

63-3-2

Handwritten symbol resembling a stylized 'R' or 'F'.

401677

CATALOGED BY ASTIA
AS AD NO.

401677

**A STUDY OF FACTORS INFLUENCING
THE JUDGMENT OF HUMAN PERFORMANCE**

Technical Report 3

**THE INFLUENCE OF UNUSUAL PERFORMANCES AND
TIME-ORDER ON PERFORMANCE JUDGMENT**

ASTIA
APR 19 1963
TISIA B

HUMAN FACTORS RESEARCH, INCORPORATED
1112 Crenshaw Boulevard • Los Angeles 19, California • WEbster 3-7356

A STUDY OF FACTORS INFLUENCING
THE JUDGMENT OF HUMAN PERFORMANCE

Technical Report 3

THE INFLUENCE OF UNUSUAL PERFORMANCES AND
TIME-ORDER ON PERFORMANCE JUDGMENT

James J. McGrath

Prepared for

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

by

Human Factors Research, Incorporated
1112 Crenshaw Boulevard
Los Angeles 19, California

April 1963
Contract No. Nonr 1241(00)
NR 153-165

U
N
I
T
E
D
S
T
A
T
E
S
G
O
V
E
R
N
M
E
N
T

Reproduction in whole or in part is permitted
for any purpose of the United States Government.

ABSTRACT

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 well 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 showed that operator B's performance was rated in the following manner: Group I > Group II < Group III, when he had an unusually good trial; and Group IV < Group V = Group VI, when he had an unusually poor trial. These 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.

ACKNOWLEDGEMENTS

The writer is indebted to the officers and men of the Service School Command, U.S. Navy Training Center, San Diego, California, for their assistance and cooperation in the conduct of the study reported here. In particular, I should like to thank LCDR Fodor, Training Officer; and Mr. Dale Lovell, Chief L. F. Biondo, and Chief E. E. Myers.

The writer is also grateful for the assistance given him by Richard L. Weis and Jon C. Rittger of the HFR staff in developing the motion picture materials, and Raymond A. Gavin for his assistance in the data-gathering phase of the study.

Table of Contents

	Page
ABSTRACT	v
ACKNOWLEDGEMENTS	vii
List of Tables and Figures	xi
INTRODUCTION	1
METHOD	3
The Experimental Task	3
Preliminary Study of Performance on the Experimental Task	4
Stimulus Materials	4
Anchoring Performances and Rating Scale	5
Performances to be Rated	5
Judges	9
Experimental Variables	11
HYPOTHESES	13
Primacy Effect	14
Recency Effect	14
Terminal Effect	15
Contrast and Assimilation Effects	15
SECONDARY EXPERIMENT	16
RESULTS	16
Preliminary Analyses	16
Time-order Effects	19
Contrast and Assimilation Effects	21
Differences Between Ratings of the Three Operators	21
CONCLUSIONS AND IMPLICATIONS	22
REFERENCES	25

List of Tables and Figures

Table		Page
I	Schedule of Performances (MRT's) Presented to the Six Groups of Judges	13
II	Analysis of Variance of Ratings of Operator A's Performance	20
III	Analysis of Variance of Ratings of Operator B's Performance	20
IV	Analysis of Variance of Ratings of Operator C's Performance	20
V	Significance of Differences Between Means and Standard Deviations of Performance Ratings of Operators A, B, and C, by All Judges Combined . .	22
Figure		
1	The experimental task	3
2	The performance rating scale	6
3	Performances of the three anchoring performers and the two control operators, <u>A</u> and <u>C</u>	7
4	The six different performance sequences of operator <u>B</u>	8
5	Experimental design	12
6	Recall of "fastest trial" of operator <u>B</u> by judges in Groups I, II, and III	17
7	Recall of "slowest trial" of operator <u>B</u> by judges in Groups IV, V, and VI	17
8	Mean rating of <u>B</u> 's performance by the six groups of judges	19
9	Means (horizontal bars) and standard deviations (vertical bars) of performance ratings of the three operators by all judges combined (N=239) .	21

THE INFLUENCE OF UNUSUAL PERFORMANCES AND TIME-ORDER ON PERFORMANCE JUDGMENT

Time-order effects have been observed in a variety of experiments on learning (e.g., Brown & Overall, 1959) and in studies of attitude and opinion formation (e.g., Luchins, 1960; Miller & Campbell, 1959); but time-order has never been demonstrated to affect the judgment of human performance. A "time-order effect," in this context, means that a judgment of a stimulus is affected by the ordinal position of the stimulus in the series to be judged. The study reported here was concerned with the effects of time-order on performance ratings.

INTRODUCTION

When performance ratings are made at periodic intervals, it is ordinarily intended that the ratings reflect the worker's average performance during the interval between ratings. If time-order effects are present, however, certain individual performances will be disproportionately weighted in the total rating, and as a result, the total rating will be invalid. The situation of concern to the present study is one in which a rater must evaluate the overall performance level of an operator, having seen him perform a task on a number of occasions. The question of interest is whether or not the operator's initial performance on the task is given a disproportionate weight by the rater in determining his overall evaluation; or conversely, whether or not the most recent (closest to the time at which the rating is made) performance is given a disproportionate weight; or alternatively, whether or not both initial and recent performances are given greater or less weight than performances in the middle of the series.

Other questions of practical interest involve the effect of unusual performances in the series on the rating of the total series. Do raters give greater weight to unusual (exceptionally good or exceptionally poor) performances than they give to the more typical

performances? Or are such unusual performances "discounted" as flukes and given less weight than they deserve?

These various effects have been observed in psychophysical research and have been given names. Although there may be no precise parallel between psychophysics and performance judgment, these traditional names will be used here for convenience. The effects under study in the research reported here were the following:

Primacy effect: Events occurring early in a sequence are given greater weight than those occurring late in the sequence in determining a judgment of (or response to) the entire sequence.

Recency effect: Events occurring close to the time at which the judgment is made are given greater weight than more remote events in determining the overall judgment.

