te technical note techn

Methods And Measurements In Real-Time Air Traffic Control System Simulation

Edward P. Buckley
B. Delano DeBaryshe
Norman Hitchner
Preston Kohn

April 1983

DOT/FAA/CT-83/26

Document is on file at the Technical Center Library, Atlantic City Airport, N.J. 08405

U.S. Department of Transportation
Federal Aviation Administration

Technical Center
Atlantic City Airport, N.J. 08405

NOTICE

This document is disseminated under the sponsorship of the U.S. Department of Transportation in the interest of information exchange. The United States Government assumes no liability for the contents or use thereof.

The United States Government does not endorse products or manufacturers. Trade or manufacturers' names appear herein solely because they are considered essential to the objective of this report.

Technical Report Documentation Page

1. Report No. DOT/FAA/CT - 83/26	2. Government Accession No.	3. Recipient's Catalog No.
4. Title and Subtitle METHODS AND MEASUREMENTS IN REAL-TIME AIR TRAFFIC CONTROL SYSTEM SIMULATION		5. Report Date April 1983
		6. Performing Organization Code
7. Author(s) Edward P. Buckley, B. Delano DeBaryshe, Norman Hitchner,* and Preston Kohn.*		8. Performing Organization Report No. DOT/FAA/CT-83/26
9. Performing Organization Name and Address Federal Aviation Administration Technical Center Atlantic City Airport, N.J. 08405		10. Work Unit No. (TRAIS)
		11. Contract or Grant No.
12. Sponsoring Agency Name and Address U.S. Department of Transportation Technical Center Atlantic City Airport, N.J. 08405		13. Type of Report and Period Covered Technical Note
		14. Sponsoring Agency Code

15. Supplementary Notes
*Computer Sciences Corporation

16. Abstract

The major purpose of this work was to asses dynamic simulation of air traffic control systems as a technique for evaluating such systems in a statistically sound and objective manner. A large set of customarily used measures based on the system mission of safe and expeditious movement of air traffic was collected by the computer generating the simulated traffic. The measures were collected during 1-hour simulation exercises. These measures were applied in two experiments involving controllers performing traffic control in single en route sectors, with coordination with simulated adjacent sectors.

Two experiments having many replications were conducted. In addition to studying the characteristics of the set of measurements, a second aim of the first experiment was to determine the effect on the measurements of surrounding circumstances, specifically sector geometry and traffic density. The results of this experiment led to a decision to conduct a much less complex experiment, confined to only one sector and geometry but with more repetitions of 1-hour runs under the same circumstances. This enabled an examination of the use of aggregation of data to improve reliability and the execution of a factor analysis in order to reduce and simplify the set of measures.

The factor analysis reduced the measure set to four factor scores. The study of aggregation of data led to the conclusion that four hours should be the minimum data point basis. The data from the first experiment was then utilized to cross validate the four factor structure which had been found. This cross validation was reasonably successful. The use of the four factor scores and their primary original scores, plus two auxiliary measures, is recommended in future air traffic control system dynamic simulations for system test and evaluation.

17. Key Words Simulation, Real Time, Air Traffic Control, System Test, Human Factors, Experimental Design and Analysis, Dynamic Simulation		18. Distribution Statement	
19. Security Classif. (of this report) UNCLASSIFIED	20. Security Classif. (of this page) UNCLASSIFIED	21. No. of Pages 176	22. Price

Form DOT F 1700.7 (8-72) Reproduction of completed page authorized

PREFACE

The authors gratefully acknowledge the assistance and support of the following people:

Richard Algeo, Bernard Goldberg, Arnold Grimes, Lloyd Hitchcock, Kenneth House, Josephine Pitale, Raymond Ratzlaff, Richard Rood, Lillian Senn, and Phillip Willoughby, FAA Technical Center.

Albert Beaton, Educational Testing Service

Thomas Higgins, Systems Engineering Service, FAA

Thomas Morgan, Computer Sciences Corporation

TABLE OF CONTENTS

	Page
EXECUTIVE SUMMARY	xi
INTRODUCTION	
Purpose	1
Background and Method	1
PROCEDURE	3
Experimental Procedure	3
Analytic Procedures	12
ANALYSES	19
SEM II Factor Analysis and Factor Cross Validation	19
Reliability Coefficients	42
Correlations with Observers' Ratings	49
Practice and Learning Effects in ATC Simulation Experiments	53
The Effects of Sector Geometry and Density on System Performance Measurements	63
Statistical Power of Real-Time ATC Simulation Experimentation	73
An Evaluation of the Index of Orderliness	82
Response to Post-Run Questionnaires	91
DISCUSSION	
CONCLUSIONS	
REFERENCES	116
APPENDICES	
A. List of System Effectiveness Measurements and Definitions: SEM Experiment I	
B. List of System Effectiveness Measurements and Definitions: SEM Experiment II	
C. Definitions and Usages	
D. Supplementary Tables	
E. Computations of Run Scores Based on the Index of Orderliness	
F. List of Terminal Area System Effectiveness Measures	

LIST OF ILLUSTRATIONS

Figure		Page
	Sector Maps, SEM I and SEM II	6
2	Observer Rating Forms, SEM I	7
3	Post-Run Rating Forms, SEM I	8
4	Observer Rating Forms, SEM II	9
5	Post-Run Rating Forms, SEM II	10
6	SEM I Schematic Laboratory Layout	11
7	Experimental Design, SEM I	13
8	Experimental Design, SEM II	14
9	Data Points, SEM I Experiment	16
10	Data Points, SEM II Experiment	17
11	Distribution of Factor Scores for SEM I and SEM II Experimental Conditions (4 Sheets)	37
12	Plot of Course of Major Measures Over Time	55
13	Plot of Day Means	61
14	The Four Basic Designs	75
15	Graph of Power	79
16	Power Table Structure	81
17	Controller Profiles in Standard Score Form	107

LIST OF TABLES

Table		Page
1	Traffic Sample Characteristics	4
2	Reliability Coefficients of Scores Based on Full Factors, Smooth Factors, and Very Smooth Factors	22
3	Linear Combination Weighting and Equal Weighting Within Each Factor	23
4	Comparison of Multiple Correlation with Judges' Rating Provided by Original Seventeen Measures, Full Factor Scores and Very Smooth Factor Scores	24
5	Cross-Validation over Days	25
6	Percent of Variance Consumed by Factors	28
7	Correlations Between SEM II Factor Scores and SEM I Sector-Density Cell-Based Factor Scores	30
8	SEM I Cell Based Factor Scores and SEM II Factor Scores in Relation to SEM I Judges' Ratings	32
9	Day Two versus Day Three Reliability of Measures Within a Factor	34
10	Correlations of Measures Within a Factor With the Factor	35
11	Reliability Coefficients	43
12	Standard Errors of Measurement	45
13	Inter-Observer Agreement	46
14	Rating Reliability	47
15	Correlations Between Measures and Ratings	50

LIST OF TABLES

Table		Page
16	Multiple Correlation (R) of Factors and Leading Measures on Ratings, SEM I	51
17	Multiple Correlation (R) of Factors and Leading Measures on Ratings, SEM II	52
18	Analysis of Variance Table: Hours	56
19	Orthogonal Analysis: Successive Simulation Hours (3 sheets)	57
20	Percent of Variance Due to Hours and Persons	60
21	Analysis of Variance Table: Days	62
22	Percent of Variance Due to Days and Persons	64
23	Day Means	65
24	Analysis of Variance Table: Sector and Density	67
25	Mean Values in Sector-Density Combinations	69
26	The Percentage of Variance Due to Sector and Density	70
27	Cross-Condition Correlations: Across Geometry at a Given Density	71
28	Cross-Condition Correlations: Across Density at a Given Geometry	72
29	Power Table Example	78
30	Comparative Statistical Power of the Four Factor Scores	80
31	Correlation Between Index of Orderliness Measures and Factor Scores and Confliction Measures (2 sheets)	83
32	Correlations Among the Three Index of Orderliness Measures	86

LIST OF TABLES

Table		
33	Run-Run Reliabilities for Index of Orderliness Measures, Factor Scores and Confliction Measures	87
34	Correlations With Ratings for Index of Orderliness, Factor Scores and Confliction Measures	88
35	Multiple Correlation To Ratings With and Without Index of Orderliness Measures	89
36	Correlations (r) Between Two Averaged Factor Scores and Index of Orderliness Measures	90
37	Mean Values of Questionnaire Item Responses- SEM I	92
38	Mean Values of Questionnaire Item Responses- SEM II	93
39	Correlations Between Questionnaire Items and other Data Items- SEM I (6 sheets)	95
40	Correlations Between Questionnaire Items and other Data Items- SEM II (3 sheets)	

EXECUTIVE SUMMARY

Proposed changes to air traffic control systems are frequently evaluated through the use of real-time system simulation. Comparative evaluation of "new" and "old" systems is often part of a cost-benefit study of possible increased productivity.

Such studies frequently yield ambiguous conclusions. In fact, the inconclusiveness of such evaluations is almost legendary, and the dissatisfaction with the results by those who need them is sometimes severe. Emotions may run high on occasions when expensively developed systems cannot be "statistically proven" to be "better than" the current (old) system, particularly when appearances and "feel" give the opposite impression.

There have been two schools of thought among those who have been close to such simulations and concerned with rendering of opinions on new or modified air traffic control systems. This issue concerns the place of the statistical treatment of the measurement data which can be collected during ATC system simulation experiments, and its utility, for making clear system evaluation conclusions.

One group favors the use of statistical inference methods, including the statement of hypotheses in advance of the experiment, and the use of statistical tests and indices to determine whether the differences found are "statistically significant". They deride those who contend that "just trying out a system" is enough to form a reasonable opinion. On the other hand, those who deride statistical methods point out the frequency of failure to find results and differences which statistical tests will allow to be called dependable enough ("significant") to rely upon. They say this sometimes occurs even when there has been large and careful experimentation and data collection, and in cases when the superiority of the new system is "obvious to the casual observer."

One factor in the debate which is sometimes ignored is the fact that every real-time simulation is a human factors experiment. In real-time simulation the results are not only a function of the systems involved, but also of the people (quite variable within and between themselves) who are performing as controllers in the simulation exercises, and of the traffic sample input given to the system to handle. It is apparent that real-time simulation exercises may be a weak tool since every exercise in which a controller or control team participates is different, even with identical traffic samples, once the first few control decisions have been made.

It could be the case that the data from dynamic simulation cannot sensibly be treated using statistical techniques such as analysis of variance. Perhaps the data are so variable that statistically repeatable conclusions are not possible without unacceptably large numbers of controllers and hours of simulation; and that to seek for them is puristic and fruitless. If this is so, we will have to be content with "gut feeling" observations of the new system at work. This approach, however, is also clearly open to criticism, particularly when it matters so much whether a newly developed costly system is successful.

In order to help resolve this dilemma, it was decided to collect empirical data through specific experiments designed to bear on the statistical and measurement issues involved in the planning and interpretation of the results of real-time simulation experiments on air traffic control systems. These experiments were named the System Effectiveness Measurement (SEM) experiments.

The FAA Technical Center's Air Traffic Control Simulation Facility (ATCSF) was utilized for the experimental work. The ATCSF is a digital computer-based air traffic control simulator in which simulated aircraft are maneuvered and corresponding radar data are presented to air traffic controllers, who are in simulated air-ground communication with the aircraft. One simulator pilot can represent up to five aircraft of various types by making digital control inputs and appropriate voice responses to the traffic controller or controllers involved.

The computer which was generating the traffic was also programmed to simultaneously collect the measurement data. A set of objective measures was assembled to represent measures of air traffic control system mission accomplishment customarily or frequently used by various air traffic control system simulation experimenters in the history of such work. These measures were collected by the computer during the control exercises. In addition, in the studies reported here, independent observers, who were qualified controllers, subjectively rated the controller performance and system performance during the same exercise session which was being objectively scored by the computer.

Two experimental evaluations were executed, and the data analyses and results are presented in this report. Both experiments worked with samples of control "teams" tested repeatedly under various circumstances, such as different sectors and traffic densities, while keeping the hardware and software system being used identical. For economy, data collected upon only single controller "teams" were utilized, although field en route sector teams generally consist of two or more people. However, various aspects of the experimental procedures were carefully designed to maintain a realistic atmosphere and situation, despite the single controller "team" data collection process. In particular, aspects of coordination with adjacent sectors were simulated by laboratory staff controllers and most of the work that is normally done by assistant controllers was accomplished in advance of each pre-designed exercise by laboratory staff personnel. But in connection with the matter of team size, as with all of system simulation, it should be remembered that only relative, not absolute, measurement can be attained in any case.

The first study, "SEM I," was aimed at examining the effects on the several system performance measurements of changes in the surrounding circumstances of sector geometry and traffic density. The second experiment, "SEM II," was aimed at specifying the effects of accumulating more data at a given data point, thus improving the dependability of the data, and at determining the impact of learning and practice in this type of measurement situation.

The effects on the system performance measurements of two extremely different en route sector geometries and three traffic levels ranging from very light to very heavy were analyzed using the data from the SEM I experiment. Using the data from the SEM II experiment, analyses were made of the repeatability and dependability of the measurements, and of the correlations among the customarily used measurements. It was concluded that a far smaller set of measures could be used without major loss in measurement adequacy and with a corresponding increase in clear interpretation of results. These new measure types were then examined to see if they could also be used to summarize the SEM I data. It was found that this smaller set of measures derived from the SEM II study provided a statistically adequate equivalent set of measures for all six of the SEM I sector geometry and traffic density combinations.

Tables for planning were derived from the data from both experiments to indicate how many subjects and runs must be used in air traffic control simulation experiments of this type to achieve statistically based conclusions of a given probability. While these tables are expressed for what is considered to be a range of sector geometries and traffic densities, they should be applied, strictly speaking, only to performance measurement during single-controller, single-sector exercises. Additional research would be required to extend the results to multi-sector, multi-person team experiments, and to terminal area control system simulation experiments. However, these tables should prove far superior to intuition for estimating resource requirements even when extrapolated to those situations. Because increased variability is possible among multi-person teams, estimates based on these tables may underestimate the resources required.

The results show that those who criticize as infeasible and impractical the use of statistical inference techniques in this field have some grounds for their criticisms, because there is much variability in the measures of air traffic performance in dynamic exercises and comparatively large amounts of data are needed for firm statistical conclusions. On the other hand, the tables resulting from this research indicate the requirements which must <u>and can</u> be met, when the occasion justifies it, to facilitate clear-cut conclusions for important experimental air traffic control system evaluations. The results of the studies are discussed in this volume and the tables will appear in a later volume.

The SEM work, then, was an approach to empirically determining (and compensating for) the strengths and weaknesses of ATC simulation experimentation as usually conducted in the past. This knowledge can provide guidance for future system evaluation experimenters both at the FAA Technical Center and at other similar laboratories. Although the focus here was on developing data which might enable more effective system test and evaluation, the work also provided a uniform basis for future experimental simulation studies of various kinds for the air traffic control system, and could also provide a basis for a controller performance criterion technique to be used for the validation of aptitude tests and other selection and training techniques.

INTRODUCTION

PURPOSE.

The purpose of this work was to determine the quality of measurement of system performance and statistical treatment that is possible and appropriate in dynamic simulation of air traffic control systems.

BACKGROUND AND METHOD OF APPROACH.

Real-time simulation of air traffic control systems is quite frequently used to evaluate new system concepts. In such studies, simulated aircraft to be controlled are fed into a system consisting of equipment, computers, and air trafffic controllers who are to use both the current and the new air traffic control systems to provide a comparative evaluation of the two systems. Thus, such system evaluations are, intrinsically, human factors experiments and the methods used should give appropriate attention to the extent and nature of individual differences and human variability. Traditionally, the design of such experiments has suffered from the lack of certain basic information which the current effort attempts to supply in order to aid and improve future system evaluators and their evaluations.

A two-experiment evaluation series provided interrelated information. In the first experiment, the aim was to discover the sensitivity of currently used system performance measurement to differing traffic levels and sector geometries. This experiment collected data on two 1-hour runs for each of 31 subjects under each of 6 sector geometry-traffic density combinations (cells). Initial analyses, involving correlations between the two runs in each cell, indicated very low correlations between the replicates. It was decided that before going further it would be best to conduct a much less complex experiment with fewer combinations of conditions involved, in order to discover the difficulty. Thus, an experiment utilizing only one of the six combinations of conditions of sector and geometry, but with several replicate runs under the same conditions, was conducted. This second experiment was aimed at studying the effects of replication and at providing a sufficient amount of data collected under the same conditions to enable a factor analysis to be done for the purpose of consolidating the measurements into a smaller meaningful set. This second experiment involved 12 1-hour runs in the same sector with the same traffic level for each of 39 controllers. The two experiments will be referred to as SEM (System Effectiveness Measurement) I and SEM II.

In both experiments, the computer which was generating the aircraft to be controlled was also collecting a set of objective measurements based on the aircraft movements traditionally assumed to be related to the success of the air traffic control being exercised. In addition to the objective measurements of performance, field-qualified journeyman air traffic control specialists provided ratings of the effectivness of the control for each

session or "run." One of the analyses later done was the examination of the relationship between these two kinds of evaluation of the same session of traffic control.

For the purpose of examining the system performance measures, three assumptions were implemented in the experiments: (1) the measures relevant to the output of an ensemble of sectors can be studied in a one-sector mini-system, (2) it is necessary for measurement purposes to use more traffic than one person would usually be expected to control in the real world, and (3) for the purpose of simply studying the measures, the staffing can be reduced and the traffic increased as long as the measures are treated as relative and not absolute.

An overview of the discussions to follow might not be amiss at this point. After explaining the experimental procedures for both experiments, the factor analysis of the SEM II data will be described. In general terms, it was found that four scores based on the factor analysis could be considered an adequate set of measures to use. It was deemed important to see if the same factors could adequately serve as the measures in other sectors and traffic levels. The SEM I data were then called back into service. The SEM I data were re-scored using the SEM II measures and examined for the presence of the same factors. It was concluded that the same factor scores could express the results of the first experiment. This made possible the analysis of sector and density effects and the effects of practice and learning in air traffic control simulation exercises using the more convenient and understandable smaller set of measures.

PROCEDURE

EXPERIMENTAL PROCEDURE.

The simulator used to conduct these experiments was the Air Traffic Control Simulation Facility (ATCSF) at the FAA Technical Center, Atlantic City, New Jersey. This is a digital computer-based simulation facility which has been described in great technical detail elsewhere (reference 1). In general terms, however, the major elements involved are the Controller Laboratory, which contains 8 air traffic control display consoles of a generic type, and the Simulator Operator Laboratory, which contains consoles that control the flight of the simulated aircraft which appear on the controller displays. A simulated air-ground communications link joins the controllers and the simulator operator "pilot." The aircraft under control are displayed to the controller with alphanumeric tags containing aircraft identity, altitude, speed, and other information. The laboratory can be configured to represent terminal or en route air traffic control. The simulation laboratory is in a constant state of improvement to increase the level of fidelity in the representation of field air traffic control, but this representation does lag behind the field. In the experiments to be discussed here, the representations of the en route system were not exact; the generic consoles were used and the conflict alert feature of the system which at the time was just beginning to enter field facilities was not available for representation.

For the SEM I experiment, two sectors were selected from the sectors at the en route air traffic control center at Leesburg, Virginia. Their designations at the time were sectors 14 and 16. They were chosen to be quite different, about as different as might be readily found. Based on examination of the sectors' traffic at the time, samples of flights were composed and programmed to fly in the simulator. The traffic samples were designed to build up the traffic for 8 minutes, and then scheduled to run for an hour with approximately the same level of traffic density, as measured by the number of targets which would usually be simultaneously present on the controller's radar scope. Three 1-hour (after buildup) samples of the traffic were composed for each of the two sectors: a low, medium and high traffic density level. As said earlier, the average level of these samples was higher than would be expected to be handled by a controller in live operations. The variable of traffic density was set so that the levels of traffic density would be approximately equal for both sectors, thus the experimental factors of sector and density would not be connected, but orthogonal (independent). The major parameters considered were the number of completable flights for the hour and the number of planned (scheduled) simultaneous aircraft present in the typical (modal) minute. As may be seen in table 1, these descriptors increase at about the same rate for both sectors. Pre-trials of the density levels indicated that while they were difficult, and would in fact be too difficult for some controllers, they were not excessively so for use in simulation exercises.

The SEM II experiment used one of the same two sectors used in the previous experiment, sector 14, which was called geometry 1. Four fresh traffic samples were generated which were generally comparable to the middle density

TABLE 1

TRAFFIC SAMPLE CHARACTERISTICS

SEM I

	Geometry 1 (Sector 14)			Geometry 2 (Sector 16)		
Density	1	2	3	1	2	3
No. Completable Flights (60 min.)	27	38	50	25	42	50
No. Arrivals Handled	17	25	30	22	36	44
No. Departures Handled	12	16	26	4	6	6
No. A/C Planned to be Under Simultaneous Control (modal value)	5	7	8	5	6	8

SEM II

Sample	A	B	C	S
No. Completable Flights (60 min.)	40	40	40	40
No. Arrivals Handled	30	30	30	30
No. Departures Handled	17	17	17	17
No. A/C Planned To Be Under Simultaneous Control (modal value)	8	8	8	8

Note: Numbers given are the planned values, i.e., as input traffic samples Minor fluctuations occurred even in the planned samples from minute to minute.

previously used. They were comparable to each other since each was constructed by slightly shifting the start times and changing the identities of the aircraft contained in reference or "seed" samples. The traffic samples were designed from the "seed" sample by means of a computer program in such a manner that the number of aircraft scheduled to be present on the scope would be the same throughout the hour of the problem. Figure 1 shows the sector maps for the two sectors. Table 1 gives the characteristics of the traffic samples for both experiments.

The computer which generated the traffic samples and presented the simulated radar signals corresponding to the aircraft postions also collected information about what was done with the aircraft by the control system. This same computer was capable of collecting data such as the position of the aircraft in the system at any given time and the clearances given by the controllers which were entered into the computer by the simulator pilots. These data were collected and reduced to the form of "run" scores, which represented sums or means of various events and types of aircraft movements which occurred in the course of the time period over which the simulation exercise ran. The list of the measures selected for the SEM I experiment appears in detail in appendix A. The list and definitions were modified in the hope of improving the measurement reliability before executing SEM II. This revised list appears in appendix B.

Some subjective measures were also taken during the two evaluations. In each experiment, additional controllers, designated as "judges," rated the performance during each 1-hour run (session). On one scale, the judges rated the technique or performance shown by the radar controller and on another scale, the overall effectiveness of the man/machine air traffic control system in handling the traffic safely and expeditiously. Also, at the end of each 1-hour run, the subject filled out a short questionnaire, the major purpose of which was to discover any equipment or procedural difficulties. The forms were changed slightly between experiments. The rating forms used in SEM I and SEM II appear in figures 2 and 3 (SEM I) and figures 4 and 5 (SEM II), respectively.

The simulation laboratory was arranged in a similar manner for both experiments. The usual way of using the simulation laboratory is with a very large team cooperating to control an entire terminal area or several cooperating en route sectors. For the purpose at hand, however, it was decided that information could be gained on the relevant topics in a much more economical way by running four separate data-independent sessions simultaneously, thus increasing the independently analyzable data by a factor of four. The essential aspects of inter-sector coordination were retained, however, by providing support controllers to represent adjacent sectors requiring coordination. In addition, the duties normally performed by assistant controllers were reduced as much as possible, as, for example, by providing preprinted flight strips. Figure 6 gives a sketch of the laboratory configuration for SEM I. The same configuration was used in SEM II with the exception that there the sector 14 map was used in all four subject stations.

In the SEM I experiment, the support controllers actively participated in lining up aircraft for handoff to the subject sector and in holding aircraft prior to handoff upon request from the subject controller. After the SEM I

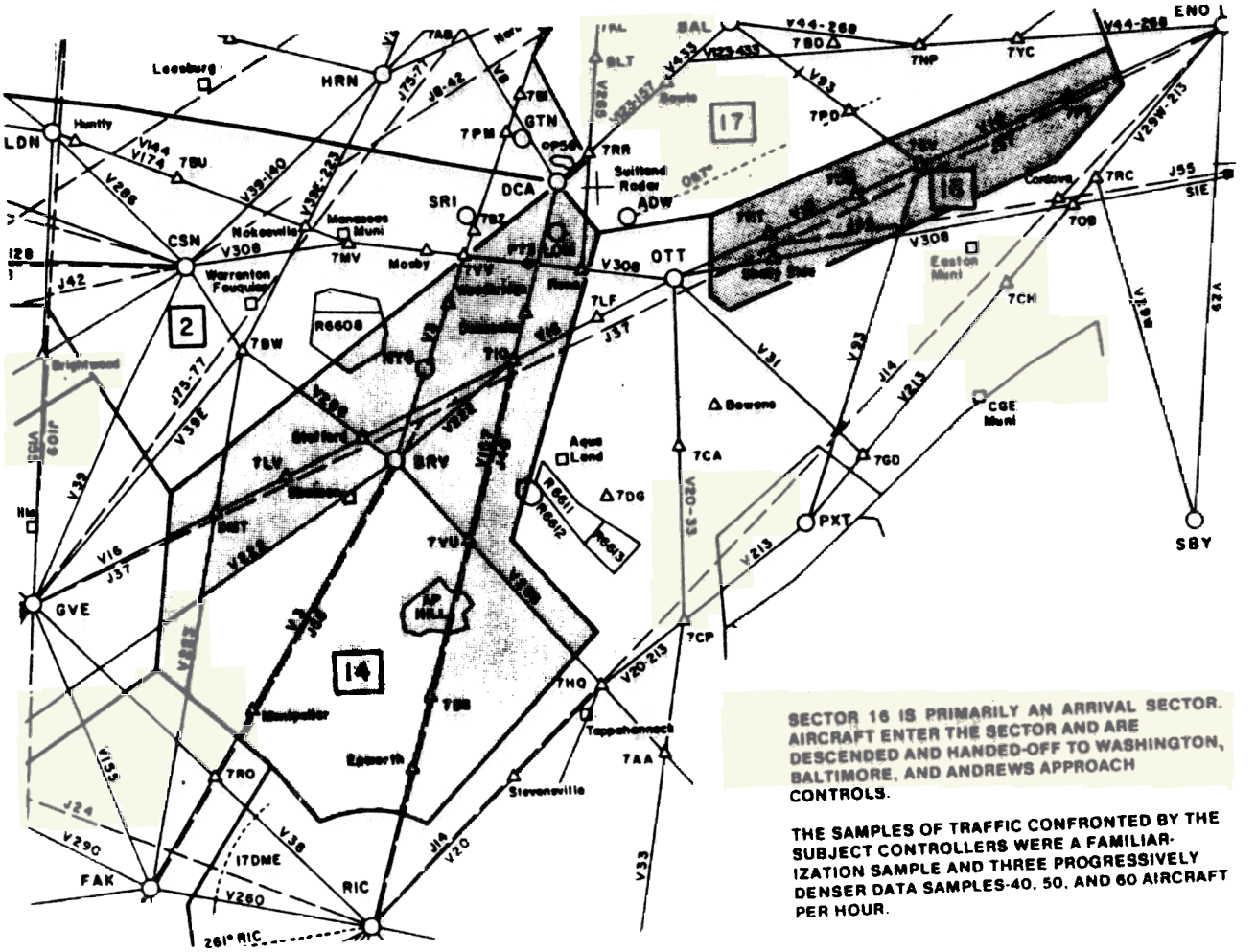

SECTOR 16 IS PRIMARILY AN ARRIVAL SECTOR. AIRCRAFT ENTER THE SECTOR AND ARE DESCENDED AND HANDED-OFF TO WASHINGTON, BALTIMORE, AND ANDREWS APPROACH CONTROLS.
THE SAMPLES OF TRAFFIC CONFRONTED BY THE SUBJECT CONTROLLERS WERE A FAMILIARIZATION SAMPLE AND THREE PROGRESSIVELY DENSER DATA SAMPLES-40, 50, AND 60 AIRCRAFT PER HOUR.
ENO
BAL
LDN
CSN
DCA
SRI
OTT
BRV
GVE
FAK
RIC
PXT
SBY
HRN
GTN
ADW
Leesburg
Huntly
Warrenton Fauquier
Manassas Muni
Easton Muni
CGE Muni
Aqua Land
Tappahannock
Stevensville
Suitland Radar
17
16
2
14
R6608
261° RIC

FIGURE 1.

RUN 1 (S-1) MONITOR # 3 PARTICIPANT # 2 SECTOR 16

SEM
POST RUN
MONITORING FORM

CONSIDERING THE LEVEL OF TRAFFIC INVOLVED, AND FROM THE VIEWPOINT OF THE PILOTS:

VERY POOR	POOR	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	GOOD	EXCELLENT
1	2	3	4	5	6	7

THE TRAFFIC RECEIVED VERY POOR HANDLING AT THE HANDS OF THIS SYSTEM: THERE WERE SEVERAL LAPSES IN SAFETY, SPEED AND SMOOTHNESS.

THE TRAFFIC RECEIVED GOOD HANDLING, ARRIVING WITH FAIR SAFETY, SPEED AND SMOOTHNESS.

THE TRAFFIC RECEIVED THE BEST HANDLING IT COULD POSSIBLY HAVE ASKED FOR, USING ANY ATC SYSTEM. ALL AIRCRAFT WERE ABLE TO SMOOTHLY FOLLOW THEIR IDEAL PATHS AND SPEEDS.

	VERY POOR	POOR	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	GOOD	EXCELLENT
0–10	☐	☐	☐	☒	☐	☐	☐
11–20	☐	☐	☐	☒	☐	☐	☐
21–30	☐	☐	☐	☒	☐	☐	☐
31–40	☐	☐	☐	☒	☐	☐	☐
41–50	☐	☐	☐	☒	☐	☐	☐
51–60	☐	☐	☐	☒	☐	☐	☐

CONSIDERING THE LEVEL OF TRAFFIC INVOLVED, WITH RESPECT TO ALL OF THE JOURNEYMEN CONTROLLERS I HAVE KNOWN, AND CONSIDERING THE PERFORMANCE OBSERVED IN THIS RUN, I FEEL THE CONTROLLER WOULD RANK IN THE:

VERY POOR	1%	X	☐
POOR	6%	XXXXXX	☐
BELOW AVERAGE	24%	XXXXXXXXXXXXXXXXXXXXXXXX	☐
	38%	XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX	☒
ABOVE AVERAGE	24%	XXXXXXXXXXXXXXXXXXXXXXXX	☐
GOOD	6%	XXXXXX	☐
EXCELLENT			☐

CONSIDERING THE LEVEL OF TRAFFIC INVOLVED, CONSIDER THE CONTROL TECHNIQUE (I. e., TAP) OF THE INDIVIDUAL CONTROLLER YOU HAVE OBSERVED DURING THIS RUN:

	VERY POOR	POOR	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	GOOD	EXCELLENT
SEPARATION	☐	☐	☐	☒	☐	☐	☐
AWARENESS MAINTENANCE	☐	☐	☐	☒	☐	☐	☐
CONTROL JUDGEMENT	☐	☐	☒	☐	☐	☐	☐
CONTROL ACTION PLANNING	☐	☐	☐	☒	☐	☐	☐
SECTOR OVERLOAD PREVENTION	☐	☐	☐	☒	☐	☐	☐

FIGURE 2. OBSERVER RATING FORMS, SEM I

RUN _____

TOR _____

ICIPANT # _____

SEM
POST RUN
CONTROLLER
OPINION SURVEY

With respect to this session (only) and considering the traffic density level involved

A. Rate your own technique this run:

VERY POOR	POOR	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	GOOD	EXCELLENT
☐	☐	☐	☐	☐	☐	☐

B. Rate the probable "feelings" of the imaginary pilots of the represented aircraft as to the smoothness of the "system's" control during this run:

VERY POOR	POOR	BELOW AVERAGE	AVERAGE	ABOVE AVERAGE	GOOD	EXCELLENT
☐	☐	☐	☐	☐	☐	☐

C. Comparing this run to a "peak" hour at your "home sector", when you are the R controller with normal support, rate this simulation run:

MUCH EASIER	EASIER	EQUAL DIFFICULTY	HARDER	MUCH HARDER
☐	☐	☐	☐	☐

D. How realistic do you feel the simulation was?

VERY POOR	POOR	ADEQUATE	VERY GOOD	EXCELLENT
☐	☐	☐	☐	☐

E. Please make note here of any technical difficulties in the equipment, etc. we should be told about:

_____

_____

_____

_____

_____

FIGURE 3. POST-RUN RATING FORMS, SEM I

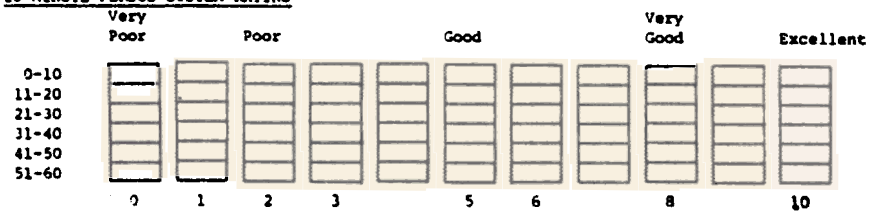
SYSTEM EFFECTIVENESS MEASUREMENT
MONITORING FORM

RUN # _____ SECTOR # _____ PARTICIPANT # _____ JUDGE # _____ DATE ___/___/___

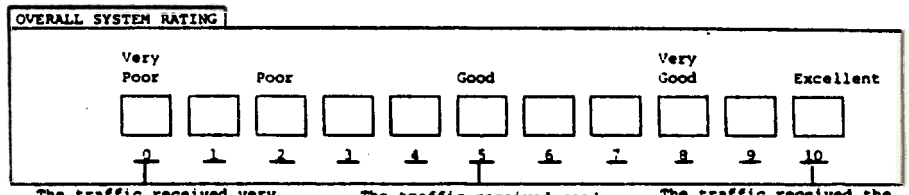
SYSTEM RATING

IN THIS RATING, FOCUS ON THE PRODUCT, ATC SYSTEM EFFECTIVENESS, NOT THE PROCESS.

10-MINUTE PERIOD SYSTEM RATING



OVERALL SYSTEM RATING



The traffic received very poor handling at the hands of this system; there were several lapses in safety, speed, and smoothness.

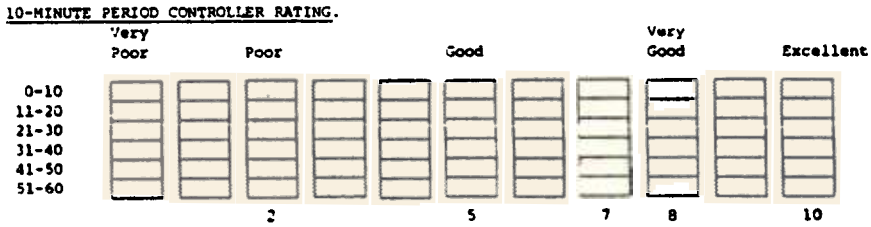
The traffic received good handling, arriving with fair safety, speed, and smoothness.

