7. International Business Machines Corporation, IBM System/360 Operating System PL/1 (F) Language Reference Manual, Order No. GC28-8201.

8. International Business Machines Corporation, IBM System/360 Operating System PL/1 (F) Programmers Guide, Order No. GC28-6594.

9. Markowitz, Karr, Hausner, and the RAND Corporation, SIMSCRIPT: A Simulation Programming Language, New Jersey: Prentice-Hall, 1969.

10. Nelson, J. G. and Jackson, T. B. F111B Systems Effectiveness and Human Factors Simulation Program: Crew Performance Simulation Model, Human Factors Engineering Systems Office Report 2-68, Johnsville, Pa.: NADC-LS-6801, 1968.

11. Pollack, S. V. and Sterling, T. D. A Guide to PL/1, New York: Holt, Rinehart, and Winston, Inc., 1969.

12. Siegel, A. I., and Wolf, J. J. Man-Machine Simulation Models, New York: John Wiley and Sons, Inc., 1969.

13. Topmiller, D. A. Mathematical Models of Human Performance in Man-Machine Systems, Wright-Patterson AFB, AMRL-TR-68-22, 1968.

MEASUREMENT OF AIR TRAFFIC CONTROLLER PERFORMANCE

DR. E. BISER
Avionics Laboratory, USAECOM, Fort Monmouth, New Jersey
MESSRS. S. BERG, W. PATTERSON, AND J. MIKULA
American Electronic Laboratories, Inc., Colmar, Pennsylvania
MR. H. MENCHER
Avionics Laboratory, USAECOM, Fort Monmouth, New Jersey

To supply technical support for the concept formulation of an Air Traffic Management System, a test vehicle was developed to evaluate certain automated enroute air traffic control concepts in a tactical environment. Designated the Semiautomatic Flight Operations Center (SAFOC), it was evaluated by its ability to control simulated Army air traffic, flying according to realistic tactical scenarios. The target simulators at the National Aviation Facility Experimental Center (NAFEC) provided the air traffic input, and automatic data collection techniques gathered the output.

One of the primary purposes of the evaluation was to test the ability of air traffic controllers to work with automated equipment while retaining the final decision on any control commands. It is felt that the data collection, reduction, and evaluation techniques to be described in this paper are of general interest in establishing and quantifying human performance measures in a semi-automated environment.

The technical objectives of the enroute test bed are:

1. To regulate Army air traffic under instrument flight rules

2. To provide flight following capability under visual flight rules

3. To improve information transfer among system elements and units being supported

4. To provide computer facilities to automatically analyze air traffic data for display and decision-making by an operator

5. To provide a means for making commanders aware of the current air traffic situation for overall tactical planning

6. To perform specific functions required for air traffic regulation

SYSTEM FUNCTIONS

SAFOC includes data processing, radar processing, display, and manual backup subsystems to provide the following capabilities:

1. Flight data processing
2. Flight following
3. Flight handoff
4. Identification assistance
5. Emergency assistance
6. Air/ground coordination
7. Ground/ground coordination

SAFOC provides the following methods of flight tracking: digital data link, radar beacon, radar skin return, and flight plan following.

SAFOC TEST CONFIGURATION

Figure 1 shows the test operations and information flow diagram. The scenario generator program generates scenarios and scripts based on realistic scenarios. The scripts are followed by pilots, who simulate actual flights, using target generators which are part of NAFEC's data link simulation.