Terminal effect: Events occurring at the beginning or end of a sequence are given greater or less weight than events occurring in the middle of a sequence in determining a judgment of the entire sequence.

Contrast effect: Unusual events in a sequence are given a disproportionately high weight compared with the more typical events in determining a judgment of the entire sequence.

Assimilation effect: Unusual events in a sequence are given a disproportionately low weight compared with the more typical events in determining a judgment of the entire sequence.

Because of the complexity of the present experiment, it is difficult to describe the experimental hypotheses without describing the method by which they were tested. Therefore, a statement of hypotheses will be postponed until the experimental method has been described.

METHOD

The Experimental Task

To study time-order effects on performance judgment, it is first necessary to devise or select a task to be performed. A desirable task for research purposes would have the following characteristics. (1) The criterion of performance on the task would be unambiguous and easily understood by the judges. (2) Rating performance on the task would be difficult enough to produce differences in the ratings made by different judges, but not so difficult that it would produce unreliable ratings. (3) To maintain motivation on the part of the judges, the task would require the judges to be directly involved, perhaps even competing with the operator whose performance is to be rated. (4) And, most important, an objective measure of the criterion would be available to the experimenter. An experimental task was devised to meet these requirements.

The experimental task (Figure 1) involved detecting and responding to signals that occurred at unpredictable times and locations. The operator stood in front of a 4' x 4' panel of 25 lights. At irregular intervals a light (any of the 25) went out. The operator's task was to detect the extinguished light, the "signal," and re-light it as quickly as possible by turning a switch located next to the light. A clock measured the elapsed time between the



extinction of any light on the panel and the turning of the proper

switch by the operator. The sole criterion of performance on the task was the mean reaction time (MRT) to all signals within a given trial period.

Preliminary Study of Performance on the Experimental Task

To obtain an indication of characteristic performance on the experimental task, 20 subjects were tested, each performing the task on five 1-minute trials. During each trial 10 signals were presented at random intervals and random locations. The results showed that the MRT for all subjects on all trials was 1.46 seconds. The range of individual scores was from 0.74 seconds for the fastest subject to 2.25 seconds for the slowest subject. These performance scores were used as guidelines in devising the stimulus materials for the experiment.

Stimulus Materials

In devising a means whereby groups of judges could observe operators performing the task, two requirements were essential. First, all judges must be able to observe precisely the same performances; second, the same set of performances must be presented to each judge, but the order of performances within the set must be re-arranged for different groups of judges. These requirements could be met adequately by motion pictures.

Silent, color, 16 mm., motion pictures were made of six operators separately performing the experimental task. The operators were all males of similar age and general appearance. Each operator wore a white laboratory coat and was positioned before the panel so that the camera viewed the scene from over his right shoulder. The motion picture film was purposely underexposed so that the operator appeared to be working in a slightly darkened room. The result was that the judges could clearly see that the operators were indeed different men, but were unable to distinguish any marked physical differences among them.

The operators were thoroughly practiced on the task until they could deliberately manipulate their performance scores with the aid of verbal alerting and feedback. Further manipulation of each operator's performance was achieved by editing the motion picture appropriately, so that it was possible to produce motion pictures of operators whose performances (MRT's) were known precisely, and had, in fact, been predetermined.

Anchoring Performances and Rating Scale

Three of the six operators were used as "anchoring performers." Their role was to provide standard performances by which the judges could rate the performances of the other three operators on a 25-point rating scale. The rating scale (Figure 2) consisted of a representation of 25 operators arranged in a hierarchy diagonally across the page. The figure at the top of the rating scale was labeled "This man is the fastest we know of." The middle figure (13th from the top) was labeled "This man is average." The figure at the bottom of the rating scale was labeled "This man is the slowest we know of." One-minute motion pictures were produced of each of the three anchoring performers. One performed at MRT = 0.50 seconds and represented the top of the scale; another performed at MRT = 1.46 seconds and represented the middle of the scale, and the third performed at MRT = 2.75 seconds and represented the bottom of the scale.

Performances to be Rated

The remaining three operators were those whose performances were to be rated. Motion pictures were produced of each operator performing the task on five 1-minute trials, with approximately a 20-second rest (blank screen) between trials. In each trial seven signals occurred at random intervals and locations. Two of these operators (hereafter to be called A and C) served as controls. Their performances varied unsystematically about a MRT of 1.46 seconds so that the performance curves, when plotted by trials,