The traffic received the best handling it could possibly have asked for, using any ATC system. All aircraft were able to smoothly follow their ideal paths and speeds.

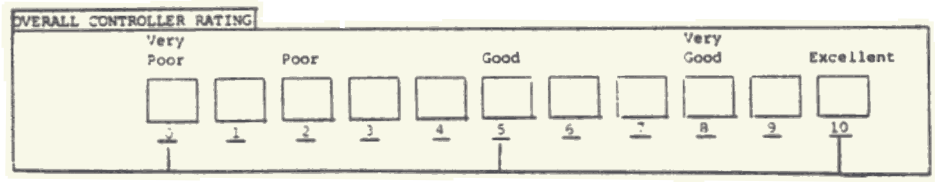
CONTROLLER RATING

IN THIS RATING, FOCUS ON THE PROCESS, CONTROLLER JUDGMENT AND TECHNIQUE, NOT THE PRODUCT

10-MINUTE PERIOD CONTROLLER RATING.



OVERALL CONTROLLER RATING



In this run, this controller performed at about the level I would have expected to see if the worst controller I have ever known were making the run.

This controller seemed about average in this run.

This controller was about as skillful and clever in handling this traffic during this run as the best controller I have ever known would have been.

FIGURE 4. OBSERVER RATING FORMS, SEM II

SYSTEM EFFECTIVENESS MEASUREMENT

PARTICIPANT SURVEY

RUN # _____ SECTOR # _____ PARTICIPANT # _____ DATE _____

Rate your technique and skill on the particular run in terms of your usual R man level of ability. Consider only your own functioning at home as an R man as a standard (ignoring other team members there and here):

I wasn't anywhere near my usual level	I could have done this run a lot better	About average for me	Very good for me	Excellent for me
☐	☐	☐	☐	☐

If you were the typical pilot of one of the aircraft you have just controlled in this run, what would be your feeling about the ATC system? For example, were many aircraft delayed or given very many vectors? Did you have a few pilots who might have had anxious moments:

This run very bumpy, exciting, and inconvenient for almost all of the pilots	Fair unsatisfactory for the majority of pilots	Neither good nor bad	Moderately safe and swift for the majority of pilots	This run gave almost all pilots a very safe and swift ride
☐	☐	☐	☐	☐

How does the level of traffic you encourtered here compare with what you usually encounter in your home sectors? Just consider the traffic as such. For this question ignore the fact that you have help there--just consider the traffic. Consider both amount and complexity of traffic here and at home.

- This traffic problem here is much heavier and more complex than what my team faces at home in an average hour. ☐
- A good bit worse here ☐
- About the same ☐
- Home is a good bit worse ☐
- Home is a lot worse ☐

How realistic do you feel the simulation technique was:

Very Poor	Poor	Adequate	Very Good	Excellent
☐	☐	☐	☐	☐

Please make note here and on the back, if needed, of any technical difficulties in the equipment or other things we should be told about

_____

_____

_____

_____

FIGURE 5. POST-RUN RATING FORMS, SEM II

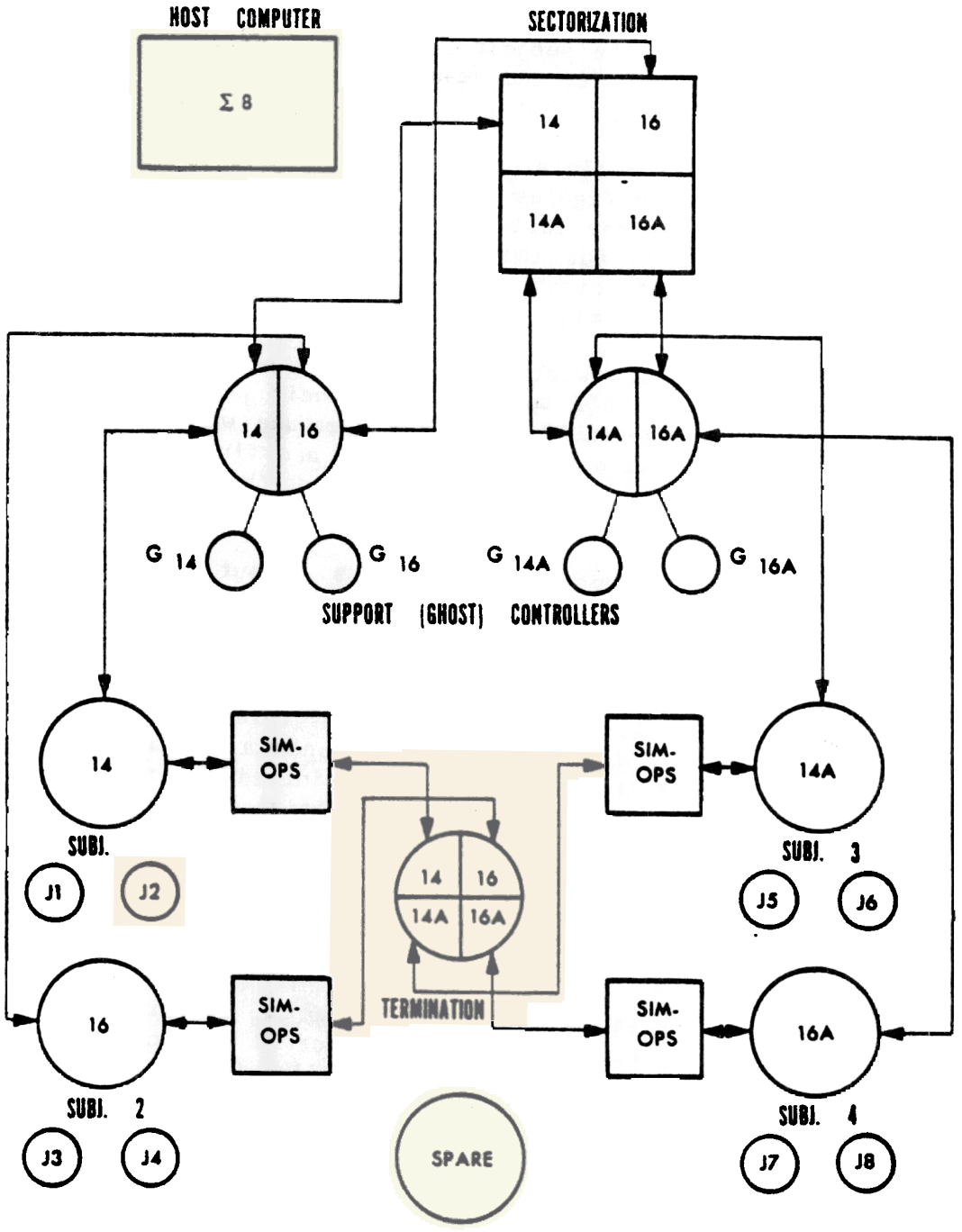


FIGURE 6. SEM I SCHEMATIC LABORATORY LAYOUT

experiment had been completed, it was suspected that this may have led to a too active participation in handling the traffic by the staff support controllers and this was changed to be a more automatic process performed by the computer. In SEM II, if the subject controller wished to have incoming traffic held, the computer held it, and resumed feeding entering traffic upon request.

The experimental designs (for definition of this term, see appendix C) for SEM I and SEM II are presented in Figures 7 and 8. Previous work (reference 2) indicated that two replicates per cell were adequate, and so that number was used in the SEM I experiment; but the results of SEM I indicated that two replicates were probably insufficient. The determination of the effect of the number of replicates was made a major aim of SEM II.

In SEM I, half the controllers worked all of their problems on one of the sectors first, and the other half worked all of their problems on the other sector first. It was considered best to have everyone work with the lowest traffic density first, then with the moderate one and finally with the heavy density. This was done by each controller, and then repeated for the replication.

In SEM II, there were in effect 12 replications. Four slightly different traffic samples were composed in an attempt to disguise the traffic, or to make it appear at least slightly different. The manner in which this was done was to designate one set of aircraft as the "seed" sample and then randomly shift the start times of the same aircraft slightly to make three other samplings of the same aircraft; aircraft call signs were also changed for the same reason. The "seed" sample was administered once a day and the order of administration of the other three samples was latinized in order to minimize and balance whatever effects the slight sample modifications might have.

The subjects in both experiments were all qualified en route journeyman controllers who came from four different FAA en route centers in four different regions. They were volunteers who had been chosen at random after volunteering. Four came at a time and stayed for 2 weeks; this was done for both experiments. Logistic and equipment problems affected the number of subjects having fairly complete data in each of the two experimental sessions; data were obtained in a rather complete manner for 31 subjects in SEM I and for 39 subjects in SEM II. The SEM I data collection was in the period January to June 1979, and the SEM II data collection was in the period January to June 1980.

ANALYTIC PROCEDURES

Standard statistical analysis techniques were implemented using the BMDP statistical software package (reference 3).

Considerable amounts of sheer data handling were involved: this is why the authors feel strongly that a reduction of the number of measures needing analysis is an important improvement.

In the SEM I evaluation, there were several equipment failures in the midst of runs, but usually at the latter part of the runs. This made for several short runs and where a run had been completely lost, or lost early in its time, it

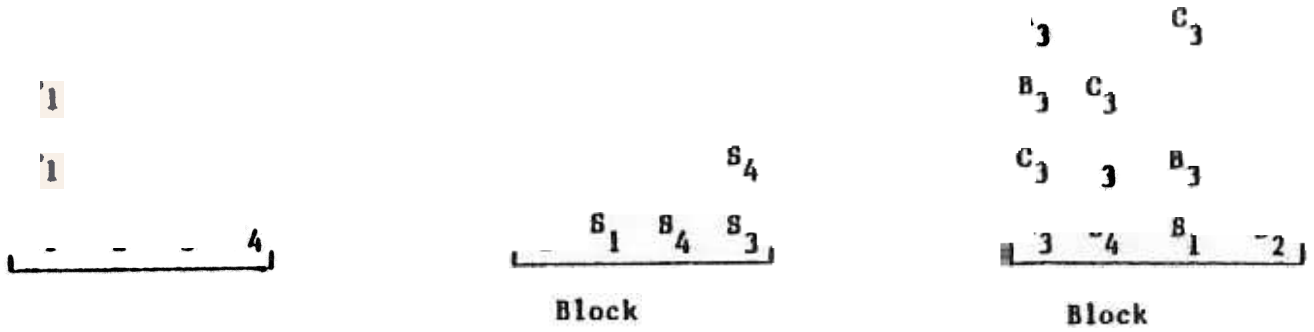
SECTOR

DENSITY	1			2		
		R_1	R_2		R_1	R_2
1	S 1 . . N			S 1 . . N		
		R_1	R_2		R_1	R_2
2	S 1 . . N			S 1 . . N		
		R_1	R_2		R_1	R_2
3	S . . N			S . . N		

FIGURE 7. EXPERIMENTAL DESIGN, SEM I

jme od

ubj 7



th aamp

i amp ( ) nd aamp ( ) prob pt ch ng i

GURE EXPE MENTAL DE GN

led to missing data in some cells. Because of this, in handling the SEM I data, the device was used of shortening the available full 60-minute runs to 50 minutes, thus increasing the number of homogeneous length runs available for analysis. These data were used as such for most analyses involving the SEM I data. In the construction of the power tables, the 50-minute runs were prorated up to 60 minutes as needed for the 1-hour unit tables.

In some runs using the very severe highest traffic density level used in the SEM I experiment, there were occasions when controllers exercised an option covered in their pre-test instructions and indicated that they had "lost the picture" which means, in controller slang, that the traffic situation had become, at that point in that particular run, too heavy for them to continue to control. There were only a comparatively few instances in which this happened, 13 out of a possible 372 (31 subjects, 12 scheduled runs each). In the event that this happened, the judges followed their previous instructions to assist the controller until the problem was over. The intention was to regard these runs as missing data runs, together with those shortened by equipment difficulties. However, through a data handling error in the analysis stage, these 13 runs remained in the data base, and by the time this situation was discovered, removal and correction was economically prohibitive.

Fewer such difficulties occured in SEM II because of improved equipment and procedures, and the lower density of traffic used in these exercises. In addition, no permission was given to the subjects to declare loss of the picture, although it probably would not have been needed. Figures 9 and 10 show where these difficulties occurred in each experiment in terms of the original experimental designs.

Various methods for handling the missing data resulting from equipment problems were explored in great depth, but none seemed any more effective than the use of the replicate run or runs to make up for the loss by allowing the available replicate or replicates to stand for the cell, either by averaging them or, in the case of only one replicate being available as in SEM I, letting the replicate stand for the cell.

There was a sequential order in the process of analysis which will be reflected in the order in which the material is presented in subsequent parts of this report. As has been mentioned, almost immediately after the execution of the SEM I experiment, it was decided that more concentrated information was needed using fewer experimental variations. Therefore, an intensive experiment (SEM II) was designed and executed. The SEM II experiment was first analyzed using factor analysis in a search for more succinct measurements. The experiment had 12, 1-hour runs per subject and, from these, 3 sets of 4 hours of data each were assembled and labeled "days," since 4 runs were usually done in a day. Each day of data was submitted to a factor analysis resulting in three sets of factor scores. The factor scores were standardized in terms of the distribution for each day separately. Some slight truncation to integer numbers was used in this scaling. Many analyses were done using this data, leading to a single set of four factor scores usable over the entire experiment (SEM II).

		Sector 14 (Geom. 1)						Sector 16 (Geom. 2)					
		Replicate 1			Replicate 2			Replicate 2			Replicate 2		
Subject	Density	1	2	3	1	2	3	1	2	3	1	2	3
1			.	.	.	.	.	.	.				
2			.	.	.	.	.	.	.	.	.	.	
3			.	.	.	.	.	.	.	.	.	.	
4			.	.	.	.	.	.	.	.	.	.	
5		.	.	T	.	.	.	.	.	.	.	.	
6		.	.	.	.	.	.	.	.	.	.	.	
7		.	.	.	.	.	.	.	.	.	.	.	
8		.	T	.	.	.	.	.	.	.	.	.	
9		.	.	.	.	.	.	S	.	U	.	.	
10		S	.	U	.	.	.	.	.	.	.	.	
11		.	.	.	.	.	.	.	.	.	.	.	.
12		S	.	U	.	.	.	.	.	.	.	.	T
13		.	S	T	.	.	T	.	.	.	.	S	
14		.	.	.	.	S	.	.	S	.	.	.	
15		.	S	T	.	T	.	.	.	.	.	S	T
16		.	.	T	.	S	.	.	S	.	.	.	
17		.	.	.	.	.	.	.	.	.	.		
18		.	.	.	.	.	.	.	.	.	.		
19		.	.	T	.	.	.	.	.	.	.		
20		A	A	A	A	A	A	A	A	A	A	A	A
21		.	.	.	.	.	.	.	.	.	.	.	.
22		.	.	.	.	.	.	.	.	.	.	.	.
23		.	.	.	.	.	.	.	.	.	.	.	.
24		.	.	.	.	.	.	.	.	.	.	.	.
25		.	.	.	.	.	.	.	.	.	U	.	.
26		.	.	.	U	.	.	.	.	.	.	.	.
27		.	.	.	.	.	.	.	.	.	U	.	.
28		.	.	.	U	.	.	.	T	.	.	.	.
29		S	.	.	.	.	.	.	.	.	.	.	.
30		.	.	T	.	.	.	S	.	.	.	.	.
31		S	.	T	.	.	T	.	.	.	.	.	.
32		.	.	T	.	.	.	S	.	T			

Key: S = short run, data deleted; U = no run; A = subject not present: T subject acknowledged loss of control prior to 50 minutes of valid data; . at least 50 minutes of valid data present

FIGURE 9. DATA POINTS, SEM I EXPERIMENT

Subject No.	Slot (Hr.) No. 1	2	3	4	5	6	7	8	9	10	11	12
	.	S	.		.				.	.		
2	.	S	.		.				.	.		
3	.	S	.		.				.	.		
4	.	S	.		.				.	.		
5	.	.	.		.				.	.		
6	.	.	.		.				.	.		
7	.	.	.		.				.	.		
8	.	.	.		.				.	.		
9	.	.	.		.				.	.		
10	.	.	.		.				.	.		
11	C	C	C		.				.	.		
12	.	.	.		.				.	.		
13	.	.	.		.				.	.		
14	.	.	.		.							
15												
16	.		.									
17	.	.	.	.	.	.	.	S				
18	.	.	.	.	.	.	.	S				
19	.	.	.	.	.	.	.	S	.	.	.	.
20	C	C	C	C	C	C	C	C	C	C	C	C
21	.	.	.	.	.	.	.	S	.	S	.	.
22	.	.	.	.	.	.	.	S	.	S	.	.
23	.	.	.	.	.	.	.	S	.	S	.	.
24	.	.	.	.	.	.	.	S	.	S	.	.
25	.	.	.	.	.	.	.	.	.	.	.	.
26	.	.	.	.	.	.	.	.	.	.	.	.
27	.	.	.	.	.	.	.	.	.	.	.	.
28	.	.	.	.	.	.	.	.	.	.	.	.
29	.	.	.	.	.	.	.	.	.	.	.	.
30	.	.	.	.	.	.	.	.	.	.	.	.
31	.	.	.	.	.	.	.	.	.	.	.	.
32	.	.	.	.	.	.	.	.	.	.	.	.
33	.	.	.	.	.	C	C	C	C	.	.	
34	.	.	.	.	.	.	.	.	.			
35		.			.	.	.	.	.			
36		.			.	.	.	.	.	.		
37		S			.	.	.	.	.	S		
38		S			.	.	.	.	.	S		
39		S			.	.	.	.	.	S		
40		S						.	.	S		

Key: S = short run; U = no run; . = data present; C = malfunction in collection of communications data, filled in with day average, except for Subject No. 20, who was dropped.

FIGURE 10. DATA POINTS, SEM II EXPERIMENT

Returning to the SEM I data, a "cross-validation" analytic effort was performed to determine whether the same factor structure could represent the data in each of the six sector-density combinations (cells). Each cell was examined separately. The cross-validation indicated the same factors were applicable.

After the cross-validation was completed, a return was made to the analysis of each of the two experiments on an individual basis. For the factor scores, it was now important to use standardization scales that covered the range involved in the particular experiment. The SEM I data standardization was against the first replicate, middle density, geometry 1 mean and standard deviation, and the factor scores were expressed on a standard score scale with a mean of 50 and a standard deviation of 1 at that point (the "first scale"). The SEM II experiment standardization used the mean and standard deviation of the fifth 1-hour run and the factor scores were expressed on a standard score scale with a mean of 500 and a standard deviation of 1 at that point (the "second scale"). Finally, it was decided to create a "third scale" in which both experiments' data were put on the same scale. Here all runs from both experiments were standardized against hour five of SEM II. The standard score distributions of the 4 factor scores were given a mean of 500 and a standard deviation of 1 at hour 5 of SEM II. This scaling was used in the power tables and to illustrate graphically the advantages of standard scores.

ANALYSES

Each of the topics listed below will be discussed in order under headings which will present the analysis of the topic and the data bearing on it, and the implications of the results:

1. SEM II factor analysis and factor cross-validation

2. Reliability coefficients

3. Correlations with observer ratings

4. Practice and learning effects in ATC simulation experiments

5. The effects on the system performance measures of enroute sector geometry and traffic density level

6. The statistical power of ATC simulation experiments

7. An evaluation of the index of orderliness

8. Subjective questionnaire replies and objective measures

SEM II FACTOR ANALYSIS AND FACTOR CROSS VALIDATION

ANALYSIS. Dynamic simulations of current and future air traffic control systems are difficult and expensive to arrange and run. They are difficult to design and analyze statistically, but worst of all they are difficult to interpret when making judgements about the desirability of air traffic control system changes. A major reason for this is the sheer cumbersomeness of the amount of data usually collected. A multitude of measures describing system performance is available, and there has been little or no evaluation as to which of the available measures is most relevant or needed. An attempt to reduce the magnitude of this problem was made here by applying a mathematical technique called factor analysis (see definition, appendix C) to see if a smaller set of measures of known relevance could be found. The second experiment (SEM II) was particularly designed to permit the use of this technique.

A factor analysis was performed on each of the three sets of "day level" data available from the SEM II experiment. Since there were 12 1-hour runs in the SEM II experiment, three 4-hour aggregates were available for each subject. These will be referred to as the first, second, and third days since each subject usually performed four runs a day. It is important to note that the factor analyses were done without the judges' ratings being involved.

Before entering the factor analysis, some of the measures in the original list of 28 which seemed not to be potentially fruitful were omitted bringing the list of measures entering the factor analysis to 17. Six (6) measures covering sub-types of delays and delay times, already represented in the summary measures of total number of delays and total delay time, were considered as redundant and dropped. These measures were the number and duration of barrier delays, the number and duration of start delays, and the number and duration of hold and turn delays. Another measure, the average aircraft time under control, was considered to be adequately represented by the measure aircraft time under control. Four (4) other measures which showed little or no variation in the data were omitted; these were the number of aircraft handled, the number of completed flights, the number of departure altitudes attained, and the number of handoffs accepted. These did not vary because of the similar traffic samples and, being essentially constants, would not have contributed to the factor analysis of the data. Two (2) further measures were dropped during the smoothing process just subsequent to the factor analysis itself because found to be non-contributing. These were the handoff acceptance delay time and the number of arrival altitudes attained.

The factor analysis was performed using varimax rotation of the principal components (see definition, appendix C) on 17 measures for 39 subjects. As has been said, a separate analysis was performed for each data day.

In the outcome, four operationally meaningful factors and quite similar factor patterns resulted from the analysis for each of the 3 days. The four factors accounted for 74.7, 67.7 and 63.3 percent of the total variance on days one, two and three respectively. The factor structures for the 3 days are shown in tables 1, 2, and 3 in appendix D, Supplementary Tables. Shown in these tables are the factor loadings, i.e., the correlations of each of the measures which had entered the process with each of the factors which had resulted. An extensive examination was conducted comparing the factor structures which had resulted on 3 days. Basically, the same four factors were identified, but the weights derived for the 3 days to generate factor scores were somewhat different.

The weighting differences among the 3 days were smoothed to 1 set of weights based on the median of the 3 days' weights. This was deemed permissible since the correlations between the scores weighted in the three different ways were generally in the .90's (see table 8, appendix D). The factor scores based on the median weights will be referred to as the "Full" factors. The Full factor weights appear in table 9 of appendix D. Further simplification was attained by rounding the weights arithmetically and zeroing out the weights for those measures which had carried factor loadings less than .15. It was during smoothing that one measure referred to earlier was dropped. The factor weights which resulted from this step will be referred to as the "smoothed" factors. These appear in table 10 of appendix D. A final rounding step and dropping of the last measure resulted in what will be called the "very smooth" factor score weights. The step involved making the remaining weights, which were in fact quite similar, equal. These appear in table 11 of appendix D. At this stage, the factor scores were computed by standardizing the measures which were to be

part of a given factor score for a given day on the day mean, applying the weights, and restandardizing the resulting factor score on the day mean. Having arrived at this point, three questions were examined about the very smooth factor score coefficients. The first question concerned the reliability of the factor scores before and after smoothing. The reliabilities appear in table 2, and clearly they were not degraded, but remained at about the middle of the range of the reliablities of the scores that made them up.

The second question concerned the statistical impact of using the very smooth factor set in which the various measures comprising the four factors were given equal weights. An analysis was done which compared, on the one hand, the simple product moment correlation of each of the factor scores (which, it will be remembered, contained the measures in equally weighted form) against the ratings and, on the other hand, the multiple correlation which resulted from mathematically optimally weighted combinations of the measures in each factor, the weights being optimized to predict the controller observer judges' ratings. These data appear in table 3. Concentrate on the "shrunken" R squared (R squared sub c) figures since they represent the percentage of variance accounted for statistically, after correcting for the the number of predictors involved. It appears that there was no essential difference in the correlations and so it is concluded that the weighting found in the factor analysis, i.e., equal, in generating the factor scores, is an acceptable weighting scheme.

The question of what weights to use in the computation of the factor scores having been decided, the next question asked concerned the ability of the factor scores, as compared with the original scores listed, to relate to the controller observers' ratings. Multiple correlations between the four factor scores in linear combination were computed with the controller observer ratings. These data are seen in table 4. Both the full factors and the very smooth factors were used. These multiple correlations were found to be at about the same level as the multiple correlations using the original 17 measures.

At this point, the cross-validation ability of the multiple regression equations based on the factor scores was investigated (table 5). Presented are the simple product moment correlations between a projected rating, based on an equation derived from data from a different day, and the actual rating given. Just as was discussed earlier, in the case of the equations using the original 17 measures, it was found that the day-to-day carryover was comparatively low. The ability of a weighting equation derived from the first day's data to predict the ratings on the second and third day was examined. The multiple correlation was found to decrease with the distance away from the day on which the weights were derived. The lesson here is that for neither factor scores nor raw scores can there be a multiple regression equation developed which will contain weights capable of carrying over to subsequent days or situations. The same system performance scores are seen as applicable

TABLE 2

RELIABILITY COEFFICIENTS OF SCORES BASED ON FULL FACTORS, SMOOTH FACTORS, AND VERY SMOOTH FACTORS

	Day-Day	Full	Smooth	Very Smooth
Confliction	1-2	.64	.65	.66
	2-3	.64	.63	.64
	1-3	.54	.53	.53
Occupancy	1-2	.59	.59	.62
	2-3	.59	.64	.62
	1-3	.27	.29	.30
Communication	1-2	.85	.86	.86
	2-3	.87	.87	.87
	1-3	.77	.76	.76
Delay	1-2	.11	.21	.19
	2-3	.27	.22	.21
	1-3	.10	.14	.12

TABLE 3

LINEAR COMBINATION WEIGHTING AND EQUAL WEIGHTING WITHIN EACH FACTOR

	Confliction Factor				Occupancy Factor			
	R*	R**	r***	r****	R*	R**	r***	r****
1 SEM	.51	.15	.44	.19	.43	.11	.29	.08
CPM	.56	.21	.49	.24	.40	.08	.27	.07
Day 2 SEM	.52	.17	.43	.18	.65	.37	.58	.34
CPM	.58	.23	.52	.27	.62	.34	.55	.30
3 SEM	.46	.10	.26	.07	.51	.19	.44	.19
CPM	.47	.11	.31	.10	.48	.16	.44	.19
	Communication				Delay			
	R	R	r	r	R	R	r	r
Day 1 SEM	.44	.15	.41	.17	.55	.29	.55	.30
CPM	.40	.12	.36	.13	.56	.29	.56	.31
Day 2 SEM	.31	.05	.25	.06	.35	.10	.25	.06
CPM	.37	.09	.22	.05	.30	.06	.26	.07
Day 3 SEM	.40	.12	.36	.13	.20	.01	.06	.00
CPM	.43	.14	.37	.14	.19	.01	.03	.00

* = R is the multiple correlation
** = the multiple correlation squared and corrected for shrinkage
*** = the product moment correlation
**** = squared product moment correlation

TABLE 4

COMPARISON OF MULTIPLE CORRELATION WITH JUDGES' RATING

PROVIDED BY ORIGINAL SEVENTEEN MEASURES, FULL FACTOR SCORES

AND VERY SMOOTH FACTOR SCORES

		Seventeen Measures				Full Factor Scores				Very Smooth Factor Scores			
		N	R*	R**	R***	N	R*	R**	R***	N	R*	R**	R***
Day 1	SEM	40	.82	.67	.42	39	.74	.55	.49	39	.73	.53	.46
	CPM	40	.83	.69	.44	39	.74	.55	.50	39	.73	.54	.47
Day 2	SEM	39	.81	.66	.39	39	.72	.51	.45	39	.69	.48	.40
	CPM	39	.87	.75	.56	39	.75	.56	.51	39	.72	.52	.45
Day 3	SEM	39	.79	.61	.29	39	.61	.38	.30	39	.60	.36	.26
	CPM	39	.79	.62	.31	39	.64	.41	.34	39	.63	.40	.31

* = the multiple R
** = the multiple R squared
*** = the multiple R squared after correction for shrinkage

TABLE 5

CROSS VALIDATION OVER DAYS (R)

		Day 1 Data	Day 2 Data	Day 3 Data
SEM	Day One Equation	.73	.60	.51
	Day Two Equation	.59	.69	.62
	Day Three Equation	.44	.53	.60
CPM	Day One Equation	.73	.61	.49
	Day Two Equation	.63	.72	.63
	Day Three Equation	.45	.55	.63

but they must be weighted (or considered) differently. An example will clarify this point. The weighting applied to the delay factor score diminished markedly on the third day. This means it had no weight in contributing to the controller observer ratings of system/controller performance on that day, whereas it had weight on the first day. But an examination of the objective data shows that there were several delays on the first day but almost none on the third day, which means the observers were right to give delay no importance on the third day. This does not mean we should not measure delay, but only that its importance may vary.

This finding is also important because it reinforces the conclusion discussed earlier that there is no possibility of joining measures into a single score, regardless of whether original measures or factor score measures of system performance are used. While the relationship between the weighted combinations of scores in the same circumstances is high, the projection of weights into different circumstances, such as in this instance, a later stage of practice, is not adequate. Therefore, a weighting equation resulting in a projected single figure of merit is not advisable.

Thus far, it has been shown that the same factors appeared in the 3 days of the SEM II experiment, that the weights of the original measures to make up the composite factor score indexes should be equal, but that assigning weights to the four factor scores to obtain a single conglomerate index was not a good idea.

A major next phase was to determine if the same four factors would appear in different traffic levels and sector structures, as represented in the six combinations of circumstances used in the SEM I experiment. It will be recalled that in the SEM I experiment there were two sectors and three traffic density levels for a total of six conditions, and that one of the six conditions was identical with that used in SEM II. It will also be recalled that the list of measures used in the two experiments was somewhat different and that there were only two replicate runs in SEM I, compared with the twelve replicates in the SEM II experiment.

The first step in determining whether the same four factors as had appeared in SEM II also would appear in the SEM I data, now that they had been discovered and seemed firm, was to re-score the SEM I data using the SEM II measurements list so that the question could be addressed. In the ATC simulator used, the most fundamental data collected are based on the aircraft movements and positions and the simulator pilots' inputs to the computer in response to the controllers' clearances. These data could be reduced in terms of either the SEM I or the SEM I list of measures. The SEM I data, then, were scored in terms of the SEM II measure list. The scoring was done up to the fiftieth minute rather than up to the sixtieth minute (as in SEM II) to overcome missing data due to equipment difficulties which had occurred in SEM I. Because of missing data, the number of data cases or subjects for SEM I was 31. For all of this analysis, the average of the two replicates in SEM I was

used. If a value for one run of the two replicates was missing, the best estimate "average" was the alternate data point.

The re-scoring having been done, the six cells in SEM I were separately subjected to factor analysis. At this stage, the factor analysis was done independently for each cell, and independently of the SEM II factor analysis. The method of factor extraction was always principal component analysis with varimax rotation, constraining the number of rotated components to four.

The next step was to utilize the SEM II factor score formulas and weights to compute the SEM II factor scores, using as input the SEM I data, scored, as mentioned above, in SEM II measures, so that these could be compared with the independently generated factor scores described above.

The results of the two operations described immediately above can be referred to, respectively, as the SEM I independent factor analysis scorings and the SEM II based factor scorings, and it is these that will be compared.

In overview, it may be said that examination of the six SEM I independent factor analysis scorings indicated that the measures had grouped similarly to those groupings which had occurred in SEM II. The factor loadings for the corresponding measures in the seven separate and independent factor analyses are similar. The percentage of variance accounted for by the SEM II-based factors is similar, and the SEM II factors predict the ratings almost as well as the SEM I factors do. There is one anomaly, it occurs in the coefficients of the delay factor, but this is capable of being understood in terms of certain difference in the definition of the details of the term delay in the two experiments. These differences will be discussed in detail later.

It is natural, of course, that the SEM I independent factors accounted for more of the variance in the data, between 73 and 80 percent, depending on which of the six conditions one examines. However, the externally based SEM II median (very smooth) factors computed for these same six conditions accounted for, in five of the six conditions, between 62 and 72 percent of the variance, and 59 percent in the remaining case. For corresponding conditions, the loss in going to the SEM II factors ranged between 6 and 12 percent, and averaged about 10 percent (see table 6).

For each of the six SEM I conditions, the SEM I-based factor structures were compared to the the SEM II-based factor structures. What is meant by this is that an examination was made of the results of the six factor analyses showing the factor loadings which had been assigned by the analysis to each of the original measures which had entered. Examined was whether the same measures clustered together as shown by their loading (correlation) with the same major factors. These data for the six SEM I combinations of conditions can be seen in tables 12 to 17 of appendix D. The SEM II factor structures are presented in tables 1 to 3 of appendix D.

A somewhat easier approach involves computing the coefficients of correlation between the factor scores resulting for the subjects as a group, computed in the two major ways described above. The correlation matrices for each of the

TABLE 6

PERCENT OF VARIANCE ACCOUNTED FOR BY FACTORS

Geometry Density	Factor Analysis for SEM-1 Data Percentage of Variance Consumed by Four Factors	Percentage of Variance Consumed by the Day 1 Loadings from the SEM-II Data when Applied to SEM-I	Percentage of Variance Consumed by the Day 2 Loadings from the SEM II Data when Applied to SEM-I	Percentage of Variance Consumed by the Day 3 Loadings from the SEM II Data when Applied to SEM-I	Median of Day 1, Day 2 Day 3*
G14 D1	73%	70%	67%	63%	67%
G14 D2	80%	75%	72%	69%	72%
G14 D3	80%	72%	70%	66%	70%
G16 D1	73%	64%	64%	60%	64%
G16 D2	73%	61%	59%	57%	59%
G16 D3	74%	68%	62%	62%	62%

*SEM-I loss from median = 6%, 8%, 10%, 9%, 12%, 12%

six combinations of conditions between the two kinds of factor scores were computed and are shown in table 7. As can be seen, the correlations are mainly in the 90's for the first three factors, but the correlations for the fourth factor, Delay, are at times negative. This is the anomaly which was mentioned earlier and it is understandable in terms of some differences in procedures and definition of delays in the two experiments. This minor discrepancy was one of the prices paid for the use of two data bases assembled under slightly different rules. Since the factor score weights ultimately go back to the correlation matrices, these were examined. Examining the correlation matrices for the six cells of SEM I and for the 3 days of SEM II showed some differences in the correlations between the measures "time in boundary" and "total delay time" between the SEM I data base and the SEM II data bases. In the case of the SEM I data there was a moderately high correlation of about minus .3 between the two measures; in the SEM II data there was a near-zero correlation between the measures for two of the original days, although there was a slightly minus correlation for the third day. This slightly minus correlation for the third day was lost in the smoothing process, but the other 2 days had virtual zero correlations and this is why the smoothed factors show this. But the more general source is probably in procedures. The negative correlation for the SEM I data would seem to indicate that, under SEM I procedures, if delay were taken before accepting the aircraft, the time in the sector would be lessened, whereas under the SEM II procedures, this made little or no difference in the amount of time in the sector.

