacked sufficient specific information for checking many items. Thus, in most instances, leading petty officers appeared to be in a better position to use performance check lists than officers.

4. Format of the Check List. The fact that satisfactory discrimination and interrater agreement were achieved within pay grade was attributed mainly to the format of the check lists which provided for direct man-to-man comparisons on all task-items. This is considered to be an important attribute for rating devices used in the Navy. Barring the development of a rating device which yields absolute scores, the frame of reference most likely to induce interrater agreement appears to be the men who are actually being rated. This means that all men should be rated on a selected task at the same time.

5. Performance Check Lists Compared to Other Criterion Measures. Insofar as generalizations are possible from this study, performance check lists appeared to be somewhat less effective than the other criterion measures studied. Specifically:

   a) The graphic man-to-man rating scale produced results that were as reliable and discriminating as the results of the check lists. However, the check lists took longer to construct than the rating scale and were applicable to only two rates; the rating scale, with the exception of one or two traits, would have been applicable to all rates.

   b) The rating scale apparently yielded measures of technical competence as meaningful as those of the more specific check lists. In addition, the rating scale yielded a measure of the personal adjustment aspects of shipboard behavior.

   c) All raters were able to rate their men on the rating scale characteristics, but the check list task-items were too specific and technical for many raters.

It was concluded that, if only one measure of shipboard performance were to be employed, the man-to-man rating scale would be the best choice since it produced measures of both technical and adjustive shipboard behavior. It appeared highly desirable, however, to employ actual tests of performance in conjunction with ratings wherever possible. While ratings can reliably establish the rank order of technical ability in a group of men, only an actual test of performance can establish the absolute level at which the group is performing. Strict comparisons of the performance level of men from different ships demand the performance test approach.

The performance check list may have advantages over other methods which were not fully explored in this study. If check lists were used regularly over a period of time, raters might sharpen their observations of men at work and, at the same time, the men might acquire a better appreciation of the variety of tasks they should be able to perform. In addition, because of the specificity of the task-items on which a man is rated, the check lists could be used to evaluate a man's progress in training more effectively than a more generalized rating scale.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part IV. A comparison between rated and tested ability to do certain job tasks, 1954.

Four types of performance measures had been developed and administered to EMs and ENs aboard submarines: performance rating scales, performance check lists, job sample performance tests, and written job knowledge tests.

The descriptions of these measures and the results of correlational analyses of total scores made on them were included in the first three reports.

Additional correlational studies were made of scores on similar items of three of the performance measures. The results of these analyses were the subject matter of this report. Below is a brief

description of the item-score correlational analyses performed:

1. Check List Task-Items vs. Job Sample Tests. Scores on selected task-items of the performance check lists were correlated with scores on individual job sample tests. The task-items selected were those that appeared to reflect the same knowledge or skills as a particular job sample test task.

2. Check List Task-Items vs. Written Job Knowledge Test Items. Scores derived from check-list ratings of men's abilities to perform a particular task were correlated with scores these same men made on written test items about the same task.

3. Written Test Items vs. Job Sample Tests. Scores made on a job sample test constructed around a particular task were correlated with scores the same men made on written test questions about that task.

The correlations obtained from these item-score analyses ranged from zero to moderately high values. While many were significantly greater than zero, on the average they were quite low indicating that ratings of men's abilities to perform a particular task (in this case check-list ratings) may not be in substantial agreement with the same men's scores on tests of their ability to perform the task.

The results of this study also revealed substantial differences between the ability to perform a particular task and the ability to answer written questions about the same task. These results support other observations that scores on written tests cannot always be accepted as valid indications of men's abilities to perform practical tasks.

Mackie, R. R., Wilson, C. L. and Buckner, D. N. Research on the development of shipboard performance measures, Part V. Interrelationships between aptitude test scores, performance in submarine school, and subsequent performance in submarines as determined by ratings and tests, 1954.

This report contained a description of a study designed to determine the extent to which the shipboard criterion measures were

predictable from Submarine School class standing and from a variety of aptitude tests administered to enlisted personnel upon entrance into the Submarine School.

In addition, the extent to which the Submarine School criterion was predictable from the same aptitude test scores was determined and a tentative identification of some of the basic variables in that criterion was made through a factor analysis.

From the results of the study, the following conclusions and implications were drawn:

1. Performance in Submarine School is highly predictable from a combination of scores on the Navy GCT, ARI, and Mechanical Knowledge tests and a test of the Direction Marking type.

