technical note technical note

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

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.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/CT - 83/26		

4. Title and Subtitle	5. Report Date
METHODS AND MEASUREMENTS IN REAL-TIME AIR TRAFFIC CONTROL SYSTEM SIMULATION	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.
	13. Type of Report and Period Covered Technical Note
12. Sponsoring Agency Name and Address U.S. Department of Transportation Technical Center Atlantic City Airport, N.J. 08405	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	18. Distribution Statement
Simulation, Real Time, Air Traffic Control, System Test, Human Factors, Experimental Design and Analysis, Dynamic Simulation	

19. Security Classif. (of this report)	20. Security Classif. (of this page)	21. No. of Pages	22. Price
UNCLASSIFIED	UNCLASSIFIED	176	

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	1
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	109
CONCLUSIONS	114
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
1	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		Page
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)	102

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 and can 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

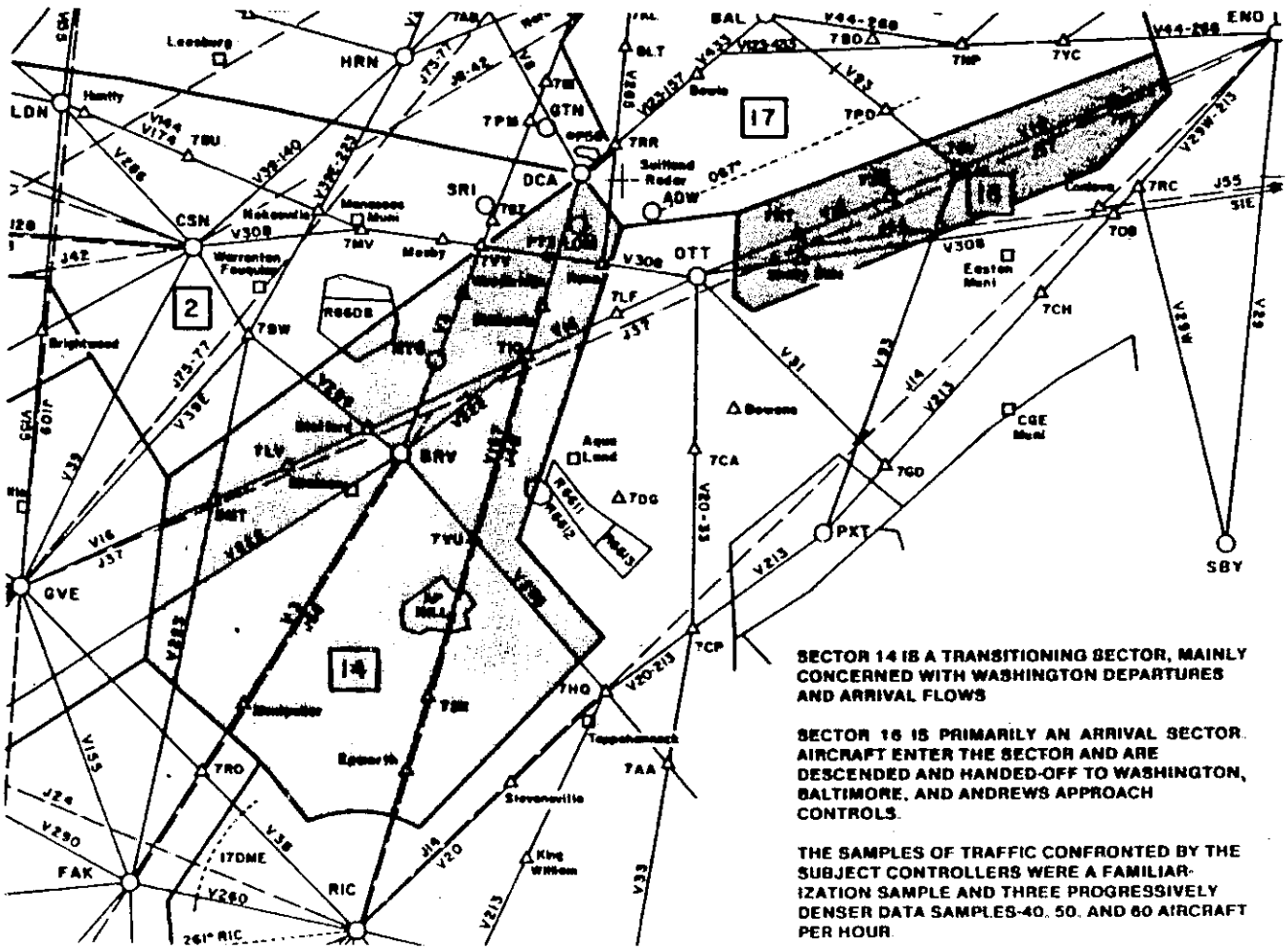


FIGURE 1. SECTOR MAPS, SEM I AND SEM II

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

SEM
POST RUN
MONITORING FORM

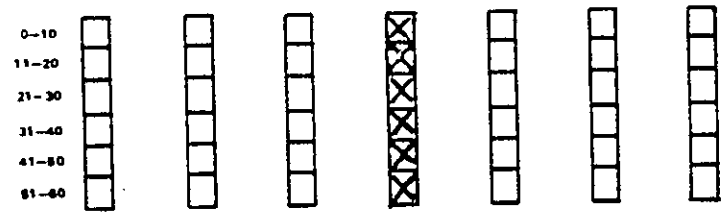
CONSIDERING THE LEVEL OF TRAFFIC INVOLVED AND FROM THE VIEWPOINT OF THE PILOTS:

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.



