ORGANIZATIONAL BEHAVIOR AND HUMAN PERFORMANCE 4, 97–111 (1969)

JUDGE: A Laboratory Evaluation[1]

L. W. MILLER AND R. J. KAPLAN
The RAND Corporation

AND

W. EDWARDS
The University of Michigan

This paper describes an experiment performed to evauate the JUDGE technique (Judged Utility Decision GEnerator). The JUDGE system is designed to dispatch aircraft on non-preplanned close air support missions, the number dispatched depending on judgments of target values made by experts at the times when targets appear.

In contrast to an earlier field study employing Air Force officers as subjects, the current experiment included an extensive training period, longer scenarios, and repeated measurements. The subjects were fourteen students from the junior and senior Army ROTC classes at UCLA; they worked for two hours a day over the eight weeks of the experiment.

For comparison, JUDGE was pitted against a second technique called DASC in the experiment, the name being taken from the Direct Air Support Center. This mode of operation is a hypothetical version of the system the Air Force currently uses, and is not a standard Air Force procedure. The subjects performed in both modes against all the situations in the simulated war.

In the JUDGE mode, the subjects assigned a value to each target as it appeared by comparing it with a "standard" target, which had a constant value of 100 throughout the experiment. A computer program then translated each subject's responses into dispatching decisions, and evaluated those decisions based on his value responses.

Operating in the DASC mode, the subject received target reports identical to those in the JUDGE mode, but also containing a graph showing how effective various numbers of aircraft would be against the target. The subject himself assigned aircraft to the target, being permitted to allocate any even number from 0 to 16 of the aircraft remaining to him.

The results clearly show the superiority of JUDGE over DASC when measured by an expected utility criterion. JUDGE performed at the 90%

[1] This research is sponsored by the United States Air Force under Project RAND —Contract No. F44620-67-C-0045—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this paper should not be interpreted as representing the official opinion or policy of the United States Air Force.

level when compared with the perfect possible performance. DASC reached a level of only 40%. We conclude from this that JUDGE is more effective in implementing a subject's value judgments than the subject is himself.

The reliability of both systems was evaluated by measuring both the intersubject and intrasubject correlations. These two measures were substantially higher for the JUDGE system than for DASC, revealing that an exceptional amount of agreement occurs within JUDGE. Examination of the data processing task in isolation from the judgmental process in the decision environment led to the conclusion that JUDGE gains its advantage by turning over the necessary mechanical calculations to a computer.

JUDGE is a decision-making technique designed to aid a commander responsible for dispatching Close Air Support missions in situations such that resource limitations do not allow all demands to be fulfilled. The technique and its underlying philosophy are described in detail by Miller, Kaplan and Edwards (1967).

As each request for close air support is received, the JUDGE system makes a dispatching decision that maximizes the difference between a return gained for sending aircraft against the target and a cost imputed to expending sorties. An estimate of the return depends on the value of destroying the target and the probability of successfully accomplishing the mission with the weapons dispatched. The cost of sorties represents loss of future capability and is derived, by a dynamic programming computation, from forecasts of the number, value and appropriate kill probability function of later targets. Expert judges provide target value estimates as inputs to JUDGE; the technique seeks to maximize total expected value over all dispatching decisions made.

This paper reports an experiment done to evaluate JUDGE. The experiment, an extension of the field study described in Miller, Kaplan, and Edwards (1967) was conducted in RAND's Logistics Systems Laboratory. This experiment was carried out in order to give each subject more time for training and for operation of the system, and thus familiarize him more thoroughly with the response modes and the stimulus environment. Another aim was to measure the test–retest reliability of JUDGE, which required presenting large portions of the experimental material to the subjects twice. Since Air Force subjects were not available for long enough periods of time, ROTC cadets were used.

EXPERIMENTAL DESIGN

The fourteen subjects were recruited from the junior and senior classes of the U. S. Army Reserve Officers Training Corps at the University of California at Los Angeles. These ROTC students had already received some training in map reading, military organization, and the tactics

employed by ground forces. The subjects were paid for their time and worked two hours a day for the eight weeks of the experiment.

The two procedures studied in the experiment are called JUDGE and DASC: JUDGE is our computer-assisted decision technique, while DASC is our version of the method currently in use. The name for the latter was taken from the Direct Air Support Center, the Air Force unit which is responsible for the decisions with which we are concerned here.