2. Factor analytic results suggest that there are at least three major factors in the Submarine School criterion that are identifiable from aptitude test scores:

    a) Mechanical knowledge or comprehension

    b) Numerical facility

    c) A Grades Factor, which may reflect academic motivation.

3. Measures of shipboard performance are also predictable to a reasonable degree, primarily by Mechanical, Numerical, Reasoning, Direction Marking, and Visual Attention test scores in various combinations.

    Tests of practical performance, such as Job Sample tests, appear to be more predictable than ratings of abilities to perform specific tasks, ratings on general traits pertaining to job knowledge, or written job knowledge tests.

4. Submarine School standing is moderately related to subsequent shipboard performance, particularly the more technical aspects of that performance.

Harris, W., Mackie, R. R. and Wilson, C. L. Research on the development of performance criteria, Technical Report VI. Performance under stress: A review and critique of recent studies, 1956.

In the course of studies aimed at the measurement of shipboard performance, the question had been raised repeatedly, by both operational and staff personnel responsible for research on the selection of submariners, about the possible effect upon performance of prolonged submergence in new type submarines. The implication was that prolonged submergence and the resulting changes in the internal atmosphere of the boats might affect the crew members in such a way that their performance would become less proficient. There was, of course, the additional implication that personality or personal adjustment problems might arise as a result of prolonged submergence.

Consequently, staff members of this project undertook a survey of the literature of recent studies that bear on the general problem of human performance under adverse conditions. The purpose was to try to achieve some systemization in the wide variety of studies purportedly concerned with performance under stress and to derive implications for future studies.

The following major conclusions were drawn from the results of the survey.

1. From what results were available on individual performance, it was apparent that there are wide differences in individual reactions to stress. The reasons for these different reactions had not been clearly identified.

2. Whether or not these individual differences are temporary or lasting had not been explored. In fact, the reliability of performance from one stress situation to another appeared not to have been investigated.

3. The majority of the studies were concerned with the effect of relatively short-term stress conditions.

4. The period during which performance was measured was also of short duration, making temporary compensatory performance readily possible.

5. Some experimental stress stimuli employed appeared to have produced artifactual effects.

6. In many of the studies the intent of the experimenter was readily apparent to the subject or the experimental situation was clearly artificial.

7. The performance tasks did not include practical tasks similar to those likely to be encountered under operational conditions. Rather, they were abstract tasks, such as intelligence tests.

8. Often, these abstract tasks were superimposed in such a way as to represent a complete interruption in operational performance.

9. The level of performance as a function of time had never been studied systematically under either long-term or short-term stress conditions.

10. The temporal relationship between the stress conditions and the performance measure had not been systematically studied, in spite of its obvious importance. It is difficult to infer anything about performance <u>during</u> actual stress conditions if measurements are made only a considerable period of time <u>after</u> stress has been discontinued.

11. A number of experimental designs had been used in stress studies, but some which are most directly analogous to typical military operating conditions had not been used.

The general conclusion from this survey was that the studies which were reviewed did not provide information which could be extrapolated satisfactorily to operational performance under stress conditions. Performance measures that had been used were often neither operationally significant nor meaningful to the subjects. Changes in the level of performance (the <u>course</u> of performance) as a function of the duration of stress had not been systematically observed. And, although a number of experimental procedures had been used, few of them seemed analogous to actual field situations.

High, W. S. and Mackie, R. R. Research on the development of performance criteria, Technical Report VII. A factor analytic study of aptitudes, interests and practical performance skills for Navy machinery repairman students, 1957.

In this study over 40 objective performance measures and 50 reference test measures plus performance ratings were obtained on 200 machinery repairmen (MR) students. A 114-variable intercorrelation matrix was factor analyzed in an effort to determine whether basic MR skills might be identified and whether they might be related to basic aptitude factors. It was found that there was virtually no overlap between the skills measured by the performance tests and the abilities measured by conventional, reference examinations.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. J. appl. Psychol., 1959, 43, 60-64.

The hypothesis tested in this study was that high agreement among scores assigned the same men by different raters does not necessarily imply predictable ratings.

Rating scores on personal adjustment and technical competence traits made by three superior officers of 100 submariners serving aboard 21 different submarines were each divided into four samples so as to achieve four levels of interrater agreement: .94, .84, .69, and .00 for the technical competence ratings and .88, .90, .61, and .12 for the personal adjustment ratings. Correlations were then computed within each sample between three predictor variables (Submarine School Class Standing and the Navy General Classification and Mechanical Aptitude Tests) and the mean of the three ratings assigned to each ratee.