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

CONSIDERING THE LEVEL OF TRAFFIC INVOLVED CONSIDER THE CONTROL TECHNIQUE (I. e. TAPI 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

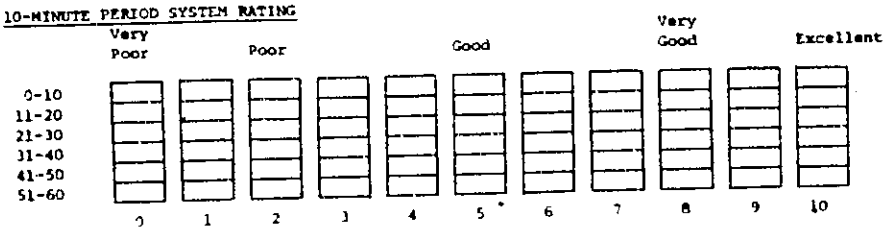
SYSTEM EFFECTIVENESS MEASUREMENT
MONITORING FORM

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

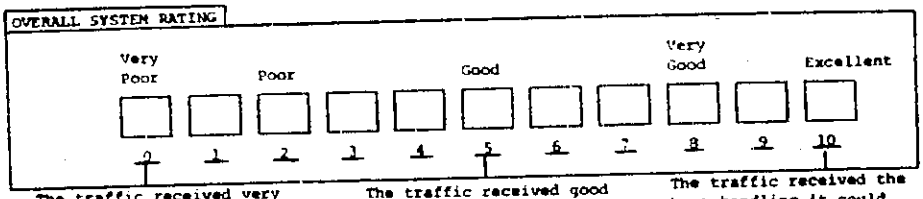
SYSTEM RATING

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.

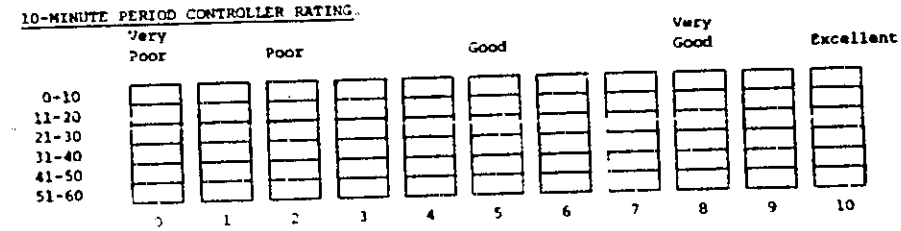
The traffic received good handling, arriving with fair safety, speed, and smoothness.

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.

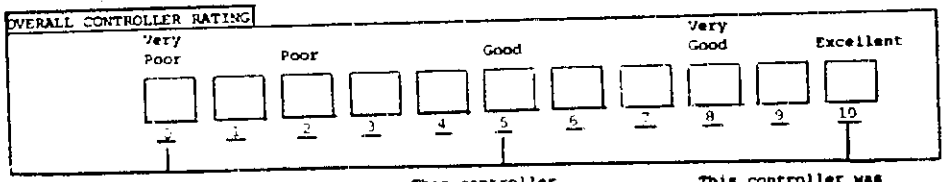
CONTROLLER RATING

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 encountered 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