This appears as something which might have occurred since under the procedures for the SEM I experiment the controller was permitted to tell the adjoining sector (the support or "ghost" controller who was a member of the experimental staff) seeking to make a handoff to him to hold or "spin" the individual aircraft. It will be remembered that the procedures were changed going into the second experiment to reduce what was perceived as the undue impact of the support controller in this and other areas.

One of the changes made for the SEM II experiment involved the method of starting aircraft into the test sector, which was now made automatic and done by the computer on schedule. As a consequence of this, the idea of "barrier delay" was seen as necessary. Under the concept of the barrier delay, if the subject wished to delay aircraft he had to impose delay on the entering stream of aircraft, and not individual aircraft one at a time. Very few barrier delays were used in SEM II (it probably being regarded by the controllers as extreme, as compared with delaying one aircraft).

The best conception of what might have happened probably is based on the idea that under SEM I procedures it seemed better to the subjects to take any delay outside the sector before accepting handoffs, and that indeed it possibly was better due to some help in lining up the aircraft provided by the ghost in his handling of the aircraft while they were still outside the sector. Thus, for SEM I data, there was a slight negative correlation between start delays and time in sector. Under SEM II procedures, the computer provided no such assistance and also the tendency probably was to minimize barrier (start) delays and take the delays if any within the sector. The small number of these would also tend to bring the correlation between start delays and any other

TABLE 7

CORRELATIONS BETWEEN SEM II FACTOR SCORES AND SEM I

SECTOR-DENSITY CELL-BASED FACTOR SCORES

Sector - Density Condition	Factor			
	Confliction	Occupancy	Communication	Delay
Geometry 1, Traffic Density 1	.75	.96	.94	.84
Geometry 1, Traffic Density 2	.96	.83	.96	.35
Geometry 1, Traffic Density 3	.96	.77	.86	.90
Geometry 2, Traffic Density 1	.98	.95	.88	-.60
Geometry 2, Traffic Density 2	.95	.95	.85	-.60
Geometry 2, Traffic Density 3	.99	.96	.80	-.64

measure down. Thus, there was a near-zero correlation for SEM II, a different correlation than that in the other data.

It appears, then, that there is probably some effect involving these procedural differences between the two experiments which caused a different relationship between the two measures mentioned and this changed relationship probably effected a difference in the delay factor between the two experiments to a sufficient extent that the weights differed enough to cause the slight negative relationship in the delay factor between the two experiments, even though, as should be remembered, the same basic factor resulted.

Another comparison between the SEM I and SEM II factors was done in terms of an index discussed by Harman (reference 4) which roughly resembles a coefficient of correlation between factor score weights in two sets of factors. It also ranges from -1.00 through zero to +1.00. It is referred to variously as the coefficient of congruence or as the index of the degree of factorial similarity or as phi.

The phi index is calculated essentially by computing a correlation between the factor weights given for the original measures by the two factor sets being compared. In this case, the phi indexes were computed for each of the six combinations of the SEM I conditions. For the logically similar factors based on the two experiments, again except for the delay factor, the correspondence was quite good. The overall picture was similar to that just given in table 7 for the correlation coefficients.

In the case of the first three factors, the phi coefficients ranged between .60 and .94 for all days and conditions. They were usually in the .70's, .80's and .90's. Of the six phi's computed for the six conditions of density and sector for the delay factor, four were negative, one was moderate (.59), and one was somewhat high (.76). In general, this phi analysis confirms the others above.

Finally, an important examination of the connection between the independent SEM I factors and the SEM II derived factors was done using the judges' scores. This analysis is important because it relates the two kinds of scoring methods to the opinions of the controller judges who were on the scene during the SEM I exercises. Multiple correlations against the opinion measurement were computed using, separately, the two kinds of factor scoring: externally based and internally based; SEM I-based and SEM II-based. Because the two ratings (SEM and CPM) were highly correlated, only one of them (CPM) was used in the computations.

In the outcome, the multiple R's were quite similar regardless of which form of weighting was used. There was only a .05 difference, in the multiple correlation, R, at most, in favor of the SEM I self-generated factor scores for any of the six sector-density combinations over the SEM II factor scorings for the same data, as seen in table 8.

Recapitulating, we may say that the evidence has shown that the four factor scores developed in the SEM II experiment are also applicable to the SEM I

TABLE 8

SEM I CELL BASED FACTOR SCORES AND SEM II FACTOR SCORES IN

RELATION TO SEM I JUDGES' RATINGS

	Using SEM-II Factor Score Coefficients to Create Factor Scores	Using SEM-I Factor Score Coefficients to Create Factor Scores	
	(Factor Scores vs. Judges' Scores)		
	R	R	N
Sector 14, Density 1	.36	.42	31
Sector 14, Density 2	.46	.52	31
Sector 14, Density 3	.57	.62	29
Sector 16, Density 1	.47	40	31
Sector 16, Density 2	.41	.33	31
Sector 16, Density 3	.59	.63	30
	(Factor Scores vs. Log of Judges' Scores)		
Sector 14, Density 1	.39	.43	31
Sector 14, Density 2	.47	.47	31
Sector 14, Density 3	.54	.61	29
Sector 16, Density 1	.46	.39	31
Sector 16, Density 2	.42	.33	31
Sector 16, Density 3	.59	.62	30

experiment's sector and geometry variations. In both experiments, the four factors acount for a majority of the variance.

There is evidence, although indirect, from other experiments which were not directly comparable for various reasons, like those of Boone (references 5,6) and Buckley (reference 2) that this factor structure has generality. In Boone's experiment, he found somewhat similar factors even though dealing with Academy trainees in early stages of training. He was, however, using the FAA Technical Center ATC simulator that was used in this experiment and the SEM I set of measures which were programmed into it. The factor analysis done by Buckley in 1969 (reference 2) used hand-collected data and combined several densities. However, there is some resemblance to the factors obtained here.

Having arrived at a small set of measures which seems to succinctly encompass the important dimensions of air traffic control system performance can be important, if it is applied. For example, if most or all simulation experiments are scored in terms of the same four factors, it may eventually be possible to conduct meaningful comparisons about results obtained at different times and in different places.

On the other hand, the basic or "raw" measures could be considered to be "buried" in the four factor scores, especially since they are necessarily of a dimensionless standard score form. However, the more specific measures, such as the number of altitude changes, can still be looked at by those with a special interest in them. There is no inherent contradiction between being interested in the specific and the general. At the very least, even if the four factor scores do not replace the many specific measures, they should be used as a short and meaningful way of summing up all of the several specific simple measures.

An avenue was examined here for minimizing any possible disadvantages of the use of standardized factor scores. An examination was made to see if one raw score could be used to represent each of the four factors. Considered in the decision were the correlation between each of the measures which entered into each of the factor scores and the factor score it entered, the comparative reliability coefficients of the measures within each factor, and whether the measure consistently appeared in the respective factor across the two experiments. The correlations between the factor scores and the observer ratings were not considered to be a major element in the choice since the purpose was to represent the already chosen factor score. As mentioned, one consideration was the reliablity of the measure, especially between Days 2 and 3. These are shown in table 9. Another main consideration, the correlation with the factor score itself, is shown for each factor in table 10.

Based on all of these considerations, then, one measure was chosen for each of the four factors to be that factor's "primary" measure, i.e., a raw score representative of the factor for those who prefer raw scores. The asterisks in Tables 9 and 10 denote the measures which were chosen as the primary measures.

Returning now, however, to the discussion of standard scores, it should be remembered that they have distinct advantages as well as potential

TABLE 9

DAY TWO VERSUS DAY THREE RELIABILITY OF MEASURES WITHIN A FACTOR

Conflict Factor

	r
Number of Four-Mile Conflicts	.69
Number of Five-Mile Conflicts	.78
Number of Three-Mile Conflicts	.41
Duration of Four-Mile Conflicts	.43
Duration of Five-Mile Conflicts	.64
Duration of Three-Mile Conflicts	.34

Occupancy Factor

Time Under Control	.66
Distance Flown Under Control	.54
Fuel Consumption Under Control	.56
Time in Boundary	.69

Communications Factor

Path Changes	.84
Number of Ground-to-Air Communications	.85
Duration of Ground-to-Air Communications	.87

Delay Factor

Total Delays	.18
Total Delay Time	.15

TABLE 10

CORRELATIONS OF MEASURES WITHIN A FACTOR WITH THE FACTOR

Conflict Factor

	Day One	Day Two	Day Three
Number of Four-Mile Conflicts	.90	.92	.87
Number of Five-Mile Conflicts	.81	.82	.87
Number of Three-Mile Conflicts	.84	.81	.79
Duration of Four-Mile Conflicts	.89	.91	.87
Duration of Five-Mile Conflicts	.87	.83	.77
Duration of Three-Mile Conflicts	.82	.79	.77

Occupancy Factor

	Day One	Day Two	Day Three
Time Under Control	.99	.94	.97
Distance Flown Under Control	.91	.74	.80
Fuel Consumption Under Control	.93	.91	.91
Time in Boundary	.69	.73	.77

Communications Factor

	Day One	Day Two	Day Three
Path Changes	.85	.89	.86
Number of Ground-to-Air Comm.	.91	.92	.89
Duration of Ground-to-Air Comm.	.90	.93	.90

Delay Factor

	Day One	Day Two	Day Three
Total Delays	.98	.91	.87
Total Delay Time	.98	.91	.87

disadvantages. They will remind us, for example, that the results from any real-time simulation are interpretable only in relative and not in absolute terms. It is possible to interpret the standard scores in terms of the percentiles they would represent in an assumed normal distribution as is often done in large scale personnel testing situations. A related approach which would not involve any assumption of normality would be interpretation in terms of the percentiles for the scores from various experiments in terms of a reference distribution, such as the SEM II data distribution. The SEM II data distribution is not large enough to be a general reference distribution and certainly not large enough to do away with the need for control groups in particular experiments. But if all experimenters used it as a distribution in terms of which to generate standard scores for the four factors, then data could be accruing for a common distribution into which all experimental data could be translated in common terms.

An example of this is given in figure 11. As part of the process of constructing the power tables, it was necessary and desirable to put the data from both experiments (SEM I and SEM II) into terms of the same scale distribution so that the power tables would be useful over a range of sectors and densities. The first step in accomplishing this was to bring the SEM I runs from a 50-minute basis to a 60-minute basis by multiplying each run score by sixty-fiftieths. This was specifically done for the power table preparation process, since it was desired that they be in hour-unit terms. It was also done for figure 11. For the data which were used in most of the SEM I analytic computations, it was felt that the prorating was not necessary. In generating this new scale, for the power tables, the factor scores for both experiments were computed using the run scores from each of the experiments after they had been converted into standard score form based on the mean and variance from the SEM II hour 5 data. They were given a mean of 500 and a standard deviation of 1 at the SEM II hour 5 point. For convenience, this was called the "third scale" to distinguish it from the standard score scales which had been used individually in SEM I and SEM II. The new scale enabled the factor score distributions from both experiments to be drawn on the same scale. This is seen in figure 11, which shows both the data from each of the six sector-density combinations of SEM I and the three days of SEM II.

From here on, the discussion will be in terms of the factor scores and the four primary scores. Two other measures, which we will call auxiliary scores, will also be carried along. These are the number of aircraft handled and fuel consumption. The number of aircraft handled measure, in the SEM II level density experiment, was very insensitive and was not entered into the factor analysis. This was due more to the particular experimental design than to the importance of the measure, and it should be kept as an auxiliary measure for reaction to traffic density variations in more general situations. The fuel consumption measure was entered into the factor analysis and formed part of one of the factors. It is of particular operational relevance and it will also be carried as a separate auxiliary measure.

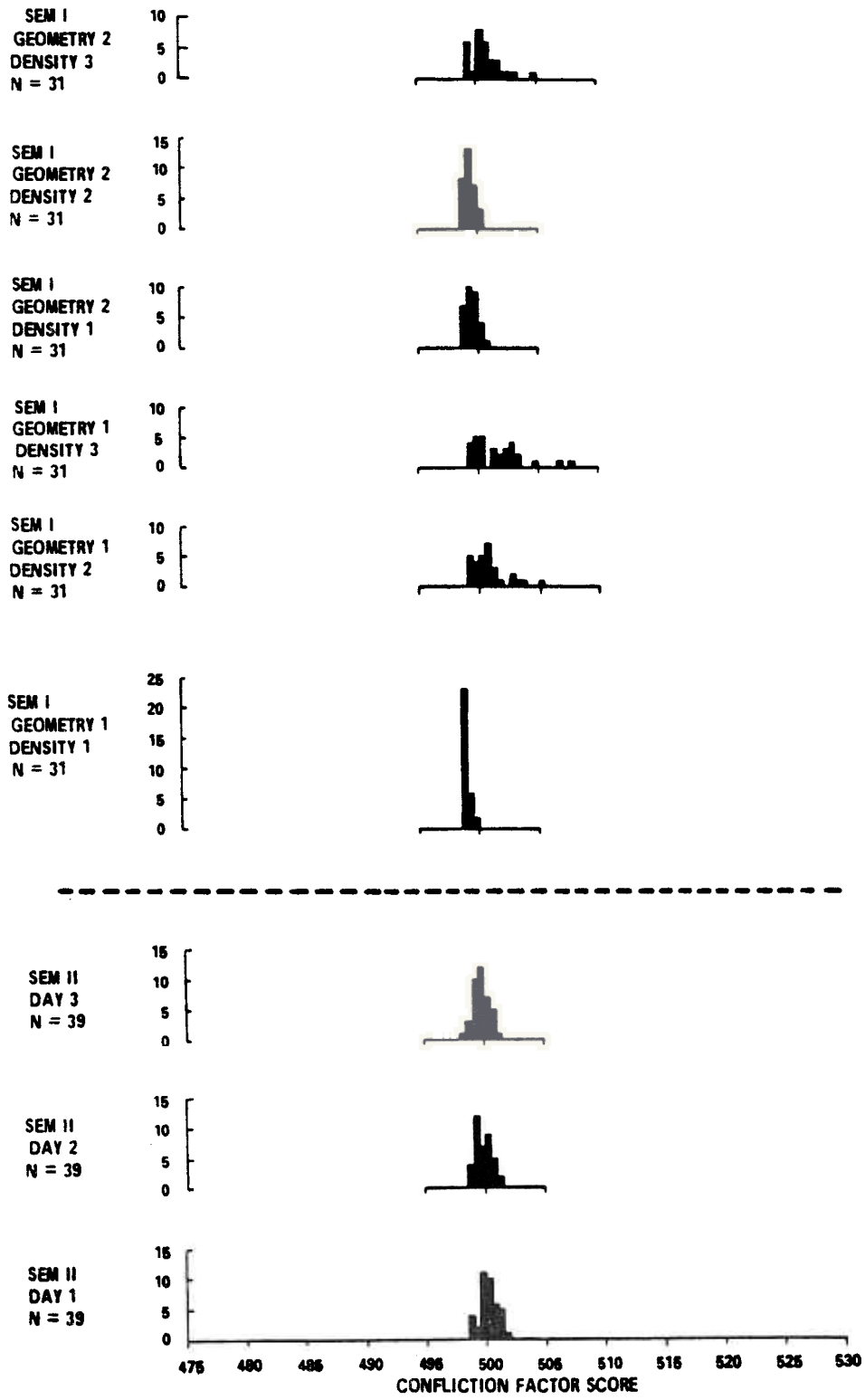


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 1 of 4)

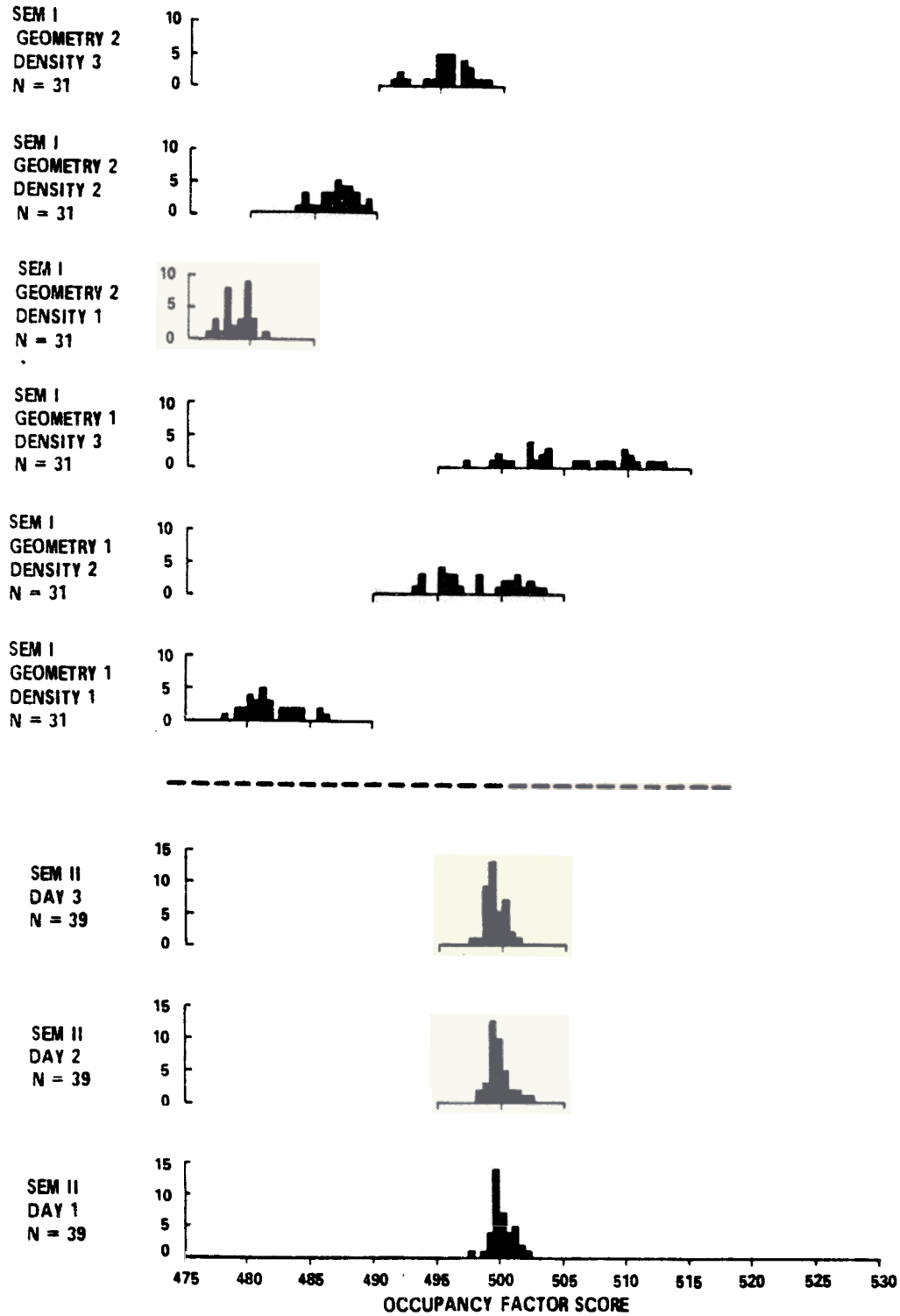


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 2 of 4)

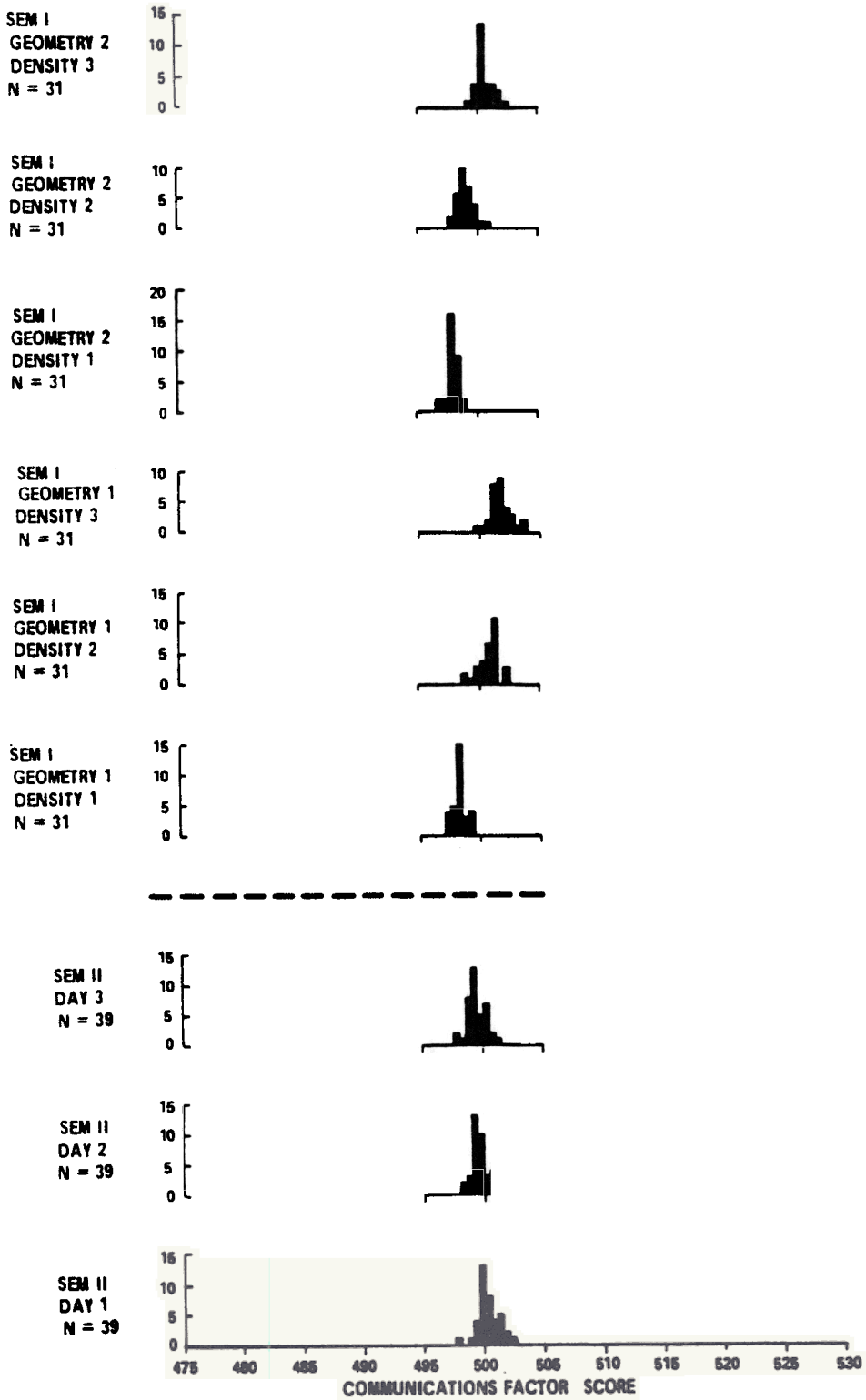


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 3 of 4)

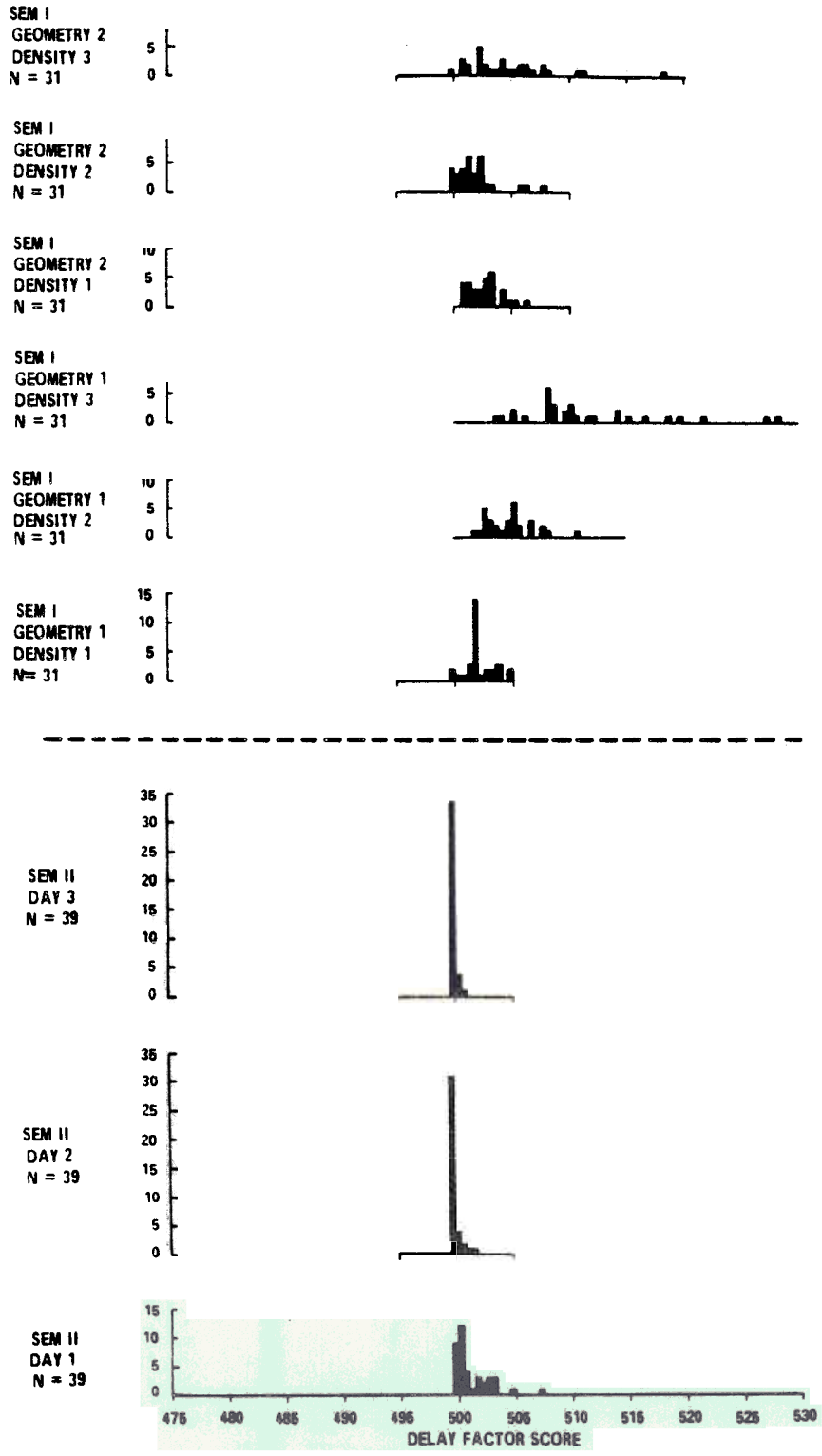


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 4 of 4)

It is important to point out that the factor scores and computations using them appearing in the tables in the balance of the report will be based on scales which standardized the entire body of data from each experiment on points within the respective experiments. In some cases there may be slight differences between these later computations done on that basis, and those appearing in the factor analytic and cross-validation sections earlier in this present section because the earlier computations are based on a day-by-day (SEM II) or a cell-by-cell separate standardization (SEM I) with occasional truncations for various purposes.

It should be pointed out here, finally, that both the four factor scores and the primary scores for each factor, and other raw scores of interest could all be used by any given experimenter. The ATC simulator data processor can immediately produce the four factor scores for any future experiment in "third scale" terms, using the SEM II hour-five data as a reference point.

IMPLICATIONS. It has been seen that:

1. The same general factors were generated by the factor analysis technique using the SEM I data and the SEM II data. The SEM II factors and weights for the measures within the factors seem adequate to characterize the SEM I data in all six combinations of sector geometry and traffic density.

2. The fact that the measures are equally weighted within the SEM II factors does not adversely impact their relationship with the controller observer judgements, as compared to the relationship generated with the same judgements by the original measures.

3. The factors found basically corresponded to those found in an independent experiment involving controller trainees working at a much lower level of difficulty (Boone, references 5,6).

4. It appears that, despite the wide range of conditions included in these two experiments, the four factors adequately summarize experimental results from ATC simulation experiments. The factors can be considered expressions of the important basic dimensions of the measurement of air traffic control system functioning in real time dynamic simulation experimentation.

5. It appears that the four factor scores may safely be used to represent all of the other measures.

6. In view of the above, it appears permissible and efficient to report experimental results in terms of the four factor scores, the four primary measures corresponding to the factors, and the two auxiliary measures, the number of aircraft handled and fuel consumption. It is suggested that all future air traffic control simulation experiments use that set of measures, as will be done in the balance of this report. Although it was not fully carried out for this report, it is further suggested that the factor scores in future work should use the "third scale standardization."

RELIABILITY COEFFICIENTS.

ANALYSIS. Reliability is defined as repeatability of measurement. To evaluate reliability, it is necessary to have repeated sessions ("runs") which, as may be seen from the experimental design (figure 8), was definitely achieved in the second experiment. There were 12 1-hour runs performed by each subject controller under essentially the same conditions except for the obvious and unavoidable one of practice.

The major index of reliability used was the product moment coefficient of correlation, or "r" (see appendix C, Definitions), between runs. This was done also for the data in the first experiment, although in that case, there were only 2 similar runs (runs by the same subject under the same conditions), not 12.

Table 11 shows the reliabiity coefficients for the set of measures which will be used from here on; the four factor scores and their corresponding primary measures, and the two auxiliary measures, the number of aircraft handled and fuel consumption. Shown are the SEM I and SEM II reliability coefficients for these measures as estimated by the correlation between 2 runs. The SEM I runs were 50 minutes in length, as discussed earlier, and as is shown in the Table. The correlations shown are those obtained when the SEM I data were scored using the SEM II measurements as defined in appendix A. In the case of the four factor scores, the SEM I computations used the first scale, and the SEM II calculations used the second scale, as will be usual.

In the case of the SEM II data, data aggregation was also possible. Table 11 shows the increase in reliability which results from the aggregation of the data into 4-hour chunks by averaging. The effect of this increased reliablililty which can be obtained by the process of averaging will be shown in a later discussion of statistical power.

A comparison of these reliability coefficient data can be made with only one other experiment in the small literature on ATC simulation, the 1969 experiment by Buckley et al. (reference 2). Another possible source, the experiment by Boone (references 5 and 6) on controller trainees which used basically the SEM I methods and measures, did not cite reliabilities. There are some data from the 1969 experiment shown in table 11, and it can be seen that moderate reliabilities were found; somewhat higher for the measures delay time and conflictions than were attained in the present work. It is interesting that the experiment was done using paper and pencil data taking, not computer data collection or target generation. In the case of the confliction count, the occurrence of a confliction was scored by the judgement of three observing controllers, and delay times were written down by the simulator pilots.

Another way of examining the repeatability of statistical data is in terms of the standard error of measurement (see appendix C, Definitions). In general terms, this index gives an error band for a single score or measurement such that the probabilities can be stated that the "true" score or value is within the stated range. The computation of the index depends on the reliability coefficient and the variance, which expresses the range of individual differences among the subjects.

TABLE 11

RELIABILITY COEFFICIENTS

Measure	1969 Exper.	SEM I Run vs Run	Run 3 vs Run 4	SEM II Run 5 vs Run 6	Day 1 vs Day 2	Day 2 vs Day 3
Confliction Factor		-.10	.48	.59	.68	.65
Occupancy Factor		.75	.46	.39	.58	.63
Communications Factor		.69	.83	.84	.85	.87
Delay Factor		-.38	.20	-.08	.20	.15
No. of 5-Mi. Conflictions	.62	.06	.48	.60	.72	.78
A/C Time Under Control	.45	.84	.45	.43	.53	.66
Duration of G/A Contacts	.56	.80	.85	.85	.87	.87
Total Delay Time	.39	-.29	-.07	-.05	.15	.15
No. of A/C Handled	.36	.27	-.04	-.04	.40	.21
Fuel Used Under Control		.73	.38	.26	.65	.56
- - - - - - - - - - - -						
Sector		14(G1	14	14	14	14
Density	med	med(D2)	med	med	med	med
No. Subjects (N), Factors		27	39	39	39	39
No. Subjects (N), Measures	36	27	39 or 38	39 or 38	39	39
Minutes of Operational Data	60	50	60	60	60	60

Standard errors of measurement computed for the factor scores and the six other measures which were listed above were computed based on 1-hour runs from both experiments, and these are given in table 12. For the scores given in the table, the probabilities are .95 that the "true" value is within the range given. Thus, for example, it may be seen that a delay time score of 78 seconds per hour based on a single 1-hour middle traffic density run in SEM I could, in fact, stand for delay time run scores ranging from 0 to 1331 seconds (22.2 minutes). For SEM II, the standard error of measurement obtained by using the first four runs aggregated is also shown. In this particular table, in order to facilitate comparisons, all calculations involving factor scores were done using the third scale. However, it might be pointed out that, in any case, the three scales are very highly correlated (.98 or higher) and differed mainly in the means.

As has been said, in addition to the objective measurements, there were also ratings made of performance. It will be remembered that there were two observers standing behind the controllers when they were controlling the simulated traffic. There were eight such observers and schedules were arranged so that they would be paired in all possible combinations. The observer/judges were qualified field controllers from facilities other than those of the subjects. The average of the two judges' opinions was used as the score for the run on this kind of data. The basic purpose of this rating process was to gain another kind of criterion against which to compare the objective measures. It was important to optimize the reliability of the ratings since they were to be used as an external criterion against which to check the objective measures. Therefore, the field controller judges received careful training in the rating process before the experiment began.

In considering the reliability of the ratings, it was possible to estimate this quality using two approaches. In one approach, the agreement between two judges observing the same occasion was considered. The inter-judge agreement was computed using the intra-class correlation (See appendix C for definition). In the other approach, the average of the two judges' ratings of a given kind (SEM or CPM) for a given run, which was always used as the rating of that kind for the session, was examined. Here, the run-to-run reliability of the average of the two ratings was examined. These two approaches were used in both experiments.

In table 13, the computed data on inter-judge agreement at a given session appear for both experiments. In table 14, the data are given for the run-to-run agreement for the average rating of a given type by the two judges watching the same runs. In the case of the SEM II data, it was also possible to examine the effects of day level aggregation as had been done with the other measures, and these day-to-day product moment correlations are also shown. Both the CPM and SEM ratings are not always shown; they were consistently found so highly correlated with each other in a given session (usually well over .85), that frequently only one of them was used in some calculations.