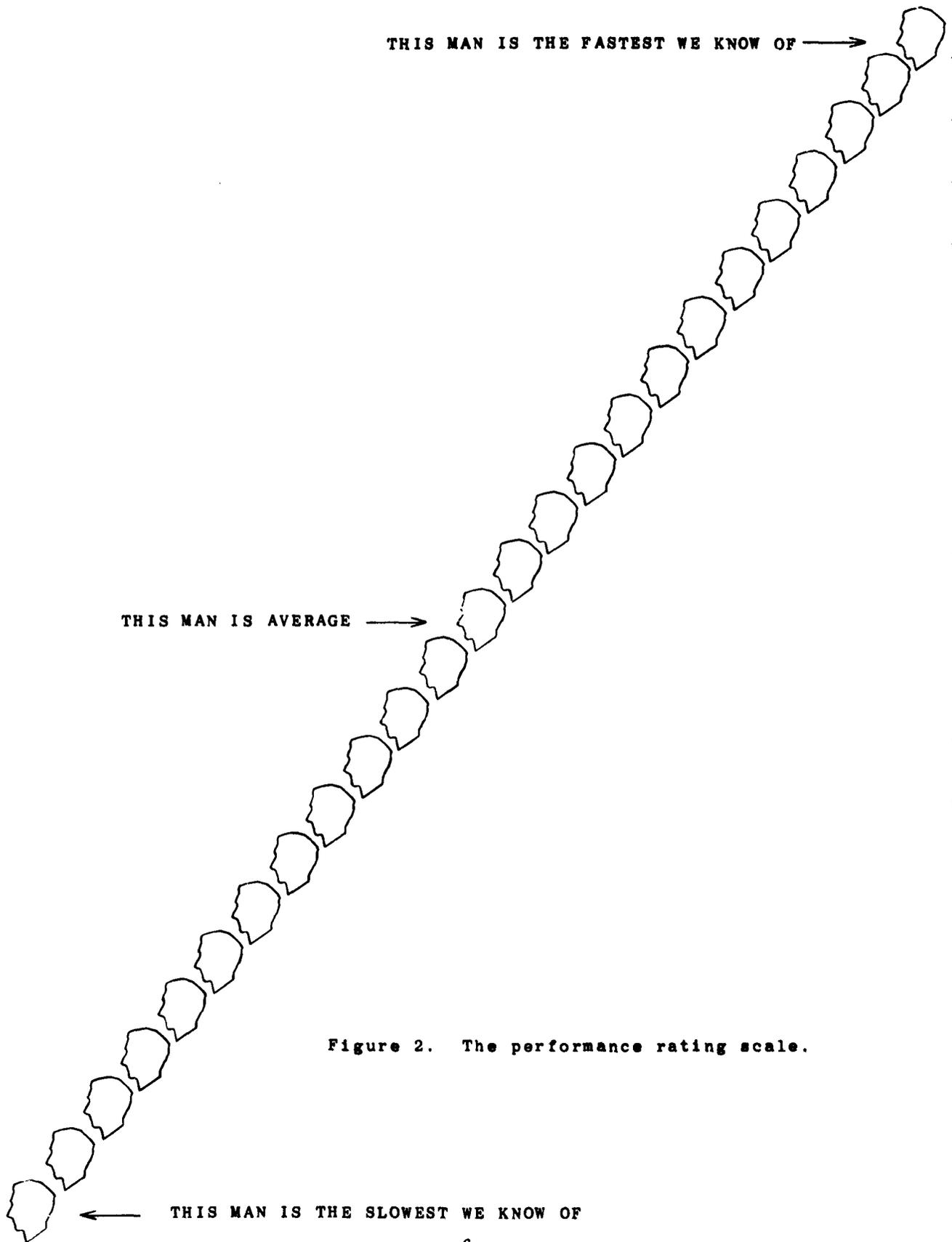


Figure 2. The performance rating scale.

gave the appearance of flat functions (Figure 3).

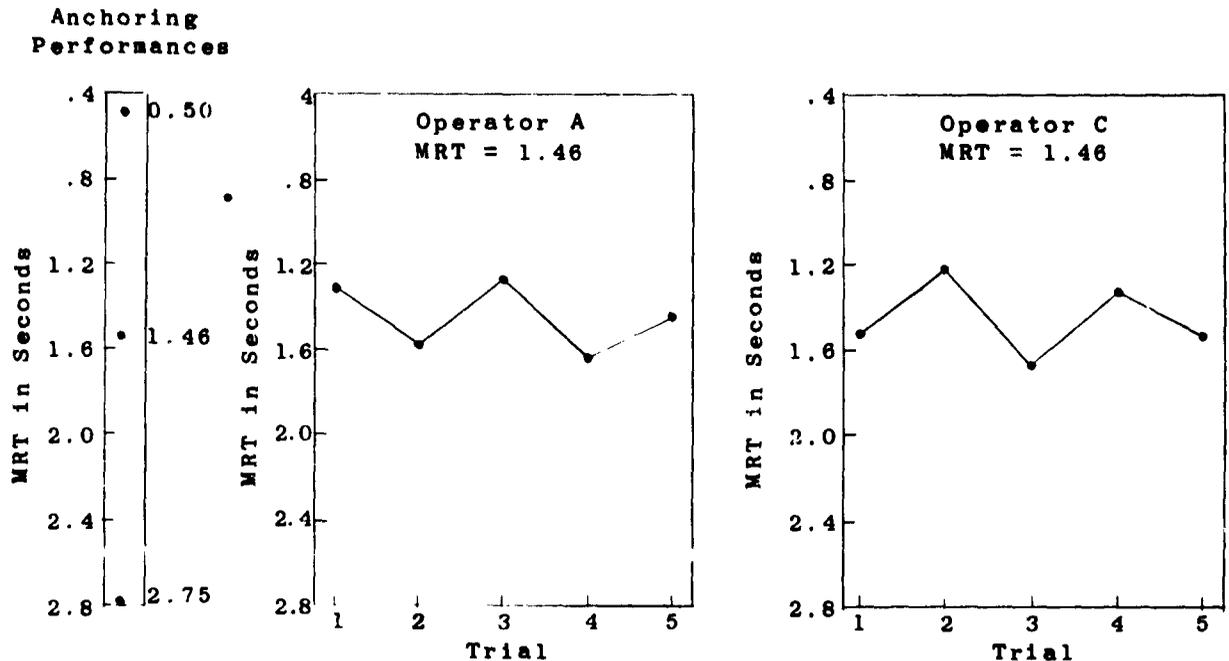


Figure 3. Performances of the three anchoring performers and the two control operators, A and C.

Experimental variations in this study concerned the performance of the remaining operator (hereafter to be called B). Two motion pictures were produced of B performing the task on five 1-minute trials. In one motion picture, B performed unusually well on one trial (MRT = 0.71), but his performance on the other four trials was such that his total MRT = 1.46. In the other motion picture, B performed unusually poorly on one trial (MRT = 2.27*), but his performance on the other four trials was such that his total MRT = 1.46. Both motion pictures were edited so that the unusual trial occurred either first, third, or fifth in the sequence of five trials. The

*To make the performance credible to the judges, B did not respond slowly to all seven signals; but rather briefly "overlooked" four of the seven signals. The actual response times to the seven signals were 1.2, 3.2, 1.2, 1.4, 3.3, 2.8, and 2.8 seconds.