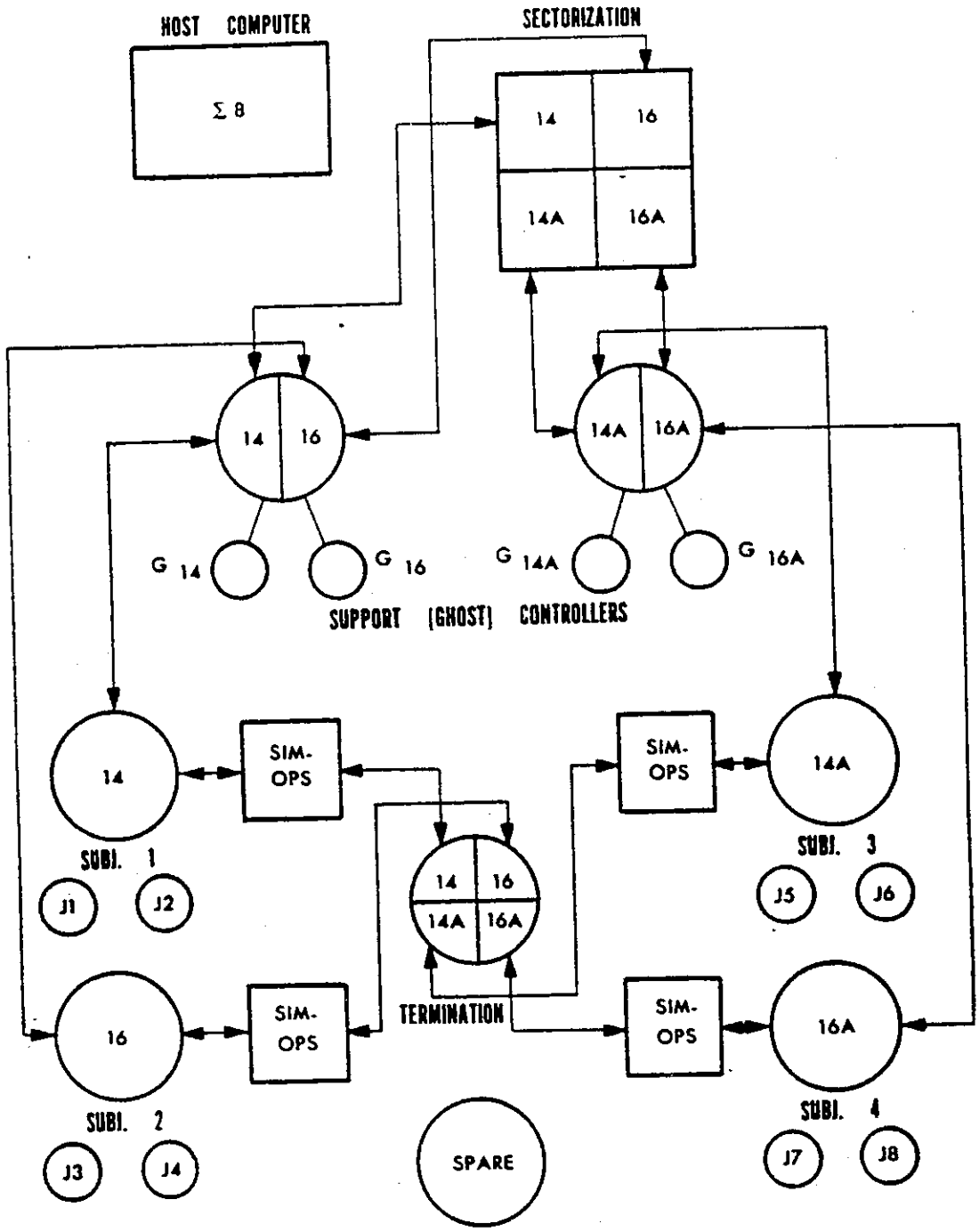


FIGURE 6. SEM I SCHEMATIC LABORATORY LAYOUT

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

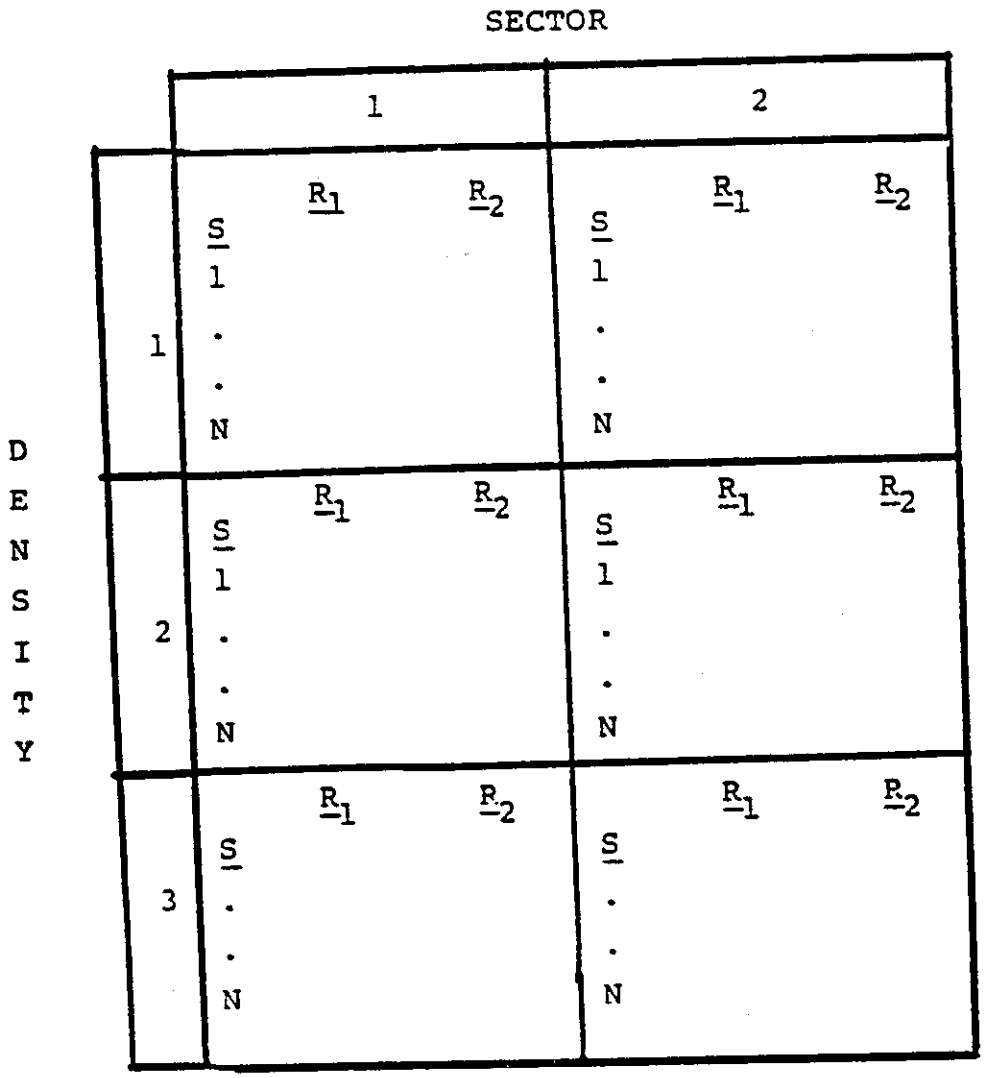


FIGURE 7. EXPERIMENTAL DESIGN, SEM I

	Time Periods											
Subjects	1	2	3	4	5	6	7	8	9	10	11	12
1	A_1	B_1	C_1	S_1	A_2	B_2	C_2	S_2	A_3	B_3	C_3	S_3
2	B_1	C_1	A_1	S_2	B_2	C_2	A_2	S_1	B_3	C_3	A_3	S_4
3	C_1	A_1	B_1	S_3	C_2	A_2	B_2	S_4	C_3	A_3	B_3	S_1
4	S_1	S_2	S_3	S_4	S_2	S_1	S_4	S_3	S_3	S_4	S_1	S_2
	Block 1				Block 2				Block 3			

S = seed, A, B, C = three generated samples

For $i \neq j$, sample (i) and sample (j) are the same problem, except for a change in tags.

FIGURE 8. EXPERIMENTAL DESIGN, SEM II

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
1	.	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****
Day 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
Day 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

	r
Conflict Factor	
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.