The hypothesis was supported. None of the 12 correlations between the predictor variables and the ratings for which the interrater agreement estimates were high (.94, .84, .88, and .90) was significantly different from zero. Six of the 12 correlations computed for the low agreement ratings (.69, .00, .61, and .12)

were significantly different from zero, two at the .01 level and four at the .05 level. Three of the six significant correlations were with the ratings for which the interrater agreement estimates were not significantly different from zero, .00 and .12.

It was concluded that high interrater agreement does not necessarily imply predictable ratings and may in some instances indicate a lack of predictability.

Mackie, R. R. and High, W. S. Research on the development of shipboard performance measures, Technical Report IX. Supervisory ratings and practical performance tests as complementary criteria of shipboard performance, 1959.

This report described a shipboard follow-up study of the performance of Navy machinery repairmen whose aptitudes, skills, interests and achievements had been thoroughly studied two years earlier while they were in Class "A" MR training.

Shipboard performance was assessed in two ways: (1) by administering a practical performance test requiring skill in the use of machinery repair equipment; (2) by securing ratings by supervising petty officers of each person's ability to perform the various aspects of the MR's shipboard job.

The results strongly suggested that performance tests and supervisory ratings were best regarded as complementary criteria of shipboard performance. While these two measures were found to be essentially uncorrelated with each other, each correlated significantly with many logical predictors, including aptitude and interest measures, scores on practical work during training, and predictions of success by Class "A" School instructors.

When the two shipboard measures were combined to form a simple composite criterion, it was estimated that over 50% of the true variance was accounted for by scores made two years earlier on a combination of: (1) mechanical knowledge tests; (2) training projects involving the use of lathe and milling machines; and

(3) predictions by school instructors as to eventual suitability as an MR.

Research on Judgments of Performance

Perhaps no other kinds of judgments are made as frequently--for economic, social, military, political, or scientific reasons--as judgments of the performance of others. The performance of both enlisted and officer personnel in the military services is periodically assessed by superiors. Similarly, the performance of personnel in industry--from laborers to managers--is periodically assessed for purposes of promotion and wage and salary adjustment. These assessments almost universally involve performance judgments. Yet it was apparent that there was very little knowledge about how such judgments are affected by numerous variables known or suspected to be operant in the performance judgment situation. Furthermore, the results of an earlier study under this contract on the predictability of ratings as a function of interrater agreement raised some doubt about the validity of the criteria commonly employed in assessing the worth of performance ratings.

In light of the ubiquity of performance judgments plus the lack of knowledge about them, research under Contract Nonr 1241(00) at this point assumed a new direction: the study of performance judgments under controlled laboratory conditions, where stimulus-response, as opposed to response-response, relations could be investigated.

Harris, W. and Buckner, D. N. A study of factors influencing the judgment of human performance, 1960.

Two studies were described in this report: an exploratory study and an extension of it.

The exploratory study was an investigation of a method by which performance judgment behavior could be observed under laboratory-like conditions. Subjects in this study judged the same stimuli,

under two conditions: (1) an "objective" condition, in which the subjects were told the stimulus was simply a flashing light which was to be judged on a physical scale; and (2) a "performance" condition, in which they were told the flashing light represented performance on a pursuit rotor which was to be judged on a psychological scale. Similar nine-category judgment scales were used for each condition. Under the objective condition, the category names on the scale were given in terms of the duration of the flashing light (e.g., "very short," "long"), and under the performance condition, they were stated in terms of the goodness of performance (e.g., "very poor," "good").

The purpose of this approach was to try to isolate what might be called a "pure" performance-judgment effect. If it were assumed that the "objective" judgments reflected only the perceptual ability of the subjects, and that this was relatively constant or varied in a determinable way, then any discrepancy between these and the "performance" judgments could be attributed to the altered meaning of the stimuli presented under the different conditions.

This exploratory study was conducted in the HFR laboratory, with clerical personnel as subjects. The first group of subjects judged the same series of stimuli on two successive occasions under the objective condition; a second group judged the series first under the objective condition and then under the performance condition. For the performance condition, the subjects were led to believe that professional staff members were performing on the pursuit rotor and that the goodness of a particular performance was indicated by the proportion of time the stimulus light was on during a 30-second judgment period. No one was actually performing the task.

The following results were obtained:

1. On the second objective session for the first group, the mean judgments "regressed toward the mean"; that is, the slope of the judgments-on-stimuli regression line became flatter as compared to that of the first objective session.