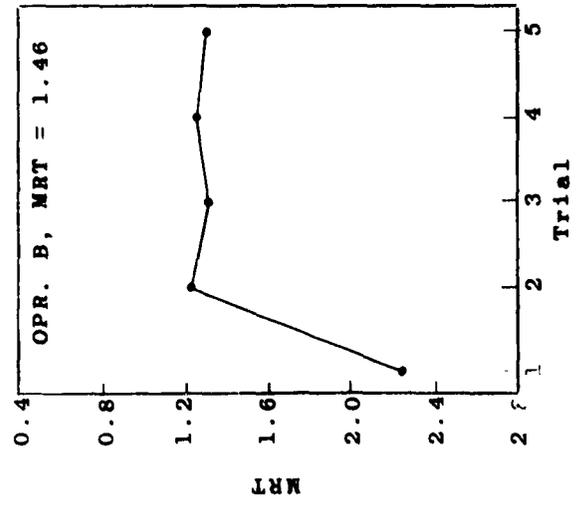
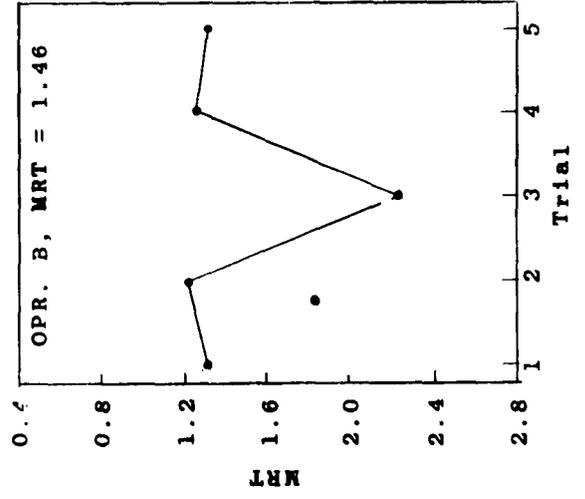
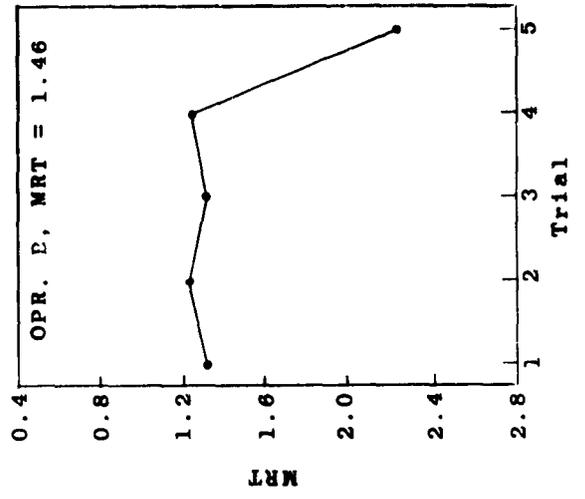
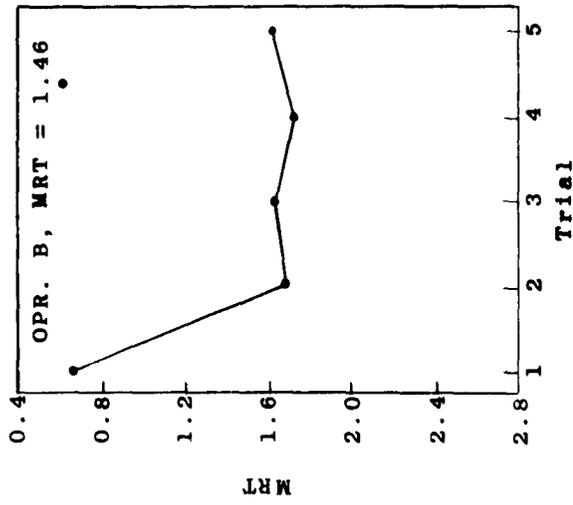
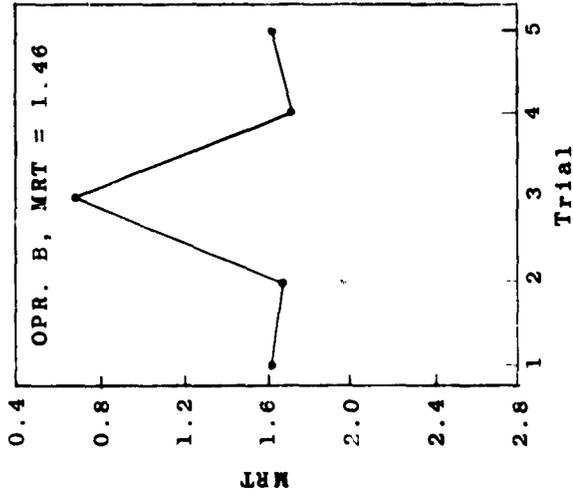
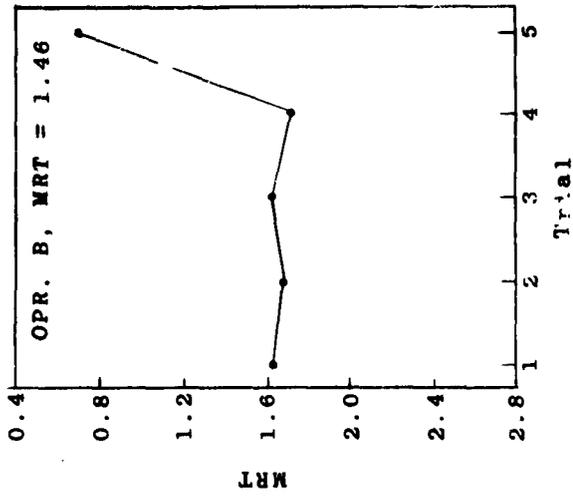


Figure 4. The six different performance sequences of operator B.

six different performance sequences that resulted from this manipulation are shown in Figure 4.

Judges

The judges were 322 Navy enlisted men, trainees at the USN Fleet Radio School in San Diego, California. Of this number, 239 judges took part in the main experiment, and 83 were used in a secondary experiment. The judges in the main experiment were divided into six groups (N's = 42, 36, 43, 43, 39, and 36) and rated the performances of A, B, and C using the following procedure.

Procedure

Each group of judges was assembled in a classroom furnished with blackout drapes. They were given these instructions:

"This is a study of how well you can judge the performances of operators on a certain task. The task requires a high degree of alertness on the part of the operator, and so you also must be alert if you are to judge his performance accurately.