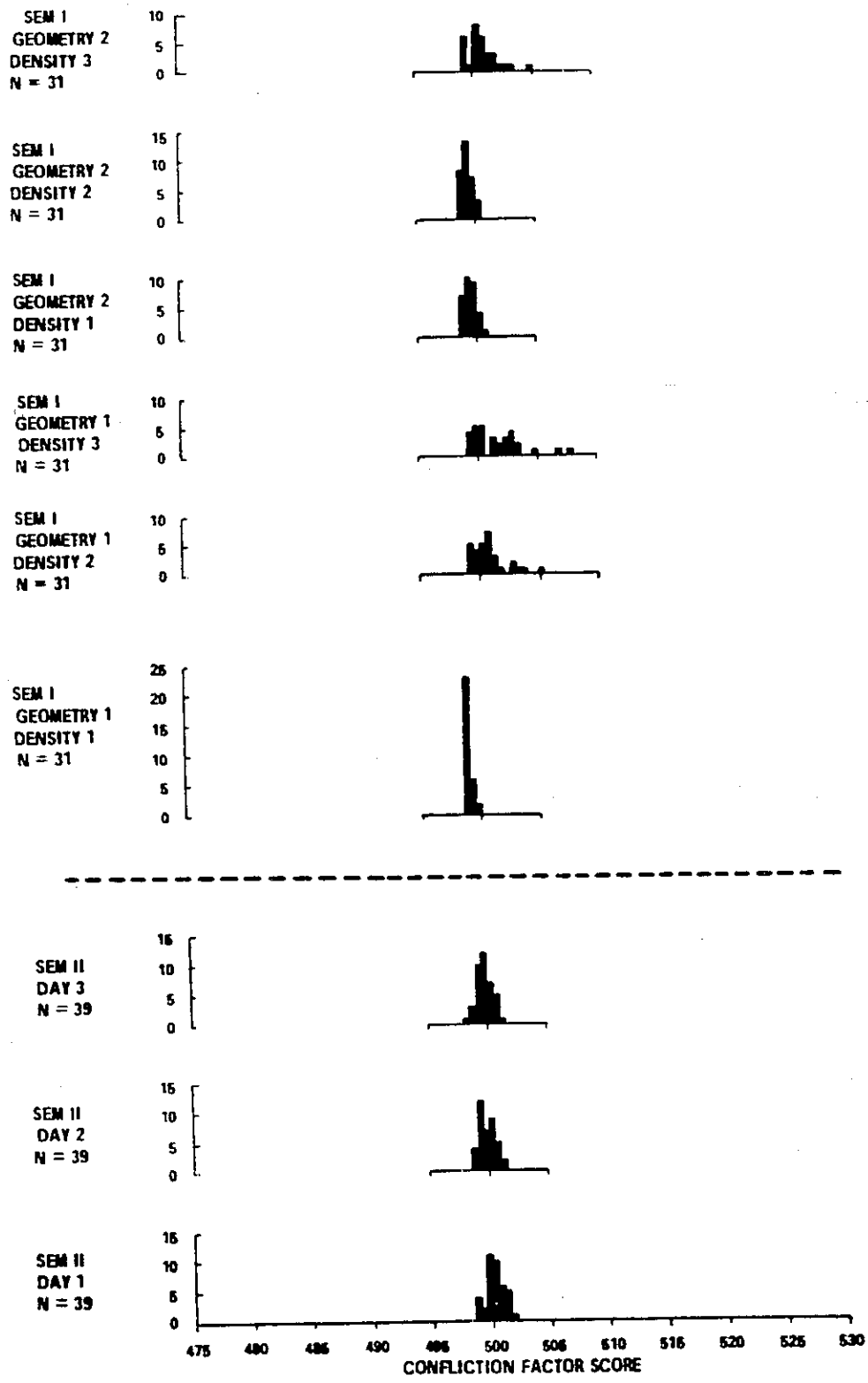


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 1 of 4)