TABLE 12

STANDARD ERROR OF MEASUREMENT

Measure	If Measured Value Were:	With 0.95 probability, the true value would lie between limits of: SEM I (G1 D2) Avg. of 2 Runs	SEMII (G1 D2) Avg. of 5th & 6th Runs	SEMII (G1 D2) Day 2 (Avg. of Runs 5-8)
Conflict. Factor	500.	495.64-504.36	498.91-501.09	499.24-500.76
Occup. Factor	500.	497.60-502.40	498.99-501.01	498.97-501.03
Comm. Factor	500.	499.27-500.73	499.42-500.58	499.34-500.66
Delay Factor	500.	495.80-504.20	498.86-501.14	499.21-500.78
No. of 5-mi. Conflictions	6 per hr.	0-14.3	1.8-10.2	3.4-8.5
A/C Time Under Control	550 min./hr.	517-583	534-566	532-568
Dura. of G/A Contacts	650 sec./hr.	572-728	570-730	565-735
Total Delay Time	78 sec./hr.	0-1331.	0-567	0-342
No. A/C Handled	47/hr.	46.4-47.6	46.3-47.7	46.7-47.3
Fuel Used Under Control	112 thousand lbs./hr.	104-120	107-117	108-116

TABLE 13

INTER-OBSERVER AGREEMENT (INTRA-CLASS CORRELATIONS)

SEM I Sector-Geometry - Replicate Cells

G1 D1		G1 D2		G1 D3		G2 D1		G2 D2		G2 D3	
R1	R2	R1	R2	R1	R2	R1	R2	R1	R2	R1	R2
.17	.06	.61	.56	.46	.44	.13	.43	.48	.44	.55	.39
.28	.32	.52	.73	.72	.45	.65	.65	.50	.45	.62	.31

SEM II

Hour

1	2	3	4	5	6	7	8	9	10	11	12
.64	.64	.40	.43	.30	.32	.58	.69	.50	.53	.44	.66
.53	.57	.43	.35	.45	.40	.42	.57	.58	.55	.43	.65

TABLE 14

RATING RELIABILITY

SEM I
(Run-Run by Cell

	G1 D1	G1 D2	G1 D3	G2 D1	G2 D2	G2 D3
SEM	.14	.27	.34	.04	.09	.48
N	24	27	24	25	27	28
CPM	.20	.37	.42	.52	.01	.38
N	25	27	28	25	27	29

SEM II
(Hours)

	1 vs 2	3 vs 4	5 vs 6	7 vs 8	9 vs 10	11 vs 12
SEM	.15	.55	.37	.31	.39	.50
CPM	.25	.57	.23	.29	.23	.55
N	31	39	39	32	31	39

SEM II
(Day-Day)

	Day 1 to Day 2	Day 2 to Day 3
SEM	.64	.64
CPM	.64	.69
N	39	39

The size of the inter-judge agreeements found here is fair but changes from time to time. In the Boone experiment, the interclass correlation expressing agreement between instructors who were rating trainees executing simulation problems was .56. In the 1969 experiment, the median interclass correlation between observers rating in a session was .53. Cobb's study (reference 7) found moderately high agreement between field supervisors of controllers.

In evaluating the rating data in the two SEM experiments, it is important to pause and discuss two things. One is the fact that these judges were well-trained and practiced in observing the same exercises and people. It is also important to discuss the intended use of these ratings. They were not an external criterion such that the value of the objective measures would stand or fall with them; they were for corroboration and for making comparative judgements as to combinations of the objective measures. The ratings were not considered to be inherently superior to the objective measures; in fact, special efforts were made to overcome the normal inferior reliability of ratings as compared to objective measures. For training the observers, there was a week set aside for the observers before each experiment in which they observed the traffic samples which were to be used in the experiment, worked this traffic themselves, rated each other, and discussed the meanings of the rating scales.

When considering the ratings, it is important to remember that these were not taken in a typical rating situation, such as, for example, the over-the-shoulder rating taken in a facility, which might show lower reliability. These ratings should be considered as special ratings for a special purpose.

IMPLICATIONS. It can be seen that

1. The reliability of the objective measures taken in these dynamic simulations was fair, considering the dynamic situation, but was found to be improved by data aggregation. When improved by aggregation, it can be brought to quite high levels. However, refinement of the initial measure collection process itself may also be needed.

2. Reliability was not appreciably better in SEM II than in SEM I even though better measure definitions and stricter procedures were used in SEM II (as was discussed under procedures). However, the use of aggregation was possible in SEM II to increase the relaiability.

3. Reliability of the judges' ratings was adequate to the purpose here, but in line with typical results with subjective ratings.

4. Later discussions will carry the matter of measure reliability into the realm of statistical power in which the reliability coefficients and the standard deviation, or variation, of the data are used in planning experimental designs.

CORRELATIONS WITH OBSERVERS' RATINGS.

ANALYSIS. Objective measures of system performance and subjective observer ratings may each be said to have their own advantages and disadvantages. On the one hand, the advantage of objectivity would be difficult to overstate. On the other hand, objective measures can sometimes turn out to be meaningless and their validity and meaningfulness must be verified by comparing them to the judgments of experienced observers.

Evaluations by people very familiar with a task can be useful for certain purposes. However, as is commonly known and accepted, a difficulty with such subjective ratings is their frequent unreliability. The ideal is objective measures which are reliable and which can be shown to be meaningful by demonstrating a strong relationship to subjective evaluations by knowledgeable persons. The demonstration of such a relationship for the objective measures of air traffic control system performance is what will be examined in this section.

We will first examine relationships between some of the individual objective measures and the ratings in the SEM I and SEM II experiments. Table 15 gives the product moment correlations between these measures and the observer ratings. For the SEM I experiment, the correlations are given separately for each sector-traffic density combination. The average of the two replicate runs in each cell was used. For the SEM II experiment, correlations based on the average of two runs are also shown. Runs 5 and 6 were chosen as occurring somewhat after an initial learning period (which will be discussed later). For all factor scores, the third scale values were used.

Also shown in table 15 is the effect of the further aggregation which was possible using the SEM II data with its many replications. The data for the first 4 runs (of the 12 runs in SEM II), the second 4 runs, and the third 4 runs have been separately aggregated into day-level aggregations. The statistical significance level for the correlations (see appendix C) is also shown in the table.

The multiple correlation (R) is the correlation between a linear combination of variables and some other variable (for an exact definition, see appendix C). Here it is the correlation between the set of the four factor scores taken in combination and one of the ratings, or, similarly, the set of the four primary measures and one of the ratings. Table 16 shows these multiple correlations for each of the six geography-density combinations in the SEM I experiment. Shown are the multiple correlations based on the averages of the 2 runs in each cell for the SEM II measure set applied to the SEM I basic data. Also shown are the effects of using the logarithmic transformation in the process.

For SEM II, the multiple correlations are shown in table 17. The SEM II multiple correlations are shown as computed using the average of 2 runs as was done in SEM I, here using runs 5 and 6, and also as computed using the day-level aggregated data. Again the effects of the logarithmic transformation are shown.

TABLE 15. CORRELATIONS BETWEEN MEASURES AND RATINGS

	SEM I *							SEM II *		
		G-1			G-2		RUNS	DAY	DAY	DAY
	D-1	D-2	D-3	D-1	D-2	D-3	5/6	1	2	3
Factor Scores:										
Conflict Factor										
SEM rating:	-.24	-.16	-.24	-.48	-.35	-.35	.28	.45	-.44	.23
CPM rating:	-.22	-.08	-.25	-.38	-.31	-.24	.35	.50	-.53	.30
Occupancy Factor										
SEM rating:	.11	-.01	.18	.05	-.16	.20	-.54	-.29	-.60	.44
CPM rating:	.04	.06	.01	-.04	-.21	.26	-.52	-.27	-.57	.44
Communication Factor										
SEM rating:	.20	-.23	-.34	.15	.31	.13	-.24	-.42	-.25	.35
CPM rating:	.08	-.19	-.22	-.05	.17	.03	-.18	-.36	-.22	.37
Delay Factor										
SEM rating:	-.36	-.25	-.58	-.23	-.20	-.37	-.16	-.56	-.26	10
CPM rating:	-.37	-.26	-.52	-.25	-.21	-.43	-.27	-.56	-.27	09
Primary Measures										
Number of Conflictions										
SEM rating:	-.23	-.29	-.33	-.35	-.28	-.28	.17	-.33	-.23	.09
CPM rating:	-.21	-.17	-.24	-.23	-.24	-.19	.18	-.37	-.31	.15
Time Under Control										
SEM rating:	.20	-.06	.14	.09	-.14	.07	-.62	-.32	-.65	.45
CPM rating:	.20	-.01	-.08	.00	-.20	.13	-.57	-.31	-.61	.44
Duration of Ground-Air Com										
SEM rating:	.13	-.45	-.37	.07	.31	-.20	-.26	-.43	-.28	.38
CPM rating:	-.02	-.44	-.41	-.04	.13	-.28	-.23	-.41	-.29	.38
Total Delay Time										
SEM rating:	-.26	-.23	-.46	-.15	-.23	-.24	-.08	-.53	-.14	.05
CPM rating:	-.20	-.17	-.46	-.12	-.28	-.33	-.19	-.54	-.17	.05
Auxiliary Measures:										
Number of Aircraft Handled										
SEM rating:	.14	.04	.37	.00	-.08	.39	31	.36	.25	.11
CPM rating:	.23	.02	.33	.03	-.13	.37	35	.34	.24	.15
Fuel Consumption										
SEM rating:	.12	-.07	.16	.09	-.08	17	-.43	-.24	-.46	46
CPM rating:	.10	-.01	-.06	-.01	-.13	24	-.45	-.23	-.47	47

N for SEM I ranges between 27 and 31; N for SEM II is usually 39. The .05 levels for r at these N's are: N=27, r/.05= .38; N=39, r/.05= .31

TABLE 16

MULTIPLE CORRELATION (R) OF FACTORS AND LEADING MEASURES ON RATINGS, SEM I

Regression	Cells (2 hours)					
	1-D1G1	2-D2G1	3-D3G1	4-D1G2	5-D2G2	6-D3G2
Factors on SEM	.40	.34	.76	.52	.50	.60
Factors on CPM	.38	.32	.64	.47	.46	.60
Measures on SEM	.39	.54	.71	.39	.41	.59
Measures on CPM	.36	.49	.62	.28	.39	.65
Log of Factors on SEM	.40	.34	.76	.52	.49	.60
Log of Factors on CPM	.37	.32	.64	.47	.45	.60
Factors on Log of SEM	.42	.33	.75	.52	.50	.58
Factors on Log of CPM	.38	.31	.62	.47	.47	.59
Log of Factors on log of SEM	.41	.33	.75	.52	.50	.58
Log of Factors on log of CPM	.38	.31	.62	.47	.46	.59
Log of Measures on SEM	.35	.57	.75	.41	.48	.48
Log of Measures on CPM	.28	.50	.65	.33	.47	.57
Measures on log of SEM	.40	.53	.69	.39	.41	.57
Measures on log of CPM	.36	.48	.61	.29	.40	.63
Log of Measures on log of SEM	.36	.56	.72	.41	.47	.46
Log of Measures on log of CPM	.29	.49	.63	.33	.48	.55
N	31	30	29	31	31	30
R for .05 Stat. Sign.	.55	.55	.56	.55	.55	.55

NOTE: Transformation used for logorithmic cases was log (X+1).

TABLE 17

MULTIPLE CORRELATION (R) OF FACTORS AND LEADING MEASURES ON RATINGS, SEM II

Regression	Hours 5 & 6 (2 hour data)	Day 1	Day 2	Day 3
		(4 hours data)		
Factors on SEM	.60	.73	.71	.59
Factors on CPM	.62	.73	.74	.62
Measures on SEM	.63	.65	.68	.58
Measures on CPM	.61	.65	.70	.59
Log of Factors on SEM	.60	.73	.71	.59
Log of Factors on CPM	.62	.73	.74	.63
Factors on log of SEM	.60	.79	.73	.61
Factors on log of CPM	.60	.79	.73	.64
Log of Factors on log of SEM	.60	.75	.73	.61
Log of Factors on log of CPM	.60	.79	.73	.64
Log of Measures on SEM	.63	.69	.65	.57
Log of Measures on CPM	.61	.68	.69	.58
Measures on log of SEM	.65	.72	.72	.62
Measures on log of CPM	.60	.73	.71	.62
Log of Measures on log of SEM	.64	.73	.72	.60
Log of Measures on log of CPM	.60	.73	.70	.61
N	39	39	39	39
R for 0.05 Stat. Sign. Level	.48	.48	.48	.48

NOTE: Transformation used for logarithmic cases was log (X+1).

The sizes of the multiple correlations vary with the conditions, such as sector and density and hour and day. The multiple correlations of the corresponding primary measures are quite similar to those for the factor scores. The SEM I multiple correlations based on 2 hours of data for the factor scores with the SEM and CPM ratings range through the .40's and .50's for the most part. The SEM II R's based on 2 hours of data are generally in the .60's. The day level R's, based on 4 hours of data, run in the 60's and 70's, and sometimes higher. The sizes of multiple correlations which meet the .05 level of statistical significance for these sample sizes and numbers of variables are shown in the tables; some of the correlations do not meet these levels, at least in the SEM I data. However, the multiple correlations can be considered good for behavioral data, particularly in the SEM II day-level data.

Let us look at some analogous results from similar experiments. In the 1969 experiment (reference 2), the 2-hour data correlated with the observer ratings at about .17 to .48, and multiple correlations (R's) were about .45. Boone (references 5,6) did not do individual correlations but found R's of about .53 between objective measures in combination and over-the-shoulder ratings by instructors.

In general, it appears that there is a good relationship between the objective measures taken in the present studies and the subjective ratings when the objective measures are taken in combination. The high relationships (around .70) for the day-level data are noteworthy.

IMPLICATIONS. The important issue here was whether there was some reasonable agreement between the objective performance measures taken in simulation and what a controller would think from watching the run. The answer is in the affirmative.

PRACTICE AND LEARNING EFFECTS IN ATC SIMULATION EXPERIMENTS

ANALYSIS. The SEM II data, in addition to fulfilling its major purpose of studying the stability of a group of measurements used to quantify simulation performance, also provided information on the effects of learning during dynamic ATC simulation experiments. The extent to which the process of familiarization and/or learning in the air traffic control simulation environment affects the measurements taken has usually been assumed to be slight since controllers are already well-trained and are "used to" air traffic control. The 12 hours of SEM II runs can be regarded as a course of training, or at least practice, since all other things were the same; system changes were not being made and the traffic samples were being changed only slightly.

The experiment was carefully designed to minimize and eliminate any effect of traffic sample differences while at the same time eliminating both actual extreme simple repetition of traffic samples and any possible sequence effects of different traffic samples.

The major techniques used to accomplish this were the design of the traffic samples and the utilization of latin square counterbalancing. There were four

traffic samples in all, and these were repeated three times by each subject. One of the samples was repeated three times without any change, except in the aircraft identities. The other three samples were based on the first and differed from it only in that the starting times of the individual aircraft were shuffled slightly (three times to make the three samples). The same basic aircraft appeared in all samples at about the same entry time and the number of aircraft scheduled to be present was kept approximately the same throughout the 1-hour planned exercise (after the traffic buildup). Aircraft identities for these latter three samples were also changed on each of the 3 days. These three samples were arranged in a latin square to counterbalance any effects they might have. The samples were given to four subgroups of the subjects in four different orders in accordance with the latin square. It was felt that since the samples were so similar and were balanced across subjects that any effects they or their order of administration might have would be nullified by the experimental design. The experimental design is shown in detail in figure 8 above.

Curves indicating the time courses of the measures over the 12 hours are shown in figure 12. Plots are presented for the means and standard deviation of the factor scores and the primary and auxiliary measures. These curves are based on the 24 subjects who missed no runs whatever. As can be seen there were large changes between the first and fourth runs, and comparative stabilization thereafter. Because of the experimental design, traffic samples and orders are balanced in these curves.

An analysis of variance confirmed that there were differences among the 12 time periods, as was seen in the graphs, for almost all measures. Prior to the analysis of variance, the test for symmetry was done and, as may be seen in the table, the conservative degrees of freedom were used when needed. The analysis appears in table 18.

An orthogonal components test was done to see at about what run levelling off occurred. This appears in table 19 for the plotted measures. For most measures, levelling off occurs by the fifth or sixth hour.

Table 20 shows the percentages of variance due to persons and hours. The technique is from Gaebelin and Soderquist (reference 8). It is of interest here in that it shows that although the variation due to practice is considerable, in most variables the variation due to individual differences among controllers is nonetheless greater, and also that individuals differ somewhat in their reaction to practice, as is indicated by the interaction variance.

The next analysis asks if the data ever did reach an asymptote. It seems from the plots of the successive hours that it did, but there is a danger that if one looks only at the day-level data, the erroneous conclusion could be reached that it is headed further down. For this reason, the plots and analysis of the data considered at the day level are of interest. The 3 day level averages are plotted in figure 13, and the analysis of variance table for these plotted means is presented in table 21. Also shown in the analysis of variance table is the critical difference for Tukey's HSD test (see appendix C for

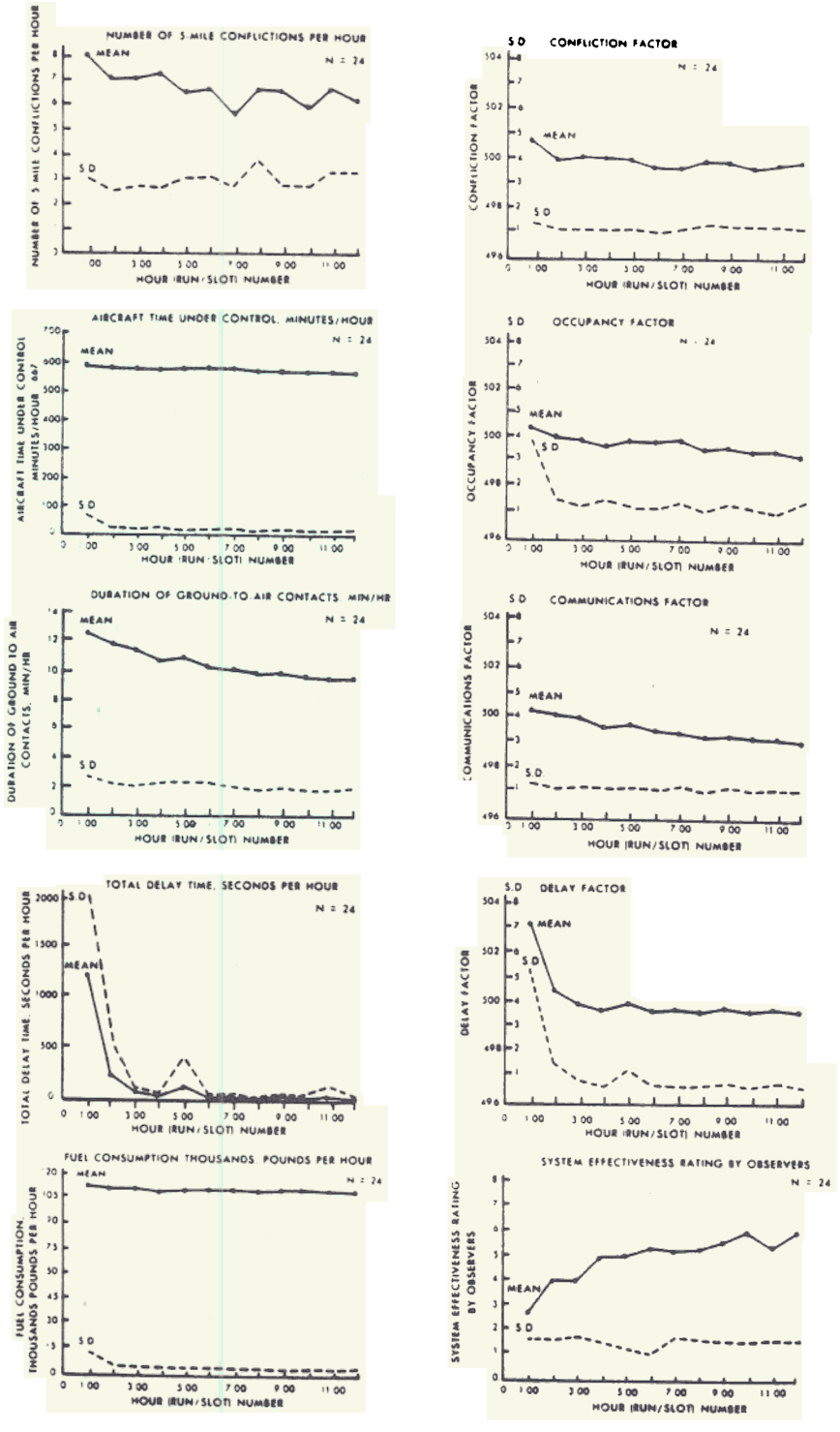


FIGURE 12. PLOT OF COURSE OF MAJOR MEASURES OVER TIME

TABLE 18. ANALYSIS OF VARIANCE TABLE: HOURS

Measure	Compound Symmetry Test Probability	F (F .05, 11/253=1.84)	df	2-Tailed Probability	Conservative Test df	Requirement and Outcome Of Conservative Test (CT) (F .05, 1/23 = 4.30)
Confliction Factor	0.1096	4.34	11/253	0.00	--	CT not required, significant
Occupancy Factor	0.0000	1.21	11/253	0.28	1/23	CT required, not significant
Communication Factor	0.0000	15.91	11/253	0.00	1/23	CT required, still significant
Delay Factor	0.0000	9.43	11/253	0.00	1/23	CT required still significant
Confliction (5 mi.)	0.3022	1.64	11/253	0.09	--	CT not required, not significant
Time Under Control	0.0000	0.92	11/253	0.52	1/23	CT required, not significant
Duration Ground-Air Contacts	0.0000	21.99	11/253	0.00	1/23	CT required, still significant
Total Delay Time	0.0000	7.50	11/253	0.00	1/23	CT required, still significant
No. of A/C Handled	0.0000	2.57	11/253	0.00	1/23	CT required, not significant
Fuel Consumption	0.0000	1.49	11/253	0.13	1/23	CT required, not significant

TABLE 19

ORTHOGONAL ANALYSIS: SUCCESSIVE SIMULATION HOURS

Comparison	Confliction Factor F	Confliction Factor P/.05	Occupancy Factor F	Occupancy Factor P/.05	Communication Factor F	Communication Factor P/.05	Delay Factor F	Delay Factor P/.05
Hour 1 vs rest	32.52	.00	6.12	.01	61.94	.00	97.43	.00
Hour 2 vs rest	0.78	.38	1.23	27	43.73	.00	4.87	.03
Hour 3 vs rest	3.67	.06	0.86	.35	33.35	.00	.42	.52
Hour 4 vs rest	3.18	.08	0.01	.93	6.24	.01	.02	.89
Hour 5 vs rest	2.34	.13	0.98	.32	19.08	.00	.65	.42
Hour 6 vs rest	.66	.42	0.98	.32	4.78	.03	.02	.88
Hour 7 vs rest	1.65	.20	2.20	.14	2.96	.09	.01	.94
Hour 8 vs rest	.55	.46	.06	.80	.09	.77	.04	.84
Hour 9 vs rest	.52	.47	.50	.48	1.66	.20	.11	.74
Hour 10 vs rest	1.50	.22	.08	.78	.42	.52	.02	.89
Hour 11 vs rest	.35	.56	.29	.59	.43	.51	.05	.82

*This test compares the first hour's value to the mean of the last 11 values the second hour's value to the mean of the last 10 values, etc. It is concluded that the values have stabilized when the difference is not significant at the .05 level.

TABLE 19 (CONTINUED)

ORTHOGONAL ANALYSIS: SUCCESSIVE SIMULATION HOURS

Comparison	Number of Conflictions		Time Under Control		Duration G/A Communications		Total Delay Time	
	F	P/.05	F	P/.05	F	P/.05	F	P/.05
Hour 1 vs rest	7.75	.01	4.07	.04	105.61	.00	79.44	.00
Hour 2 vs rest	1.02	.31	0.58	0.45	57.24	.00	2.46	.12
Hour 3 vs rest	1.45	.23	0.28	0.60	37.74	.00	.06	.81
Hour 4 vs rest	3.22	.07	0.05	.83	9.63	.00	.01	.93
Hour 5 vs rest	.16	.68	.74	.39	22.44	.00	.49	.48
Hour 6 vs rest	.55	.46	1.16	.28	4.71	.03	.00	.48
Hour 7 vs rest	2.07	.15	2.46	0.12	2.54	.11	.00	.97
Hour 8 vs rest	.26	.61	0.08	0.77	.28	.60	.01	.92
Hour 9 vs rest	.22	.64	0.26	0.61	1.55	.21	.01	.94
Hour 10 vs rest	.81	.37	0.07	0.79	.13	.72	.01	.93
Hour 11 vs rest	.52	.47	0.36	0.55	.01	.91	.02	.90

TABLE 19 (CONTINUED)

ORTHOGONAL ANALYSIS: SUCCESSIVE SIMULATION HOURS

Comparison	No. A/C Handled F	P/.05	Fuel Consumption F	P/.05
1 vs rest	24.82	.00	7.89	.01
Hour 2 vs rest	.47	.49	1.66	.20
Hour 3 vs rest	.00	.97	1.94	.17
Hour 4 vs rest	.58	.45	0.0	.97
Hour 5 vs rest	1.60	.21	0.66	.42
Hour 6 vs rest	.05	.81	1.79	.18
7 vs rest	.02	.89	1.56	.21
8 vs rest	.03	.87	0.06	.80
9 vs rest	.05	.83	0.35	.55
10 vs rest	.60	.44	0.42	.52
Hour 11 vs rest	.07	.79	0.04	.85

TABLE 20. PERCENT OF VARIANCE DUE TO HOURS AND PERSONS

	PERCENT OF VARIANCE DUE TO:		
	Persons	Hours	Interaction
Factor Scores:			
Conflict Factor	37	8	55
Occupancy Factor	30	1	69
Communication Factor	66	12	22
Delay Factor	7	24	69
Primary Measures:			
Number of Conflictions	39	2	59
Time Under Control	30	0	70
Duration of Ground-Air Com	63	17	20
Total Delay Time	7	19	73
Auxiliary Measures:			
Number of Aircraft Handled	9	6	85
Fuel Consumption	33	1	66

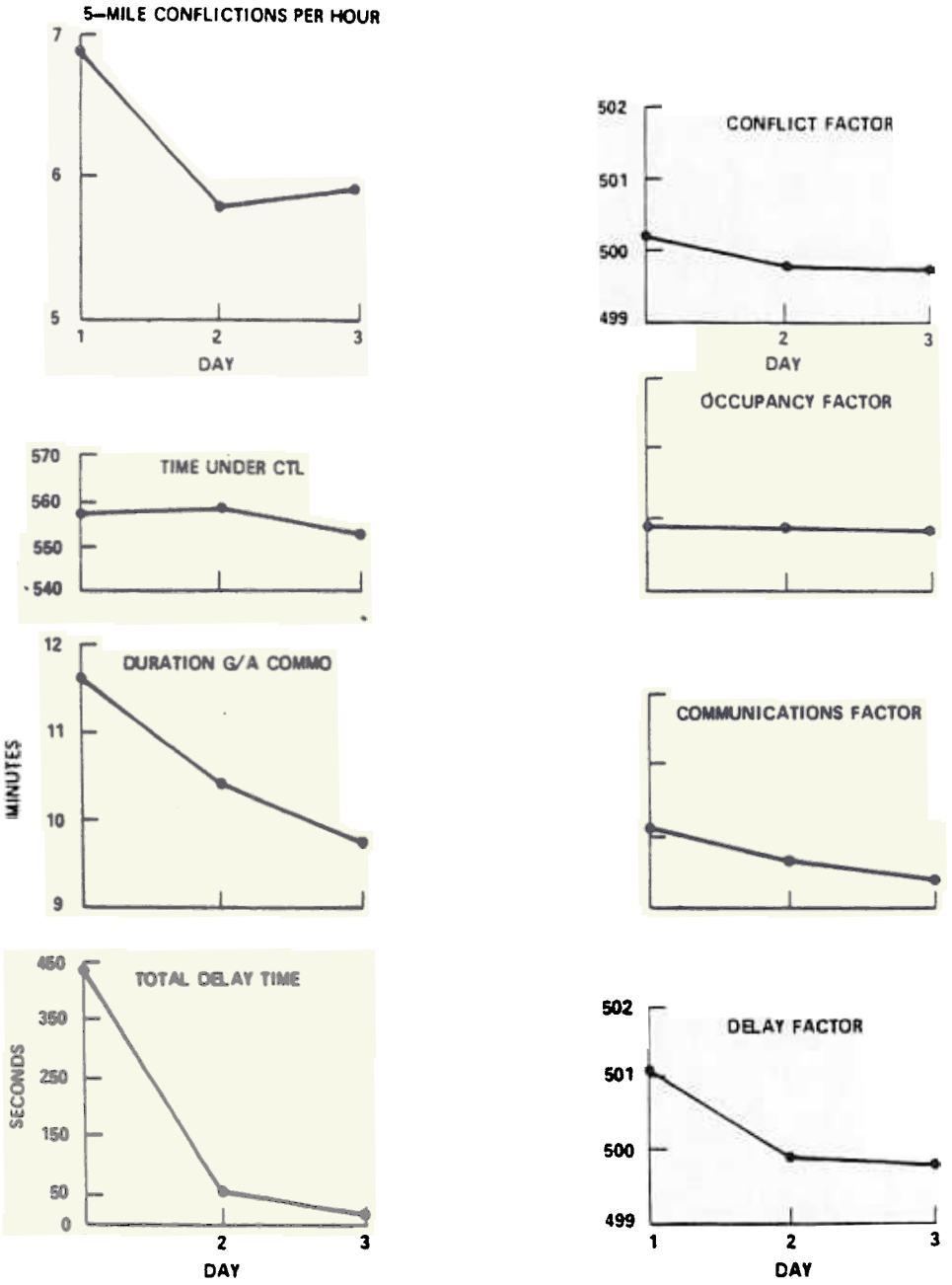


FIGURE 13. PLOT OF DAY MEANS

TABLE 21. ANALYSIS OF VARIANCE TABLE: DAYS

	MS Days	MS Subjects	MS Error	df Days	F *** Days	P Days	Tukey's HSD	Day Differences ** 1-2	1-3	2-3
Confliction Factor	2.313	1.017	0.170	2/76	13.62	0.00	0.224	0.394	0.446	0.052
Occupancy Factor	1.300	1.948	0.635	2/76	2.05	0.136	*	*	*	*
Communication Factor	5.399	1.799	0.120	2/76	44.89	0.00	0.189	0.459	0.737	0.279
Delay Factor	20.278	1.105	0.907	2/76	22.36	0.00	0.619	1.199	1.293	0.094
Confliction (5 mi.)	13.778	14.391	2.000	2/76	6.89	0.002	0.770	1.077	0.974	0.103
Time Under Control	1.18×10^6	1.87×10^6	5.97×10^5	2/76	1.98	0.146	*	*	*	*
Duration Ground-Air Contacts	1.32×10^5	3.23×10^4	2.00×10^3	2/76	66.17	0.00	24.327	73.97	114.83	40.86
Total Delay Time	2.13×10^6	1.51×10^5	1.34×10^5	2/76	15.92	0.00	199.037	384.9	421.7	36.77
No. of A/C Handled	1.488	0.181	0.133	2/76	11.19	0.00	0.199	0.316	0.357	0.041

* = Basic test not significant; not to be tested for HSD.
** = Underlined differences are significant at .05 level according to HSD test.
*** = Ordinary df: 2/76, F/.05= 3.15; Conservative df: 1/38, F/.05= 4.08. No F's affected.

explanation). The underlined differences are significant at the five percent level. From the Tukey test, it is apparent that the differences involving the first day are those which result in significant differences between days, whereas in most measures the differences between the second and third days are not significant. This would seem to indicate that stabilization occurs after the first day in most cases. Table 22 gives the percent of variance attributable to days (not hours this time) and persons, and, finally, the day means themselves are shown in table 23.

IMPLICATIONS. It has been shown that:

There is in general a massive learning effect of the first 4 runs in this type of experiment. The best procedure, then, for the usual simulation experiment, would be the provision of 2 hours of familiarization plus about 4 runs in each experimental condition of importance before beginning to save data.

THE EFFECTS OF SECTOR GEOMETRY AND DENSITY ON SYSTEM PERFORMANCE MEASUREMENTS.

ANALYSIS. One of the persistent problems in approaching the planning and execution of an experiment utilizing real-time simulation to compare systems or concepts for the en route air traffic control system is the selection of a particular sector and traffic density level to use in the experiment. These two aspects of the stimulus situation which the system, however large or small, will face may have some impact on the outcome of the experiment. Unless we have some knowledge of their effects, we have an area of ignorance which will impede our planning, execution, and interpretation of all experimental system evaluations required in the future.

Frequently, for example, it is necessary to repeat experimental sessions with the same controllers. If we could say that the geometric shape of the sector chosen had no real impact, then we could use sectors interchangeably in the various experimental system modifications, thus avoiding boredom and extreme practice effects. If the level of difficulty of different sector-density combinations did not differ much, then these could be considered as parallel forms of a test and used interchangeably, or one standard sector could be used for all experiments, and sampling several sectors need not be considered.

The SEM I experiment was designed to explore these issues, among others. Its design (figure 7) involved two sectors and three traffic densities. The sectors were chosen to represent two extremely different geometries; one was quite long and narrow, the other was almost circular. Controllers were asked to select two contrasting sector shapes. The traffic levels were chosen such that the planned number of aircraft present to the controller at all times was the same over the time course of the problems, and the same in both sectors. The three density levels were defined in terms of the number present at all times, in the planned traffic sample. Three density levels, roughly representing, in controller opinion, low, medium and high difficulty levels for

TABLE 22

PERCENT OF VARIANCE DUE TO DAYS AND PERSONS

	Percent of Variance Due to:		
	Persons	Days	Interaction
Factors:			
Confliction Factor	60	10	30
Occupancy Factor	50	1	49
Communication Factor	70	16	14
Delay Factor	21	28	51
Primary Measures:			
No. of Conflictions (5)	68	4	28
Time Under Control	50	1	48
Duration of G/A Comm.	67	21	12
Total Delay Time	22	22	57
Auxiliary Measures:			
No. of A/C Handled	26	15	58
Fuel Consumption	51	2	47

TABLE 23. DAY MEANS

	DAY 1	DAY 2	DAY 3
Factor Scores:			
Conflict Factor	500.200	499.807	499.755
Occupancy Factor	499.916	499.946	499.616
Communication Factor	500.189	499.730	499.452
Delay Factor	501.113	499.913	499.820
Primary Measures:			
Number of Conflictions/hr.	6.9	5.8	5.9
Time Under Control, sec./hr.	33,410	33,508	33,170
Duration of Ground-Air Com., sec./hr.	699	625	584
Total Delay Time, sec./hr.	443	58	21
Auxiliary Measures:			
Number of Aircraft Handled/hr.	46	47	47
Fuel Consumption, lbs./hr.	13,403	111,770	10,922

our planned single controller "teams" were chosen. Each controller began on one of the two sectors after considerable verbal orientation and one or two practice runs. Half of the subjects began with one of the sectors and half began with the other sector. Each did a low, medium and high density traffic hour, repeated that sequence in the same sector, and then went to other sector and did the same. About four 1-hour runs were done each day.