A subject operating in the DASC mode received target reports similar to those used in the field study. "DASC" here refers to our experimental mode of operation, not any real Air Force doctrine or procedure. The report included the time, the position and description of the target, and a graph showing how effective various numbers of aircraft will be against the target. The subject's task was to assign aircraft to the target; he could allocate any even number from 0 to 16 of the aircraft he had remaining.

In the JUDGE mode, the subject received for each target a report identical to the DASC-mode report except that it contained neither the mission-effectiveness graph nor the time. The JUDGE task was to assign a value to the target. The subject did so by comparing the target at hand to a standard target having an arbitrary value of 100, which was carefully defined and kept constant throughout the experiment.

A narrative description provided the political and economic background for the simulated battle. For the scenario, we expanded and modified a war game from an unclassified lesson plan used at the U. S. Army Command and General Staff College. The simulated battle was divided into four situations representing six-hour periods on four successive days. Each situation was broken into three 2-hour parts or horizons, and known quantities of sorties became available at the start of each horizon. Each situation was introduced by a short narrative describing the action leading up to it. A map overlay for each situation showed the position of friendly and enemy forces. When he worked in the DASC mode, the subject was told how many additional aircraft he had available at the beginning of each horizon. He knew that the expected arrival rate of requests was ten per hour, so that he could expect an average of twenty target reports during each horizon. Table 1 shows the actual numbers of targets presented and the aircraft available.

Each experimental situation was contained in a single loose-leaf binder; each subject worked individually at his own pace. A form was provided to the DASC operators to keep track of the number of undispatched aircraft, and this form had to be updated after each target and after the beginning of each new horizon, when additional aircraft were made available. The JUDGE operators made their value estimates

TABLE 1
AIRCRAFT AVAILABLE AND NUMBER OF TARGETS, BY HORIZON AND SITUATION

	No. of aircraft available; horizon:			No. of targets presented; horizon:		
Situation	1	2	3	1	2	3
1	60	50	40	15	28	16
2	50	40	30	30	19	19
3	44	36	30	19	17	16
4	40	30	24	24	21	18

on a separate form for each target by putting a mark on a scale. The scale was about six inches long, open-ended at the top, with the numeral 0 at the bottom and the number 100 two inches from the bottom. The word "Standard" appeared to the left of the 100 mark to remind subjects constantly that each judgment was a comparison of the target under consideration with the standard target.

TRAINING

The subjects received a series of lectures about Division-level Organization and Deployment, Squad and Platoon Tactics, Reconnaissance, Joint Air/Ground Operations, and the Tactical Air Control System.[2] A discussion period followed each lecture.

Materials from the field study, identical in form to those to be used in the experiment, were used to train subjects in both the JUDGE and DASC procedures. Responses of the Air Force officers in the earlier experiment were given to the subjects as guidance in making their judgments, but they were told that differences of opinion were allowed and even desirable. A total of ten hours was spent in training.

RUNNING ORDER

The fourteen subjects were randomly divided into two equal-sized groups to counterbalance learning effects over time. One group performed in the sequence JDDJ (JUDGE-DASC-DASC-JUDGE), and the other group was given the DJJD sequence. The first two situations were replicated for each group, and then the last two situations were handled

[2] The authors wish to express their appreciation to Brig Gen L. L. Wheeler, USA, Ret., Col J. H. Hayes, USA, Ret., Col G. C. Reinhardt, USA, Ret., Lt Col C. B. East, USAF, Ret., and Lt Col J. T. Hanton, USAF, members of the RAND staff, for their skilled assistance in this portion of the subjects' training.

in the same way. The complete running order for the scenario portion of the study is given below.

Order Position	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Group 1	D-1	D-2	J-1	J-2	J-1	J-2	D-1	D-2	D-3	D-4	J-3	J-4	J-3	J-4	D-3	D-4
Group 2	J-1	J-2	D-1	D-2	D-1	D-2	J-1	J-2	J-3	J-4	D-3	D-4	D-3	D-4	J-3	J-4

The sixteen order positions are indicated in the top line, and the body of the diagram shows the system by the letter (D = DASC, J = JUDGE) and the situation by number. Each subject, it can be seen, operated twice against each situation using each system, making a total of four exposures to the same material.

THE DATA PROCESSING

The numerical results reported in the next section were obtained from a series of four computer programs written in the SIMSCRIPT language (Markowitz, Hausner, and Karr, 1962). Breaking the data processing down in this way resulted in a convenient set of short programs, allowing us to expand and modify the scope of the analysis as new ideas occurred. The functions of these four programs are outlined below.