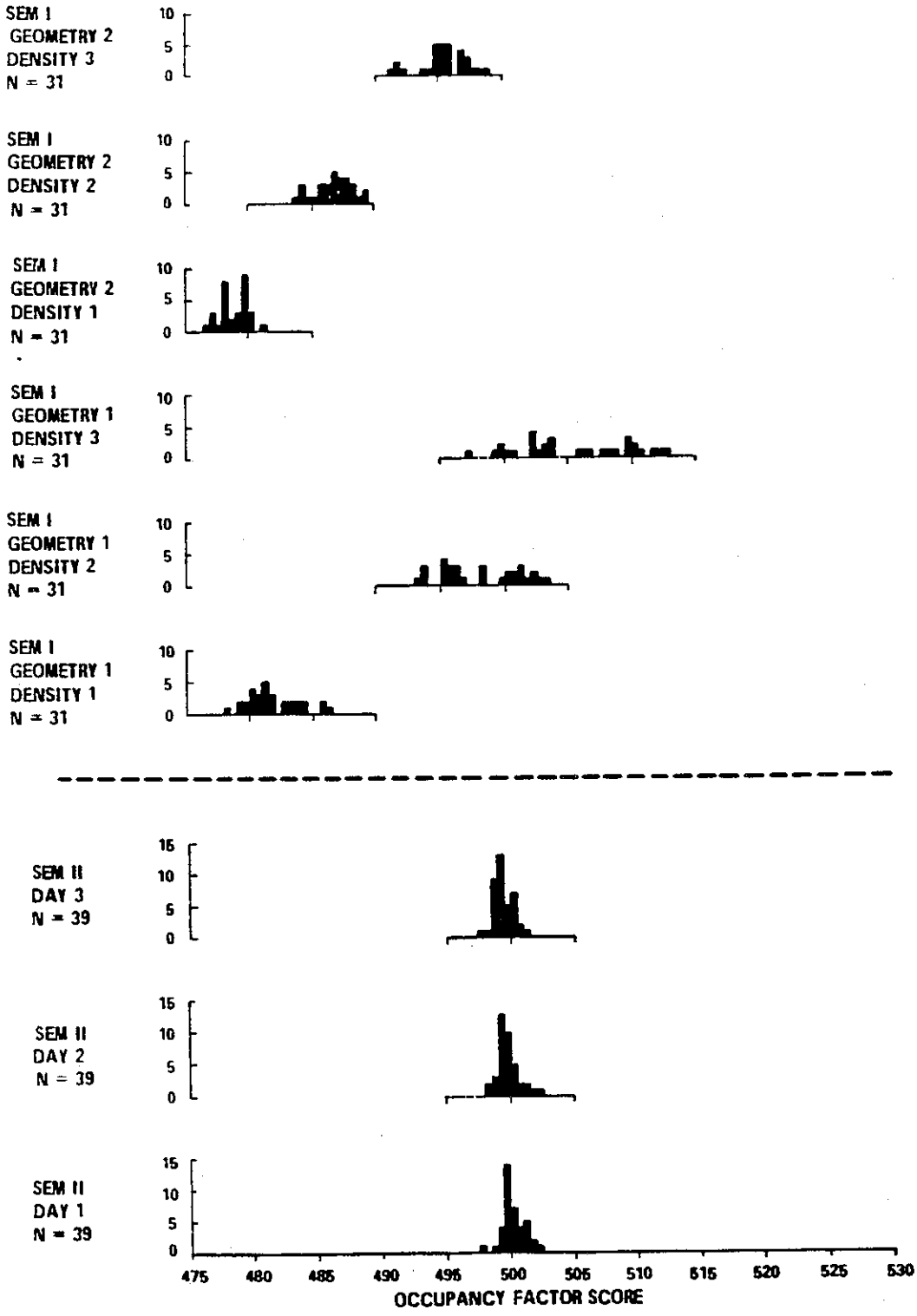


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 2 of 4)

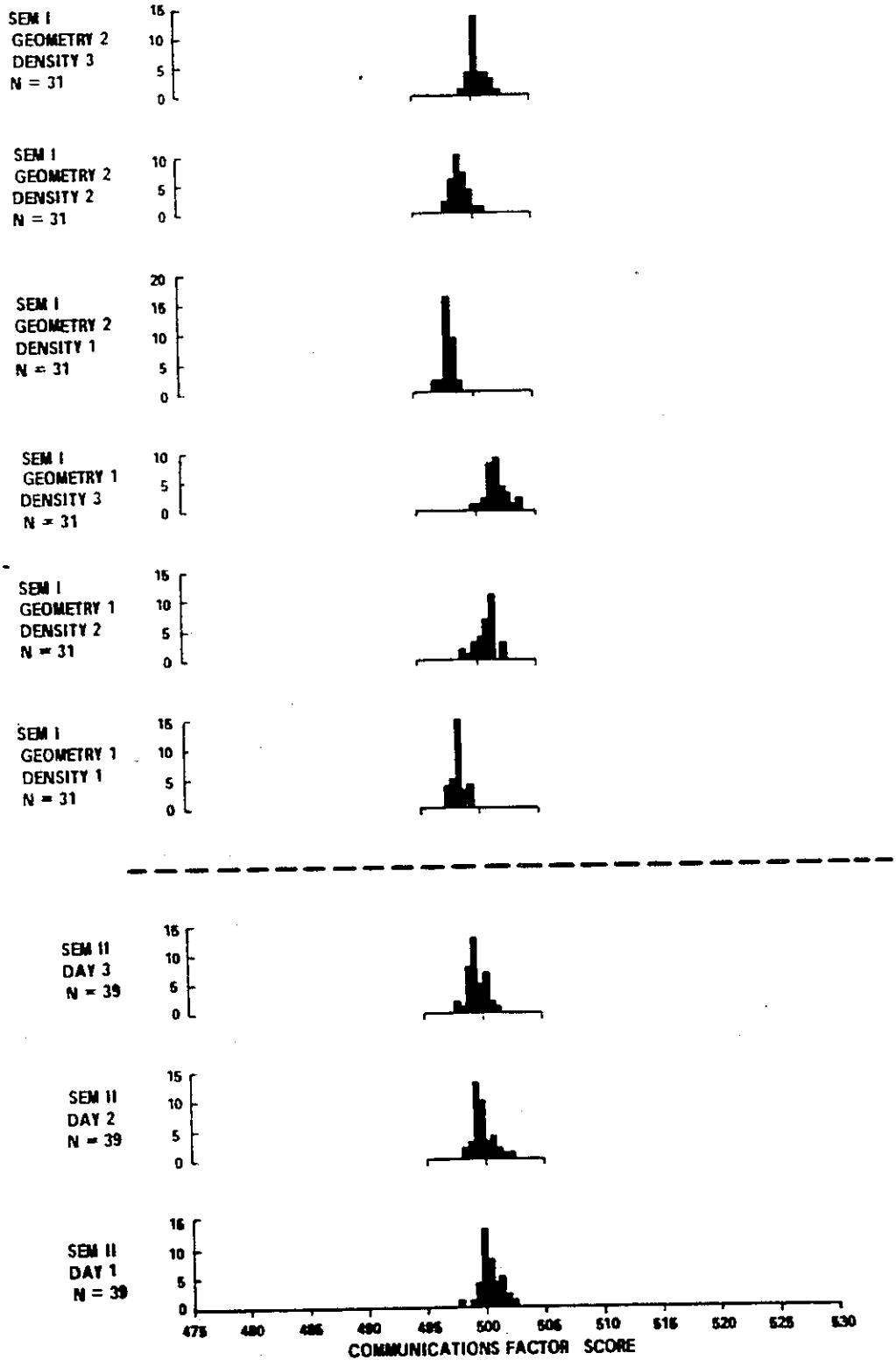


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 3 of 4)