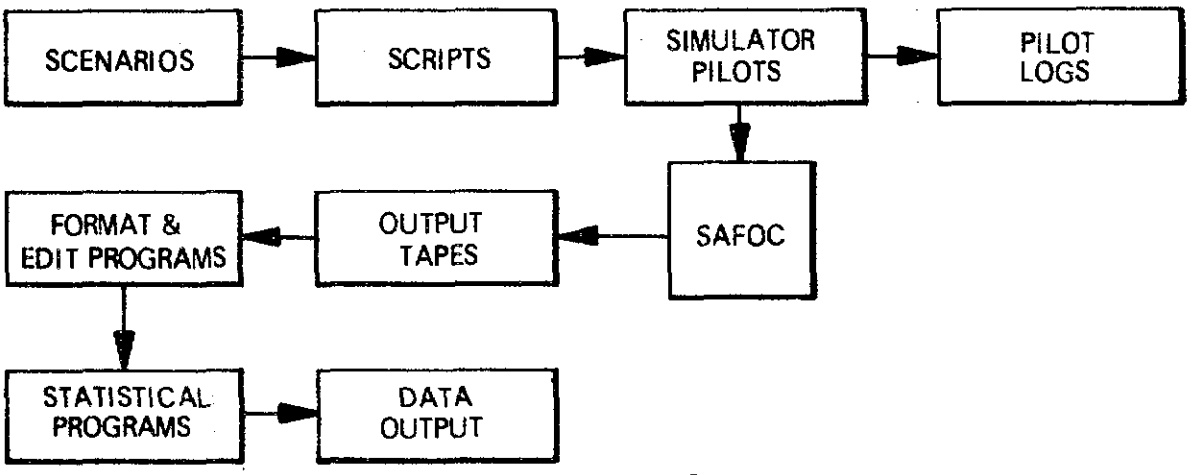

Figure 1. SAFOC Test Operations and Information Flow Diagram

Using a predetermined, operational-procedural mode, the system controls the simulated flights and produces exhaustive time histories on magnetic tape. These histories include all actions performed by the equipment or by the controller. The raw output data tape and the target generator history tape are processed using a series of formatting and editing programs, after which statistical programs generate the desired data output.

Evaluation of the SAFOC consists of a series of tests including Preliminary testing for familiarization with the equipment and training of controllers; Phase I testing to determine the best operational method, to rank controller performance, and to find the system performance measures and effectiveness measures; and Phase II testing to evaluate system and controller performance, using realistic tactical scenarios, and to recommend changes to optimize the SAFOC.

PHASE I TEST PLAN*

To evaluate controller performance and to determine the best method for operating the SAFOC, a series of tests was planned and conducted using the NAFEC simulation facility. The tests consist of a series of scenarios of three different traffic levels. Each of four controller teams operate the SAFOC according to four different operational-procedural combinations. The outputs, consisting of system effectiveness measures, are ranked to determine the best operational-procedural mode. Figure 2 shows this experimental design.