2. On the performance session for the second group, the mean judgments went in an opposite direction; that is, the slope of the regression line became steeper as compared to that of the previous objective session. An analysis of variance, using the slope of individual regression lines as the dependent variable, showed a significant groups-by-sessions interaction.

When the stimuli represented performance of other persons, the value of poor "performance" was generally underestimated and that of good "performance" generally overestimated. By contrast, when the stimuli were simply various durations of "light or " low stimulus values were judged to be higher, and higher ones lower.

The two groups were also given differential performance training on the pursuit rotor, before making additional judgments of the same stimuli under the performance condition. The purpose of giving the raters training in performing the task was to determine the effect of the raters' own task proficiency upon their judgments of others' performances. It was found that subjects with brief experience on the task, a single 30-second period, systematically overestimated all stimulus values, as compared to their pre-training judgments. Because the performances of subjects with brief experience were very poor, this was interpreted as a "contrast" effect. The judgments of subjects with greater experience, five 30-second periods, were not appreciably different from their pre-training judgments.

A second study was designed as a replication and extension of the exploratory one, but differed from it in two respects:

1. Navy enlisted men, ASW School students, were subjects instead of HFR clerical personnel;

2. A session in which performers were present and ostensibly performing the task while the stimuli were being judged was included to permit the test of additional hypotheses.

The control and display apparatus, the performance task, the

stimulus values judged, and the general method followed were the same as in the exploratory study.

Two of the hypotheses tested were drawn from the results of the exploratory study:

1. As compared to initial judgments under the objective conditions, judgments during a second objective session would be less variable, and regress toward the mean, judgments during a second session that involved "performance" would be more variable.

2. Raters with brief experience in performing the task would overestimate the values of stimuli representing performance, as compared to pre-training estimates; judgments of raters with greater experience would not be affected.

Hypothesis 1 was only partially confirmed in an analysis of all judgments: judgments of subjects on a second objective session were less variable, but judgments of subjects on a second session under the performance condition were not more variable. However, the results of an analysis of more stable judgments, those of the last occurrence of each stimulus, fully confirmed the hypothesis: there was a significant groups-by-sessions interaction.

The "contrast" effect, the overestimation of stimulus values as a function of brief experience on the task, which was so strikingly displayed in the exploratory study, was not observed in the ASW School study.

Another hypothesis tested was that judgments would be less valid when performers were present during a judgment session than when they were absent; and that judgments would be influenced by the raters' knowledge of the performers' school achievement. The men ostensibly performing the task during the performance judgment sessions were classmates of the raters, selected on the basis of class standing--from the top, middle, and bottom of the distribution of all subjects' school achievement scores.

The first part of this hypothesis was not confirmed: there was no decrease in the validity of the average judgments as a result of performers being present. There was, however, a significant correlation (rho = .82) between the mean judgment values assigned the performers and their class standing.

Harris, W. and Buckner, D. N. Rater performance skill and attitudes toward performers, 1962.

This was a study of the effects on performance judgments of two independent variables: rater performance skill and rater attitudes toward performers. The major hypotheses tested were

1. The more proficient the rater in performing the task, the more valid, reliable, and discriminating would be his judgments of the performance of others.

2. The less proficient the rater in performing the task, the more likely his judgments would be influenced by his attitudes toward performers and the more he would overestimate the performance of others.

3. The longer a rater had known a performer and the closer their friendship, the better the rater would judge his performance to be.

4. Rater attitudes, particularly likes and dislikes, toward performers would be reflected in the judgments of performance.

The subjects were 33 college men, living together in a fraternity house. Nine subjects were selected as performers: the three most popular men, the three least popular men, and three men of "indifferent" popularity. The remaining 24 subjects were raters.

The performance task was a two-hand tracking task. The performer was required to move a stylus around a closed, narrow track. His performance score was determined from an equal weighting of total tracking time and total error time, a cumulation of the time the stylus was off the track during the performance period. A buzzer sounded whenever the stylus was off the track.

Each rater performed a training trial and a test trial on the two-hand task. The raters did not observe one another perform, but they were given their performance scores and rank positions.

The raters judged the absolute and relative positions of the performers on five traits: Physical Skill, Social Skill, Leadership Ability, Likeableness, and Contribution to the Fraternity. The performers were ranked on 21-category graphic scales having extreme and midpoint anchors. The raters also indicated how long they had known each performer and their degree of friendship.