RULE. The program RULE produces tables used to transform the subjects' value responses into dispatching decisions. Inputs to this program are the following:

(1) Beginning and end times for each part (horizon) of the situation.

(2) Number of targets presented in each part.

(3) Number of aircraft to become available at the beginning of each part.

(4) Time of presentation of each target.

(5) Type index (indicating the applicable mission success function) for each target.

(6) Forecasted distribution of target types.

(7) Parameters p and α for each of the six mission success functions.

(8) Forecasted rate at which requests are expected to arrive.

(9) Forecasted value distributions. (These were introduced as modifications to subroutines that calculated the distribution function and conditional means of the assumed value distribution.)

The output is a table for each part, indexed by number of aircraft

remaining and the serial number of the target. An entry in the table represents the expected future value of having a given number of sorties remaining at a time corresponding to the time of presentation of the target. The tables are punched into cards so that they may be used as input to the VALUE program, which performs the actual calculation of aircraft dispatchings.

These calculations are similar to those performed for the evaluation of the field study. The difference here lies in the treatment of boundary values needed to connect the three parts of each situation.

As an example, consider the boundary conditions for the middle horizon of situation 1. The value equation was solved for the last part with the time parameter running between zero and two hours, and for even numbers of sorties up to 80, even though only 40 sorties were to be made available at the beginning of that part. The additional levels of n (number of sorties remaining) would be necessary in case sorties are left over from the preceding part and are needed to obtain the boundary values for the part immediately preceding.

Let W_n be the expected value of n sorties at the beginning of the last part. Since 40 additional sorties were to be made available at that time, the value of m sorties remaining at the end of the middle part would be $W_{m+40} - W_{40}$, for $m = 0, 2, 4, \ldots, 40$. It is also desirable to calculate curves for m up to 80 in the middle part, and the extra boundary values were obtained by extrapolation based on constant third differences.

For each situation the forecasted value distribution was uniform. In situation 1, the mean was set to 100. This guess was based on our experience in the field study and the consideration that the standard target in the present experiment was likely to appear less valuable to subjects than that used previously. But the average of the JUDGE responses in situation 1 turned out to be 123, so that a mean of 120 was used in the computation of the dispatching rule for the remaining three situations. A comparison between the forecasted means and the actual means is given below.

Item	Situation				Average
	1	2	3	4	
Forecast	100	120	120	120	
Group 1	114	112	112	140	120
Group 2	132	124	122	163	135
Average, all subjects	123	118	117	152	128

Our prediction given to RULE was that all target types[3] would appear with equal frequency. The actual distribution, however, taken over all situations, was as shown in the following tabulation:

Target type	1	2	3	4	5	6
Relative frequency	.079	.145	.240	.244	.170	.124

In all four situations, the expected request rate was set equal to ten requests per hour, yielding an expectation of 60 targets in each situation. The actual numbers of targets presented were 59, 68, 52, and 63.

VALUE. The program VALUE is used to translate each subject's responses into dispatching decisions, and to evaluate the dispatching decisions based on his value responses according to the expected value criterion. The inputs to this program are:

Table of expected values by number of remaining sorties and target index (the output of RULE).
Number of targets presented in each part.
Number of aircraft to become available at the beginning of each part.
Type index for each target.
Mission success functions for each target type.

For each subject, and for each replication of each situation, the program is given the sequence of DASC responses and of JUDGE responses.

In addition to translating JUDGE responses into aircraft dispatches, the program also produces a sequence of dispatches based on the FCFS (first-come, first-served) rule and a sequence of perfect dispatches with knowledge of all target types and JUDGE value responses made by the subject. The FCFS method simply dispatches four sorties to each target until the supply is exhausted. The perfect system is identical to that used in the previous study: pairs of aircraft are assigned to targets in order of decreasing marginal utility, regardless of the ordering of targets. Aircraft becoming available at the beginning of the second part are available for targets in the last two parts, and finally the initial supply of aircraft may be used on any targets in any of the three parts of the situation.

[3] The six target types were the same as those used in the field study reported in Miller, Kaplan, and Edwards (1967), and were defined by a mission effectiveness formula of the form $\eta(x) = 1 - (1 - p)^{x^{\alpha}}$. The target types ranged from very "hard" (Type 1), for which even a large number of aircraft sent against it produced little effect, to very "easy" (Type 6) for which even a small number of aircraft produced nearly 100% mission effectiveness.