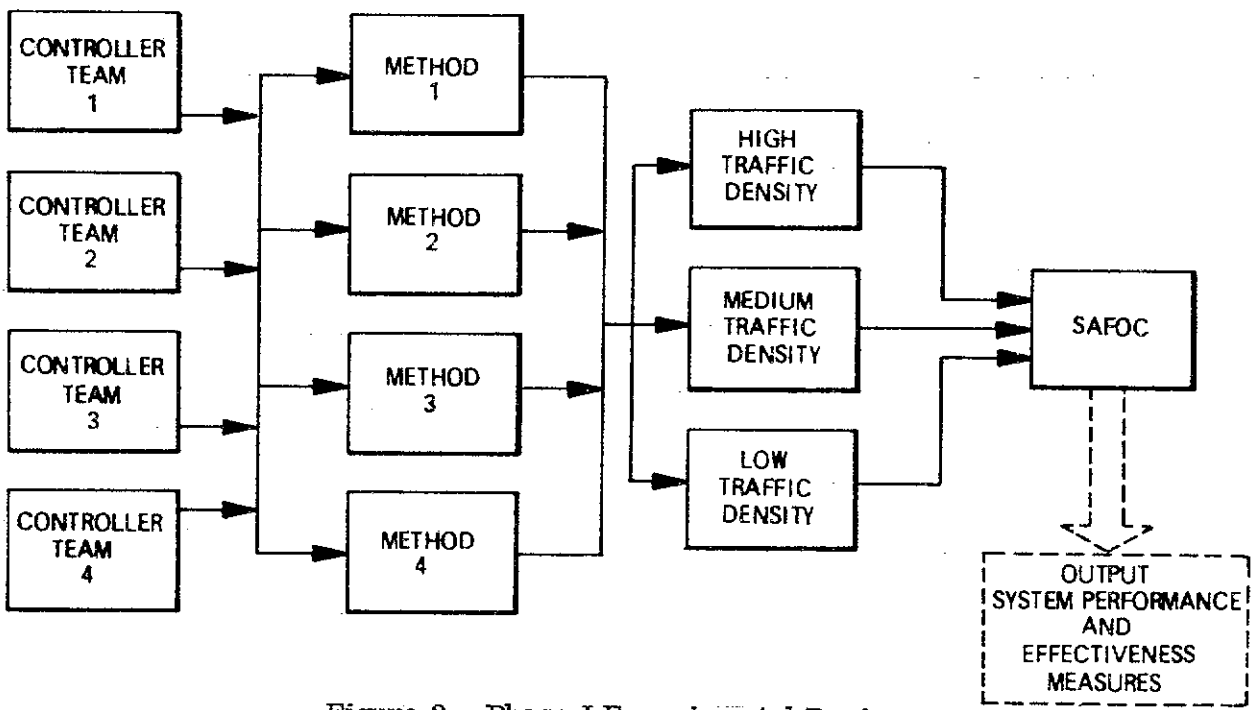

Figure 2. Phase I Experimental Design

* "Phase I Test Plan for a Semiautomatic Flight Operations Center (Design of Experiments Program)", July 1, 1970, Contract No. DAAB07-69-C-0040.

SYSTEM PERFORMANCE MEASURES. Evaluate the individual items which, in total, influence the system effectiveness. The performance measures represent, for example, the actual contributors to total workload, and it is through improvements in the performance measures that the effectiveness measures can be improved.

1. Time to perform each service
2. Service rate
3. Waiting time for service
4. Event time history
5. False dismissal probability
6. Actual density history
7. Queue lengths
8. Typewriter errors
9. Near miss history
10. Communication time history
11. Number of impossible requests
12. Altitude change history
13. Closest approach history

SYSTEM EFFECTIVENESS MEASURES. (SEMs) are used to provide relative rankings of the operational-procedural modes and to evaluate relative controller team performance. The following measures are chosen because they represent the characteristics most important to the user:

Safety is the number of near misses per aircraft mile flown.
Controller workload is the total time for all flight servicing.
Communications workload is the total time spent in communications.
Delays are the actual departure time delay from the planned departure time. Delays are important in a tactical situation.
Throughput is defined as the actual number of flights entered during steady state divided by the number of entries planned in that time.
Capacity is the peak flight density safely handled by the system.
Uncontrolled time is the total time of flights within the control area without being controlled by the system.

MATHEMATICAL MODEL AND CONTROLLER TEAM PERFORMANCE

It is possible to rank the controllers on some measurable characteristics. For this ranking the C teams in R replications using different but equivalent scenarios are used. In essence, random trials of the controllers' ability to handle repeated scenarios are performed to determine whether the controllers are significantly different in their abilities. If they are different, they are ranked in order of their abilities to determine those needing additional training.

Table 1 shows the symbology of the effects (effectiveness measures) resulting from replications of different scenarios by each controller, and indicates the sums to be performed. Table 2 shows the usual analysis of variance for a two-way classification based on a mixed-model of fixed teams and random samples from the hypothetical population of replication observations.

When the scenarios are run, the data (such as the workload times of each controller) consisting of the X_{ij}'s shown in table 1 is operated on by performing the column and row sums, followed by the operations shown in table 2. Then the sums of squares and mean squares shown in table 2 are computed. The F ratio is computed:

$$(1)\quad F = \frac{S_2\,(R-1)}{S_3}$$

TABLE 1. EFFECTS OF REPLICATIONS

TEAMS						
Replications	1	2	.	.	C	Row Sum
1	X_{11}	X_{12}	.	.	X_{1C}	$\sum_{j}^{C} X_{ij} = C\bar{X}_{1.}$
2	X_{21}	X_{22}	.	.	X_{2C}	$\sum_{j}^{C} X_{2j} = C\bar{X}_{2.}$
.	.	.	.	.	.	
.	.	.	.	.	.	
R	X_{R1}	X_{R2}	.	.	X_{RC}	$\sum_{j}^{C} X_{Rj} = C\bar{X}_{R.}$
Column Sum	$R\bar{X}_{.1} = \sum_{i}^{R} X_{i1}$	$\sum_{i}^{R} X_{i2} = R\bar{X}_{.2}$			$R\bar{X}_{.C} = \sum_{i}^{R} X_{iC}$	$\sum_{i}^{R}\sum_{j}^{C} X_{ij} = RC\bar{X}_{..}$
C = Number of teams R = Number of replications						