Entering the evaluation, the expectation was that sector geometry as such would make little difference, because the number of aircraft simultaneously present in each of the two sectors had been set to be about the same. This, it was thought, especially since very extreme geometries had been chosen in the first instance, would allow acceptance of the principle that sector geometry as such made very little difference, if traffic level were controlled. Establishment of this principle, it was felt, would simplify the decisions to be made by future experimenters in arranging traffic samples for system evaluations.

The reduction, which was discussed earlier, of the number of measures to be examined makes the task of examining the data considerably more feasible and bearable than it would have been without that reduction.

The analysis used followed the experimental design and was a repeated measures analysis of variance performed on each of the measures to be examined. These were the four factor scores, the four primary scores, the number of aircraft handled, and the fuel consumption model index. The data for 27 subjects were available for use in this particular analysis.

The analysis of variance table is presented in table 24 for the ten measures mentioned above. The major fact to note is that in all ten measures the interaction between sector and density is statistically significant, at the .05 level. It is plain that traffic density always is a significant factor, as was clearly expectable. Also, in all but two of the ten measures, there is a significant effect of sector geometry, and even these two measures approach significance, being significant at the .09 and .11 levels. The Greenhouse-Geiser (see appendix C) conservative degrees of freedom, which probably are appropriate here, were examined and it was seen that their use would not impact the interpretation of significance.

The major factor worthy of attention is the interaction which we have seen. While this was not the expected outcome, it can be just as useful in assisting the planning of system tests. The interaction can be seen visually by looking back at figure 11. In that figure, it can be seen that for the measure sector occupancy, for example, scores were rather similar as to location of their distributions on our common scale for Geometry 1-Density 2 and Geometry 2-Density 3. Similar equivalence points could be empirically found for other measures. This means that a way has been shown, although not fully developed, to generate problems of equivalent, and thus interchangeable, difficulty.

TABLE 24

ANALYSIS OF VARIANCE TABLE: SECTOR AND DENSITY

Test Measure	Geometry F	Geometry df	Geometry P	Density F	Density df	Density P	Geometry by Dens. F	Geometry by Dens. df	Geometry by Dens. P
Confliction Factor	5.51	1/26	.027	46.09	2/52	.00	11.65	2/52	.00
Occupancy Factor	462.28	1/26	.00	2846.90	2/52	.00	206.67	2/52	.00
Communications Fac.	89.51	1/26	.00	511.52	2/52	.00	61.02	2/52	.00
Delay Factor	39.51	1/26	.00	82.64	2/52	.00	46.41	2/52	.00
Confliction (5 mi.)	3.12	1/26	.085	82.48	2/52	.00	13.91	2/52	.00
Time Under Control	71.98	1/26	.00	1313.51	2/52	.00	71.68	2/52	.00
Duration Ground-Air Contacts	54.85	1/26	.00	503.20	2/52	.00	66.60	2/52	.00
Total Delay Time	2.72	1/26	.11	43.26	2/52	.00	15.35	2/52	.00
No. of A/C Handled	117.25	1/26	.00	6785.20	2/52	.00	73.15	2/52	.00
Fuel Consumption	532.62	1/26	.00	1858.60	2/52	.00	302.92	2/52	.00

The sector-density interaction was significant in all of the measures. For this reason, the averages for the six cells rather than for the two sectors and the three densities, separately considered, are given in table 25. For the factor scores, the averages are given on the common scale and are given in raw score form for the other major measures.

Table 26 presents similar information but in a different way. It presents the percentage of variance due to the major dimensions of the analysis of variance. In this case, these source dimensions are sector and density, their interaction, and the individual differences due to controllers.

As to the sources of variance generation, the obvious expectation was that the extremes of traffic density used here would generate the most difference in the scores, with individual differences in the performance of the sample of controllers being the next largest source, and geometry coming last. Of course, the facts are not that simple. There is complex interaction involved, and the results are not the same for all of the measures. It is true, for example, that the traffic density levels used here do generate between 20 and 60 percent of the variance or more in the cases of most of the ten measures. About as often as not, however, geometry outweighs the effect of individual differences among controllers. Again, the interaction between geometry and density is seen to be very important, and the overall interaction is also seen to contain a great deal of the variance.

Another approach to the disentanglement of this area was attempted by examining the correlations between the scores obtained on the various measures by the individual controllers in the several circumstances. It was the thought that the effects of sector and density could be more legitimately minimized in planning experiments if individuals performed about the same in the several sector-density combinations which had been tested. For example, it was thought that the correlation would be higher between geometries at the same traffic density level, than between traffic density levels controlled in the same sector geometry. The data on these two types of correlation: between geometries at a given density and between densities at a given geometry, are presented in tables 27 and 28 respectively.

It is clear that the data again did not follow expectations: the correlations are higher across densities for the same geometry. This might lead us to wonder if geometry should not be considered somewhat more powerful than indicated in the other analyses. However, there may be another explanation. It will be remembered from the discussion of procedure that the subjects did all of their runs on one of the geometries before shifting to the other. Considering the finding of the other (SEM II) experiment about how the correlation between runs decreases with their distance apart in time, it appears possible that this correlation is due to the sequence of executing the runs. At the time SEM I was planned, the sequence seemed the best way to run the experiment, but it probably is responsible for this finding.

There is a more positive aspect to this result, however. This is the fact that these correlations do exist and in some cases are fairly substantial between the performances under different circumstances by the controllers. For example,

TABLE 25

MEAN VALUES IN SECTOR-DENSITY COMBINATIONS

	G1			G2		
	D1	D2	D3	D1	D2	D3
Measures						
Cfl. Factor	49.26	49.77	49.92	49.41	49.37	49.74
Occ. Factor	45.14	49.82	52.04	44.29	46.44	48.99
Com. Factor	47.60	50.03	51.02	47.31	48.24	49.71
Delay Factor	49.06	49.77	51.39	49.22	49.02	49.71
No. of 5-M:Confl./Hr.	1.98	8.82	11.84	4.28	4.64	10.36
Time Under Control Min./Hr.	304.7	507.7	588.3	283.5	392.5	512.9
Dur. A/G Com. Sec./Hr.	476.8	793.7	908.4	483.6	598.8	764.4
Total Delay Time, Sec./Hr.	141.4	658.6	2216.7	442.8	483.6	974.4
No. A/C Handled/Hr.	33.6	49.0	55.1	32.8	49.7	59.1
Fuel Consumption lb./Hr.	59,428	106,645	141,062	46,861	64,266	87,091

NOTE: Data based on 50 minute samples, reduced to hourly rate for measures.

TABLE 26

THE PERCENTAGE OF VARIANCE DUE TO SECTOR AND DENSITY

Measure	Persons	Geometry	Density	Geom. X Dens.	Remaining Interaction
Conflict Factor	7	3	20	11	59
Occupancy Factor	2	23	65	7	3
Communication Factor	8	16	57	8	11
Delay Factor	7	14	28	20	31
No. of 5-Mile Conf.	9	1	34	11	45
Time Under Control	2	11	75	5	7
Dura. G/A Contacts	17	10	53	9	11
Total Delay Time	12	2	17	11	58
No. A/C Hold	0	1	96	2	1
Fuel Consumption	1	27	69	2	1

TABLE 27

CROSS-CONDITION CORRELATIONS: ACROSS GEOMETRY AT A GIVEN DENSITY*

	D-1 G-1/G-2	D-2 G-1/G-2	D-3 G-1/G-2
Factor Scores:			
Conflict Factor	.20	-.09	-.14
Occupancy Factor	.67	.55	.65
Communication Factor	.36	.39	.54
Delay Factor	.04	.02	.30
Primary Measures:			
Number of Conflictions	.41	.10	.14
Time Under Control	.67	.58	.62
Duration of Ground-Air Com.	.64	.61	.71
Total Delay Time	-.15	.01	.01
Auxilary Measures:			
Number of Aircraft Handled	-.02	-.06	.32
Fuel Consumption	.59	+.54	.57

*Two run average

TABLE 28

CROSS-CONDITION CORRELATIONS: ACROSS DENSITY AT A GIVEN GEOMETRY*

	G-1		
	D-1/D-2	D-2/D-3	D-1/D-3
Factor Scores:			
Conflict Factor	-.01	.38	.02
Occupancy Factor	.69	.89	.71
Communication Factor	.73	.82	.64
Delay Factor	.10	.61	-.04
Primary Measures:			
Number of Conflictions	.01	.40	.16
Time Under Control	.87	.93	.86
Duration of Ground-Air Com.	.88	.90	.81
Total Delay Time	-.17	.45	-.18
Auxiliary Measures:			
Number of Aircraft Handled	.29	-.10	-.20
Fuel Consumption	.83	.86	.83

	G-2		
	D-1/D-2	D-2/D-3	D-1/D-3
Factor Scores:			
Conflict Factor	.50	.64	.34
Occupancy Factor	.78	.78	.79
Communication Factor	.71	.63	.49
Delay Factor	.69	.44	.41
Primary Measures:			
Number of Conflictions	.42	.56	.21
Time Under Control	.83	.87	.87
Duration of Ground-Air Com.	.78	.75	.74
Total Delay Time	.86	.60	.51
Auxiliary Measures:			
Number of Aircraft Handled	.03	.11	.28
Fuel Consumption	.74	.79	.79

*Two run average

In the data in table 27 it can be seen that the correlations of the occupancy factor score from sector to sector are .67, .55 and .65 at each of the three traffic densities, and some other correlations are of fair sizes. In table 28, the correlations of the performance scores between the middle and high density levels of traffic are quite high, often above the 50's, for both sectors.

It appears possible that, in a new experiment with more replicates and more care for order effects, there would appear a consistently high correlation between performance scores obtained in several different sector geometries and traffic levels, thus demonstrating a general controller ability factor which could be considered to be independent of specific sector geometry and traffic density level.

IMPLICATIONS. The implications of these data for the design of system tests involving different sectors and traffic densities are:

1. Sector and density are, as expected, important factors in determining the results which will occur in a given experiment, but they interact in a complex way. The nature and extent of this interaction depends on the measures involved. While, on the one hand, this is obviously not startling news, it should make us aware, when reading the reports of system evaluations, that there is no such thing as two traffic density levels which can be called comparable in any terms if they exist in different sector geometries.

2. On the other hand, it appears possible to empirically develop pairs or sets of particular combinations of sector and density that are of equivalent difficulty and so are usable interchangeably in experimentation.

3. There may be a policy implication for controller training if it can be confirmed in further experimentation along these lines that there is a generalized controller ability factor which is measurable and carries across sector geometries and traffic densities. The indication would be that a greater proportion of controller training could be done in a general manner, not bound to a particular sector geography.

STATISTICAL POWER OF REAL TIME ATC SIMULATION EXPERIMENTATION

ANALYSIS. The major purpose of these two experiments was to evaluate the measures used in dynamic air traffic control simulation for their statistical power. Evaluation is used here to mean determining what is necessary for statistically sound conclusions to be made using the data from such experimentation.

The main determinants of statistically sound conclusions are the repeatability of the measures and the extent of individual differences among the subjects serving in the tests. Formulas have been developed to enable the estimation, given the above inputs, of the power of a given kind of experimentation to provide conclusions of a desired level of statistical dependability. Calculations based on the data from the two SEM evaluations have been performed and tables prepared of the statistical power involved in air traffic control simulation using the four factor scores, the four primary measures, the number of aircraft handled and fuel consumption.

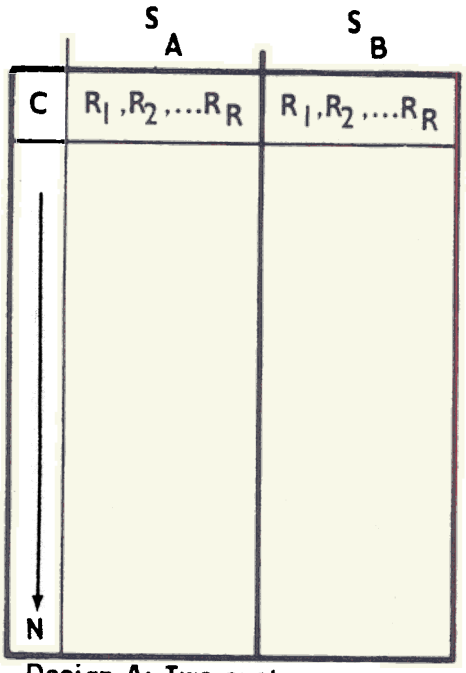
It is not appropriate in this report to go into a detailed basic orientation on the matter of statistical hypothesis testing as it particularly applies in the unique field of real-time simulation testing of air traffic control man-machine systems for effectiveness. In very general terms, it is important to avoid rejecting a system which is an improvement over the present system and accepting a system as the system of the future when it is really not an improvement. It is a matter of dispute as to which is worse, and it varies with the situation. Put slightly differently, if one accepts the hypothesis of no difference between two systems and does so mistakenly, this is a beta error. If one asserts that two systems are different, and does so mistakenly, this is an alpha error. Appendix C gives a further explanation of these error types and references for further reading. A major reference on this subject is the book by Cohen (reference 9).

The power tables can be found in a separate volume, published as an adjunct to this report. Tables are given for the four factor scores and the primary measures. The tables present data on a 1-hour unit run basis. An example of the use of tables in planning tests appears below.

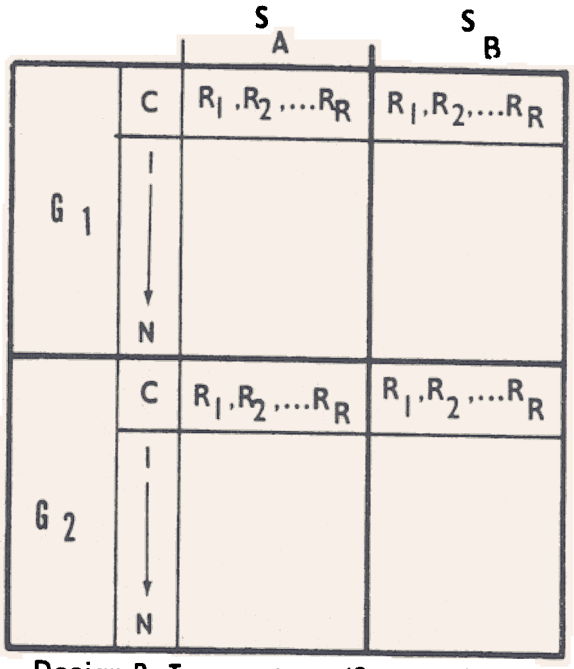
The power tables must be entered with two parameters: (1) the size of the difference in each of the measures which is considered worthwhile detecting in each measure as a meaningful or important difference between systems, and (2) the alpha and beta error probabilities it is felt important to protect against.

The tables are constructed in the case of the factor scores in terms of the previously mentioned third scale. For developing the tables, the data for the SEM I and SEM II factor scores (generated using the SEM II weights) were put on a common scale (based on the SEM II fifth time period's mean and standard deviation) and given a mean of 500 and a standard deviation of 1. The primary measures remained in raw score terms. It will be remembered, though, that because of SEM I data losses, 50 minutes of data were used per run. At this point, these raw measures' run scores were multiplied by 6/5 to bring the 50 minute data to a 1-hour equivalent for the raw scores themselves. The tables used the data from the SEM II runs (60 minutes) for the middle density level table. For the two other densities (very low and very high), the data from both of the SEM I sectors were examined, and worst case values, for example, the sector with the larger standard deviation, were used to estimate the parameters which were used to generate the tables. A separate table is presented for these three cases, and adjustments are presented for combinations of low, medium and high density conditions.

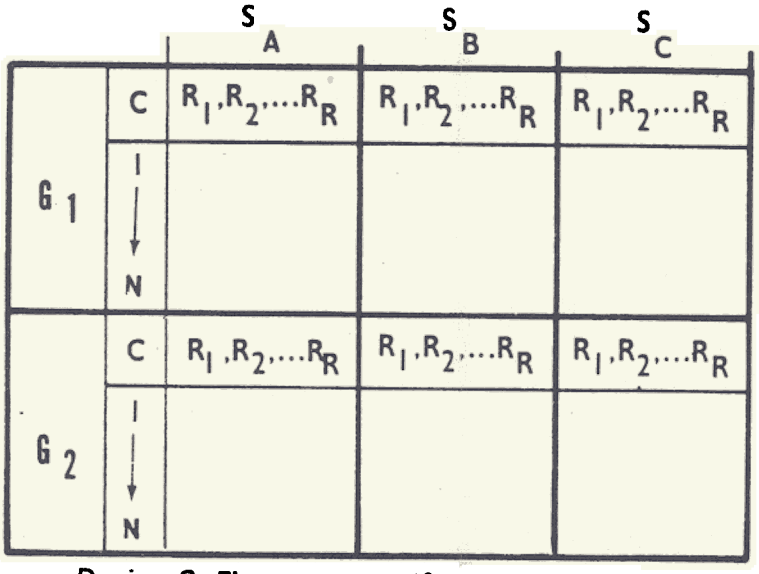
The tables were formulated to be specific to four statistical experimental design (a technical term, see appendix C) types which might be expected to be frequently applicable to system testing. Design A is a paired, or correlated t test design, in which the same controllers are used in both systems at a given density. Design B is a 2 x 2 repeated measures analysis of variance design in which, for example, two types of systems are used in two sectors. Design C is a 2 X 3 repeated measures analysis of variance design in which, for example, three system arrangements might be used in two operational sector geometries. Design D is a design in which the repeated measures (same subjects) approach is not used, but different subjects serve in the two different system arrangements. The four basic designs are shown in figure 14.



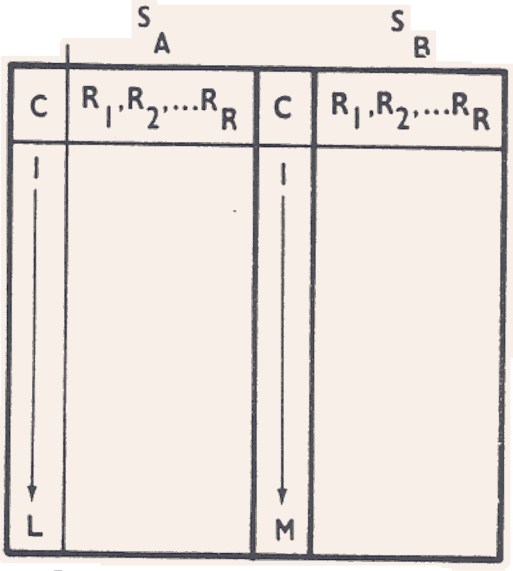
Design A: Two systems (S_A and S_B), several repetition runs (R), and several controllers (or crews) C, same C's work in both systems.



Design B: Two systems (S_A and S_B), several repetition runs (R), two geometries (G_1 and G_2), several controllers or crews (C). Same C's work in all four conditions.



Design C: Three systems (S_A, S_B, S_C), several repetition runs (R), two geometries (G), same controllers work in all six conditions.



Design D: Two systems (S_A and S_B), two different groups of controllers, one group assigned to each system. Several repetitions for each group with their system.

FIGURE 14. THE FOUR BASIC DESIGNS

Obviously, since the tables have been assembled on the basis of the data from the two SEM experiments reported here which were based on single controller sectors, the application of the tables is strictly speaking limited to single controller experiments. However, it is assumed that many important questions can be attacked effectively and efficiently using only one sector, particularly with reference to human factors and man-machine interface issues, and not with a requirement for "a cast of thousands." This can be done if the functions and interactions with adjacent sectors are adequately and efficiently represented, in a manner similar to that used in the SEM experiments.

On the other hand, it is important to point out that the power tables can also be useful in a more limited way for planning simulation evaluations involving multi-person teams operating a single sector and in multi-sector system situations. In such cases, the main difference which would affect the tabled values would probably be a larger extent of differences among multiple-person teams (the variance), as distinguished from individual controller "teams," and an even larger variance among multi-person teams working in multiple sector systems. The effect of these presumably larger variances would be that the power of the measures would be less than that appearing in the tables, as they are based on smaller variance parameters. And so the tables in their current form can be used to get an optimistic estimate of the experimental power that must be reckoned with in the planning process.

The following example is presented to illustrate the method of use of the power tables in planning single sector air traffic control simulation experiments (as described above).

Suppose an experimenter plans to compare two ATC systems in two sector geometries at the middle traffic density. For the sake of discussion, the assumption is made that ATC system A is the present sector arrangement or computer functional role assignment and that ATC system B is a proposal which is claimed to reduce the number of conflictions. The experimenter establishes the null hypothesis to be tested as that the number of conflictions finally occurring will be equal for the two systems, that is, there will be no statistically dependable (significant) difference. (Also considered in other hypotheses will be the effects of traffic density and of the interactions involved.)

The experimenter will now proceed to study the following variables:

alpha: the probability of Type I error, that is the error wherein the null hypothesis is rejected when in fact System A = System B.
beta: the probability of Type II error, that is the error wherein the null hypothesis is accepted when in fact System A is different from System B. (The power of the test is the obverse of the beta error (1-) that is, the probability that the null hypothesis will be correctly rejected. The tables involve power in that they ask the planner of an experiment to choose a beta error level appropriate to the test situation.)
delta: the minimum difference it is felt necessary to detect in the measure under study between the two systems.
N: the number of subjects.

Power calculations are a systematic method of analysing the trade-offs of these four variables. The experimenter may choose to set the acceptable chance of

alpha and beta error at .05 and .10, respectively. Then, the major analysis is between the minimum detectable difference required to reject the null hypothesis and the number of experimental runs and subjects (N) required to detect this difference between the systems.

The appropriate design for this example is a 2 x 2 repeated measures analysis of variance with alpha = .05 and beta = .10. The table for this design and these probabilities and for the confliction measure at middle density is given as table 29. If the experimenter wishes to detect a difference between systems of 2 or more conflictions, the number of subjects needed will depend on the number of hours of testing that can economically be conducted using the same people. For example, travel and other economic considerations may come into this decision. The determination of the tradeoff between repetitions (also called replicates, shown between 1 and 4 hours of running in the table) and the number of subjects (N) would be made using the table in the manner summarized below.

If alpha= .05, beta= .10, delta= 1.9, then:

		Number of Subjects
	1	14
Number of	2	11
Replicates	3	10
	4	10

Having made this calculation the experimenter would now know the subject hours and simulator hours necessary to meet his goals. The alternatives are to guess and have either too many hours of testing or too few to meet the goals.

Figure 15 shows how the detectablity of differences varies as a function of the number of subjects, the amount of replication, and the error levels set for one of the measures. This differs with the design used and with the particular measure involved. Table 30 points out the fact that the four factor scores differ in power and not always in direct proportion to their reliability. Figure 16 gives the overall structure of the power tables.

IMPLICATIONS. There are some critical implications of this rather academic discussion:

1. The estimates of power given in the tables depend on the input data from the SEM experiments. If further work can improve the estimates of the parameters, such as the reliability coefficients over the current values as estimated by the SEM experiments, more economical experimentation would be possible.

2. If some approach resembling this one is not taken, then one is left to fall back on operational judgement as to what is to be the system decision taken as the outcome of a system test, and opinions differ. An even worse alternative, though, is experimentation wherein objective measures are duly collected but interpreted as if they were physical data with no variability and rather perfect repeatability. This, in fact, depends upon sheer chance. Another alternative has happened at times which is equally painful for those involved.

TABLE 29. POWER TABLE EXAMPLE

CB-05 TSA B5D - TARGET SPACING ANALYSIS (5 MILES, 950 FEET)

POWER TABLES FOR THE MAIN EFECTS IN A 2X2 FACTORIAL REPEATED MEASURES ANOVA

ALPHA=.05

	BETA=.20				BETA=.10				BETA=.05				BETA=.01			
	1	2	3	4	1	2	3	4	1	2	3			2	3	4
N= 5	3.7	3.2	3.0	2.9	4.3	3.7	3.5	3.3	4.8	4.1	3.9	3.7	5.8	5.0	4.7	4.5
N= 6	3.2	2.7	2.5	2.4	3.7	3.1	2.9	2.8	4.1	3.5	3.3	3.1	4.9	4.2	3.9	3.8
N= 7	2.8	2.4	2.2	2.2	3.2	2.8	2.6	2.5	3.6	3.1	2.9	2.8	4.3	3.7	3.4	3.3
N= 8	2.5	2.2	2.0	2.0	2.9	2.5	2.3	2.3	3.2	2.8	2.6	2.5	3.9	3.3	3.1	3.0
N= 9	2.3	2.0	1.9	1.8	2.7	2.3	2.2	2.1	3.0	2.6	2.4	2.3	3.5	3.0	2.8	2.7
N=10	2.2	1.9	1.7	1.7	2.5	2.1	2.0	1.9	2.8	2.4	2.2	2.1	3.3	2.8	2.6	2.5
N=11	2.1	1.8	1.6	1.6	2.4	2.0	1.9	1.8	2.6	2.2	2.1	2.0	3.1	2.6	2.5	2.4
N=12	1.9	1.7	1.6	1.5	2.2	1.9	1.8	1.7	2.5	2.1	2.0	1.9	2.9	2.5	2.3	2.2
N=13	1.9	1.6	1.5	1.4	2.1	1.8	1.7	1.6	2.4	2.0	1.9	1.8	2.8	2.4	2.2	2.1
N=14	1.8	1.5	1.4	1.4	2.0	1.7	1.6	1.6	2.2	1.9	1.8	1.7	2.6	2.3	2.1	2.0
N=15	1.7	1.5	1.4	1.3	2.0	1.7	1.6	1.5	2.2	1.8	1.7	1.7	2.5	2.2	2.0	2.0
N=16	1.6	1.4	1.3	1.3	1.9	1.6	1.5	1.5	2.1	1.8	1.7	1.6	2.4	2.1	2.0	1.9
N=17	1.6	1.4	1.3	1.2	1.8	1.6	1.5	1.4	2.0	1.7	1.6	1.5	2.4	2.0	1.9	1.8
N=18	1.5	1.3	1.2	1.2	1.8	1.5	1.4	1.4	1.9	1.7	1.6	1.5	2.3	1.9	1.8	1.8
N=19	1.5	1.3	1.2	1.1	1.7	1.5	1.4	1.3	1.9	1.6	1.5	1.4	2.2	1.9	1.8	1.7
N=20	1.4	1.2	1.2	1.1	1.7	1.4	1.3	1.3	1.8	1.6	1.5	1.4	2.1	1.8	1.7	1.7

CB-05 TSA B5D - TARGET SPACING ANALYSIS (5 MILES, 950 FEET)

POWER TABLES FOR THE MAIN EFECTS IN A 2X2 FACTORIAL REPEATED MEASURES ANOVA

ALPHA=.01

	BETA=.20				BETA=.10				BETA=.05				BETA=.01			
	1	2	3	4	1	2	3		1	2	3	4	1	2	3	4
N= 5	5.8	5.0	4.7	4.5	6.7	5.7	5.4		7.4	6.4	6.0	5.7	8.9	7.6	7.1	6.8
N= 6	4.7	4.0	3.7	3.6	5.3	4.6	4.3		5.9	5.0	4.7	4.5	7.0	5.9	5.6	5.4
N= 7	4.0	3.4	3.2	3.1	4.5	3.9	3.6		5.0	4.3	4.0	3.8	5.9	5.0	4.7	4.5
N= 8	3.5	3.0	2.8	2.7	4.0	3.4	3.2		4.4	3.8	3.5	3.4	5.1	4.4	4.1	4.0
N= 9	3.2	2.7	2.6	2.5	3.6	3.1	2.9		4.0	3.4	3.2	3.1	4.6	3.9	3.7	3.6
N=10	2.9	2.5	2.4	2.3	3.3	2.8	2.7		3.6	3.1	2.9	2.8	4.2	3.6	3.4	3.3
N=11	2.7	2.3	2.2	2.1	3.1	2.6	2.5		3.4	2.9	2.7	2.6	3.9	3.4	3.1	3.0
N=12	2.6	2.2	2.1	2.0	2.9	2.5	2.3		3.2	2.7	2.5	2.4	3.7	3.1	2.9	2.8
N=13	2.4	2.1	1.9	1.9	2.7	2.3	2.2		3.0	2.6	2.4	2.3	3.5	3.0	2.8	2.7
N=14	2.3	2.0	1.9	1.8	2.6	2.2	2.1		2.8	2.4	2.3	2.2	3.3	2.8	2.6	2.5
N=15	2.2	1.9	1.8	1.7	2.5	2.1	2.0		2.7	2.3	2.2	2.1	3.1	2.7	2.5	2.4
N=16	2.1	1.8	1.7	1.6	2.4	2.0	1.9		2.6	2.2	2.1	2.0	3.0	2.6	2.4	2.3
N=17	2.0	1.7	1.6	1.6	2.3	2.0	1.8		2.5	2.1	2.0	1.9	2.9	2.5	2.3	2.2
N=18	2.0	1.7	1.6	1.5	2.2	1.9	1.8		2.4	2.1	1.9	1.9	2.8	2.4	2.2	2.1
N=19	1.9	1.6	1.5	1.5	2.1	1.8	1.7		2.3	2.0	1.9	1.8	2.7	2.3	2.1	2.1
N=20	1.8	1.6	1.5	1.4	2.1	1.8	1.7		2.3	1.9	1.8	1.7	2.6	2.2	2.1	2.0

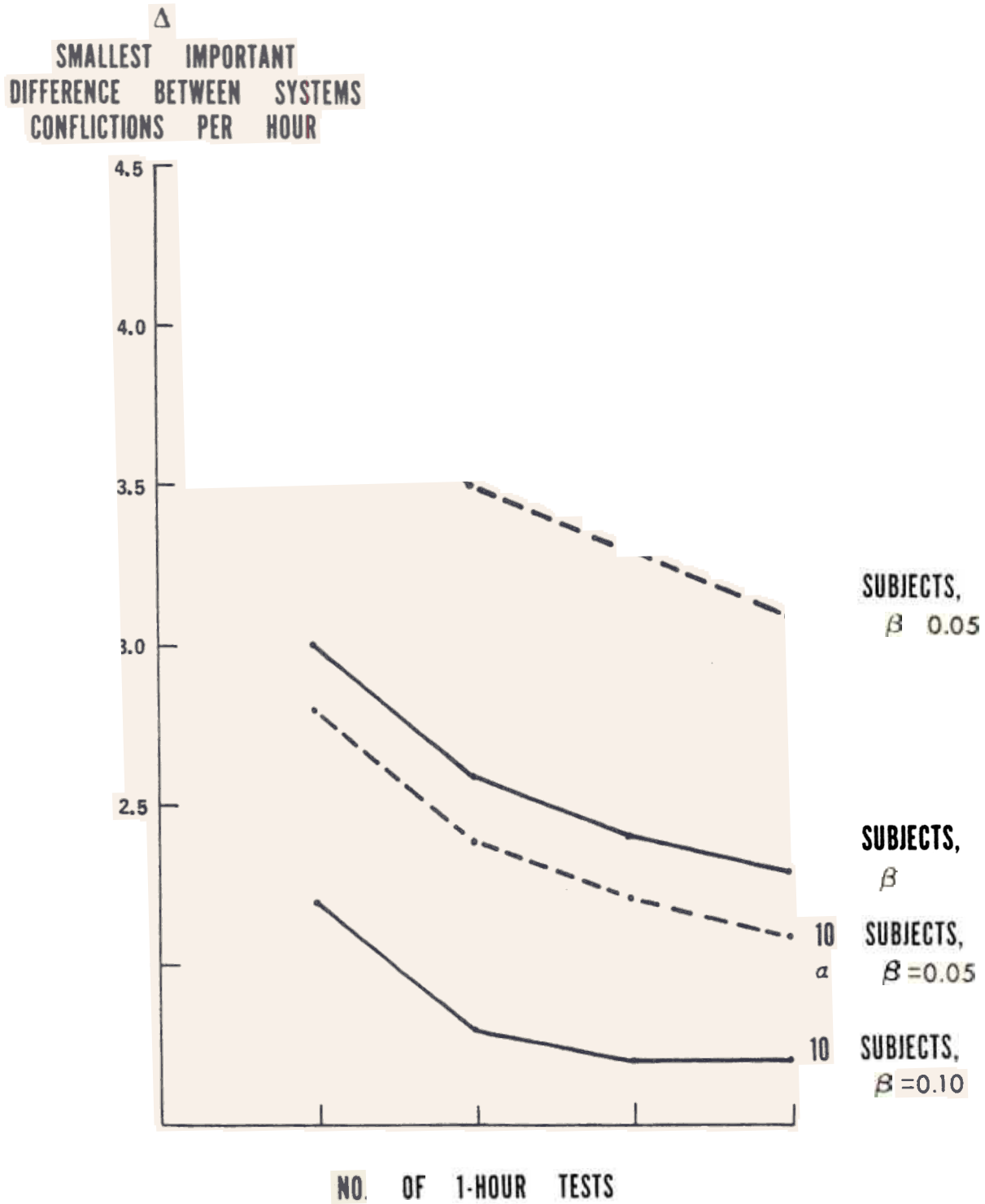


FIGURE 15. GRAPH POWER

TABLE 30. COMPARATIVE STATISTICAL POWER OF THE FOUR FACTOR SCORES

Paired t test (alpha=.10,beta=.01) one hour runs

NUMBER OF CASES	CONFLICT FACTOR	OCCUPANCY FACTOR	COMMUNICATION FACTOR	DELAY FACTOR
5	2.1	2.3	1.3	1.8
6	1.8	2.0	1.1	1.6
7	1.6	1.8	1.0	1.4
8	1.4	1.6	0.9	1.3
9	1.3	1.5	0.8	1.2
10	1.2	1.4	0.8	1.1
11	1.2	1.3	0.7	1.0
12	1.1	1.3	0.7	1.0
13	1.1	1.2	0.7	0.9
14	1.0	1.1	0.6	0.9
15	1.0	1.1	0.6	0.9
16	0.9	1.1	0.6	0.8
17	0.9	1.0	0.6	0.8
18	0.9	1.0	0.6	0.8
19	0.8	1.0	0.5	0.7
20	0.8	0.9	0.5	0.7
correlation *	.39	.44	.77	.07
standard dev. *	.7413	1.0296	.7742	0.3781

* based on running averages.