"You will view a series of motion pictures of operators performing a simple reaction-time task. The operator will stand in front of a large panel of 25 lights. Occasionally a light will go out and the operator must turn it back on as quickly as possible by turning a switch next to the light. Your judgment of his performance will be based on the average speed with which he detects extinguished lights and turns them back on. You will indicate your judgment of his speed by marking one of these rating sheets. The rating sheet consists of a series of men arranged from the fastest to the slowest. The man in the middle of the scale indicates the average performer on this task. We will now show you what the task is like, and then show you performances that represent the middle, top, and bottom points on the rating scale."

At this point the judges were shown an introductory motion picture in which the experimenter was shown demonstrating how the task was to be performed. Following the demonstration film, the

anchoring performers were shown, preceded by the film titles, "Average operator," "Fastest operator," and "Slowest operator." The narratives that accompanied the anchoring performances were the following:

"You will now see an operator whose performance is exactly average. If you were rating his performance on the scale you would mark the middle point of the scale."

"You will now see the fastest operator we have ever tested. If you were rating his performance you would mark the very top of the scale."

"Now you will see the slowest operator we have ever tested. If you were rating his performance you would mark the very bottom of the scale."

After the anchoring performances were shown, the following instructions were given:

"Now will you please write your name in the upper left-hand corner of each of the three rating sheets. You will be asked to rate the performances of three operators. Each operator will perform the task five times. Each of the five trials will last about one minute. At the end of the fifth trial you will rate the operator on his total performance for all five runs. Do not make any mark on the rating sheet until after the fifth run. At that time make your rating by writing the operator's identification letter (either A, B, or C) in the appropriate figure on the rating sheet. Remember you must consider his total performance on all five runs in making your ratings. Now here is the first operator whose performance you will judge, Operator A. Remember he will have five trials on the task and then you will rate his total performance by marking an "A" in the appropriate figure. Do not make any mark on your rating sheet until after the fifth trial, and do not make any comments to your neighbor."

The motion picture of A was shown, and the judges rated his performance at the end of the fifth trial. The rating sheets for A were collected and the motion picture of B was shown. After the judges rated B, the rating sheets were collected, and the procedure was repeated for C.

At this point a new response sheet was distributed which contained the special box shown below:

	Fastest Trial	Slowest Trial
Operator "A"	_____	_____
Operator "B"	_____	_____
Operator "C"	_____	_____

The judges were told:

"Write your name in the upper left-hand corner of this new sheet. Now try to remember the performances you have seen and in the box indicate which of the five trials was the fastest and which was the slowest for each operator. If you cannot remember, just make a guess, but in any case write in the number indicating the fastest and slowest trial for each operator."

The purpose of having the judges recall the fastest and slowest trials of each operator was to obtain a quantitative indication of whether the "unusual" trial of B was distinctive enough to be recalled by the judges.

Experimental Variables

Each of the six groups of judges went through the same procedure as described in the preceding section. All conditions were the same for each group with the exception of the performance of B. A representation of the experimental design is shown in Figure 5. The independent variables were (1) the type of unusual performance that occurred on one of B's five trials—either an unusually good performance or an unusually poor one, and (2) the ordinal position of the unusual trial in the sequence of five trials—either the first, third, or fifth position. All groups saw exactly the same performance sequences by the control operators, A and C. The performances rated by the six groups of judges are shown in Table I and in

Figures 3 and 4. It will be noted that the total MRT for each of the three operators was 1.46 seconds and that this was equivalent to the performance of the "average" anchoring performer. The dependent variable was the mean rating given B by each of the six groups of judges.

		ORDINAL POSITION OF OPERATOR B'S DEVIANT TRIAL		
		1st	3rd	5th
TYPE OF DEVIANT PERFORMANCE BY OPERATOR B	GOOD	<p>Group I</p> <p>A B C</p>	<p>Group II</p> <p>A B C</p>	<p>Group III</p> <p>A B C</p>
	POOR	<p>Group IV</p> <p>A B C</p>	<p>Group V</p> <p>A B C</p>	<p>Group VI</p> <p>A B C</p>

Figure 5. Experimental design.

Table I

Schedule of Performances (MRT's) Presented to
the Six Groups of Judges¹

	<u>Group</u>					
	<u>I</u>	<u>II</u>	<u>III</u>	<u>IV</u>	<u>V</u>	<u>VI</u>
<u>Anchoring performers</u>						
"Average operator"	1.46					
"Fastest operator"	0.50			(Same as Group I)		
"Slowest operator"	2.75					
<u>Operator A</u>						
1st trial	1.36					
2nd "	1.58					
3rd "	1.30			(Same as Group I)		
4th "	1.62					
5th "	1.43					
<u>Total Performance</u>	<u>1.46</u>					
<u>Operator B</u>						
1st trial	(0.71)	1.62	1.62	(2.27)	1.30	1.30
2nd "	1.66	1.66	1.66	1.21	1.21	1.21
3rd "	1.62	(0.71)	1.62	1.30	(2.27)	1.29
4th "	1.69	1.69	1.69	1.25	1.25	1.25
5th "	1.62	1.62	(0.71)	1.29	1.29	(2.27)
<u>Total Performance</u>	<u>1.46</u>	<u>1.46</u>	<u>1.46</u>	<u>1.46</u>	<u>1.46</u>	<u>1.46</u>
<u>Operator C</u>						
1st trial	1.53					
2nd "	1.21					
3rd "	1.67			(Same as Group I)		
4th "	1.36					
5th "	1.51					
<u>Total Performance</u>	<u>1.46</u>					

¹The MRT's in this table were derived from a frame-count of the edited movies, where one frame equals 1/24 second.

HYPOTHESES

Since overall performance was the same for each of the three operators as seen by each of the six groups of judges, the performance ratings, if completely valid, should be the same (and should, in fact, be equivalent to the middle point on the rating scale).