Summary statistics for the four systems are calculated and the sequence of JUDGE dispatching decisions is punched for further analysis.

COR. The program COR accepts as data the set of responses made by a group of subjects within one order-position and produces a table of interpersonal correlation coefficients (Pearson product moment). The mean and standard deviations of these are then calculated.

TERET. The program TERET (TEst–RETest) is used to assess the agreement between replications of the experiment. For all subjects within a group, the responses from the two replications of the same system and situation are given as input. For each subject, the correlation coefficient between his two sequences of responses is calculated. In addition, an analysis of variance is performed in which the sources are Targets, Subjects, Target by Subject interaction, and error. The interaction component may be interpreted as being caused by disagreement among different subjects, while the error component measures the degree of unreliability within individual subjects.

RESULTS

System Performance

Our first concern is with the relative performance of the two systems being studied in their allcoation of aircraft against the targets presented. The criterion of performance is the product of two elements: (1) the utility of the mission against a target, calculated for each subject using the value he himself assigned to the target while operating in the JUDGE system, and (2) the probability of mission success associated with the number of aircraft dispatched either by the subject himself in the DASC mode of operation or by the dispatching rule in the JUDGE system. In the same way, a single subject's score for an individual decision point in the DASC system is obtained by multiplying these two elements. The sum of these values for all decision points in a situation is his score for that condition.

To get an upper and a lower bound on system performance against which to compare our experimental systems, two hypothetical decision systems were created. The "Perfect" system enjoys perfect foreknowledge of all the targets to appear in the situation, and therefore methodically assigns aircraft two at a time to targets in order of decreasing marginal utility until it has spent all the aircraft. The score achieved in this way represents the maximum obtainable for the particular set of values under consideration over that target set.

As a lower limit from which to compare the performance of other

systems, zero is not a fair value since even a very bad dispatcher would pick up at least a few points on the expected utility scale. Consequently, the performance of a first-come, first-served system was chosen as the lower limit. The FCFS system merely assigns four aircraft to the targets in the order in which they appear until the supply of aircraft is exhausted. Properly speaking, this is not a true decision system, of course, since it dispatches aircraft without consulting any of the input information available. Any system, to be considered useful, would have to surpass the performance of the FCFS system.

Table 2 presents the average of system performance scores over all subjects for each situation and replication. The first four columns show the mean over subjects of the total expected utility for each of the four systems. The FCFS and the optimum columns represent the lower and upper bounds, respectively, and the data for the two experimental systems are presented under their appropriate labels.

Since the raw scores are difficult to interpret meaningfully, they have been translated into percentages of the perfect score. This measure is referred to as the "Efficiency" of the system, and the values for the three dispatching methods are shown in the next three columns of the table. The last two columns further reduce the number of scores to two, leaving only the experimental systems. The values under the label "Effectiveness" are the percentages of the distance from the FCFS (taken to be 0) toward the Perfect (at 1.00) that each of the systems achieves. These figures are the means of the effectiveness numbers for the individual subjects and therefore are not calculable from the utility measures provided in the corresponding rows of this summary table. It can be seen from these two columns that the JUDGE system consistently performs at about the 90-% level, while the DASC only rarely reaches the 40-% mark.

It is obvious from inspection of this table that, using expected utility as a criterion, the JUDGE system outperforms DASC in every case. The JUDGE system dispatches aircraft far more effectively than does the subject himself; or in other words, given the value system supplied by the subject, JUDGE can implement this set of judgments under the constraints of the situation better than the subject himself can.

Intrasubject Agreement

One of the main reasons for conducting the second study in a controlled laboratory environment was the opportunity for collecting repeated observations on the same individual. Whether or not a particular judge of values agrees with other judges is of some concern, especially

TABLE 2
SYSTEM PERFORMANCE

Situation	Expected utility				Efficiency			Effectiveness	
	FCFS	DASC	JUDGE	Optimum	FCFS	DASC	JUDGE	DASC	JUDGE
Replication 1									
1	1670	2104	2524	2773	.610	.770	.923	.406	.813
2	1311	1477	2001	2109	.626	.706	.952	.198	.865
3	1224	1408	1697	1833	.673	.765	.967	.262	.902
4	971	1342	1785	1932	.521	.686	.940	.300	.885
Replication 2									
1	1614	2035	2408	2576	.630	.792	.943	.433	.855
2	1310	1498	2036	2123	.621	.714	.962	.226	.897
3	1217	1400	1798	1843	.662	.757	.975	.254	.927
4	968	1319	1777	1924	.521	.683	.934	.305	.873

to the degree that all the judges can be considered expert in the area under consideration. If the value judgments are to be regarded as competent measures of the individual's opinions, however, his judgments must be reasonably stable over time as he deals with the same item. For this reason the entire set of stimulus material was presented to each subject twice.