The performers were given a training trial on the task and then performed individually before the entire group of raters on two different occasions. Each performer was judged immediately after he performed on separate 21-category graphic scales. Raters were instructed to weight total tracking time and total error time equally in arriving at a performance judgment.

Three rater groups of eight men each were identified on the basis of how skillful they were in performing the task: a "Good Performing" (GP) group, and "Average Performing" (AP) group, and a "Poor Performing" (PP) group.

The following scores were computed for each performer: a Performance Score, derived from an equal weighting of tracking and error times, which was the objective measure of his performance; Rating Scores, the mean ratings assigned the performer by all raters and by the GP, AP, and PP rater groups; and a Mean Trait Score, the mean of all the trait ratings (the individual trait ratings were highly intercorrelated, so only this score was used in the analyses).

The results of the study were

1. On Session I, the judgments of the GP raters were significantly more valid than those of the less proficient groups: the correlations between Performance Scores and Rating Scores were .90 (GP), .83 (AP), .80 (PP), and .86 (All Raters). On Session II, there was no

difference between the groups in judgment validity: .88 (GP), .93 (AP), .91 (PP), .90 (All Raters). The judgment validity of the GP group remained high and that of the AP and PP groups improved markedly.

2. On Session I, interrater agreement was highest for the GP group, second highest for the PP group, and lowest for the AP group. On Session II, the order was PP, GP, and AP. Contrary to the hypothesis, the least proficient rater group (PP) produced the most reliable ratings overall, if interrater agreement is the measure of reliability. There was a marked increase in interrater agreement from Sessions I to II of all rater groups.

3. On Session I, the GP raters were significantly more discriminating than were the other raters. The mean regression line slopes were .808 (GP), .609 (AP), .677 (PP), and .698 (All Raters). On Session II, the GP raters were still the better discriminators but not significantly so: .869 (GP), .794 (AP), .804 (PP), .822 (All Raters). The discrimination of all raters improved markedly from Sessions I to II.

4. Performances of the poorer performers were overestimated most by the less proficient rater groups. The mean discrepancies between Performance Scores and Rating Scores for the bottom three performers were, on Session I, 0.4 (GP), 2.1 (AP), 1.1 (PP), 1.2 (All Raters), and on Session II, 0.7 (GP), 1.7 (AP), 2.0 (PP), 1.5 (All Raters).

5. Performances of disliked performers were markedly underestimated on both sessions. The mean discrepancies for the three least-liked performers were -3.8 (Session I) and -3.9 (Session II).

6. Length of acquaintance and degree of friendship influenced performance judgments on both sessions. For example, on Session I, the mean Rating Scores by length of acquaintance were 12.4 (3-12 months), 13.0 (13-30 months), and 15.5 (31-48 months); and by friendship, 11.8 (acquaintance), 12.8 (friend), 13.6 (close friend).

7. The judgments of the PP raters were more influenced by their attitudes toward performers than those of other groups. The correlations between Rating Scores and Mean Trait Scores were, on

Session I, .44 (GP), .41 (AP), .78 (PP), .52 (All Raters); and on Session II, .42 (GP), .20 (AP), .61 (PP), .40 (All Raters).

The influence of the extraneous variables, rater performance skill and rater attitudes, was more pronounced on Session I than on Session II. Apparently, extraneous variables are more likely to influence performance judgments when the judgment task is difficult for the raters or unfamiliar to them. But even though the rating skill of most raters improved substantially from Sessions I to II, they still underestimated, strikingly, the performance value of men they disliked.

It is apparent that if performance judgments are to be interpreted meaningfully, several factors have to be taken into account. Perhaps the most important is the rater's attitude toward the performer he is judging. If a performer is disliked by a rater, his performance value may be greatly underestimated--even though the rater is a trained and objective judge of performances on a given task. In this study, the judgment task was probably not too difficult for the raters by the second session, witness the .90 validity coefficients, but still the disliked performers were judged to be much poorer performers than they actually were. Rater attitudes must be a great deal more influential in more difficult performance judgment situations.

If their attitudes toward performers are taken into account, it seems highly probable that raters who are themselves highly skilled performers will be the most accurate and valid raters of the performances of others. The more proficient raters seem to be especially good judges of the poorer performances.

McGrath, J. J. The influence of unusual performances and time-order on performance judgment, 1963.

This was a study of the effects of the occurrence of an unusual performance and of time-order on the judgment of a sequence of performances.