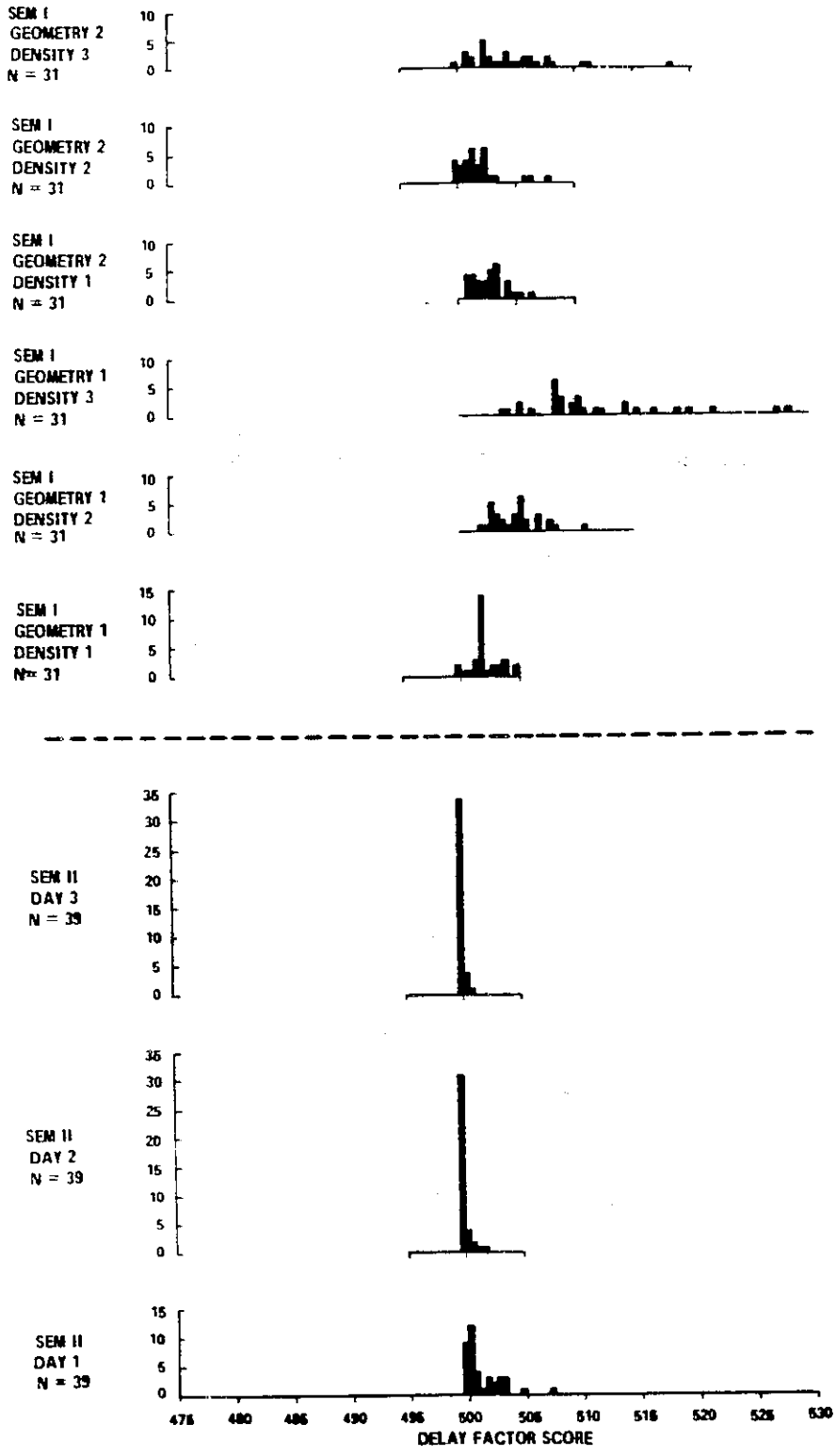


FIGURE 11. DISTRIBUTION OF FACTOR SCORES FOR SEM I AND SEM II EXPERIMENTAL CONDITIONS (Sheet 4 of 4)

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
CPM	.17	.06	.61	.56	.46	.44	.13	.43	.48	.44	.55	.39
SEM	.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
CPM	.64	.64	.40	.43	.30	.32	.58	.69	.50	.53	.44	.66
SEM	.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	-.54	-.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	-.35	-.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