TABLE 2. ANALYSIS OF VARIANCE FOR A TWO-WAY CLASSIFICATION

SOURCE	SUMS OF SQUARES	DEGREES OF FREEDOM	MEAN SQUARE
Replications	$S_1 = \sum_{i}^{R} (\bar{X}_{i.} - \bar{X}_{..})^2$	R-1	$S_1/(R-1)$
Teams	$S_2 = \sum_{j}^{C} (\bar{X}_{.j} - \bar{X}_{..})^2$	C-1	$S_2/(C-1)$
Error	$S_3 = \sum_{i}^{R}\sum_{j}^{C} (X_{ij} - \bar{X}_{i.} - \bar{X}_{.j} + \bar{X}_{..})^2$	(R-1) (C-1)	$S_3/(R-1)$ (C-1)
TOTALS	$S_4 = \sum_{i}^{R}\sum_{j}^{C} (X_{ij} - \bar{X}_{..})^2$	RC-1	---

where S_2, S_3 and R are defined in table 2. Reject the null hypothesis of no team difference if (1) exceeds the α point of the F distribution with (C-1) and (R-1) (C-1) degrees of freedom. For example, if α equals 0.05, C equals four and R equals two, then the critical value is:

(2) $F_{0,\ 0.05,\ 3,\ 3} = 9.28$

If the computed value of F was larger than 9.28 one could say that the controllers performed their tasks with significantly different capabilities and that similar tests would repeat this conclusion with a 5 percent risk of being in error.

If we wish to rank each of the teams with respect to every other team, Tukey's multiple comparison procedure can be used. This consists of taking the difference between every pair of team performance means computed during the test:

(3) $B_1 - B_2 = Z_1$

$B_1 - B_3 = Z_2$

$B_1 - B_4 = Z_3$

$B_2 - B_3 = Z_4$

$B_3 - B_4 = Z_6$

where B_ℓ is the mean workload performance parameter of ℓth team

$$B_\ell = \frac{1}{R} \sum_{i}^{R} X_{i\ell}$$

We then compute:

(4) $$Z_1 - \frac{S}{\sqrt{2}}\, q_\alpha \leq \delta \leq Z_1 + \frac{S}{\sqrt{2}}\, q_\alpha$$

where $S = \sqrt{S_3 / (C-1)\ (R-1)}$

q_α = critical value for the Studentized Range.

The value of q_α is extracted from the indicated table at some value of α (say 0.05) and for the error degrees of freedom (3 in the example). If (4) contains zero (that is, if the left hand term is negative and the right hand term is positive) then the difference Z_1 is not significant. This is performed for the remaining Z's and rankings are achieved on the basis of the relative values of (4). For example,

(5) $\underline{B_2\ B_1}\ B_4\ B_3$

where the underlines might indicate no significance between differences in B_2 and B_1 but significant differences in B_2 and B_4. We would then conclude that Team 2 is not significantly better than Team 1 (at the 5 percent level of significance) but that the difference in performance betweem Team 2 and Team 4 is significant. This result could be used to indicate that Teams 3 and 4 needed additional training to bring them up to Team 2 and Team 1.