If time-order effects are present, however, the ratings of A and C should be the same for all groups, but there should be significant differences between the ratings of B by the six groups of judges. These differences should appear as a significant interaction between the two independent variables, i.e., between the type of deviant performance by B and the ordinal position of that performance. The direction of the interaction would indicate the particular type of time-order effect that is present. The specific predictions associated with the various effects described on page 2 are the following:

Primacy Effect

If there is a primacy effect the initial trial will receive a greater weight than later trials in determining the overall rating. In the case of the control operators, A and C, this would not produce any significant differences among the mean ratings of the six groups of judges because these operators performed at a relatively constant level throughout. Therefore the prediction regarding the ratings of A and C would be the following (where the letter indicates the operator rated and the subscript indicates the group of judges who rated him.)

$$A_I = A_{II} = A_{III} = A_{IV} = A_V = A_{VI}$$

and $C_I = C_{II} = C_{III} = C_{IV} = C_V = C_{VI}$

But a primacy effect would produce differences in the ratings of B. Group I saw B perform unusually well on the first trial and Group IV saw him perform unusually poorly. If the first-trial performances were overly weighted the following effects on performance ratings would occur:

$$B_I > B_{II} \Rightarrow B_{III} \quad \text{and/or} \quad B_{IV} < B_V \Leftarrow B_{VI}$$

Recency Effect

If there is a recency effect, the fifth trial, since it was the most recent one, will be overly weighted in determining the

performance rating. Again this should have no effect on ratings of A and C, but should affect the rating of B. A recency effect would produce differences among the groups of judges that would be opposite to those predicted from a primacy effect. If the last-trial performances were overly weighted the following effects on performance ratings would occur:

$$B_I \leq B_{II} < B_{III} \quad \text{and/or} \quad B_{IV} \geq B_V > B_{VI}$$

Terminal Effect

If the terminal trials (first and fifth) are weighted more than the middle trials (second, third, and fourth), there will again be no differences among the ratings of A and C, but predictable differences among the ratings of B. If the first-trial and last-trial performances were overly weighted the following effects on performance ratings would occur:

$$B_I > B_{II} < B_{III} \quad \text{and/or} \quad B_{IV} < B_V > B_{VI}$$

Contrast and Assimilation Effects

If there are contrast or assimilation effects, they will be reflected by a significant main effect of the first independent variable—the type of deviant performance by B—on ratings of his performance. The null hypothesis is: $B_{I+II+III} - B_{IV+V+VI} = 0$. That is, there would be no significant difference between the mean ratings of B by the three groups who saw him perform unusually well on one occasion (Groups I, II, and III) and the three groups who saw him perform poorly on one occasion (Groups IV, V, and VI). If, however, the unusual performance were given an unduly high weight (contrast effect), the first three groups would rate B higher than the last three groups. And if the unusual performance were given an unduly low weight (assimilated to the plateau level), the opposite result would occur.

SECONDARY EXPERIMENT

To test the contrast and assimilation hypotheses, it was necessary to assume that the plateau level of B's performance (i.e., the MRT for the four remaining trials when the deviant trial is excluded) as seen by the first three groups was discriminably different from the plateau level as seen by the last three groups. To test this assumption, the remaining 83 judges were divided into two groups (N = 41, 42) and were shown the demonstration film and the anchoring performances. They then rated the performance of B on two occasions. On one occasion they saw only the four trials that represented B's plateau level as seen by Groups I, II, and III and where total MRT = 1.65 seconds. On the other occasion they saw only the four trials that represented B's plateau level as seen by Groups IV, V, and VI and where total MRT = 1.26 seconds. The order of rating was counterbalanced between the two groups of judges.

RESULTS

Preliminary Analyses

The first concern of the data analysis was whether or not the parametric statistics could be used in testing the hypotheses. Simple numerical values were assigned to each of the 25 points on the rating scale, ranging from 25 for the highest point on the scale to 1 for the lowest point. An inspection of the distributions of ratings of each operator by each group of judges showed that they were normal in shape, and by Hartley's test (Winer, 1962) homogeneous in variance, thereby satisfying the major requirements for parametric statistics.

The next preliminary analysis was performed to test the assumption that the "unusual" trial by B was noticeably unusual to the judges. Specifically, could Groups I, II, and III correctly identify B's fastest trial, and could Groups IV, V, and VI correctly identify his slowest trial? Figure 6 shows the percentages of judges

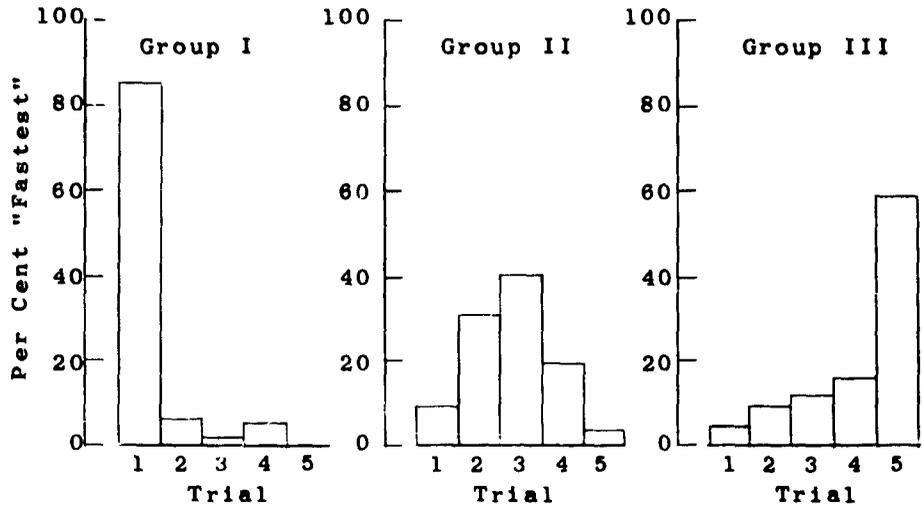


Figure 6. Recall of "fastest trial" of operator B by judges in Groups I, II, and III.