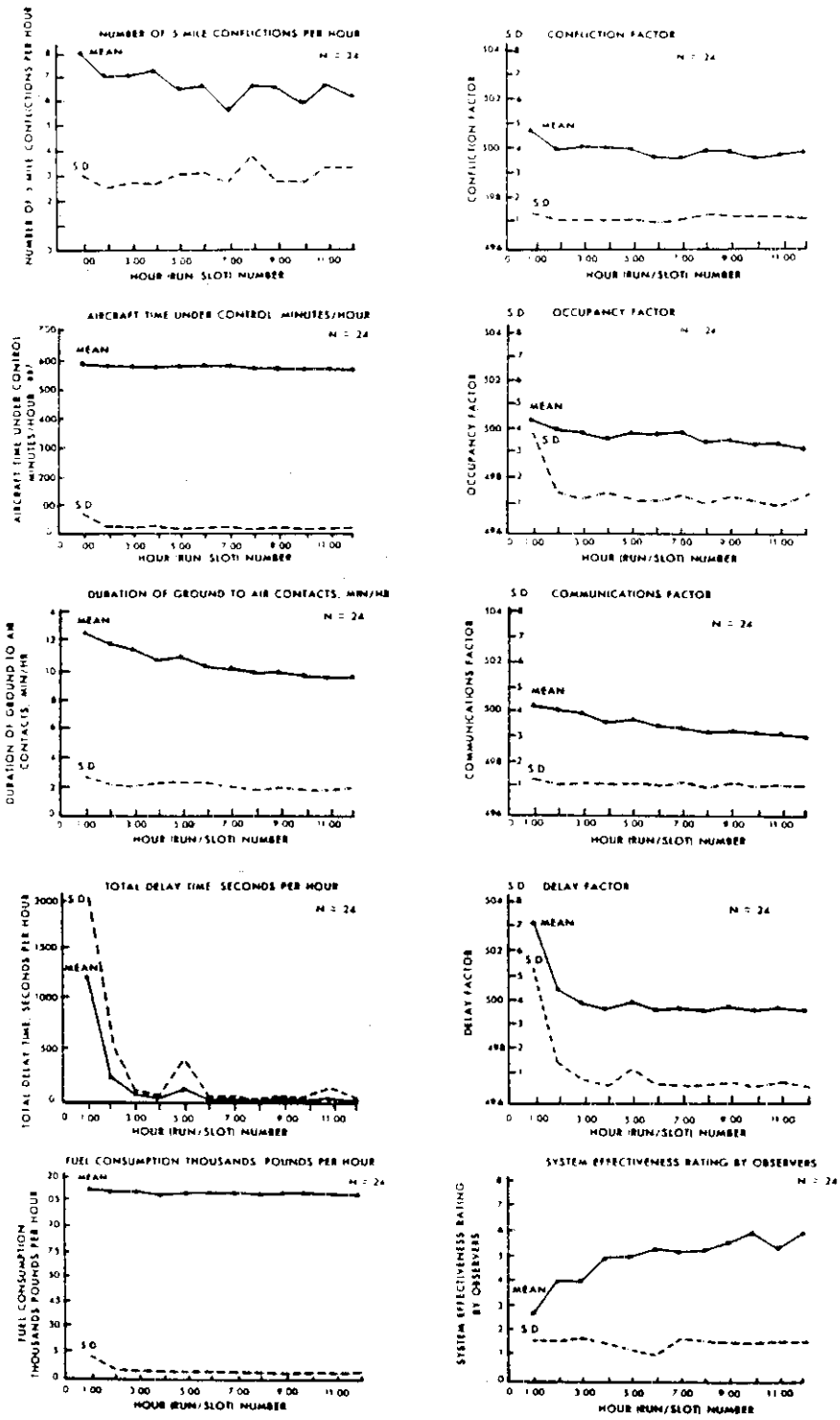


FIGURE 12. PLOT OF COURSE OF MAJOR MEASURES OVER TIME

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	0.1006	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 F	Number of Conflictions P/.05	Time Under Control F	Time Under Control P/.05	Duration G/A Communications F	Duration G/A Communications P/.05	Total Delay Time F	Total Delay Time 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
Hour 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
Hour 7 vs rest	.02	.89	1.56	.21
Hour 8 vs rest	.03	.87	0.06	.80
Hour 9 vs rest	.05	.83	0.35	.55
Hour 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

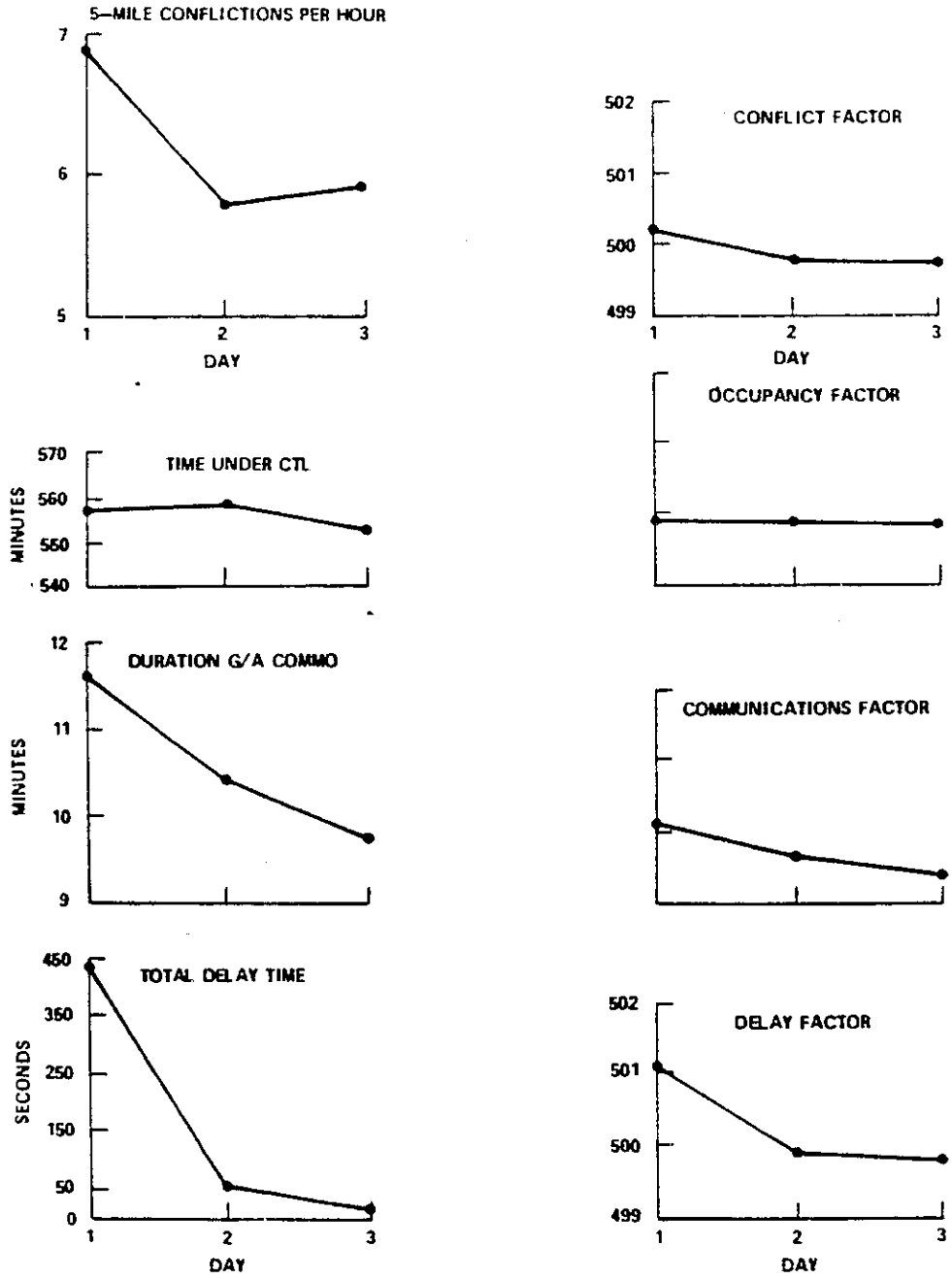


FIGURE 13. PLOT OF DAY MEANS

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	34.327	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	199.232	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.	113,403	111,770	110,922

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	df	P	Density F	df	P	Geometry by Dens. F	df	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.