SAFOC DATA COLLECTION

A write-only magnetic tape unit records data directly from the SAFOC computer memory using an I/O channel. The data directly accessible to the tape unit is raw data, not necessarily the data to be processed by analysis or statistical programs. To demonstrate the difference between the two types of data, the following is a partial list of data variables of interest in statistical processing:

- Number of conflicts versus flight density
- Departure queues
- Handoff queues
- Alert queues
- Number of impossible requests
- Departure rate
- Arrival rate
- Alert rates
- Service request rates
- Waiting times in queues

- Communication channel occupancy
- Total fly-through time
- Simulation fly-through time
- Mean service times

None of the above are directly available from the computer memory. The form of most data in computer memory involves the data related to flight plans; the activities at the console and typewriter, such as illuminated or blinking pushbuttons; and the SAFOC status, such as the display modes and content.

The method of obtaining data from the SAFOC is to record events on magnetic tape as they occur. For example, consider the measure of alert waiting time. This is measured by the time between the setting and the resetting of an alert blink bit. When an alert blink bit is set, the type, the time, and the flight number, for which it is being set, are recorded on magnetic tape. When the blinking alert is answered by a controller, the alert blink bit is reset and this event is again recorded on magnetic tape. In order to obtain the alert waiting time, the data reduction program must extract these events from the magnetic tape and compute the difference in time between set and reset of the blink bit for that particular flight.

Alert service queues are also obtained in the data reduction program by extracting the times of set and reset of an alert blink bit. Whenever a blink bit is set, a counter is incremented by one. Whenever a blink bit is reset, the counter is decremented by one. The time history of this counter is a measure of alert service queue length throughout the run.

The Magnetic Tape Synchronizer (MTS) is used to collect SAFOC data by interfacing with one to four Magnetic Tape Units (MTU) and with one of the I/O channels of the SAFOC General Purpose Computer. The MTS provides control of four MTU's through programmed instructions received from the SAFOC computer. The MTS controls data transfers between the computer and the MTU's in both directions; it controls tape positioning, and it supplies status to the computer on itself and any MTU. MTS operations are initiated by programmed instructions received from the computer. Once an operation has been initiated, communication between the computer and the MTS is accomplished by input data requests, output data requests, interrupt requests, and acknowledge signals. This allows transfer of data into and out of memory without impeding program operation.

In order to obtain the SAFOC data, "bugs" are inserted into the operational SAFOC program at the appropriate locations causing data to be recorded on magnetic tape related to particular events; such as:

1. Event occurring (handoff blink, conflict alert, flight hook, etc.)
2. Time event occurred

3. Console initiating event

4. Flight number associated with event (more than one flight number for conflicts).

DATA EXTRACTION

When both the SAFOC history and target generator history tapes are ready for processing, tape data is edited and converted so that the information which will be used for statistical processing is directly available. The handling of this data is performed by the Format and Edit programs. These programs convert the data and store it in appropriate lists. From these lists magnetic tape files are prepared. For the purpose of eliminating data, which may prevent proper statistical processing, a printout of these data lists is also produced. After examining the printout and determining which is "bad data", the Format and Edit programs are rerun, the "bad data" removed, and new magnetic tape files generated.

These files provide the input data for the Analysis and Statistical Programs, which use "packaged" subroutines such as those provided in the Biomed Statistical Package. The elements of such a package are used as subroutines which are "called" as required for the specific statistical processing to be performed. Outputs are histograms and other statistical data such as means, standard deviations, etc.

PHASE I TEST RESULTS

Because a lesser number of controllers were available, the planned testing of four teams was changed to test individual controllers. The results of the controller performance analysis indicate no significant differences in the system effectiveness measures with the results of all operational methods combined. That is, an F test shows that the variation among controllers is not significantly larger than the variation among the test results for the same controller.

A comparison of the incomplete events for each controller was performed. An incomplete event is an alert not answered or a service not completed. An F test indicates that there were no significant differences in the number of incomplete events among the controllers at the 0.1 level of significance.

In addition to the effectiveness comparisons, tests were made to determine if the differences in controller errors per run were significant. The average numbers of controller errors per Phase I test were found to be significant, and the following ranking of the controllers (designated as C, D, L, M, Sk, St, W) was performed on this basis:

Phase I Tests

Ranking: L M W St C Sk

Since controller errors exhibited significant differences among controllers, regression analyses on controller errors versus controller experience, education, and aptitude scores were tried. These analyses did not indicate any significant relationships between controller errors and the other factors.