FOUR BASIC DESIGNS	Paired t Test 2 x 2 Repeated Measures ANOVA 2 x 3 Repeated Measures ANOVA Separate t Test	
TEN MEASURES	Confliction Factor Occupancy Factor Communication Factor Delay Factor	Conflictions (5 mi.) Aircraft Time Under Control Duration of Ground Air Com. Total Delay Time Number of Aircraft Handled Fuel Consumption Under Control
THREE DENSITY LEVELS	Low Medium High	
ALPHA ERROR LEVELS	.20 .10 .05 .01	
BETA ERROR LEVELS	.20 .10 .05 .01	
NUMBER OF SUBJECTS	6 to 20	
NUMBER OF REPLICATIONS	1 to 4	

DELTA (DETECTABLE INCREMENT) - in respective measures above

FIGURE 16. POWER TABLE STRUCTURE

These are cases in which important and expensive systems are tested, but because the power has not been adequately considered and thought about, the results which seem like clear improvements are found to be not significantly different from existing systems. This is likely if no allowance is made for the beta error and if the alpha level selected is too stringent for this purpose, leading to the erroneous finding of no significant difference.

AN EVALUATION OF THE INDEX OF ORDERLINESS

ANALYSIS. Frequently, new ideas for ATC system measures are suggested. It would be useful to have a method for evaluating such ideas. It is suggested here that a data base like the SEM data can be useful for this purpose. As an example of how that might be done, a brief examination is made of the measure "the index of orderliness" which had been omitted from the original list of measures. This measure was developed by Halvorsen at the FAA National Aviation Facilities Center (reference 10) and has been studied in various places, but has rarely been used in dyamic simulation studies of en route systems. It was examined as a way to evaluate air traffic control systems by Gent at the Royal Radar Establishment (RRE) (reference 11), and was applied in a U.S. Transportation Systems Center study (reference 12) cited by Horowitz in connection with his study of the ARTS III system (reference 13). The RRE thought it was a promising measure, and the Horowitz study group found it was highly related to time duration of the state of confliction.

As has been explained earlier, it was possible to re-score the basic data tapes containing the records of the simulation exercises. For scheduling reasons, it was decided to re-score only the SEM I data to obtain the index of orderliness for that experiment's runs. To be consonant with the other data from the simulation runs, it was necessary to develop some summary statistics to represent the run as a whole. Three such measures were generated. The basic form of the index of orderliness which was used and how the run scores were composed is discussed in detail in appendix E. The basic approach was to generate an index for each aircraft at each second of the problem, average these for the minute, and then average these over the hour. One of the three measures was this average, and another was the variance computed over the minutes for the hour, and the third was developed into what was called the "probablity expression of the index values." These will be referred to as "ORD 1," ORD 2," and "ORD 3."

Several criteria were used to evaluate these index of orderliness measures: the reliability of the three indexes, their correlations with other measures which might be expected to be similar, their correlations with the judges' ratings, and their multiple correlations with the judges ratings. As was mentioned above, Horowitz (reference 13) cited some work at TSC (reference 12) as indicating that there was a strong correlation with the confliction measures, notably the time two aircraft spent in a state of confliction, and the index of orderliness type of measure. This finding was confirmed. Table 31 presents the correlations for each of the six sector-density cells between the three versions of the index of orderliness and the four factor scores and the two major confliction measures, the number and duration of 5-mile (separation standard) conflictions. The correlations between the first two index of orderliness scores and two of the factor scores (confliction and occupancy) and the confliction measures are sometimes quite high, at least in one of the two sectors.

TABLE 31

CORRELATIONS BETWEEN INDEX OF ORDERLINESS MEASURES

AND FACTOR SCORES AND CONFLICTION MEASURES*

Sector 14, Density 1

	Ord 1	Ord 2	Ord 3
Confliction Factor	.44	.54	-.13
Occupancy Factor	.65	.29	.20
Communication Factor	.24	.09	.25
Delay Factor	.03	.28	.10
No. 5 Mile Conflicts	.34	.45	.12
Duration 5 Mile Conflicts	.36	.42	-.09

Sector 14, Density 2

	Ord 1	Ord 2	Ord 3
Confliction Factor	.79	.70	.08
Occupancy Factor	.77	.60	.19
Communication Factor	.29	.17	-.06
Delay Factor	.11	.13	.10
No. 5-Mile Conflicts	.78	.73	.08
Duration 5-Mile Conflicts	.77	.66	.13

Sector 14, Density 3

	Ord 1	Ord 2	Ord 3
Confliction Factor	.72	.78	.00
Occupancy Factor	.83	.77	.00
Communication Factor	.11	.01	.00
Delay Factor	-.42	-.28	.00
No. 5-Mile Conflicts	.55	.55	.00
Duration 5-Mile Conflicts	.78	.87	.00

CORRELATIONS BETWEEN INDEX OF ORDERLINESS MEASURES

AND FACTOR SCORES AND CONFLICTION MEASURES (CONTINUED)

Sector 16, Density 1

	Ord 1	Ord 2	Ord 3
Confliction Factor	.02	.19	-.01
Occupancy Factor	.23	.11	.37
Communication Factor	.12	.04	-.09
Delay Factor	-.16	-.08	-.03
No. 5-Mile Conflicts	-.15	.04	-.15
Duration 5-Mile Conflicts	.05	.29	.05

Sector 16, Density 2

	Ord 1	Ord 2	Ord 3
Confliction Factor	.38	.63	-.08
Occupancy Factor	.30	.19	.09
Communication Factor	-.37	-.29	+.24
Delay Factor	+.17	+.01	+.20
No. 5-Mile Conflicts	.36	.58	-.01
Duration 5-Mile Conflicts	.49	.46	.21

Sector 16, Density 3

	Ord 1	Ord 2	Ord 3
Confliction Factor	.27	.46	.00
Occupancy Factor	.52	.49	.00
Communication Factor	-.38	-.57	.00
Delay Factor	-.11	-.12	.00
No. 5-Mile Conflicts	.30	.43	.00
Duration 5-Mile Conflicts	.33	.56	.00

* Data based on two-run aggregates; N is generally 27-31.

Table 32 presents the correlations among the three index of orderliness scores for each of the six cells. The first two index of orderliness scores (ORD 1 and ORD 2 in the table) seem well correlated with each other, but ORD 3 seems only occasionally related to the others.

In table 33, the run-to-run reliabilities based on the correlations between two similar runs are shown. The reliability coefficients are shown for the index of orderliness variables in comparison to the four factor scores and the two conflict measures. The first two index of orderliness measures are not better than the other measures, and the third index of orderliness measure is somewhat worse. The general inadequacy of 1-hour runs as to reliability has been discussed earlier; in addition, it will be recalled that the SEM I runs were reduced by 10 minutes to adjust for computer data losses by maximizing the number of runs of the same length.

In table 34 are shown the relationships of these measures to the observer ratings. These are not remarkably stronger than others, and they differ somwhat in the two sectors.

Thus far, it is seen that the index of orderliness measures are highly correlated with each other, highly correlated with two of the four factor scores, and have nothing in particular to add in the way of reliabiliity. In one final analysis, let us examine them in the light of whether they can add anything to our already available prediction of the judges' ratings by the four factor scores. These multiple R's are shown in table 35, compared to the multiple R's found <u>without</u> these measures added in. The index of orderliness measures add very little.

The fact that these new measures add very little to the prediction of the judges' scores suggests that much of the variation these new measures carry is already accounted for by the four factor scores. If this is true, then perhaps the two factor scores which are most highly correlated with the indexes can, taken together, allow us to dispense with the index scores. Using this approach, the two factor scores for confliction and occupancy were averaged and the resulting average was correlated with the index meeasures. These correlations are shown in table 36.

As was just speculated, the two factor scores combined do account for a great deal of the two main index of orderliness measures' variance in several of the conditions studied, but again there is a marked difference in the correlations depending on the sector involved. This sector difference raises a question beyond the scope of the present exploration of the index of orderliness measures.

<u>IMPLICATIONS</u>. The index of orderliness measurement type seems to have some puzzling but interesting qualities. It is suggested that it is still worth further examination. Its examination here was not complete. The primary purpose of its examination here was to exemplify this method of using a data base to study measures other than those that had been included in the original study.

TABLE 32

CORRELATIONS AMONG THE THREE INDEX OF ORDERLINESS MEASURES*

		Density 1			Density 2			Density 3		
		ORD 1	ORD 2	ORD 3	ORD 1	ORD 2	ORD 3	ORD 1	ORD 2	ORD
Geom.1	ORD 1	1.00			1.00			1.00		
	ORD 2	.79	1.00		.89	1.00		.80	1.00	
	ORD 3	.45	.37	1.00	.21	.18	1.00	.00	.00	1.00
Geom.2	ORD 1	1.00			1.00			1.00		
	ORD 2	.89	1.00		.78	1.00		.69	1.00	
	ORD 3	.68	.54	1.00	.41	.31	1.00	.00	.00	.00

*Data based on two-run aggregates; N is generally 27-31

TABLE 33

RUN-RUN* RELIABILITIES FOR INDEX OF ORDERLINESS

MEASURES, FACTOR SCORES AND CONFLICTION MEASURES

	Confl. Fac.Sc.	Occ. Fac.Sc.	Comm. Fac.Sc.	Delay Fac.Sc.	Confl. Count (5 mi.)	Confl. Dura.	ORD 1	ORD 2	ORD 3
G1 D1	-.03	+.44	+.68	-.52	+.24	-.15	+.30	+.09	-.07
G1 D2	-.10	+.75	+.69	-.38	+.06	+.08	-.05	-.09	.00
G1 D3	+.47	+.83	+.63	+.41	+.37	+.56	+.59	+.50	.00
G2 D1	-.13	+.61	+.53	+.04	-.04	+.12	-.28	-.32	.02
G2 D2	+.29	+.64	+.52	+.68	+.52	-.13	+.12	+.21	.00
G2 D3	+.42	+.66	+.52	+.07	+.44	+.43	+.34	+.17	.00

* N is generally 25-29. Data based on one 50-minute run vs. another. These data are for comparative purposes within this table. Negative coefficients can be taken as due to low reliability fluctuations.

TABLE 34

CORRELATION WITH RATINGS FOR INDEX OF ORDERLINESS

MEASURES, FACTOR SCORES, AND CONFLICTION MEASURES*

	Confliction Factor		Occupancy Factor		Comm. Factor		Delay Factor Score		Confliction Count (5 mi.)	
	SEM	CPM	SEM	CPM	SEM	CPM	SEM	CPM	SEM	CPM
G1 D1	-.23	-.22	+.11	+.04	+.21	+.08	-.36	-.37	-.23	-.21
G1 D2	-.16	-.08	-.01	+.06	-.23	-.19	-.25	-.26	-.29	-.17
G1 D3	-.24	-.25	+.18	+.01	-.34	-.22	-.58	-.52	-.33	-.24
G2 D1	-.48	-.38	+.05	-.03	+.15	-.05	-.23	-.25	-.35	-.23
G2 D2	-.35	-.31	-.16	-.21	+.32	+.17	-.20	-.21	-.28	-.24
G2 D3	-.35	-.24	+.20	+.26	+.13	+.03	-.37	-.43	-.28	-.19

	Conflicts Duration		ORD 1		ORD 1		ORD 3	
	SEM	CPM	SEM	CPM	SEM	CPM	SEM	CPM
G1 D1	-.29	-.24	+.22	+.24	+.02	+.06	+.04	+.04
G1 D2	-.17	-.07	-.23	-.11	-.15	+.01	-.03	+.05
G1 D3	+.04	-.03	+.11	+.08	-.03	+.10	.00	.00
G2 D1	-.20	-.19	+.11	-.04	-.01	-.06	+.07	-.09
G2 D2	-.15	-.19	-.26	-.23	-.34	-.29	+.03	-.12
G2 D3	-.14	.00	-.08	+.04	-.11	.00	.00	.00

* Data based on two-run aggregates; N is generally 27-31.

TABLE 35

MULTIPLE CORRELATION TO RATINGS WITH AND WITHOUT INDEX OF ORDERLINESS MEASURES*

	R vs SEM	R vs CPM	N
Density One Sector 14 (D1 G1)			
Factors	.40	.38	31
Factors and "ORD 1"	.46	.51	31
Factors and "ORD 2"	.42	.43	31
Density Two Sector 14 (D2 G1)			
Factors	.34	.32	30
Factors and "ORD 1"	.43	.37	30
Factors and "ORD 2"	.35	.32	30
Density Three Sector 14 (D3 G1			
Factors	.76	.64	29
Factors and "ORD 1"	.76	.66	29
Factors and "ORD 2"	.77	.64	29
Density One Sector 16 (D1 G2)			
Factors	.52	.47	31
Factors and "ORD 1"	.53	.48	31
Factors and "ORD 2"	.52	.47	31
Density Two Sector 16 (D2 G2)			
Factors	.50	.46	31
Factors and "ORD 1"	.50	.46	31
Factors and "ORD 2"	.50	.46	31
Density Three Sector 16 (D3 G2)			
Factors	.60	.60	30
Factors and "ORD 1"	.61	.60	30
Factors and "ORD 2"	.60	.60	30

* Data based on two-run aggregates.

TABLE 36

CORRELATIONS (r) BETWEEN TWO AVERAGED FACTOR SCORES AND INDEX OF ORDERLINESS MEASURES*

		D1	D2	D3
ORD 1	G1	.67	.86	.88
	G2	.23	.37	.56
ORD 2	G1	.34	.71	.85
	G2	.18	.33	.63
ORD 3	G1	.18	.17	.00
	G2	.35	.07	.00

* Data based on two-run aggregates; N is generally 27-31.

RESPONSES TO POST-RUN QUESTIONNAIRES.

ANALYSIS. Questionnaires were given to the subjects of the two experiments in order to obtain their opinions on the realism of the simulation, any difficulties with the equipment, and their own opinion on the difficulty of the task and how well they were doing.

These data are of interest in that they provide an opportunity to examine the topics above, but also they provide an opportunity to examine some questions involving the relationships between these responses and other data in the experiment.

Similar questions were asked after each run in both experiments. The first question requested the controller to give a self-rating of the quality of the control technique which had been applied in the run just finished. The second question was meant to be an inquiry into system performance and was phrased as a question about the controllers' estimate of the feelings of the hypothetical pilots flying through the sector about how the system handled the traffic during the run. These two questions were on 7-point scales where the fourth box represented the average value. The third question asked for a comparison of the traffic level in the experimental run compared to the home sector. The fourth question asked about the realism of the simulator. These last two questions were on 5-point scales. When the data was coded for data reduction, numerical values were assigned to the rating scale positions. The questionnaires used in the two experiments, which were slightly different in phrasing although basically the same, are presented in Figures 2 to 5, in the discussion of procedures.

Tables 37 and 38 present the basic information about the questionnaire replies given by the average subject, for SEM I and SEM II, respectively.

In the SEM I experiment, the average controller thought technique was better in Geometry 2 than in Geometry 1, and better at lower densities than at higher denstities, although one should hasten to add that an interaction between sector and density is again apparent. A similar tendency is seen in the relative ratings given to what we have called above their rating of system performance. In these two items, the coding was such that a high number means the "good" end of the scale.

The SEM I question about traffic asked for a comparison between the traffic level in the simulation problem just completed and the difficulty in a peak hour at the home sector when serving as the radar controller having normal team support. Here, "much easier" was coded as a "1" in the data reduction and "much harder" here was coded as "5." Of course, the answers varied with sector and density. The difficulty of the highest SEM I traffic density was rated as somewhat higher than that they faced at home at peak hours, and the middle density as about the same, or slightly easier than, peak hour work with the assistance of the team. There was about a half's rating point difference between the two sectors in the middle density rating, indicating a slight

TABLE 37

MEAN VALUES OF QUESTIONNAIRE ITEM RESPONSES - SEM I

Item	Cell 1 G1 D1	Cell 2 G1 D2	Cell 3 G1 D3	Cell 4 G2 D1	Cell 5 G2 D2	Cell 6 G2 D3
1. Technique (1)	4.4	3.9	3.6	4.2	4.5	4.1
2. System (1)	4.5	3.9	3.4	4.4	4.5	3.8
3 Traffic Comparison (2)	1.7	2.8	3.4	1.7	2.4	3.3
4 Realism (2)	3.3	3.1	3.1	3.0	3.2	3.2

NOTES: (1) Rating scale 1 to 7
(2) Rating scale 1 to 5

TABLE 38

MEAN VALUES OF QUESTIONNAIRE ITEM RESPONSES - SEM II

Item	Day 1	Day 2	Day 3
1. Technique (1)	2.4	2.8	2.9
2. System (1	3.3	3.8	3.9
3. Traffic Comparison (2)	2.8	3.0	3.0
4. Realism Comparison (2)	3.2	3.3	3.3

NOTES: (1) Rating scale 1 to 7
(2) Rating scale 1 to 5

feeling that geometry 2 was easier. Finally, in SEM I, the realism of the simulation process was considered adequate. In an open-ended question about the equipment, daily problems with the equipment were picked up and remedied. There were some complaints about the input devices on the radar consoles being different from those the controllers were used to in the field; this is now being remedied in a re-design of the simulator's controller positions.

For the SEM II experiment, the phrasing of three of the four rating questions was revised, although seeking similar information. In the first two questions, about the controller's own performance and the pilots' feelings about system performance, the wording was made more concrete, but the 7-point scales remained. Again, the poorer end of the scale was coded as "1" for the data reduction and the better end as "7." In responding to these first two items, the controllers generally regarded their performance in the runs about average for themselves, and felt that the system had performed at about an average level.

The rating item about the traffic was worded somewhat differently in the second experiment. The first experiment questionnaire had asked for a comparison of difficulty in the simulation hour exercise just completed with the difficulty in a peak hour in the home sector with the usual support; the second experiment items asked for a comparison of the traffic level just run to the traffic level which was usually encountered in the home sector, regardless of the team support used there. The direction of the scale and the coding were changed; a "1" in the second experiment's coding meant the traffic was considered heavier in the simulation and a "5" meant the traffic was heavier at home. Neither group of subjects expressed much difficulty with using these items.

On the first day, the SEM II traffic was rated somewhat heavier than the home sector traffic, where teams usually operate, as may be seen by the mean rating of 2.8 for day 1 in table 38. It will be remembered that this was approximately the same traffic level as had appeared in SEM I's geometry 1, density 2. There they had said it was about equal to the home sector's peak hour. On the second and third days, the traffic was rated at 3.0, or about the same as the traffic in the home sector.

In general, despite the differences in wording in the items, it can be said that they thought the traffic in these experiments was at least equal to the usual sector load in the field and somewhat higher and harder at times, as had been intentionally arranged, as was explained earlier under the topic of procedures and experimental design.

Turning now from the original purposes of the subject questionnaires of seeing how the subjects felt about the experimental runs as they proceeded, and of collecting information about equipment functioning, these data now might also be used to shed some light on some other questions of general interest.

In a general way, we might consider that there are four kinds of data here which might show interesting and informative relationships to one another,

TABLE 39

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS

AND OTHER DATA ITEMS

SEM I - CELL 1 Geometry (1), Density (1)

	Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.37	0.09	0.29
System	0.37	1.00	0.06	0.74
Traffic Comparison	0.09	0.06	1.00	0.17
Realism Comparison	0.29	0.74	0.17	1.00
Observer Ratings				
SEM	0.24	0.38	-0.08	0.30
CPM	0.00	0.25	-0.12	0.17
Factors				
Confliction	-0.03	0.00	0.01	-0.05
Occupancy	0.31	0.44	0.32	0.38
Communications	-0.10	0.17	0.11	0.09
Delay	-0.05	-0.30	-0.09	-0.39
Measures				
N5C	0.03	-0.04	-0.22	-0.18
A/C Time Under Ctl.	0.36	0.46	0.32	0.40
Dur. G/A Contacts	-0.21	0.08	0.12	0.12
Total Delay Time	0.15	-0.16	-0.18	-0.21
# A/C Hdld	-0.04	0.08	-0.09	0.25
Fuel	0.36	0.42	0.31	0.32
N3C	-0.12	-0.09	0.22	0.05
# Delays	-0.08	-0.31	-0.07	-0.39

TABLE 39 (CONTINUED)

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS

AND OTHER DATA ITEMS

SEM I - CELL 2

Items	Geometry (1), Density (2) Self-Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.58	-0.19	0.47
System	0.58	1.00	0.26	0.79
Traffic Comparison	-0.19	0.26	1.00	0.35
Realism Comparison	0.47	0.79	0.35	1.00
Observer Ratings				
SEM	0.50	0.38	-0.13	0.41
CPM	0.48	0.38	-0.18	0.39
Factors				
Confliction	-0.27	0.12	0.21	0.04
Occupancy	-0.19	-0.04	0.24	0.06
Communications	-0.20	-0.04	0.17	-0.11
Delay	-0.38	-0.24	0.11	-0.13
Measures				
N5C	-0.29	0.06	0.12	-0.03
A/C Time Under Control	-0.20	-0.04	0.26	0.07
Duration G/A Contacts	-0.26	0.03	0.40	-0.08
Total Delay Time	-0.41	0.00	0.36	0.11
# A/C Hdld	0.04	-0.05	0.09	0.03
Fuel	-0.21	-0.06	0.25	0.02
N3C	-0.25	0.16	0.22	0.06
# Delays	-0.17	-0.36	-0.18	-0.29

TABLE 39 (CONTINUED)

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS
AND OTHER DATA ITEMS

SEM I - CELL 3	Geometry (1), Density (3)			
	Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.51	0.06	0.02
System	0.51	1.00	0.16	0.43
Traffic Comparison	0.06	0.16	1.00	0.10
Realism Comparison	0.02	0.43	0.10	1.00
Observer Ratings				
SEM	0.56	0.63	-0.01	0.32
CPM	0.52	0.59	-0.11	0.28
Factors				
Confliction	-0.27	-0.10	0.24	0.10
Occupancy	0.01	0.03	0.20	0.13
Communications	-0.39	-0.47	-0.10	-0.05
Delay	-0.37	-0.40	0.21	-0.18
Measures				
N5C	-0.24	-0.03	0.11	0.02
A/C Time Under Ctl.	-0.02	0.01	0.18	0.18
Dur. G/A Contacts	-0.27	-0.38	0.30	-0.13
Total Delay Time	-0.35	-0.31	0.21	0.01
# A/C Hdld	0.29	0.24	0.08	-0.01
Fuel	-0.05	-0.01	0.14	0.07
N3C	-0.38	-0.20	0.16	0.06
# Delays	-0.36	-0.47	0.18	-0.39

TABLE 39 (CONTINUED)

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS
AND OTHER DATA ITEMS

SEM I - CELL 4	Geometry (2), Density (1			
	Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.41	-0.38	0.11
System	0.41	1.00	-0.04	0.24
Traffic Comparison	-0.38	-0.04	1.00	0.01
Realism Comparison	0.11	0.24	0.01	1.00
Observer Ratings				
SEM	0.14	0.05	-0.36	-0.01
CPM	0.21	0.21	-0.44	0.12
Factors				
Confliction	-0.21	-0.11	0.28	-0.19
Occupancy	0.05	0.25	0.12	0.33
Communications	-0.23	-0.04	0.12	0.31
Delay	-0.05	0.03	0.10	-0.45
Measures				
N5C	-0.19	-0.12	0.23	-0.16
A/C Time Under Ctl.	0.02	0.23	0.07	0.31
Dur. G/A Contacts	-0.20	-0.10	0.04	0.29
Total Delay Time	0.16	0.27	0.12	-0.10
# A/C Hdld	-0.06	0.28	0.23	0.43
Fuel	-0.01	0.22	0.10	0.19
N3C	-0.11	-0.15	0.24	-0.14
# Delays	-0.22	-0.20	0.02	-0.53

TABLE 39 (CONTINUED)

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS
AND OTHER DATA ITEMS

SEM I - CELL 5	Geometry (2), Density (2) Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.47	-0.09	0.10
System	0.47	1.00	0.01	0.42
Traffic Comparison	-0.09	0.01	1.00	-0.08
Realism Comparison	0.10	0.42	-0.08	1.00
Observer Ratings				
SEM	-0.19	-0.13	-0.12	-0.06
CPM	-0.10	-0.03	-0.37	0.06
Factors				
Confliction	-0.15	-0.06	0.20	-0.35
Occupancy	-0.11	-0.00	0.31	0.26
Communications	-0.09	0.11	0.24	0.18
Delay	-0.04	-0.03	-0.13	0.16
Measures				
N5C	-0.04	0.02	0.13	-0.33
A/C Time Under Ctl.	-0.14	-0.01	0.30	0.28
Dur. G/A Contacts	-0.07	-0.12	0.11	-0.04
Total Delay Time	-0.06	0.08	0.02	0.23
# A/C Hdld	0.33	0.11	0.02	-0.17
Fuel	-0.22	-0.01	0.32	0.26
N3C	-0.22	-0.18	0.20	-0.20
# Delays	-0.04	-0.23	-0.31	-0.09

TABLE 39 (CONTINUED)

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS

AND OTHER DATA ITEMS

SEM I - CELL 6	Geometry (2), Density (3) Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.57	-0.16	0.07
System	0.57	1.00	0.05	0.25
Traffic Comparison	-0.16	0.05	1.00	-0.09
Realism Comparison	0.07	0.25	-0.09	1.00
Observer Ratings				
SEM	0.58	0.51	-0.20	0.20
CPM	0.56	0.46	-0.26	0.13
Factors				
Confliction	-0.43	-0.21	0.16	-0.49
Occupancy	0.07	0.25	0.09	-0.02
Communications	0.05	0.03	0.05	0.25
Delay	-0.12	-0.30	-0.21	-0.17
Measures				
N5C	-0.43	-0.24	0.19	-0.51
A/C Time Under Ctl.	0.02	0.18	0.09	-0.06
Dur. G/A Contacts	-0.06	-0.01	0.09	0.14
Total Delay Time	-0.09	-0.11	-0.05	-0.02
# A/C Hdld	0.12	0.24	0.17	0.12
Fuel	0.00	0.20	0.05	-0.05
N3C	-0.51	-0.23	0.18	-0.41
# Delays	-0.11	-0.36	-0.29	-0.24

omitting the rating on the simulation realism. The four kinds of data are:

a. Performance; Own opinion (subject)
b. Performance; Judge's opinion
c. Performance; Measured
d. Workload felt by subject (traffic level reply)

If this were merely a set of variables being intercorrelated, there would be ten possible inter-relationships here; but with four or more performance measures, depending on whether only the four factor scores or some others are used, there would be a considerably larger number of correlations. For this reason, the number of measures of each type will be restricted.

In SEM I, one such intercorrelation table was done for each cell (sector-density combination). In SEM II, one intercorrelation table was done for each day. The SEM II day data should be more informative since it is based on twice as many runs (four per day as compared to two per cell in SEM I). These tables appear as table 39 for the six SEM I sector-density cells and as table 40 for the three SEM II days.

Possibly the best way to approach this is by means of a series of single simple questions, all of which apply to both SEM I and II. Some questions of interest are:

a. What is the relationship between self-judged performance and other-judged (by observers) performance?

b. What is the relationship between self-judged performance and objectively measured performance?

c. What is the relationship between self-judged performance and self-judged workload?

d. What is the relationship between other-judged (by observers) performance and objectively measured performance?

e. What is the relationship between other-judged (by observers) performance and self-judged workload?

f. What is the relationship between self-judged workload and objectively measured performance?

Let us now examine these questions in an exploratory way, mainly to suggest hypotheses for other experimenters. The number of cases used for the correlations for the SEM I data is usually 29 to 31; and in the SEM II data, 39. The correlation value tabled as statistically significant (See appendix 3 for explanation) at the .05 level for 29 cases is approximately .37, for 39 cases is approximately .30.; only correlations above .30 will be looked at here.

The first question is: What is the relationship between self-judged performance and other-judged performance?

TABLE 40

CORRELATIONS BETWEEN SUBJECT QUESTIONNAIRE ITEMS
AND OTHER DATA ITEMS

SEM II - DAY 1	Geometry (1), Density (2) Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.44	0.18	0.35
System	0.44	1.00	0.35	0.10
Traffic Comparison	0.18	0.35	1.00	-0.12
Realism Comparison	0.35	0.10	-0.12	1.00
Observer Ratings				
SEM	0.20	0.47	0.29	-0.22
CPM	0.12	0.37	0.24	-0.23
Factors				
Confliction	-0.28	-0.10	-0.18	-0.11
Occupancy	-0.08	-0.07	-0.15	0.12
Communications	0.13	-0.24	0.00	0.24
Delay	-0.28	-0.53	-0.22	0.11
Measures				
N5C	-0.30	-0.22	-0.23	-0.12
A/C Time Under Ctl.	-0.08	-0.10	-0.16	0.10
Dur. G/A Contacts	0.06	-0.19	0.06	0.22
Total Delay Time	-0.25	-0.48	-0.23	0.14
# A/C Hdld	0.28	0.47	0.18	-0.14
Fuel	-0.11	-0.05	-0.20	0.02
N3C	-0.21	-0.10	-0.20	0.02
# Delays	-0.30	-0.58	-0.21	0.05

TABLE 40 (CONTINUED)

CORRELATIONS BETWEEN QUESTIONNAIRE ITEMS AND OTHER DATA ITEMS
SEM II

SEM II - DAY 2	Geometry (1), Density (2) Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.36	-0.05	0.48
System	0.36	1.00	0.16	0.34
Traffic Comparison	-0.05	0.16	1.00	-0.04
Realism Comparison	0.48	0.34	-0.04	1.00
Observer Ratings				
SEM	0.18	0.20	0.29	0.04
CPM	0.10	0.18	0.33	0.13
Factors				
Confliction	-0.13	-0.08	-0.12	-0.13
Occupancy	-0.23	-0.44	-0.43	0.19
Communications	0.01	-0.16	-0.29	0.16
Delay	0.07	-0.10	-0.24	0.35
Measures				
N5C	-0.09	-0.10	0.05	-0.08
A/C Time Under Ctl.	-0.16	-0.40	-0.45	-0.09
Dur. G/A Contacts	0.05	-0.11	-0.28	0.10
Total Delay Time	0.17	-0.02	-0.27	-0.25
# A/C Hdld	0.10	0.25	0.32	0.22
Fuel	-0.29	-0.39	-0.25	-0.35
N3C	-0.16	-0.15	-0.26	-0.15
# Delays	-0.04	-0.16	-0.17	-0.38

TABLE 40 (CONTINUED)

CORRELATIONS BETWEEN QUESTIONNAIRE ITEMS AND OTHER DATA ITEMS
SEM II

SEM II - DAY 3	Geometry (1), Density (2) Self Ratings			
	Technique	System	Traffic Comparison	Realism Comparison
Self Ratings				
Technique	1.00	0.42	-0.32	0.34
System	0.42	1.00	-0.10	0.56
Traffic Comparison	-0.32	-0.10	1.00	-0.02
Realism Comparison	0.34	0.56	-0.02	1.00
Observer Ratings				
SEM	-0.01	0.10	0.28	0.07
CPM	0.04	0.10	0.28	0.03
Factors				
Confliction	-0.25	-0.20	0.01	-0.23
Occupancy	0.21	-0.02	-0.44	-0.11
Communications	0.20	-0.05	-0.41	0.20
Delay	-0.02	-0.15	-0.11	-0.20
Measures				
N5C	-0.17	-0.26	0.04	-0.30
A/C Time Under Ctl.	0.24	-0.00	-0.45	-0.12
Dur. G/A Contacts	0.18	-0.04	-0.34	0.15
Total Delay Time	-0.06	-0.06	-0.05	-0.19
# A/C Hdld	-0.04	-0.30	-0.03	-0.16
Fuel	0.10	0.14	-0.38	-0.15
N3C	-0.13	-0.15	0.08	-0.14
# Delays	0.00	-0.17	-0.12	-0.19

If we consider the answer to this question to be obtainable from the relationship between the self-rating questions on technique and systems performance, on the one hand, and the observers' two ratings on the other, we can attempt an answer. The correlations between these subjective ratings by the observer and by the observed are sometimes encouraging, but fluctuate rather widely with the conditions, or are perhaps simply fluctuating on a sampling basis.

In the SEM I experiment, there is evidence of the expectable relationship, at least at the middle and high density levels, although somewhat more clearly in one sector rather than the other. In the middle density of Geometry 1, for example, there are correlations of .50 and .48 between the SEM and CPM ratings by the judges and the self-ratings of technique. Also, there are two positive correlations of .38 of the two observer ratings with the self-rating of system performance. Similar level correlations appear in two other cells, such as Geometry 1, Density 3, and Geometry 2, Density 3, and on Day 1 in SEM II for the system rating only.

The second question is: What is the relationship between self-judged performance and objectively measured performance?

Let us consider this question by examining the four factor scores and the self-ratings of technique and system performance. It is to be expected that these relationships will be negative, since a high self-rating should reflect a low number of conflictions, i.e., the scales run in opposite directions. In most cases, the correlations are indeed negative in sign. However, there are only a few correlations above .30. The primary and auxiliary raw score measures follow the factor scores in this, as usual.

The third question is: What is the relationship between self-judged performance and self-judged workload?

To answer this question, an examination was made of the correlation between the subject's rating of own technique and the rating of the traffic level faced. There are a few high correlations, but there seems to be no consistent pattern although there is a tendency to rate technique lower when the traffic seems heavier. It should be remembered that high number ratings in SEM I meant the subject felt the traffic was heavier in the simulation than at home, but in SEM II this scale was numerically reversed and a low coding number meant higher traffic.

The fourth question is: What is the relationship between other-judged (by observers) performance and objectively measured performance?

This one has already been answered at length in a previous section devoted to the subject. There it was found, at least when multiple correlations were used, that the relationships between objective scores and rated controller ability were substantial. Here, however, let us pause further over this question to simply illustrate a more graphic approach to the question of the relationhip between performance scores and controller ability, which might be examined further in the future.

Using SEM II day-level data, the controller judges' ratings for controller performance were arranged from lowest to highest. The four factor scores associated with those ratings were assembled into profiles for each individual, which was possible because they were on the same scale, as was discussed in the earlier discussion of both experiments' data having been put on the "third" scale. In figure 17, it could be said that it appears that the high performance controllers and the lower performance controllers may show different types of profiles. This constitutes a suggestion for further examination; much further work might be done in the realm of cluster analysis and profile analysis to explore such questions as the number of unique controller profiles of performance there might be.