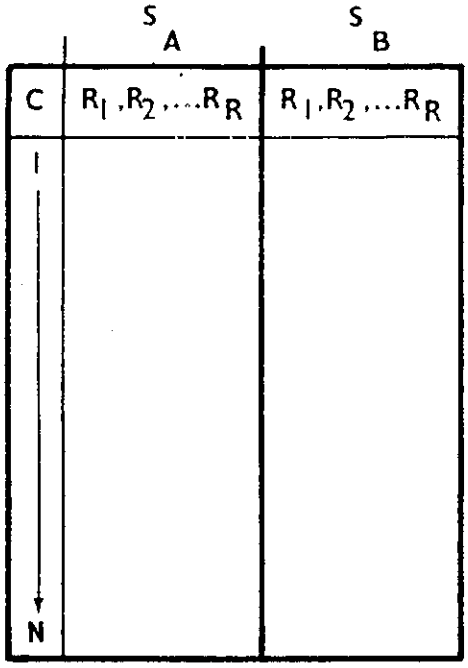
It is not appropriate in this report to go into a detailed basic orientation on the matter of statistical hypothesis testing as it particularly applies in the unique field of real-time simulation testing of air traffic control man-machine systems for effectiveness. In very general terms, it is important 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.

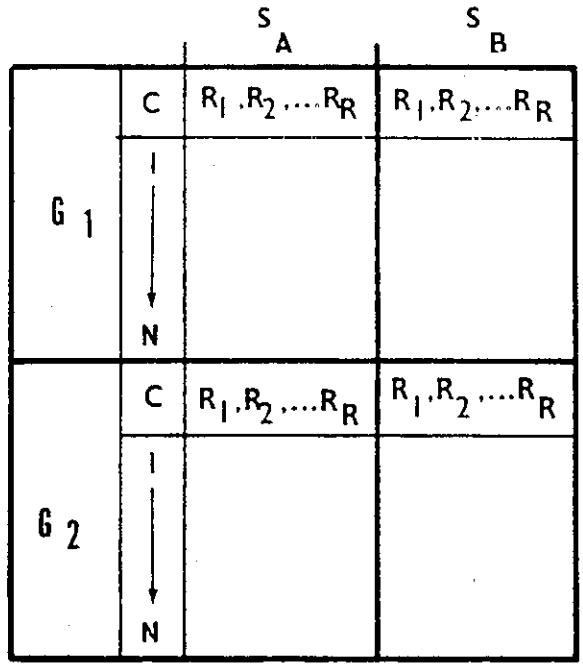
The power tables must be entered with two parameters: (1) the size of the difference in each of the measures which is considered worthwhile detecting in each measure as a meaningful or important difference between systems, and (2) the alpha and beta error probabilities it is felt important to protect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

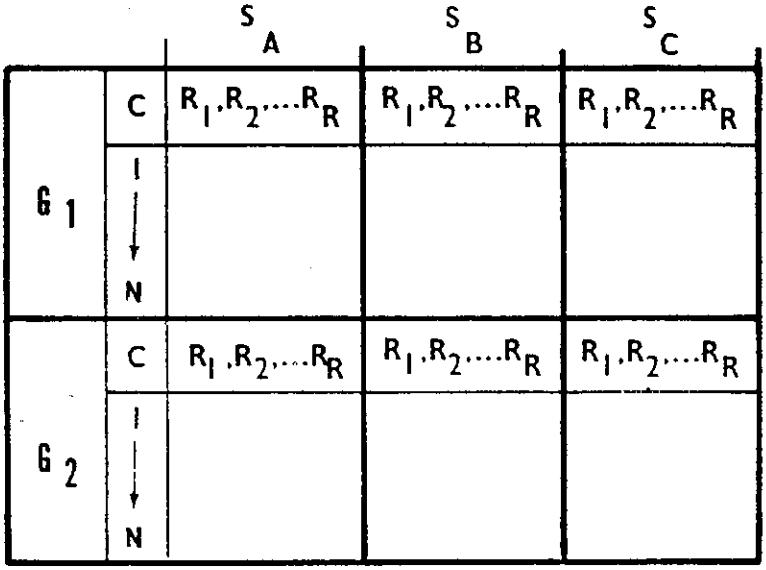
The tables were formulated to be specific to four statistical experimental design (a technical term, see appendix C) types which might be expected to be frequently applicable to system testing. Design A is a paired, or correlated t test design, in which the same controllers are used in both systems at 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.



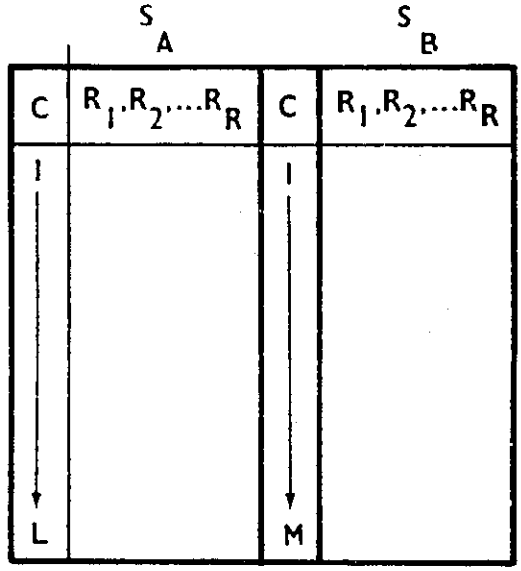
Design A: Two systems (S_A and S_B), several repetition runs (R), and several controllers (or crews) C, same C's work in both systems.



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
Number of Replicates	1	14
	2	11
	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.