Silent, color movies were made of six male operators performing a simple reaction-time task. The operators had been thoroughly practiced until they could deliberately manipulate their mean reaction times (MRT). Three operators were used as "anchoring performers" to illustrate the top, bottom, and middle performance levels of a rating scale. The other three, operators A, B, and C, each performed five 1-minute trials. A and C produced relatively constant performance levels (MRT = 1.46 sec.) Operator B, however, had one unusually "good" trial or he had one unusually "poor" trial. The four remaining trials were such that his overall MRT was also 1.46 seconds.

Six groups of raters (total N = 239) viewed the movie. They saw the three anchoring performers, then separately rated A, B, and C on their overall performance. The movie was edited so that Group I saw operator B perform <u>well</u> on the first trial; Group II on the third trial; and Group III, on the fifth trial. Group IV saw B perform poorly on the first trial; Group V, on the third trial; and Group VI, on the fifth trial.

The results indicated that an unusually good performance was overly weighted in the final rating when that performance occurred on the first trial or on the last trial, while an unusually poor performance was overly weighted only when it occurred on the first trial. The results also showed that the judges gave significantly different mean ratings to the three different operators in spite of the fact that their performances were objectively equivalent. Operator C, the last man rated, was given a lower rating than either operator A or B.

It was concluded that "first impressions" of a worker being rated (and in some instances "last impressions") can significantly bias a performance judgment and produce invalid ratings.

Index of Reports Prepared Under Contract Nonr 1241(00)

Buckner, D. N. Research on the development of performance criteria, Technical Report VIII. The predictability of ratings as a function of inter-rater agreement, 1957.

Buckner, D. N. The predictability of ratings as a function of interrater agreement. J. appl. Psychol. 1959, 43, 60-64.

Harris, W. and Buckner, D. N. A methodological study of the judgment of human performance. Paper presented at the Western Psychological Association Convention, San Diego, 1959.

Harris, W. and Buckner, D. N. A study of factors influencing the judgment of human performance, Technical Report 1, 1960.

Harris, W. and Buckner, D. N. A study of factors influencing the judgment of human performance, Technical Report 2. Rater performance skill and attitudes toward performers, 1962. Paper presented at the Western Psychological Association Convention, San Francisco, 1962.

Harris, W., Mackie, R. R. and Wilson, C. L. Research on the development of performance criteria, Technical Report VI. Performance under stress: A review and critique of recent studies, 1956.

High, W. S. and Mackie, R. R. Research on the development of performance criteria, Technical Report VII. A factor analytic study of aptitudes, interests and practical performance skills for Navy machinery repairman students, 1957.

High, W. S. and Mackie, R. R. A factor analysis of the practical performance of machinery repairmen. Paper presented at the Western Psychological Association Convention, Monterey, 1958.

High, W. S. and Mackie, R. R. A follow-up study of the predictability of machinery repairman shipboard performance two years after training. Paper presented at the Western Psychological Association Convention, San Diego, 1959.

High, W. S. and Mackie, R. R. Ratings and practical performance tests as complementary criteria of performance (performance criterion development). Paper presented at the Western Psychological Association Convention, San Jose, 1960.

Mackie, R. R. Inter-judge and retest reliabilities for products made in a machinery repair shop. Paper presented at the Western Psychological Association Convention, Berkeley, 1956.

Mackie, R. R. and High, W. S. Research on the development of shipboard performance measures, Technical Report IX. Supervisory ratings and practical performance tests as complementary criteria of shipboard performance, 1959.

Mackie, R. R., Wilson, C. L. and Buckner, D. N. Research on the development of shipboard performance measures Part V: Interrelationships between aptitude test scores, performance in submarine school, and subsequent performance in submarines as determined by ratings and tests, 1954.

McGrath, J. J. A study of factors influencing the judgment of human performance, Technical Report 3. The influence of unusual performances and time-order on performance judgment, 1963.

McGrath, J. J. First impressions and performance judgment. Paper presented at the Western Psychological Association Convention, Santa Monica, 1963.

Wilson, C. L. and Mackie, R. R. Research on the development of shipboard performance measures, Part I. The use of practical performance tests in the measurement of shipboard performance of enlisted naval personnel, 1952.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part II. The use of a performance rating scale in the measurement of shipboard performance of enlisted naval personnel, 1954.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part III. The use of performance check lists in the measurement of shipboard performance of enlisted naval personnel, 1954.

Wilson, C. L., Mackie, R. R. and Buckner, D. N. Research on the development of shipboard performance measures, Part IV. A comparison between rated and tested ability to do certain job tasks, 1954.