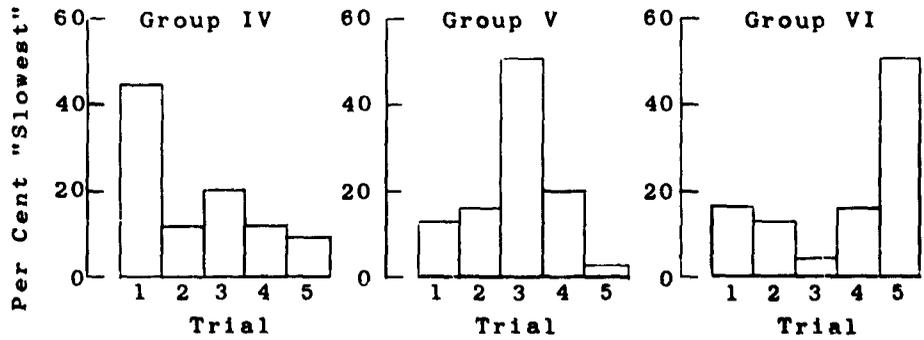


Figure 7. Recall of "slowest trial" of operator B by judges in Groups IV, V, and VI.

in the first three groups who selected each of the five trials as being the "fastest" of the sequence. Figure 7 shows the percentages of judges in the last three groups who selected each of the five trials as being the "slowest" of the sequence. In every instance the highest percentage occurred for the correct trial. Although there was always a plurality of correct judgments, there was not always a majority of correct judgments. It appeared that the ordinal position of an unusually good performance was more accurately recalled than the ordinal position of an unusually poor performance. The good performance appeared to be more noticeable when it occurred at either the beginning or the end of the sequence than when it occurred in the middle, while the poor performance was about equally well recalled whether it occurred at the beginning, middle, or end of the sequence. Performance on the five trials by A and C were evidently perceived as being homogeneous, since the judges in all groups gave judgments of "fastest trial" and "slowest trial" about equally often to each of the five trials.

The final preliminary analysis concerned the discriminability of the plateau performance levels of B as seen by the first and last three groups of judges. The ratings obtained in the secondary experiment were used to test the assumption that judges could accurately discriminate the difference between a plateau performance level of MRT = 1.65 seconds and a plateau performance level of MRT = 1.26 seconds. The results showed that 100% of the 83 judges rated B's 1.65-second mean performance as being poorer than his 1.26-second mean performance. The former performance was given a mean rating of 10.1, while the latter was given a mean rating of 15.8 on the 25-point scale. (Although a significance test of the difference between mean ratings was probably superfluous, it should be recorded that $t = 16.3$). Clearly, the two performance plateaus were discriminably different, thereby allowing contrast and assimilation effects to be tested.

Time-order Effects

The hypothesis was that time-order effects, if present, would produce significant differences among the six groups of judges in their ratings of B as a result of an interaction between the type of deviant performance by B and the ordinal position of that performance. No such differences should occur among the ratings of A and C. This hypothesis was tested by analyses of variance of the ratings of the three operators by the six groups of judges.

The results of the analyses of variance are shown in Tables II, III, and IV. As predicted, there were no significant effects on the ratings given the control operators (Tables II and IV), but there was a significant interaction ($F = 7.81, p < .01$) between the two independent variables affecting the ratings given B. This interaction is shown graphically in Figure 8, which shows the mean ratings of B's performance by the six groups of judges. The direction of

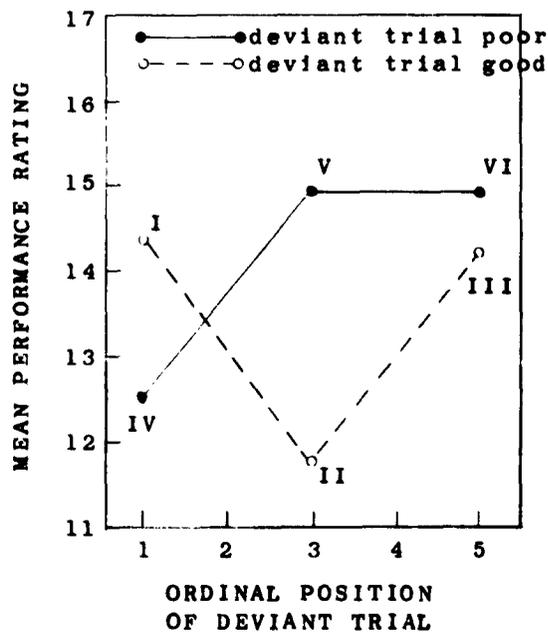


Figure 8. Mean rating of B's performance by the six groups of judges.

Table II
Analysis of Variance of Ratings
of Operator A's Performance

Source	SS	df	MS	F
A. Type of deviant performance by <u>B</u> . (Good vs. poor)	1.2	1	1.2	--
B. Ordinal position of deviant performance by <u>B</u> (1st, 3rd, or 5th).	23.8	2	11.9	1.49
A X B	34.8	2	17.4	2.17
Residual	1872.2	233	8.0	
Total	1932	238		

Table III
Analysis of Variance of Ratings
of Operator B's Performance

Source	SS	df	MS	F
A. Type of deviant performance (good vs. poor).	21.4	1	21.4	1.23
B. Ordinal position of deviant performance (1st, 3rd, or 5th).	77.6	2	38.8	2.44
A X B	248.3	2	124.2	7.81**
Residual	3702.7	233	15.9	
Total	4250	238		

Table IV
Analysis of Variance of Ratings
of Operator C's Performance

Source	SS	df	MS	F
A. Type of deviant performance by <u>B</u> . (Good vs. poor)	2.4	1	2.4	--
B. Ordinal position of deviant performance by <u>B</u> (1st, 3rd, or 5th).	59.4	2	29.7	2.18
A X B	37.6	2	18.8	1.38
Residual	3166.6	233	13.6	
Total	3266	238		

**Sig. .01 level.

the interaction and subsequent t-tests indicated that a terminal effect occurred when B had an unusually good trial ($B_I > B_{II} < B_{III}$), and a primacy effect occurred when B had an unusually poor trial ($B_{IV} < B_V = B_{VI}$).