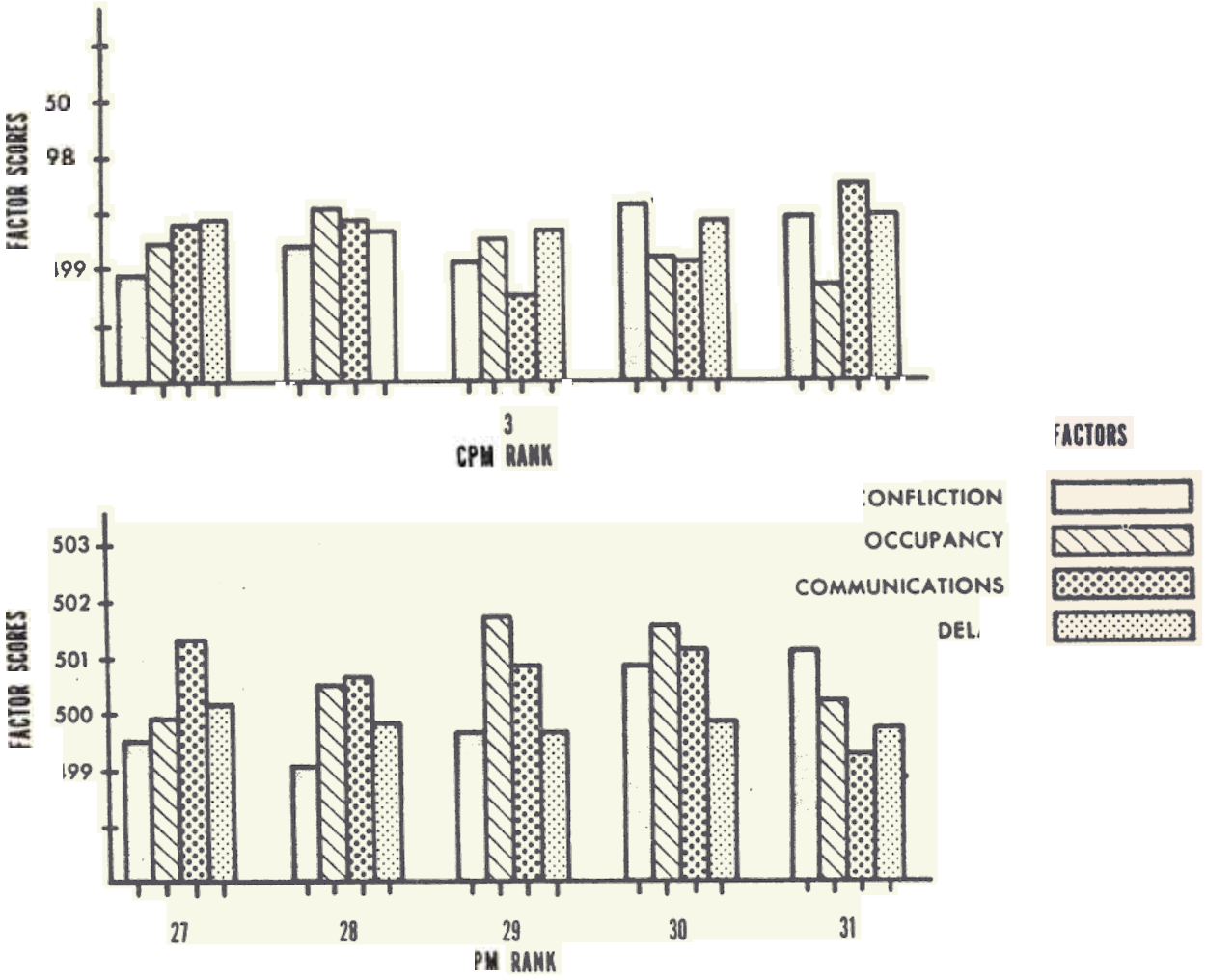
The fifth question is: What is the relationship between other-judged (by observers) performance and self-judged workload?

In examining the SEM I correlations between the traffic question and the two obsevers' ratings, a few correlations in the negative thirties appear, -.36 and -.44 in the case of Geometry 2, Density 1, and -.37 in Geometry 2, Density 2. Apparently those who are functioning well in the opinion of the judges, at least, feel that the workload is lighter than others do. In the same data for the day level in SEM II, the correlations are close to thirty, but they are positive. This is probably a manifestation of the same phenomenon; the change in sign is understandable in that it may be remembered that the SEM II rating scale of traffic ran in the opposite direction from the SEM I rating scale.

The sixth question is similar to the fifth and is the following: What is the relationship between self-judged workload and objectively measured performance?

While there are not many correlations over .30 here, their directionality is appropriate. In SEM I if the controller felt that the traffic was heavy it would receive a higher numerical rating. In heavy traffic, most of the performance scores would naturally get higher (like delays). Therefore, positive correlations between the traffic ratings and the performance scores would be expected in the SEM I data, and this is generally the case. Because the SEM II scale on traffic ran in the opposite direction, essentially from "lighter here" coded as "1" to "heavier here" coded as "5," the SEM II correlations on this point would be expected to be opposite in sign and they usually are.

Finally, a word should be added here about an interesting relationship with the realism rating which was omitted from the earlier main discussion. There were some cases of positive correlations, some fairly high, between the subject's opinion of the realism of the simulation and the opinion held on the goodness of own-technique and system performance.



GURE CONTROLL LE STANDARD CORE FORM

IMPLICATIONS. The implications are

1. The subjects felt that they did an average job, were not disturbed by any lack of realism, and felt that the traffic samples were tough; equal to "peak hour" with a full sector team helping them. The main purpose of this questionnaire, as has been said, was to check on daily experience, equipment functioning, and so on, and this purpose was fulfilled.

2. The data were adapted to make some explorations into the relationships between workload and performance, even though not ideally suited for the purpose. About all that can be said here is that such relationships, if they exist, are weak and situation-dependent.

DISCUSSION

There is no question as to whether real-time air traffic control system simulation will be used in the future. It seems an eminently worthwhile, albeit expensive, thing to do. Although people feel that they get information out of it about air traffic control system problems and issues, the real question is whether they get information or misinformation.

Here enters a true philosophical issue. Are impressions information? If someone watches a controller use a proposed system, and thinks it functions better than the current system in use, is that information? If the controller is asked for an opinion and gives it, is that information? Suppose that the traffic mix or level or procedures are somewhat different from those which the controller or the observer are used to. Are their impressions dependable enough to base huge expenditures for new systems on them? Suppose the designer of the new system is giving his observation, is that information?

These are the kinds of considerations that make objective measurement and statistical techniques desirable. It is because grave errors can be caused by subjectivity in interpreting what is seen, and sometimes even in interpreting what has been genuinely measured, as, for example, when the hypothesis about what measures shall be considered important has not been stated in advance. However, measurement of the joint performance of human and machinery in accomplishing the mission of an information processing and decision making system is not a simple task. To develop methods and measures for such a purpose is a difficult, time consuming and risky effort. It must be remembered that the performance under study is not rote or mechanical but very dynamic. The thing to be surprised about is not that the measurement process may be discouraging, but that there is anything encouraging about it at all.

There may, in fact, be a middle ground possible between sheer impressionism and strict empiricism. This might consist of carefully controlled and administered observation and rating forms being given to trained, impartial and fresh observers. But even this would be in need of an evaluation and refinement process.

The worst case of all, though, is the one that appears to be more frequent and customary than even those who engage in it acknowledge. Simply stated, these are studies in which the investigators, in all good faith, use objective measurements that can be obtained from a simulator apparently without realizing that such measures, even though numerical, are behavior and performance measures and have a wide band of error around them.

On the other hand, only the most crucial system evaluations, perhaps, need to be conducted using strict inferential rules. There are times, as Stammers and Bird (reference 14) say, using the Sinaiko and Belden term (reference 15), when the proper thing to do is the "indelicate experiment." The work by Stammers and Bird concerned a data transfer and display system for airport controllers and was carried out for the Royal Radar Establishment. It is a fine example of

such an effort. Another type of brief and uncomplicated simulation being a good idea is when it is done for the purpose of exploring concepts as part of a long continuing examination. What appears to be an example of this is the work of Tobias and O'Brien on RNAV (area navigation) for NASA (reference 16).

In working on evaluating human factors aspects of computer aiding for air traffic controllers, Whitfield, Ball and Ord (reference 17) achieved a good integration of the best features of the "indelicate" experiment and the more traditional experiment.

The topics of methods and measurements in the air traffic control system have been discussed at length by Hopkin (references 18 and 19) and the general topic of systems experimentation involving performance measurement has been discussed in a book by Parsons (reference 20).

While admitting that various degrees of indelicacy may be permissible depending on the circumstances, it is still important to pursue the ideals of classic experimentation where possible and appropriate. That being the case, let us review some of the "lessons learned", which might be of use in pursuing both the delicate and indelicate experiment.

The first and most important lesson was also and first pointed out by Horowitz (reference 13), and it is to consider the beta error. As Horowitz pointed out, people in medicine and medical research do this all the time and people in other practical fields should do so too. What he had encountered was the tendency in some statistically minded people to set the level of the alpha error they will accept at the traditional .05 level, and to ignore the beta error. Especially with difficult data such as is found in dynamic simulations, this leads to frequent, if not continual, failure to reject the null hypothesis. In a practical sense, that sort of uncritical application of statistical techniques could lead to the rejection of many fine system concepts. This is what Horowitz rightly pointed out.

The data from these experiments and the power tables based on them can reduce the likelihood of that kind of error by asking that the levels of alpha and beta that will be used and the amount of difference it is sought to detect be specified in advance. It is possible to compensate for the lack of statistical robustness in the measurement process by choosing moderate levels of these parameters.

A second major lesson learned is the importance of the practice and familiarization factors in system experiments and evaluations. The learning curves sought and found in the SEM II experiment were quite dramatic. For this reason, careful thought must be given to practice effects and hence sequence effects in the design of such experiments. However, it should be of some assistance to know how long these effects last, as indicated by the curves.

Related to the question of the statistical power of simulation data, is the question of the reliability (repeatability) of such data. While one can compensate for such unreliability as was found here, it was found to be lower than expected based on the only other experiment having such data. It was, in fact, expected to be some amount higher since now the data was being collected

by computer instead of by paper and pencil. This did not turn out to be the case. While the reliability was not totally discouraging, due to the fact that it can be compensated for by means of considering the setting of the alpha and beta levels actually needed and by data aggregation, it is puzzling. This is a topic which deserves some careful work and thought. The lesson to be learned here is that new ideas for system measures should be sought out on a continuing basis.

This same unreliability should caution those who wish to run simulations to the effect that if a single sector system is comparatively unreliable, then a multi-sector simulation's data are almost certainly much more unreliable, because of the additional sources of variance introduced. While the reliability and power calculations made here do not apply to multi-sector simulations, they can be regarded as an optimistic estimate of what would occur in a larger simulation.

While on the subject of the single-sector, single-controller system, it did, of course, include simulated conversation and coordination with adjacent sectors and even terminal areas, and, while we obtained no evidence on this topic beyond the subject controllers' ratings of simulation realism, it seemed quite satisfactory as a method for simulating the essence of the controller's job. It would seem to recommend itself as a rather economical way of studying many man/machine interface problems or plans, and even as a way to evaluate individual controller training progress.

Another lesson learned here was that we only need to analyze a comparatively small number of measures: the four factor scores, the four primary measures, and the two additional auxiliary scores. This makes an enormous difference in the sheer feasibility of data handling chores and interpreting this kind of data. This set of scores should be accepted as an operating base for all enroute simulations, at least until something better comes along, and programmed into the simulation data collection system. A bonus from this practice would be that after some time all ATC system simulations would be interpretable in common data distribution terms.

Excessive reliance on ratings by judges is not recommended even though the judges here performed with some reliability. It must be remembered that they were carefully and deeply trained, and were constantly observing the same exercises.

Another lesson which should be learned is that there is available a way to accumulate a set of traffic problems which are extremely different, thus reducing practice effects, but which can be shown to be of a comparable level of difficulty. The interaction between sector geometry and traffic density could be used to generate a library of traffic samples whose level of difficulty, as indicated by score distributions obtained in small experiments, could be considered interchangeable. Another way of handling the traffic sample "same but different" requirement was also demonstrated here, the shuffling of start times in the same level-profile trafffic sample.

The main lesson to be learned from the experience with the index of orderliness was not a clear-cut lesson about that index, which did not emerge, but, nonetheless, a demonstration that, given a data base like that used here, many investigations about different and novel measures might be conducted.

A major question which arises is that of whether there is additional work in this area which should be done. There are at least three study efforts which should be undertaken, and it should be pointed out at the outset that accomplishing them will be considerably easier because of that which has been done so far since, in subsequent investigations, even of methodology, the power estimates which are available will enable careful planning of the required size of the experiments which are to be conducted.

The first and most obvious follow-on work would involve continuing to work with the available data bases from these experiments in order to seek for refined measures. It should be remembered that the focus in this effort so far was evaluative, not developmental. As a next step, various ideas for novel measures could be computed in these data bases and their relationships to one another and to the standard measures already present could be examined.

The second step would be to extend the method to a multiple-controller sector team and to a multi-sector system of reasonable size, say three sectors. The goal here would be the comparatively simple one of determining the change in variance, power and reliability which would be caused by working with these more complex system spaces. This would probably be desirable to do even though, on the one hand, it would be hoped that the need people feel for duplicating complex system spaces in simulation would be diminishing, and, on the other hand, that the present power estimates could be used as approximations (albeit optimistic ones for large systems).

The third possible direction would be to make a start into the study of terminal area simulation methodology and measurements. A beginning on this had been made, but has since been postponed. In basic outline, the approach that had been tentatively decided upon was as follows. First, there was to be an assembly of the customary classic measures for terminal area air traffic control system functioning. Next, these measures would be administered at three levels of traffic density and with several replications to a large number of control "teams." The first attempt would be to try to reduce the number of measures by searching for the basic dimensions of measurement, and having found those, to examine the data to estimate the parameters needed to plan experiments of desired levels of statistical power for system evaluations.

However, the terminal area air traffic control system is nowhere near as simple as the en route system. It is easy and clearly legitimate to represent the en route system in microcosm; but the terminal system does not readily lend itself to such simplification. The terminal team is composed of several individuals working not on the same airspace but on different parts of the airspace. While the smallest en route team groups around one radar picture, the smallest terminal team might consist of an arrival controller, a departure controller, and a local controller and ground controller. While ground and local control have rarely been simulated, they could be by use of some simplifying assumptions and rough presentations. Specifically, it would probably not be too unrealistic to use the simulator to show the airport surface as if on radar for the purpose of running a complete simulation. Doing

such a simulation was considered. Also considered was running, with the same people, a terminal area simulation in which one controller was looking at the entire terminal area and performing the total control function alone, at a much reduced level of traffic, of course. One major purpose would be to determine if the same measures were statistically important in both team and microcosm (single controller) terminal simulations. Another purpose would be to determine if, when similar conditions (systems, geometries, etc.) were compared in team and microcosm simulations, similar outcomes resulted. This would render many terminal area issues investigatable by simulation which are now almost prohibitive in the amount of effort required to accomplish them. Progress was made in developing the list of measures which was to be evaluated and it is presented as Appendix F for the use of those who might be engaged in terminal area simulation work.

There is one last comment it seems important to make about possible future research that this experience has suggested. This, briefly, has to do with the application of the methodology developed here to a related field, as a training progress criterion measure device for the individual controller. While, as said earlier, the reliability needs considerable improvement for such a purpose, such improvement does not seem impossible.

CONCLUSIONS

These experiments provided a statistical and methodological baseline for quantitative system assessment using real-time air traffic control simulation testing. In particular, the following conclusions have been reached:

1. The en route measure set as presently constituted forms recognizable operationally meaningful clusters of measures. These are confliction, occupancy, communication and delay.

2. The four factor measures produce as valid an assessment of system performance as do the original many raw measures.

3. The acquisition of stable data requires six hours of preliminary familiarization and training in the experimental environment.

4. The same four factors were tried in another experiment with another sector geometry, two additional traffic densities, and a different group of controllers, the factors still held up as being adequate basic dimensions of measurement.

5. System evaluation using real-time ATC simulation in an objective manner is only possible in a technically sound way if account is taken in planning experiments of the relatively low statistical power of the measurement which can be accomplished in the dynamic exercises. Tables of the statistical power of the basic factor scores have been assembled based on the data here collected and analyzed. Failure to assure adequate power will in most system evaluations lead to the rejection of actually promising system ideas.

It is to be emphasized that the above conclusions were reached during tests where one person, serving as the radar controller for the sector, was responsible for all the traffic in the sector. Also, the traffic density was held at a relatively constant level throughout a given session. However, adequate provision for the exercise of adjacent sector coordination was included, and some of the assistant controller duties were pre-performed. It seems certain that the "one-person team" procedure would not have affected the basic dimensions of measurement found for system effectiveness; although the estimates of inter-team variation which entered into the power calculations might possibly have been affected.

REFERENCES

1. Bona, L. J. and Pszczolkowski, S. R. The FAA's Air Traffic Control Simulation Facility. in: Technology in air traffic control training and simulation., Proceedings, Vol. II, Second International Learning Technology Congress and Exposition on Applied Learning Technology., February 14-16, 1978. Society for Applied Learning Technology, Warrenton, Virginia.

2. Buckley, E.P., O'Connor, W. F. and Beebe, T. A Comparative Analysis of Individual and System Performance Indices for the Air Traffic Control System, Federal Aviation Administration, National Aviation Facilities Experimental Center, Report NA-69-40, 1969.

3. Dixon, W. J. Biomedical Computer Programs. University of California at Angeles Press, Berkeley, California, 1975.

4. Harman, H. H. Modern Factor Analysis. 2d edition. University of Chicago Press, Chicago, Illinois. 1979.

5. Boone, J. O. et al. The Federal Aviation Administration's Radar Training Facility and Employee Selection and Training. FAA-AM-80-15. CAMI, FAA, Oklahoma City, Okla., Sept. 1980.

6. Boone, J. O. and Steen, J. A. A Comparison Between Over the Shoulder and Computer-Derived Measurement Procedures in Assessing Student Performance in Radar Air Traffic Control. Aviation, Space and Environmental Medicine, 52(10): 589-593, 1981

7. Cobb, B.B. The Relationships Between Chronological Age, Length of Experience, and Job Performance Ratings of Air Route Traffic Control Specialists. FAA-CAMI Report AM 67-1, June, 1967.

8. Gaebelein, J.W. and Soderquist, D.R. The Utility of Within-Subject Variables Estimates of Strength. Educational and Psychological Measurement, 1978, 38, p. 351-9.

9. Cohen, J. Statistical Power Analysis for the Behavioral Sciences York. Academic Press. 1969

10. Halvorsen, A. G. Definition and Description of System Performance Measure Entitled Index of Orderliness for Use in Digital ATC Simulation Studies. National Aviation Facilities Experimental Center: Working Paper, September, 1970.

11. Gent, H. A Measuring Rod for ATC Systems: The Index of Orderliness. Symposium, Boston, MA, 1975.

12. Mengert, P. H. and Englander, T. Preliminary Statistical Validation Index of Orderliness ARTS III Test Data. U. S. D.O.T. Transportation Systems Center: Final Report, April 19,1971.

13. Horowitz, Seymour M. An Evaluation of the ARTS III Level of Automation.(third lot procurement). July, 1972. FAA Office of Aviation Economics.Final Report.
Not numbered.

14. Stammers, R. B. and Bird, J. M. Controller Evaluation of a Touch Input Air Traffic Data System: An "Indelicate" Experiment. Human Factors, 1980, 22(5), 581-589.

15. Sinaiko, H. W. and Belden, T. G. The Indelicate Experiment. in: J. Speigel and D. E. Walker (Eds.), Second Congress on the Information System Sciences. Washington, D. C.: Spartan Books, 1965, 343-348.

16. Tobias, L. and O'Brien, P. Real-Time Manned Simulation of Advanced Terminal Area Guidance Concepts for Short-Haul Operations. NASA, Washington DC, August 1977. NASA Technical Note NASA TN D-8499.

17. Whitfield, D., Ball, R. and Ord, G. Some Human Factors Aspects of Computer-Aiding Concepts for Air Traffic Controllers. Human Factors, 1980,22(5), 569-580.

18. Hopkin, V. D. Conflicting Criteria in Evaluating Traffic Control Systems. Ergonomics, 1971, 1971, 14,557-564

19. Hopkin, V. D. The Measurement of the Air Traffic Controller. Human Factors, 22(5), pp.547 to 560.

20. Parsons, H. M. Man-Machine System Experiments. Baltimore: The Johns Hopkins Press, 1972.

21. Mc Nemar, Q. Psychological Statistics. 3rd edition. New York, Wiley 1962.

22. Winer, B. J. Statistical Procedures in Experimental Design. 2d edition. New York, Wiley, 1971.

23. Kirk, R. E. Experimental Design: Procedures for the Behavioral Sciences. Brooks/Cole, Belmont, CA. 1968.

24. Guilford, J. P., Fundamental Statistics in Psychology and Education. Fourth Edition. New York, Wiley, 1965.

APPENDIX A

LIST OF SYSTEM EFFECTIVENESS MEASUREMENTS AND DEFINITIONS:

SEM EXPERIMENT I

1. TGT Spacing Analysis (A)

   A count of the number of instances two aircraft violate the separation allowance of 950 feet vertically and 4 miles horizontally.

2. TGT Spacing Analysis (B)

   Same as above with 5 mile horizontal separation allowance.

3. TGT Spacing Analysis (C)

   Same as above with 3 mile horizontal separation allowance.

4. Number of Start Delays

   A count of the number of instances an aircraft entered the system at a time greater than its scheduled time (plus two minutes).

5. Start Delay Time

   The duration of the start delays (Measure 4).

6. Number of Hold and Turn Delays

   A count of the number of holding delays plus a count of the number of turn delays lasting more than 100 seconds.

7. Hold and Turn Delay Time

   The duration of the hold and turn delays (Measure 6).

8. Number of Arrival Delays

   A count of those start delays of arriving aircraft.

9. Arrival Delay Time

   The duration of arrival delays.

10. Number of Departure Delays

A count of those start delays of departing aircraft

11 Departure Delay Time

The duration of departure delays.

Time in System

The number of active aircraft controlled by the subject, incremented each second that control was exercised.

13. Number Aircraft Handled

Total number of aircraft under subject's control.

Number of Completed Flights

The number of flights terminated by a handoff.

15. Number of Arrivals Achieved

A count of enroute traffic transferred to the termination frequency.

Number of Departures Achieved

A count of active departures.

17. Arrival Altitudes not Attained

A count of enroute arrivals not transferred to the termination frequency at an altitude greater than was predetermined, plus 100 feet.

18. Departure Altitudes not Attained

A count of enroute departures not transferred to the termination controller at an altitude less than was predetermined minus 100 feet.

19. Number of Contacts

A count of ground to air microphone contacts.

20. Communication Time

The duration of ground to air contacts (Measure 19).

21. Number of Altitude Changes

A count of pilot messages to alter aircraft altitude.

22. Number of Heading Changes

A count of pilot messages to change heading.

23 Number of Speed Changes

A count of pilot messages to revise aircraft speed.

24. Number of Handoffs

The number of acknowledged handoffs to the subject.

25. Handoff Delay Time

The time between a handoff and the subject's acceptance of that aircraft

26. Re-idents

A count of beacon identity requests.

APPENDIX B

LIST OF SYSTEM EFFECTIVENESS MEASUREMENTS AND DEFINITIONS:

SEM EXPERIMENT II

Given below is a list of measures used in this experiment with definitions and commentary They generally consist of event counters with their respective duration. All duration measures are in seconds.

Unl s noted to the contrary, all measures are keyed to the following rule to determine if an aircraft is under the control of the subject

CONTROL RULE

An aircraft is under control if it is within the sector boundary or on the frequency of the subject

That is to say, in order for an aircraft not to be under control it must be both outside the sector and off the subject's frequency. When under control, an aircraft is considered the subject's responsibility and all events relative to that aircraft are charged to the subject

DA - 01 Number of Path Changes (PTHCHAD)

The number of altitude, heading, and speed change messages sent to aircraft under control

DA - 02 Number of Barrier Delays (BRNDELD)

The number of instances a subject asks that all entering traffic be halted.

DA - 03 Duration of Barrier Delays (BRDURAD)

The cumulative time that barrier delays remain in effect The beginning of a barrier delay is referred to as a STOP message and its termination as a START message.

DA - 04 Number of Start Delays to Aircraft (NSTADLD)

The number of instances that an aircraft was scheduled to enter the problem while a STOP message was in effect

Note that STOP and START messages can occur without any start delays accumulating (If, for instance, no aircraft were scheduled to enter during this interval.)

DA - 05 Duration of Start Delays to Aircraft (DSTADLD)

The cumulative duration of start delays. For each affected aircraft, the start delay equals the difference between its scheduled start time and the time a START message is entered.

When traffic is stopped and then restarted all aircraft have their problem entry time adjusted to keep the original spacing intact

DA - 06 Number of Hold and Turn Delays to Aircraft (NOHTDLD)

The number of occasions that aircraft are put into a hold or a turn lasting more than 100 seconds This is counted aircraft under control

DA - 07 Duration of Hold and Turn Delays to Aircraft (TMHTDLD)

The cumulative time of hold and turn delays.

Note that hold and turn delays occur only within the sector, and that turn delays are counted only after 100 seconds. This is to allow course changes to be counted as such

DA - 08 Number of Handoffs Accepted (NOHDFAD)

The number of aircraft handed off and accepted by the subject controller

DA - 09 Handoff Acceptance Delay Time (HDFDELD)

The cumulative time between a handoff and the acceptance of that aircraft by the subject

DA - 10 Number of Contacts (Ground to Air) (NOCTCSD)

The number of times microphone transmission is made by the subject

DA - 11 Duration of Contacts (Ground to Air) (DUCTCSD)

The cumulative time of ground to air contacts

DA - 12 Total Delays (Hold + Turn + Start) (TODLYND)

(DA - 04) + (DA - 06)

DA - 13 Total Delay Time (TODLTID)

(DA - 05) + (DA - 07)

DB - 01 Number of Aircraft Handled (NACHDLD)

The number of aircraft that are accepted onto the subject frequency, or enter the sector. (See Control Rule above.)

DB - 02 Aircraft Time Under Control (ACTUC D)

The amount of time aircraft are under control, summed for all aircraft handled.

DB - 03 Average Aircraft Time Under Control (ACTUCAD)

(DB - 02) divided by (DB - 01)

DB - 04 Target Spacing Analysis - A (TSA A4D)

The number of instances that aircraft violate the separation standard of 4 miles horizontal spacing and 950 feet vertical spacing. At least one of the aircraft involved must be under control (see Control Rule above).

The measure is also referred to as 4 mile conflicts

DB - 05 Target Spacing Analysis - B (TSA B5D)

Same as above with 5 mile horizontal separation.

DB - 06 Target Spacing Analysis - C (TSA C3D)

Same as above with 3 mile horizontal separation

DB - 07 Duration of TSA-A (DURTSAD)

The cumulative duration of 4 mile conflicts

DB - 08 Duration of TSA-B (DURTSBD)

The cumulative duration of 5 mile conflicts.

DB - 09 Duration of TSA-C (DURTSCD)

The cumulative duration of 3 mile conflicts.

DB - 10 Aircraft Distance Flown (ACDSTFD)

The cumulative distance in miles flown by aircraft while under control.

DB - 11 Fuel Consumption (FUELCOD)

The cumulative fuel in pounds consumed by aircraft under control.

DB - 12 Number of Completed Flights (NCPFTSD)

The number of aircraft accepted by the subject that reach their destination and are transferred by frequency change. Control, as defined by the Control Rule, must be relinquished at the destination point to be counted as a completed flight.

Note that flights under control when the data period begins are completable.

DB - 13 Arrival Altitudes Attained (ARVLATD)

The number of arrival aircraft whose flight is completed within 100 feet of their goal altitude.

DB - 14 Departure Altitudes Attained (DPTRATD)

Same as above for departure aircraft.

DB - 15 Aircraft Time in Boundary (ACBTM D)

The cumulative time that aircraft under control are within the test sector.

SEM System Effectiveness Measure

(See Appendix C)

CPM Controller Performance Measure

(See Appendix C)

APPENDIX C

DEFINITIONS AND USAGES

DEFINITIONS AND USAGES

A. Definitions: Defined here are technical terms from the area of statistics. An attempt has been made here to define the terms in a non-technical manner. Also given are sourcebooks, usually textbooks, where more detailed definitions can be gleaned.

TERM	SOURCEBOOK	DEFINITIONS (not necessarily direct quotes from Sourcebook)
Experimental design	Kirk, Winer	A plan for the collection of data which includes the form of analysis which will be applied including the hypothesis to be tested.
Statistical significance	Mc Nemar	An outcome of a test is said to be statistically significant when the calculations and tests indicate that the probability is small (of a certain pre-determined small size) that such an outcome could have occurred by chance.
Coefficient of correlation	Mc Nemar	The coefficient of correlation expresses the degree of linear relationship between two variables. It ranges from - 1.00 (an inverse perfect relationship) to zero relationship through + 1.00 (positive perfect relationship). Signified by r.
Standard err of me irement	Mc Nemar	If it were possible to take a large number of repetitions of the same run in the exact same circumstances, their scores would normally distribute themselves around the "true" value. Using this, it is possible to say the chances that the "true" value is within certain bounds around the obtained value.

TERM	SOURCEBOOK	DEFINITIONS
Intra-class correlation	Mc Nemar	A correlation form used in cases where there is no prior reason to assign a score to one of the two distributions being correlated; in this case, the members of the pairs of judges.
Multiple correlation coefficient	Mc Nemar	A correlation between a linear combination of variables on the one hand and another single variable on the other hand. Signified by R.
Logarithmic transformation	Winer	Taking the logarithm of each score in a distribution of scores sometimes results in a distribution more closely resembling the normal distribution.
Coefficient of determination	Mc Nemar	The square of the correlation coefficient (simple or multiple) expresses the common variance between the two variables, is the variance in one accountable for by variance in the other.
Factor analysis	Guilford	A statistical technique which uses the correlation among measures to find the minimum set of measures which adequately expresses the same information in a more condensed manner.
Correction for shrinkage	Mc Nemar	In multiple correlation, this formula can be applied to correct R for the number of predictors. If the number of predictors (n) approaches N (the number of cases) there is a bias. The formula corrects for that bias and makes allowance for a decrement in R which occurs when applied to new samples of subjects. The formula is $R' 1.23..n = \sqrt{1-(1-R^2)\left(\frac{N-1}{N-M}\right)}$

TERM	SOURCEBOOK	DEFINITIONS
Phi coefficient	Harman	A coefficient expressing the relationship or resemblance between factor analysis weights from two sources.
Statistical significance, r, R	Guilford	A size of correlation coefficient (simple or multiple) which is large enough to be sure with a probability of 5/100 chances of error that there is a relationship between the variables (or combinations of variables).
Analysis of Variance	Winer	A statistical technique used to test statistical inference hypotheses. Basically it compares the variance in scores across conditions (systems) with the natural variance among peoples' performances.
Standard scores	Mc Nemar	A standard score distribution is created by expressing each score in a distribution as a deviation from the mean of the original distribution and dividing this difference by the standard deviation of the distribution.

B. Usages: Defined or explained here are words having particular usages in this report.

Day-level - In the SEM II experiment, four exercises were usually run in a day. Thus, a four-run aggregate is usually referred to as a "day-level" figure.

Traffic sample - A group of aircraft with flight plans which are scheduled to enter the simulated air traffic control system at designated entry times.

System Effectiveness (SEM) Rating - A rating of the effectiveness of the simulated air traffic control system under study. These were made by the controller observer-judges. By effectiveness is meant here the degree to which the system achieves its missing of the safe and expeditious movement of aircraft.

Controller Performance (CPM) Rating - A rating of the skill and technique of the subject controller performing the air traffic control exercise.

Measure - An objective quantitative recording of some aspect of a phenomenon. The basic measures used here were believed to be indicators of the quality of air traffic control such as the number of aircraft handled, the number of aircraft delayed, etc.

Factor score - A composite score based on several measures as was indicated by the factor analysis of the initial set of measures.

Subject - An air traffic control specialist who controls traffic in the simulated ATC system. To be distinguished from a "ghost" or support controller who plays the role of adjacent air traffic control facilities.

Primary score - One of a set of four of the original measures which was chosen as capable of being a good representative of one of the four factors.

Auxiliary score - Two of the original set of measures which were not picked out by the factor analysis, but it was decided were important to keep in the final measure set; these were the number of aircraft handled and full consumption.

Common scale - A common scale is one which is in the same units (such as is the case with such units as feet, pounds, etc., as examples). While not so readily visualizable, standard score scales adjusted to have the same mean and standard deviation have the same desirable quality (see standard scores under "Definitions."