Table 3 contains the test-retest correlations broken down by system and group on one dimension of the table and by situation on the other. The relative positions of the correlation magnitudes are the same here as they are in the intersubject data, and the same explanations apply to this set of numbers. The main point to stress here is that these numbers, especially those associated with the value judgments, are exceptionally high. A subject exhibits the greatest agreement with his own previous opinions when he is making value judgments, and the least when he is dispatching aircraft on missions.

TABLE 3
TEST–RETEST CORRELATIONS

System	Group	Situation 1	2	3	4
DASC	1	.519	.510	.580	.644
	2	.639	.545	.688	.480
Values	1	.830	.806	.865	.857
	2	.687	.673	.849	.913
JUDGE	1	.767	.718	.714	.752
	2	.743	.669	.762	.765

One set of test-retest correlations among value judgments—that for group 2 in the first two situations—is much lower than the rest of the correlations. This group had the judgmental task to perform first, and the novelty of the response requirements made the performance much more variable. Table 4 shows the correlations arranged in the order in which the situations were presented to the subjects for both DASC and the value judgments. It is encouraging to note that the figures in the row for the values are increasing, indicating that the responses are stabilizing over time, or in other words, learning is taking place. No such tendency can be observed in the DASC data.

INTERSUBJECT AGREEMENT

The concern for the reliability of any system often comes to focus on the agreement that results when a group of individuals operate the

TABLE 4
TEST–RETEST CORRELATIONS IN THE ORDER OF ADMINISTRATION OF SITUATIONS TO SUBJECTS

	Order position			
System	1	2	3	4
DASC	.59	.51	.61	.58
Values	.68	.82	.86	.88

system independently. This reliability question is discussed in Miller, Kaplan, and Edwards (1967), which reports the first JUDGE experiment. The product-moment correlations presented in Table 5 are the commonest measure used to describe the reliability of a set of measures. It is to be remembered that an upper bound is placed on these correlations by the intra-subject agreement measures discussed in the preceding section.

As can be seen from the table, the correlations associated with the DASC system are the lowest, averaging about .40. The intercorrelation of the value judgments among the subjects is considerably higher, about .60. There is a tendency to agree more about value judgments than about dispatching aircraft. Contrary to the finding in the first study, however, the dispatchings from the JUDGE system based on the values do not correlate as well as do those values themselves. This result is explained on the basis of the difference in the coarseness of scale between the values and the dispatchings, which had a chance to operate in the expanded environment of the second experiment. The value judgments of two individuals, for example, might have been 115 and 120 for a particular target. These, being fairly close together, would tend to increase the correlation of judgments between these two individuals. Should this difference in judgment encompass a threshold in the dispatching rule, however, there would be a marked difference in the output of the JUDGE system, say from 2 to 4 aircraft being sent on the mission. At critical points in the value scale, therefore, small differences in judg-

TABLE 5
AVERAGE CORRELATION OF SUBJECTS WITH ALL OTHER SUBJECTS

	Situation				
System	1	2	3	4	Mean
DASC	.50	.32	.46	.35	.41
Values	.58	.52	.68	.74	.61
JUDGE	.64	.41	.54	.42	.50

ments will have relatively large impacts on the resulting differences in decisions, and when enough of these points are encountered in the course of operation, the correlation of the outputs among subjects will suffer accordingly.

Sources of Variability in Value Responses

An analysis of variance provides another way of looking at the questions of intersubject and intrasubject agreement. Assuming random-effects models for both targets and subjects, each value response can be viewed as having the form

$$V_{ijk} = T_i + S_j + U_{ij} + e_{ijk},$$

where the terms on the right-hand side refer respectively to mean target value, subject bias, interaction between targets and subjects, and a random deviation by a subject responding to a target on different occasions. A more elaborate model including group and situation effects could be proposed, but it is instructive to display these effects by tabulating eight different analyses, as in Table 6.