PHASE II TEST RESULTS

In Phase II, the man/machine system was evaluated using both special purpose scenarios and realistic tactical scenarios supplied by the Army.

Statistical tests indicate that there were no significant differences in controller performance observed in the effectiveness measures used for the tests. Significant differences were observed in the average number of controller errors per run, however, resulting in the following rankings of the controllers:

Phase II Special Purpose Scenarios

Ranking: L C D Sk St W M

Phase II Realistic Scenarios

Ranking: L D W C M Sk St

(not ranked)

EXPERIENCE EFFECTS

C and M were the most experienced controllers while L, W and D were the least experienced. The results indicate that previous air traffic control and radar experience does not necessarily guarantee the best performance. D entered the program more than a year later than all other controllers, and had less time on the system; yet his performance was near the top in terms of controller errors. Thus, it cannot be said in this case that either experience on the system or related air traffic control experience has a great bearing on controller performance.

APTITUDE EFFECTS

Differential Aptitude Tests (DAT) were conducted in the training program. Regression analyses were performed on controller errors versus experience, education and DAT scores. The only significant fit was found for average DAT score and the realistic scenario controller errors. The results indicate that higher aptitude test scores were correlated with better performance.

LEARNING EFFECTS

The tests were run in the following order:

Phase I tests
Phase II tests (special purpose scenarios)
Phase II tests (realistic scenarios)

The data does not indicate any improvement with time attributable to learning. In fact, the data indicates performance degradation with time. The degradation in performance is probably a result of the differences in the scenarios and motivation effects, rather than any negative learning effect.

MOTIVATION EFFECTS

Though motivation effects could not be quantified, it was the opinion of the test conductor that motivation was the primary factor effecting controller performance.

During the tests, an independent subjective evaluation of controller motivation was made by the test conductor:

Phase I Subjective Analysis of Motivation

Ranking: L St M W C Sk

Phase II (Special Purpose Scenarios)

Ranking: L Sk St D C W M

Phase II (Realistic Scenarios)

Ranking: L D C W M

These ranks were compared with the rankings by average number of errors per run and the Hotelling and Pabst's Spearman Rank-Order Correlation Test* applied. Ties were broken based upon the nature of the errors. Controllers with least serious errors were given better ranks.

The rank correlation test is as follows:

Compute Spearman's rank difference correlation coefficient

$$r_s = 1 - \frac{6D}{n(n^2 - 1)}$$

where $D = \sum_{i=1}^{n} d_i^2$

d_i = difference between the rankings for controller i

n = number of controllers

The null hypothesis to be tested is H_o: Ranks are independent versus the alternate hypotheses of positive correlation.

If $D \leq D_\alpha$ where D_α is obtained from tables of the critical lower-tail values of D for Hotelling and Pabst's Spearman Rank-Order Correlation Test for a level of significance α then reject the hypothesis of independence.

The results of this test were:

Phase I: $r_s = 0.886$, $D = 4$, reject H_o at $\alpha = 0.025$
Phase II: $r_s = 0.75$, $D = 14$, reject H_o at $\alpha \cong 0.035$
Realistic: $r_s = 1.0$, $D = 0$, reject H_o at $\alpha = 0.01$

These results indicate that there is no reason to reject the hypotheses of positive correlation between motivation and average errors per run for any of the test situations.

CONCLUSIONS

The results of this evaluation indicate a significant consideration which must be made in the design and testing of any semi-automated system, where a human operator is expected to interface closely with data processing and display equipment.

To attain the level of operator performance necessary to accurately measure system performance, operator motivation must be maintained. In this evaluation, a high frequency of controller errors was attributed to deteriorating motivation, based on the judgment of the test conductor. As the frequency of controller errors rose, the evaluation of system effectiveness was impaired.

Alternatively, the system can be designed in such a way as to reduce the dependency of system effectiveness on the variability of operator performance.

---

* Bradley, James V., Distribution-Free Statistical Tests, 1968, Prentice-Hall Inc., Englewood Cliffs, N.J., pp. 91-96.