Contrast and Assimilation Effects

There was no significant difference between the ratings given B by the first three groups of judges and the ratings given by the last three groups. This difference was tested in the analysis of variance by the main effect of the first independent variable, good vs. poor deviant performance, on the ratings of B's performance. As shown in Table III, no significant effect ($F = 1.23$) was found.

Differences Between Ratings of the Three Operators

Since by objective measurement of their reaction times A, B, and C were equal in overall performance (see Table I), it was of interest to note whether they were given equal performance ratings by the judges. The means and standard deviations of the performance ratings given the three operators by all groups of judges combined ($N = 239$) are shown in Figure 9.

As indicated in Table V, C was rated significantly lower than either A or B, but there was no significant difference between the mean ratings of A and B. Further, variability of ratings was significantly less for A than it was for either B or C.

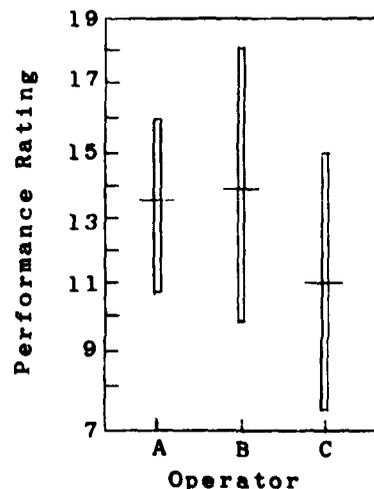


Figure 9. Means (horizontal bars) and standard deviations (vertical bars) of performance ratings of the three operators by all judges combined ($N=239$).

Table V
 Significance of Differences Between
 Means and Standard Deviations of Performance Ratings
 of Operators A, B, and C, by All Judges Combined

	Operator		
	A	B	C
M	13.5	13.8	11.3
σ	2.9	4.2	3.8
Difference Tested		t	p
M _a - M _b		0.76	--
M _a - M _c		7.01	.001
M _b - M _c		6.80	.001
σ_a - σ_b		5.53	.001
σ_a - σ_c		4.07	.001
σ_b - σ_c		1.55	

CONCLUSIONS AND IMPLICATIONS

It was concluded from the results of the experiment that:

1. An unusually good performance in a sequence was weighted more heavily by judges in determining the rating of the entire sequence when that performance occurred either at the beginning or end of the sequence than when it occurred in the middle of the sequence. This was called a "terminal effect."

2. An unusually poor performance in a sequence was weighted more heavily by judges in determining the rating of the entire sequence when that performance occurred at the beginning of the sequence than when it occurred later in the sequence. This was called a "primacy effect."

3. When the ordinal position of an unusual performance was disregarded, the unusual performance was given neither more nor

less than its due weight by judges in determining the overall performance rating. That is, there were no significant contrast or assimilation effects.

4. Even though performances were objectively equal, judges in this experiment gave different ratings to different operators.

The first two conclusions refer to the time-order effects observed under the conditions of this experiment. It is noted that a "terminal effect" is simply a combination of both primacy and recency effects. Therefore, the results showed both primacy and recency effects when an unusually good performance occurred in the sequence, and only a primacy effect when an unusually poor performance occurred. It might be concluded then, that a primacy effect consistently occurred, while a recency effect occurred only in the presence of an unusually good performance. This suggests the possibility that a primacy effect is a general phenomenon in performance judgment, while a recency effect is specific to unusually good performance.

It is possible that the terminal effect occurred because the unusually good performance was more noticeable when it occurred at the beginning or end of the sequence than when it occurred in the middle. The recall data presented in Figure 6 would support such an explanation. Since a primacy effect occurred when an unusually poor performance was present, it would be expected that recall of the unusually poor performance would be better when it occurred on the first trial than when it occurred on later trials. But, the recall data presented in Figure 7 show that the unusually poor performance was recalled equally well whether it occurred first, third, or fifth in the sequence. Therefore, the time-order effects that were observed in this experiment cannot consistently be attributed to the "noticeability" of the unusual performances.

The consistent occurrence of a primacy effect in the present experiment seems to emphasize the importance of "first impressions"

in determining performance ratings. The man who bungles a job at the outset of his tenure may continue to be given inappropriately low ratings even though his performance later improves. And in a similar manner, the man who gives a strikingly good performance when he first comes under the observation of raters may continue to be given inappropriately high ratings even though his performance later deteriorates.

It is interesting to note that the last man rated, operator C, was given a significantly lower mean rating than either of the first two operators, in spite of the fact that performances were objectively equal for all three. This result could be attributed to the characteristics of the man, himself, or to the fact that he was the last to be rated. Although the experimental design did not allow the assessment of the separate effects of individual operators and the ordinal position of an operator, it seems unlikely that the lower rating given operator C was attributable to his personal characteristics. As noted earlier, the operators were of similar appearance and the motion picture films were deliberately underexposed, making it difficult for the judges to perceive physical differences among the operators. It is most probable that the significant differences between the mean performance rating given C and those given A and B were a result of another type of time-order effect. Thus, the rating given an individual worker may depend not only upon the sequence of his own performances, but also upon his ordinal position among the other workers to be rated.

Further study of time-order effects in the context of performance judgment should be directed toward the study of the following variables: (1) the number of performances observed before rating (i.e., rating after every performance vs. rating after a sequence of performances); (2) the duration of the time interval between performances; and (3) the duration of the time interval between the last performance observed and the time at which the rating is made. Until the influence of these variables becomes known, it may not be possible to assess accurately the validity of ratings of human performance.

REFERENCES

- Brown, W.L. & Overall, J.E. Implications of recency effects for probability learning theories. J. gen. Psychol., 1959, 61, 243-251.
- Luchins, A.S. Influence of experience with conflicting information on reactions to subsequent information. J. soc. Psychol., 1960, 51, 367-385.
- Miller, N. & Campbell, D.T. Recency and primacy in persuasion as a function of the timing of speeches and measurement. J. abnorm. soc. Psychol., 1959, 59, 1-9.
- Winer, B.J. Statistical principles in experimental design. New York: McGraw-Hill, 1962.