APPENDIX D

SUPPLEMENTARY TABLES

TABLE 1. SORTED ROTATED FACTOR LOADINGS (PATTERN), DAY

	Conflict Factor	Occupancy Factor	Communication Factor	Delay Factor
Duration Target Spacing Analysis A	.926	*	*	*
Duration Target Spacing Analysis C	.864	*	*	*
Duration Target Spacing Analysis B	.854	*	*	*
Target Spacing Analysis A	.836	*	*	*
Target Spacing Analysis C	.824	*	*	*
Target Spacing Analysis B	.708	*	*	*
Time Under Control	*	.968	*	*
Fuel Consumption Under Control	*	.950	*	*
Aircraft Distance Flown Under Control	*	.924	*	*
Aircraft Time in Boundary	.311	.588	*	*
Handoff Accept Delay Time	*	.583	*	.515
Number of Ground to Air Contacts	*	*	.887	*
Duration of Ground to Air Contacts	*	*	.874	*
Path Changes	*	*	.791	*
Total Delay Time	*	*	*	.926
Total Delays	*	*	*	.924
Arrival Altitude Attained Completed Flights	*	*	*	*
Variance Accounted For	4.464	3.531	2.406	2.306

The above factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 5000 appear first Loadings less than 2500 have been replaced by *

TABLE 2. SORTED ROTATED FACTOR LOADINGS (PATTERN), DAY 2

	Conflict Factor	Communication Factor	Occupancy Factor	Delay Factor
Target Spacing Analysis A	.935	*	*	*
Target Spacing Analysis B	.892	*	*	*
Duration Target Spacing Analysis B	.859	*	*	*
Duration Target Spacing Analysis A	.834	*	*	.313
Target Spacing Analysis C	.724	*	*	.337
Duration Target Spacing Analysis C	.630	*	*	.620
Arrival Altitude Attained Completed Flights	.508	.363	*	*
Number of Ground to Air Contacts	*	.898	*	*
Duration of Ground to Air Contacts	*	.868	.284	*
Path Changes	*	.861	*	*
Time Under Control	*	*	.851	*
Time in Boundary	*	.336	.850	*
Fuel Consumption Under Control	*	*	.665	*
Total Delay Time	*	*	*	.869
Total Delays	*	*	*	.856
Aircraft Distance Flown Under Control	*	*	.346	*
Handoff Accept Delay Time	*	*	*	*
Variance Accounted For	4.370	2.657	2.305	2.174

The above factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors The rows have been rearranged so that for each successive factor, loadings greater than 5000 appear first Loadings less than .2500 have been replaced by

TABLE 3. SORTED ROTATED FACTOR LOADINGS (PATTERN), DAY 3

	Conflict Factor	Occupancy Factor	Communication Factor	Delay Factor
Duration Target Spacing Analysis A	.853	*	*	*
Target Spacing Analysis C	.847	*	*	*
Target Spacing Analysis A	.845	*	*	*
Target Spacing Analysis B	.829	*	*	*
Duration of Target Spacing Analysis C	.819	*	*	*
Duration Target Spacing Analysis B	.701	*	*	*
Time Under Control	*	.964	*	*
Time in Boundary	*	.865	*	*
Fuel Consumption Under Control	*	.853	*	*
Aircraft Distance Flown Under Control	*	.677	*	*
Duration of Ground to Air Contacts	*	*	.906	*
Number of Ground to Air Contacts	*	*	.885	*
Path Changes	*	*	.851	*
Total Delay Time	*	*	*	.868
Total Delays	*	.385	*	.828
Arrival Altitudes Attained Completed Flights	*	*	*	*
Handoff Accept Delay Time	*	*	*	*
Variance Accounted For	4.097	3.134	2.403	1.639

The above factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than .5000 appear first. Loadings less than 2500 have been replaced by

TABLE 4. CONFLICT FACTOR

Factor Score Coefficients

	Day 1	Day 2	Day 3
Path Changes	-.02	.03	-.03
Handoff Accept Delay Time	.03	-.02	00
Number of Ground to Air Contacts	.03	.01	-.01
Duration of Ground to Air Contacts	.04	.00	.01
Total Delays	-.06	-.07	-.03
Total Delay Time	-.07	-.04	-.03
Time Under Control	.01	.01	.01
Target Spacing Analysis A	.17	.23	19
Target Spacing Analysis B	.10	.24	.18
Target Spacing Analysis C	.19	.15	.26
Duration Target Spacing Analysis A	.25	.17	.20
Duration Target Spacing Analysis B	.20	.22	.13
Duration Target Spacing Analysis C	.24	.19	25
Aircraft Distance Flown Under Control	-.04	-.05	.00
Fuel Consumption Under Control	-.03	-.05	.02
Arrival Altitudes Attained Completed Flights	-.03	.16	-.06
Time in Boundary	08	-.01	-.03

TABLE 5. OCCUPANCY FACTOR

Factor Score Coefficients

	Day 1	Day 2	Day 3
Path Changes	.01	-.15	.00
Handoff Accept Delay Time	-.18	.20	.07
Number of Ground to Air Contacts	-.03	-.04	-.02
Duration of Ground to Air Contacts	-.05	.03	-.00
Total Delays	.00	-.06	.11
Total Delay Time	.06	-.06	-.13
Time Under Control	.28	.39	.32
Target Spacing Analysis A	.03	-.05	.01
Target Spacing Analysis B	.04	-.03	.02
Target Spacing Analysis C	-.03	-.04	-.06
Duration Target Spacing Analysis A	-.03	.08	-.00
Duration Target Spacing Analysis B	.03	.08	.10
Duration Target Spacing Analysis C	-.06	.00	-.04
Aircraft Distance Flown Under Control	.27	.03	.16
Fuel Consumption Under Control	.28	.25	.25
Arrival Altitudes Attained Completed Flights	-.04	-.28	-.02
Time in Boundary	.15	.46	.34

TABLE 6. COMMUNICATIONS FACTOR

Factor Score Coefficients

	Day 1	Day 2	Day 3
Path Changes	.34	.37	.36
Handoff Accept Delay Time	.06	-.04	.00
Number of Ground to Air Contacts	.38	.36	.37
Duration of Ground to Air Contacts	.39	.33	.38
Total Delays	-.02	.03	-.05
Total Delay Time	-.05	.00	.02
Time Under Control	-.01	-.04	-.01
Target Spacing Analysis A	-.07	-.02	.00
Target Spacing Analysis B	-.11	.01	-.06
Target Spacing Analysis C	-.00	.01	.02
Duration Target Spacing Analysis A	.09	-.03	-.02
Duration Target Spacing Analysis B	.01	.02	-.05
Duration Target Spacing Analysis C	.10	-.03	.06
Aircraft Distance Flown Under Control	-.05	-.04	-.02
Fuel Consumption Under Control	-.04	-.09	.02
Arrival Altitudes Attained Completed Flights	.04	.23	.03
Time in Boundary	.09	.00	-.01

TABLE 7. DELAY FACTOR

Factor Score Coefficients

	Day 1	Day 2	Day 3
Path Changes	-.06	-.00	.07
Handoff Accept Delay Time	.20	.06	.09
Number of Ground to Air Contacts	-.03	-.01	-.01
Duration of Ground to Air Contacts	-.03	.01	-.09
Total Delays	.43	.43	.53
Total Delay Time	.45	43	.55
Time Under Control	.03	-.07	-.04
Target Spacing Analysis A	.04	-.07	-.06
Target Spacing Analysis B	.16	-.15	-.10
Target Spacing Analysis C	.00	.10	-.01
Duration Target Spacing Analysis A	-.13	.07	.10
Duration Target Spacing Analysis B	-.08	-.10	-.05
Duration Target Spacing Analysis C	-.11	25	-.00
Aircraft Distance Flown Under Control	.08	.02	.05
Fuel Consumption Under Control	-.01	.04	.03
Arrival Altitudes Attained Completed Flights	.05	-.05	.01
Time in Boundary	-.12	-.07	-.13

TABLE 8. FACTOR SCORE CROSS VALIDATION CORRELATION (SHEET 1 of 2)

Conflict Factor

Factor Scores Computed Using Standard Scores of Day One Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9337	.9802
Day 2 Coefficients		1.0000	.9259
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Two Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9377	.9752
Day 2 Coefficients		1.0000	.9293
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Three Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9272	.9769
Day 2 Coefficients		1.0000	.9261
Day 3 Coefficients			1.0000

Throughput Factor

Factor Scores Computed Using Standard Scores of Day One Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.8146	.9438
Day 2 Coefficients		1.0000	.9141
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Two Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.6924	.8797
Day 2 Coefficients		1.0000	.8991
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Three Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.7401	.8954
Day 2 Coefficients		1.0000	.9184
Day 3 Coefficients			1.0000

TABLE 8. FACTOR SCORE CROSS VALIDATION CORRELATION (SHEET 2 of 2)

Communications Factor

Factor Scores Computed Using Standard Scores of Day One Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9282	.9789
Day 2 Coefficients		1.0000	.9425
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Two Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9429	.9855
Day 2 Coefficients		1.0000	.9535
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Three Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.9316	.9802
Day 2 Coefficients		1.0000	.9559
Day 3 Coefficients			1.0000

Delay Factor

Factor Scores Computed Using Standard Scores of Day One Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.8462	.9404
Day 2 Coefficients		1.0000	.9411
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Two Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.8650	.9440
Day 2 Coefficients		1.0000	.9610
Day 3 Coefficients			1.0000

Factor Scores Computed Using Standard Scores of Day Three Data

	Day 1 Coefficients	Day 2 Coefficients	Day 3 Coefficients
Day 1 Coefficients	1.0000	.8096	.9329
Day 2 Coefficients		1.0000	.9251
Day 3 Coefficients			1.0000

TABLE 9. FACTOR SCORE COEFFICIENTS FOR FULL FACTORS

Factor Analysis of SEM II Data

		Conflict	Occupancy	Communication	Delay
DA01	Path Changes	-.02	.00	.36	.00
DA09	Hand-off Accept Delay Time	.00	.07	.00	.09
DA10	Number Ground-to-Air Contacts	.01	-.03	.37	-.01
DA11	Duration Ground-to-Air Contacts	.01	.00	.38	-.03
DA12	Total Delays	-.06	.00	-.02	.43
DA13	Total Delay Time	-.04	-.06	.00	.45
DB02	Time Under Control	.01	.32	-.01	-.04
DB04	TSA-4 (Number of 4 Mile Conflicts)	.19	.03	-.02	-.06
DB05	TSA-5 (Number of 5 Mile Conflicts)	.18	.02	-.06	-.10
DB06	TSA-3 (Number of 3 Mile Conflicts)	.19	-.04	.01	.00
DB07	Duration TSA-4 (Duration of 4 Mile Conflicts)	.20	-.03	-.02	.07
DB08	Duration TSA-5 (Duration of 5 Mile Conflicts)	.20	.08	.01	-.08
DB09	Duration TSA-3 (Duration of 3 Mile Conflicts)	.24	-.04	.06	.00
DB10	Aircraft Distance Flown Under Control	.00	.16	-.04	.05
DB11	Fuel Consumption Under Control	-.03	.25	-.04	.03
DB13	Arrival Altitude Attained Completed Flights	-.03	-.02	.04	.01
DB15	Time in Boundary	-.01	.34	.00	-.12

TABLE 10. FACTOR SCORE COEFFICIENTS FOR VERY SMOOTH FACTORS

Factor Analysis of SEM II Data

		Conflict	Occupancy	Communication	Delay
DA01	Path Changes	*	*	.36	*
DA09	Hand-off Accept Delay Time	*	*	*	.09
DA10	Number Ground-to-Air Contacts	*	*	.37	*
DA11	Duration Ground-to-Air Contacts	*	*	.38	*
DA12	Total Delays	*	*	*	.43
DA13	Total Delay Time	*	*	*	.45
DB02	Time Under Control	*	.32	*	*
DB04	TSA-4 (Number of 4 Mile Conflicts)	.19	*	*	*
DB05	TSA-5 (Number of 5 Mile Conflicts)	.18	*	*	*
DB06	TSA-3 (Number of 3 Mile Conflicts)	.19	*	*	*
DB07	Duration TSA-4 (Duration of 4 Mile Conflicts)	.20	*	*	*
DB08	Duration TSA-5 (Duration of 5 Mile Conflicts)	.20	*	*	*
DB09	Duration TSA-3 (Duration of 3 Mile Conflicts)	.24	*	*	*
DB10	Aircraft Distance Flown Under Control	*	.16	*	*
DB11	Fuel Consumption Under Control	*	.25	*	*
DB13	Arrival Altitude Attained Completed Flights	*	*	*	*
DB15	Time in Boundary	*	.34	*	*

(* = .00)

TABLE 11. FACTOR SCORE COEFFICIENTS FOR VERY SMOOTH FACTORS

Factor Analysis of SEM II Data

		Conflict	Occupancy	Communication	Delay
DA01	Path Changes	*	*	.37	*
DA09	Hand-off Accept Delay Time	*	*	*	*
DA10	Number Ground-to-Air Contacts	*	*	.37	*
DA11	Duration Ground-to-Air Contacts	*	*	.37	*
DA12	Total Delays	*	*	*	.44
DA13	Total Delay Time	*	*	*	.44
DB02	Time Under Control	*	.26	*	*
DB04	TSA-4 (Number of 4 Mile Conflicts)	.20	*	*	*
DB05	TSA-5 (Number of 5 Mile Conflicts)	.20	*	*	*
DB06	TSA-3 (Number of 3 Mile Conflicts)	.20	*	*	*
DB07	Duration TSA-4 (Duration of 4 Mile Conflicts)	.20	*	*	*
DB08	Duration TSA-5 (Duration of 5 Mile Conflicts)	.20	*	*	*
DB09	Duration TSA-3 (Duration of 3 Mile Conflicts)	.20	*	*	*
DB10	Aircraft Distance Flown Under Control	*	.26	*	*
DB11	Fuel Consumption Under Control	*	.26	*	*
DB13	Arrival Altitude Attained Completed Flights	*	*	*	*
DB15	Time in Boundary	*	.26	*	*

(* = .00)

TABLE 12. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 14 DENSITY 1

	FACTOR	FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB02	15	.988	.000	.000	.000
DB10	23	.975	.000	.000	.000
DB11	24	.971	.000	.000	.000
DB15	28	.604	.564	.000	.000
DA13	13	.000	.800	.000	.000
DB05	18	.000	.757	.000	.000
DB04	17	.000	.695	.330	.000
DB08	21	.378	.693	.320	.000
DA12	12	-.251	.679	.268	.000
DB06	19	.000	.000	.953	.000
DB09	22	.000	.000	.937	.000
DB07	20	.000	.296	.824	.000
DA11	11	.000	.000	.000	.815
DA10	10	.000	-.262	.000	.802
DA01	1	.000	.000	.000	.795
DA09	9	.251	.000	.000	.698
DB13	26	.000	.000	.259	.337
	VP	3.596	3.283	2.888	2.702

The above factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

TABLE 13. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 14 DENSITY 2

		FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB06	19	.934	.000	.000	.000
DB04	17	.929	.000	.000	.000
DB07	20	.922	.000	.000	.000
DB08	21	.914	.318	.000	.000
DB09	22	.911	.000	.000	.000
DB05	18	.850	.000	.000	.000
DA13	13	.822	.000	.255	-.300
DB02	15	.358	.909	.000	.000
DB10	23	.347	.895	.000	.000
DB11	24	.372	.886	.000	.000
DA12	12	.000	-.589	.000	-.258
DA10	10	.000	.000	.917	.000
DA01	1	.000	.000	.751	.000
DA11	11	.000	.254	.747	.000
DB13	26	.000	.000	.000	.866
DB15	28	.387	.000	.000	.820
DA09	9	.310	.000	.441	-.351
	VP	6.357	3.108	2.336	.865

The above factor loading matrix has been rearranged so that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

TABLE 14. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 14 DENSITY 3

		FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB04	17	.936	.000	.000	.000
DB06	19	.922	.000	.000	.000
DB05	18	.906	.000	.000	.000
DB07	20	.887	.347	.000	.000
DB08	21	.873	.341	.000	.000
DB09	22	.801	.407	.000	.000
DB02	15	.356	.883	.000	.000
DB10	23	.365	.876	.000	.000
DB11	24	.362	.859	-.289	.000
DA09	9	.000	.527	.356	.264
DA12	12	.000	-.291	.878	.000
DA13	13	.000	.000	.861	.000
DB15	28	.421	.000	-.707	.338
DA10	10	.000	.000	.000	.853
DB13	26	.268	.000	.000	.677
DA01	1	.000	.000	.000	.660
DA11	11	.000	.356	.490	.615
	VP	5.425	3.278	2.735	2.238

The above factor loading matrix has been rearranged to that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

TABLE 15. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 16 DENSITY 1

		FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB04	17	.916	.000	.000	.000
DB08	21	.876	.000	.000	.000
DB06	19	.858	.000	.000	.000
DB07	20	.764	.295	-.432	.000
DB05	18	.713	-.387	.256	.000
DB10	23	.000	.967	.000	.000
DB02	15	.000	.952	.000	.000
DB11	24	.000	.948	.000	.000
DA10	10	.000	.000	.757	.000
DA01	1	.000	.436	.753	.000
DB09	22	.528	.306	-.663	.000
DA11	11	.000	.000	.627	.330
DB13	26	.000	.000	.000	.939
DA13	13	.000	.000	.000	-.836
DB15	28	.000	.508	.000	.640
DA09	9	-.391	.000	.000	.350
DA12	12	.000	.000	-.298	.000
	VP	3.934	3.665	2.481	2.277

The above factor loading matrix has been rearranged to that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

TABLE 16. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 14 DENSITY 2

		FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB11	24	.961	.000	.000	.000
DB02	15	.949	.000	.000	.000
DB10	23	.938	.000	.000	.000
DA12	12	-.692	.000	.000	.000
DA01	1	.542	.000	.486	.000
DB06	19	.000	.874	.000	.000
DB07	20	.000	.839	.000	.000
DB04	17	.000	.817	.000	.000
DB09	22	.000	.776	.000	.000
DA10	10	.000	.000	.852	.000
DA11	11	.000	.000	.836	.000
DA09	9	.000	.000	.806	.000
DB05	14	.000	.505	-.537	.000
DB13	26	.000	.000	.266	.897
DA13	13	.000	.000	-.264	-.839
DB15	28	.000	.000	.000	.741
DB08	21	.284	.416	-.409	-.374
	VP	3.686	3.305	3.031	2.432

The above factor loading matrix has been rearranged to that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

TABLE 17. SORTED ROTATED FACTOR LOADING, SEM I, SECTOR 16 DENSITY 3

		FACTOR 1	FACTOR 2	FACTOR 3	FACTOR 4
DB07	20	.931	.000	.000	.000
DB09	22	.926	.000	.000	.000
DB05	18	.911	.000	.000	.000
DB04	17	.909	.000	.000	.000
DB06	19	.894	.000	.000	.000
DB08	21	.803	.340	-.258	.000
DB11	24	.000	.951	.000	.000
DB10	23	.000	.947	.000	.000
DB02	15	.000	.928	.000	.000
DA12	12	.000	-.623	.317	.000
DA09	9	.000	.000	.787	.000
DA11	11	-.364	.000	.718	.000
DA10	10	-.380	.000	.617	.000
DA01	1	.000	.386	.507	.000
DA13	13	.000	.000	.000	-.869
DB13	26	.000	.000	.000	.850
DB15	28	-.255	.446	.000	.536
	VP	5.361	3.718	2.051	1.984

The above factor loading matrix has been rearranged to that the columns appear in decreasing order of variance explained by factors. The rows have been rearranged so that for each successive factor, loadings greater than 0.500 appear first, loadings less than 0.250 have been replaced by zero. For explanation of the abbreviations used in the first column, see appendix B.

APPENDIX E

COMPUTATIONS OF RUN SCORES BASED ON THE INDEX OF ORDERLINESS

THE INDEX OF ORDERLINESS

A. George Halverson derived the index of orderliness as a measure of the risk of collision of an air traffic control situation. References 10, 11, 12, and 13 of the main body of this report contain the technical background. Halverson's original work, particularly as described in an unpublished technical note of August 1971, "Index of Orderliness: Proposed Measure of ATC System Performance", contains many alternative formulations. Some of these allow for accelerated motion, turns, etc. In many cases the index values are not constrained to lie between zero and one. Some of these indices are inversely proportional to the miss circle or miss volume with or without a time-dependent exponential damping term. Halverson discussed several means of obtaining an overall rating, including frequency analyses and use of autocorrelation functions.

In the Air Traffic Control Simulation Facility (ATCSF) the instantaneous index of orderliness for two targets has been implemented in the form:

(1) $ADD = z_n \cdot r_n \cdot e^{-tm}$

where: t_m is the time to minimum horizontal separation, in minutes
r_n is the normalized horizontal separation at minimum horizontal separation (CPA)
and: z_n is the normalized vertical separation at CPA.

This version of the index of orderliness is essentially a measure of the risk (probability) of a confliction occurring if no control action is taken and all targets continue on straight, unaccelerated flight paths. The index is roughly proportional to the ratio of a) the volume of a cylinder with height equal to the altitude separation at CPA (ZRMIN) and radius of the horizontal miss distance (RMIN) to b) the volume of a cylinder of height equal to the critical altitude separation, ZRCR, and radius of the critical horizontal separation, RCR. The negative exponential term discounts potential conflictions in terms of their distance in time.

In the ATCSF the value of the risk index, ADD, is calculated every simulation time step (normally every second) for all active targets, pairwise. The calculations performed during data reduction and analysis (DR&A) in the ATCSF are as follows:

Consider two targets (1) and (2), with coordinates (x(1),y(1),z(1)) and (x(2), y(2),z(2)). Define their respective velocity components as (XDOT(1),YDOT(1), ZRATE(1)) and (XDOT(2),YDOT(2),ZRATE(2)).

Then:
Separation between the two targets is
in X coordinate, XR=X(1)-X(2)

in Y coordinate, YR=Y(1)-Y(2)

in Z coordinate, ZR=Z(1)-Z(2)

and the square of the horizontal separation, $RSQ = XR^2+YR^2$

Relative velocity components-
in X, XRDOT = XDOT(1)-XDOT(2)

in Y, YRDOT = YDOT(1)-YDOT(2)

in Z, ZRATER = ZRATE(1)-ZRATE(2)

Horizontal speed, $SPEED = (XRDOT^2+YRDOT^2)^{\frac{1}{2}}$

Relative distance to CPA, PATHL = XR·XRDOT+YR·YRDOT

Horizontal separation at closest point of approach (CPA) -
RMIN = |YRDOT·XR - XRDOT·YR| /SPEED

Time to CPA -
TRMIN = -PATHL/($SPEED^2$)

Vertical separation at CPA
ZRMIN = |ZR+ZRATE·TRMIN|

For SEM the critical horizontal separation, RCR, was set at 10.0 nmi, and the critical vertical separation, ZRCR, was set at 1,000 feet.

Equation (1) becomes:

$$(2)\ ADD = e-(TRMIN/60).\ \frac{(ZRCR-ZRMIN)}{ZRCR}\ .\frac{(RCR^2-RMIN^2)}{RCR^2}$$

where ADD is the instantaneous index for two targets.

The instantaneous (or every second) risk index, ADD, was subjected to a set of constraints. ADD was set to zero if-

a) ADD calculated is less than 0.01

b) Range is not closing, i.e.: $RSQ_{t=i} \geq RSQ_{t=i-1}$

c) The minimum range at CPA, RMIN, is greater than RCR

d) The minimum altitude separation at CPA, ZRMIN, is greater than ZRCR

e) Time to CPA,TRMIN, is greater than 420 seconds (7 minutes

f) Either target is a departure flying below 1,000 feet.

g) Targets are locked onto parallel ILS courses

h) Either target has landed or is inactive (during the one-minute interval)

The risk measure for a pair of targets for a minute is taken as the maximum value of ADD for that pair for that minute. The risk for a controller (subject) for a minute is the risk of at least one confliction occurring during that minute. This is equal to 1.0 less the risk of no conflictions, which is the product over the pairs of 1.0 minus the risk of confliction.

$$(3)\quad IOO = 1.-(1.-ADD_1)\cdot(1.-ADD_2)\cdot(1.-ADD_3)\cdots(1.-ADD_n) = 1.-\prod_{i=i}^{i=n}(1.-ADD_i)$$

A single value is needed to express the index of orderliness for a run. Three different cumulation methods were evaluated for obtaining a measure compatable with the SEM measure set and the SEM experimental conditions. These were the arithmetic mean,ORD1, the variance, ORD2, and the cumulative probability function,ORD3, of the index.

For a run of n minutes duration, the minute-by-minute values of the index,IOO (equation 3) are cumulated by:

$$(4)\quad ORD1 = \frac{\sum_{i=i}^{i=n} IOO_i}{n} = \overline{IOO}$$

$$ORD2 = \frac{\sum_{i=i}^{i=n} (IOO_i-\overline{IOO})^2}{(n-1)}$$

$$(6)\quad ORD3 = 1.0-\prod_{i=i}^{i=n}(1.-IOO_i$$

Note that ORD3 will be identically 1.0 if at any instant during the simulation the risk of confliction is 1.0. In addition, the maximum value of ORD3 would be 1.0, no matter what else occurred in the balance of the run.

APPENDIX F

LIST OF TERMINAL AREA SYSTEM EFFECTIVENESS MEASURES

Given below is a list of proposed measures for SEM experiments in the terminal environment. The major feature of these measures is their division into groups as follows.

Group A - System measures (Delays, Throughput, Communications)

Group B - System measures (Conflicts)

Group C - Radar Advisory Aircraft

Group D - IFR Aircraft

Group E - VFR Aircraft

All data measures will be calculated for the controller team as well as the North and South controllers individually.

Group A - (System Elements)

1. Number of Aircraft Handled - The number of aircraft entering the boundary of the sector, defined as being within the sectors vertical and horizontal limits (10,000 feet by 38 nautical miles from the radar center.

2. Number of Completed Flights - Flights entering the boundary and reaching ultimate points; arrivals - the middle marker; departures - the system boundary (horiz. or vert.) at or above a specified altitude, over or within 5 miles of a specified fix. A fix passage plus or minus 5 miles will be sensed even though passage may be well above the sector horizontal boundary.

Altitudes -	IFR Types 1-8	> 3000 ft.
	IFR Types 9-12	> 6000 ft.
	VFR All	> 2500 ft.

3. Aircraft Time Under Control - The amount of time aircraft are within the boundary, summed over all aircraft.

4 Number of Start Delays to Aircraft - The number of instances that an aircraft was scheduled to enter the problem while a STOP message was in effect

5. Turn and Hold Delays - The number of occasions aircraft within the boundary are put into a hold or a turn lasting more than 70 seconds.

6 Total Delays - Turn and Hold Delays plus Start Delays.

7 Start Delay Duration - The cumulative duration of Start Delays. For each affected aircraft, the start delay equals the difference between its scheduled start time and the time a start message is entered

8. Turn and Hold Duration - The cumulative duration of Turn and Hold Delays within the boundary.

9. Total Delay Duration - The cumulative duration of Start Delays as well as Hold and Turn Delays within the boundary.

10. Number of Path Changes - The number of altitude, heading, and speed changes issued to aircraft within the boundary.

Number of Path Changes Outside Boundary - The number of altitude, heading, and speed changes issued to aircraft outside the boundary

Number of Handoffs Accepted - The total number of aircraft handed off and accepted by the subject controller (inside the boundary, outside the boundary, and north to south within the boundary

13 Hand-off Accept Delay Time - The cumulative time between a handoff and the acceptance of that aircraft by the subject controller.

14 Number of Handoffs Outside the Boundary - The total number of aircraft handed off and accepted by the subject controller outside the boundary

15 North-South Hand-offs Accepted - The total number of aircraft handed off between the two members of the controller team.

16. North-South Hand-off Delay Time - The cumulative duration of North-South Hand-offs Accepted.

17. Aircraft Distance Flown - The distance flown by aircraft within the boundary summed over all aircraft

18. Aircraft Fuel Consumption - The cumulative fuel in pounds consumed by aircraft within the boundary computed using the ATCSF fuel consumption model

19. Number of Arrivals - The number of completed arrivals for both IFR and VFR aircraft.

20. Number of Departures - The number of departures for both IFR and VFR aircraft.

21 Departure Altitude Not Attained - The number of departing aircraft which do not climb above:

IFR (Category 1-8) - 3000 feet

IFR (Category 9-12) - 6000 feet

VFR - 2500 feet

22 Missed Approaches - The number of system generated missed approaches. Aircraft misaligned with the ILS are spontaneously sent into missed approach status

23 Ground-to-Air Contacts - The number of times microphone transmission is made by the subject or team.

24 Ground-to-Air Contacts Duration - The cumulative time of ground-to-air contacts.

25 Arrival Interval (Seconds) - The average number of seconds between completed arrivals.

Arrival Interval Variance (Seconds) - The variance in the distribution of arrival intervals.

27 Arrival Interval (Miles x 100) - The average number of miles between an arrival and the next arrival for all arrivals in the 60 minute test period times 100.

28 Arrival Interval Variance (Miles x 100) - The variance in the distribution of Arrival Intervals for miles x 100.

29 ILS Clearances - The number of aircraft cleared to the Instrument Landing System (ILS).

30 Control Actions After ILS Approach Clearance - Aircraft cleared for ILS approach will complete that approach unless another clearance, other than a speed control is given. These actions, after the approach clearance, are counted and shown under this heading.

Missed approaches The ATCSF already provides an automatic missed approach if an aircraft which has been cleared for an ILS approach is physically positioned such that it is impossible to perform the approach. The controller has the option of requiring vectors for spacing after an approach clearance.

31 Number of Barrier Delays - The number of instances a subject asks that all entering traffic be halted.

32. Barrier Delay Duration - The cumulative time that barrier delays remain in effect. The beginning of a barrier delay is referred to as a STOP message and its termination as a START message.

33 Aircraft Displayed - The total number of aircraft displayed on the CRT.

34. Aircraft Time Displayed - The cumulative duration of time in which active aircraft are displayed regardless of their position or classification

35. Total Fuel Consumption - The cumulative fuel consumption of all active aircraft in the problem regardless of their position or classification.

36. Total Distance Flown - The cumulative distance flown by all active aircraft in the problem regardless of their position or classification.

37. Uncontrolled Aircraft Displayed - The number of uncontrolled aircraft displayed.

38. Uncontrolled Aircraft Time Displayed - The cumulative duration in which uncontrolled aircraft are displayed.

39. Controller Keyboard Errors - Keyboard errors which are detectable as such through the baseline ATCSF software.

40. Pilot Keyboard Errors - Keyboard errors by simulator operators which are detectable as such through the baseline ATCSF software.

Group B - (System Elements)

41. Target Spacing Analysis 4.0 for IFR Aircraft (TSIFR 4.0-950 ft.) - The number of instances that IFR aircraft violate the separation standard of 4 miles horizontal spacing and 950 feet vertical spacing. Both aircraft involved must be under IFR control and within the boundary.

42. Target Spacing Analysis 3.0 for IFR Aircraft (TSIFR 3.0-950 ft.) - Same as TSIFR 4.0 except horizontal separation is 3 miles.

43 Target Spacing Analysis 2.5 for IFR Aircraft (TSIFR 2.5-950 ft.) - Same as TSIFR 3.0 except horizontal separation is 2.5 miles.

44 Target Spacing Analysis 2.0 for IFR Aircraft (TSIFR 2.0-950 ft.) - Same as TSIFR 2.5 except horizontal separation is 2.0 miles.

45 Target Spacing Analysis 1.0 for IFR Aircraft (TSIFR 1.0-950 ft.) - Same as TSIFR 2.0 except horizontal separation is 1.0 mile.

46 Duration TSIFR 4.0 - The cumulative duration of 4.0 mile conflicts for IFR aircraft.

47 Duration TSIFR 3.0 - The cumulative duration of 3.0 mile conflicts for IFR aircraft.

48 Duration TSIFR 2.5 - The cumulative duration of 2.5 mile conflicts for IFR aircraft.

49 Duration TSIFR 2.0 - The cumulative duration of 2.0 mile conflicts for IFR aircraft.

50 Duration TSIFR 1.0 - The cumulative duration of 1.0 mile conflicts for IFR aircraft.

51 Target Spacing Analysis 2.0 for VFR Aircraft (TSVFR 2.0-450 ft.) The number of instances that VFR aircraft violate the separation standard of 2.0 miles horizontal spacing and 450 ft. vertical spacing below a height of 6,500 feet and within a radius of 10 miles of the radar center At least one aircraft must be VFR

52 Target Spacing Analysis 1.5 for VFR Aircraft (TSVFR 1.5-450 ft.) Same as TSVFR 2.0, but with horizontal separation of 2.0 miles.

53 Target Spacing Analysis 1.0 for VFR Aircraft (TSVFR 1.0-450 ft.) - Same as TSVFR 1.5, but with horizontal separation of 1.0 mile.

54 Duration TSVFR 2.0 - The cumulative duration of 2.0 mile conflicts for VFR aircraft.

55 Duration TSVFR 1.5 - The cumulative duration of 1.5 mile confl for VFR aircraft.

56. Duration TSVFR 1.0 - The cumulative duration of 1.0 mile conflicts for VFR aircraft.

57 Target Spacing Analysis 6.0 for Aircraft on the ILS (TSILS 6.0) The number of instances that appropriate categories of aircraft violate the 6.0 mile separation standard in the table below.

Conflict Separation Parameters

Index No.	Trailing A/C Size	Lead A/C Size	Horizontal Separation (Pillbox Radius)
1.	Small	Small	3 miles
2.	Small	Large	4 miles
3	Small	Heavy	6 miles
4	Large	Small	3 miles
5	Large	Large	3 miles
	Large	Heavy	5 miles
	Heavy	Small	3 miles
8.	Heavy	Large	3 miles
9	Heavy	Heavy	4 miles

58. Target Spacing Analysis 4.0 for Aircraft on the ILS (TSILS 4.0) - The same as TSILS 5.0, except separation is 4.0 miles.

59. Target Spacing Analysis 4.0 for Aircraft on the ILS (TSILS 4.0) - The same as TSILS 5.0, except horizontal separation is 4.0 miles.

60. Target Spacing Analysis 3.0 for Aircraft on the ILS (TSILS 3.0) - The same as TSILS 4.0, except horizontal separation is 3.0 miles.

Duration of TSILS 6.0 - The cumulative duration of 6.0 mile conflicts for aircraft on the ILS.

Duration of TSILS 5.0 - The cumulative duration of 5.0 mile conflicts for aircraft on the ILS.

63. Duration of TSILS 4.0 - Same as above, but for 4.0 mile conflicts.

64. Duration of TSILS 3.0 - Same as above, but for 3.0 mile conflicts.

65. ARTS Conflict Alert - The number of ARTS conflict alerts.

66. IFR (3 mile) Conflicts Outside Boundary - The number of three mile conflicts occurring outside the boundary for IFR aircraft.

Group C - (Radar Advisory Aircraft)

list of measures below is defined here only for Radar Advisory Aircraft. The counts and durations of these measures are computed for Radar Advisory Aircraft only. In every other respect their definition is identical to the analogous system elements in Group A.

67. Number of Aircraft Handled (RA)

68 Aircraft Time Under Control (RA)

Number of Start Delays to Aircraft (RA)

70 Turn and Hold Delays (RA)

71. Total Delays (RA)

Start Delay Duration (RA)

Turn and Hold Duration (RA)

74 Total Delay Duration (RA)

75 Number of Path Changes (RA)

76. North-South Handoff Accepts (RA)

77. North-South Handoff Accept Delay Time (RA)

78. Aircraft Distance Flown (RA)

79, Aircraft Fuel Consumption (RA)

Group D - (IFR Aircraft)

The list of measures below is defined here only for IFR aircraft. The counts and durations of these measures are computed for IFR aircraft only. In every other respect their definition is identical to the analogous system elements in Group A.

80 Number of Aircraft Handled (IFR)

81 Number of Completed Flights (IFR)

82 Aircraft Time Under Control (IFR)

83 Number of Start Delays to Aircraft (IFR)

84. Turn and Hold Delays (IFR)

85. Total Delays (IFR)

86. Start Delay Duration (IFR)

87 Turn and Hold Duration (IFR)

88. Total Delay Duration (IFR)

89 Number of Path Changes (IFR)

90. Number of Handoffs Accepted (IFR)

91. Handoff Accept Delay Time (IFR)

92. North-South Handoff Accepts (IFR)

93. North-South Handoff Accept Delay Time (IFR)

94. Aircraft Distance Flown (IFR)

95. Aircraft Fuel Consumption (IFR)

96. Arrivals (IFR)

97. Departures (IFR)

98. Departure Altitude Not Attained (IFR)

99. Missed Approaches (IFR)

Group E - (VFR Aircraft)

The list of measures below is defined here only for VFR aircraft. The counts and durations of these measures are computed for VFR aircraft only. In every other respect their definition is identical to the analogous system elements in Group A.

100. Number of Aircraft Handled (VFR)

101. Number of Completed Flights (VFR)

102. Aircraft Time Under Control (VFR)

103. Number of Start Delays to Aircraft (VFR)

104. Turn and Hold Delays (VFR)

105. Total Delays (VFR)

106. Start Delays Duration (VFR)

107. Turn and Hold Duration (VFR)

108. Total Delay Duration (VFR)

109. Number of Path Changes (VFR)

110. North-South Handoff Accepts (VFR)

111. North-South Handoff Accept Delay Time (VFR)

112. Aircraft Distance Flown (VFR)

113. Aircraft Fuel Consumption (VFR)

114. Arrivals (VFR)

115. Departures (VFR)

116. Departure Altitude Not Attained (VFR)

☆U.S. GOVERNMENT PRINTING OFFICE: 1993 - 704-078/80059