TABLE 6
Sources of Variability in Value Responses

Group	Source	Situation 1	2	3	4
1	Targets	.54	.34	.54	.46
	Subjects	.07	.04	.04	.08
	Target by subject	.22	.45	.31	.34
	Error	.17	.17	.11	.09
2	Targets	.42	.34	.57	.55
	Subjects	.03	.12	.10	.12
	Target by subject	.12	.19	.17	.21
	Error	.43	.35	.15	.12

With the assumption implied in the random-effects model, it is possible to obtain estimates of variances associated with each of the four factors in the model. Table 6 shows the proportion of total variance attributed to each source by group and situation.

Except for the early situations with group 2, the largest component of variance is due to target difference, as one would hope. Variability due to subject biases is small. The target-by-subject interaction is a measure of the disagreement among subjects over the importance of particular targets. The error term measures unreliability within subjects and is low except for group 2 in the early portions of the experiment.

That this measure is low compared to the interaction term indicates that the less-than-perfect correlations of Table 4 are due more to differences of opinion than to random behavior.

The Data-Processing Task

One of the fundamental principles in the JUDGE technique is the separation of the judgmental portion of the decision task from the data processing portion. To get an idea of the relative contribution of these separate portions to the difficulty of the entire task, we constructed a situation in which the subject had to do only the data processing portion of the task. The entire experimental situation was run through for each subject, using as stimulus material not the target reports previously given, but rather the values each subject had himself given while operating in the JUDGE mode. His task in this part of the study was to dispatch aircraft against these values with the aim of maximizing expected utility. Additional training was given to the subjects before this part of the experiment, explaining the expected utility criterion and giving complete details on how the effectiveness score would be calculated. The instructions were to dispatch aircraft so as to make that score as high as possible.

Table 7 compares the effectiveness scores for each group of subjects on the DAVO task (Dispatching Against Own Values) with the corresponding effectiveness of the DASC and the JUDGE tasks. In every case, the DAVO scores fell between the DASC and the JUDGE scores accounting for about 25 to 50% of the difference between these two. This result is interpreted to mean that somewhat less than half of the improvement in decision-making brought about by the JUDGE technique is attributable to the separation of the decision tasks, which

TABLE 7
Effectiveness of Dispatchings against Own Values (DAVO) Compared with DASC and JUDGE Dispatchings

Situation	Replication	DASC	DAVO	JUDGE	$\frac{\text{DAVO} - \text{DASC}}{\text{JUDGE} - \text{DASC}}$
1	1	.41	.63	.81	.55
	2	.43	.56	.86	.30
2	1	.20	.54	.87	.51
	2	.23	.46	.90	.34
3	1	.26	.46	.90	.31
	2	.25	.43	.93	.26
4	1	.30	.62	.88	.55
	2	.30	.53	.87	.40

allowed the expected utility calculations to be made in isolation from the other portions. The man in the system can perform the data processing task to a rather limited extent, but the larger improvement in performance comes when this mechanical operation is turned over to a machine.

CONCLUSIONS

The purpose of the present experiment was to evaluate the JUDGE technique in a laboratory, where more control could be exerted over the subjects' responses than was possible in the field study reported in Miller, Kaplan, and Edwards (1967). The results of this study confirm the vast superiority of JUDGE over a conventional system in dispatching close air support missions found in the previous investigation.

In terms of the performance measure used—total expected utility attained—JUDGE performed at the 90% level, whereas DASC, our version of the current system, reached a level of only 40% when compared with perfect performance.

The reliability question was attacked by collecting data on both intrasubject and intersubject agreement. Test-retest correlations showed that there was an exceptionally high degree of agreement within individuals when operating in the JUDGE mode, but only modest agreement when subjects were performing in the DASC portion of the experiment. The agreement among subjects appears respectably high for the JUDGE system when considered in the light of the upper bound placed on it by the intrasubject agreement, while the DASC system shows only moderate agreement. The correlations of both kinds increases over time for JUDGE, but not for DASC, indicating that learning operates to advantage in the former but not in the latter.

Data were collected on the subjects' performance on the data-processing portion of the decision task in isolation from the judgmental part, and it was found that they were relatively poor at it. The real advantage of the JUDGE technique comes when the assigned values are turned over to a machine for calculation, so that the man in the system need not perform the entire task.

REFERENCES

MARKOWITZ, H. M., HAUSNER, B., AND KARR, H. W., SIMSCRIPT: A Simulation Programming Language. The RAND Corporation, RM-3310-PR, November 1962.

MILLER, L. W., KAPLAN, R. J., AND EDWARDS, W., JUDGE: A value-judgment-based tactical command system. *Organizational Behavior and Human Performance*, 1967, **2**, 329–374.

RECEIVED: June 24, 1968