AD-787 594

TRAINEE PERFORMANCE MEASUREMENT DEVELOPMENT USING MULTIVARIATE MEASURE SELECTION TECHNIQUES

Donald Vreuls, et al

Manned Systems Sciences, Incorporated

Prepare_ for:

Naval Training Equipment Center
Advanced Research Projects Agency

September 1974

DISTRIBUTED BY:

NTIS
National Technical Information Service
U. S. DEPARTMENT OF COMMERCE
5285 Port Royal Road, Springfield Va. 22151

Technical Report: NAVTRAEQUIPCEN 73-C-0066-1

TRAINEE PERFORMANCE MEASUREMENT DEVELOPMENT USING MULTIVARIATE MEASURE SELECTION TECHNIQUES

ABSTRACT

A study was conducted to extend a descriptive structure for measuring human performance during training to a fixed-wing, high-performance aircraft simulation, and to develop measure selection statistical techniques. The effort required: (1) definition of candidate performance measures for the simulated flight task, (2) development of computer programs to acquire raw data and produce candidate measures for 18, one-hour training sessions with four participants, and (3) most especially, to develop methods to reduce the resulting candidate measures to a small and efficient set which reflects the skill change that occurs as a function of training.

It was desired that the resultant measurement have the capability of: (1) discriminating between different levels of proficiency and (2) predicting later performance based on measures of current performance. Therefore, two measure selection methods were developed. One was based *in part* on a multiple discriminant analysis model. The second was based *in part* on a canonical correlation model.

The multiple discriminant procedure was able to reduce measures to an efficient set which could discriminate between early and later training performance, and produced weights for the summation of individual measures into one composite score. Minor improvements in the method were suggested.

The canonical correlation procedure to choose measures which predict later performance worked also, but the data revealed the need for additional criteria in the selection of predictive measures. More comprehensive algorithms were suggested.

It was concluded that additional data should now be collected to verify the results with a large number of participants. Real-time, or near-real-time production of measures while training is in progress should be attempted in an automated flight trainer.

UNCLASSIFIED
Security Classification

AD787594

DOCUMENT CONTROL DATA - R & D

(Security classification of title, body of abstract and indexing annotation must be entered when the overall report is classified)

1. ORIGINATING ACTIVITY *(Corporate author)* Manned Systems Sciences, Inc. Northridge, California 91324	2a. REPORT SECURITY CLASSIFICATION Unclassified 2b. GROUP

3. REPORT TITLE

TRAINFE PERFORMANCE MEASUREMENT DEVELOPMENT USING MULTIVARIATE MEASURE SELECTION TECHNIQUES

4. DESCRIPTIVE NOTES *(Type of report and inclusive dates)*

Final Report (Dec 72 - Dec 73)

5. AUTHOR(S) *(First name, middle initial, last name)*

Donald Vreuls, Richard W. Obermayer, and Ira Goldstein

6. REPORT DATE September 1974	7a. TOTAL NO. OF PAGES 53	7b. NO. OF REFS 15
8a. CONTRACT OR GRANT NO. N61339-73-C-0066 b. PROJECT NO. NAVTRAEQUIPCEN Task No. 3754-01P01 c. d.	9a. ORIGINATOR'S REPORT NUMBER(S) 9b. OTHER REPORT NO(S) *(Any other numbers that may be assigned this report)* NAVTRAEQUIPCEN 73-C-0066-1	

10. DISTRIBUTION STATEMENT

Approved for public release; distribution unlimited.

11. SUPPLEMENTARY NOTES Reproduced from best available copy.	12. SPONSORING MILITARY ACTIVITY Naval Training Equipment Center Orlando, Florida 32813

13. ABSTRACT

A study was conducted to extend a descriptive structure for measuring human performance during training to a fixed-wing, high-performance aircraft simulation, and to develop measure selection statistical techniques. The effort required (1) definition of candidate performance measures for the simulated flight task, (2) development of computer programs to acquire raw data and produce candidate measures for 18, one-hour training sessions with four participants, and (3) most especially, to develop methods to reduce the resulting candidate measures to a small and efficient set which reflects the skills change that occurs as a function of training. It was desired that the resultant measurement have the capability of ~~(1)~~ discriminating between different levels of proficiency and ~~(2)~~ predicting later performance based on measures of current performance. ~~Therefore,~~ T two measure selection methods were developed. One was based in part on a multiple discriminate analysis model. The second was based in part on a canonical correlation model. The multiple discriminate procedure was able to reduce measures to an efficient set which could discriminate between early and later training performance, and produced weights for the summation of individual measures into one composite score. Minor improvements in the method were suggested. The canonical correlation procedure to choose measures which predict later performance worked also, but the data revealed the need for additional criteria in the selection of predictive measures. More comprehensive algorithms were suggested. It was concluded that additional data should now be collected to verify the results with a large number of participants. Real-time, or near-real-time production of measures while training is in progress should be attempted in an automated flight trainer.

UNCLASSIFIED
Security Classification

(and the results of applying them are discussed.)

14. KEY WORDS	LINK A		LINK B		LINK C	
	ROLE	WT	ROLE	WT	ROLE	WT
Performance measurement						
Training performance						
Automated flight training						
Measure selection						
Computer measurement system						
Performance discriminate analysis						
Multiple discriminate analysis						
Canonical correlation						

FOREWORD

This report documents the current status of ongoing man-machine training performance measurement method development. A previous report (Vreuls, Obermayer, Lauber, and Goldstein, 1973) emphasized the development of a descriptive structure for obtaining measurement in a man-machine training situation. This report emphasizes the development and current status of measurement selection techniques based on multivariate analyses, which were explored as a means of selecting measures, rather than the more traditional use as a means of personnel selection and classification. Further work on measure selection techniques is necessary, and is ongoing under the direction of NAVTRAEQUIPCEN and the sponsorship of the Advanced Research Projects Agency.

The views and conclusions contained in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the Advanced Research Projects Agency or the United States Government.

Ira Goldstein

IRA GOLDSTEIN
Scientific Officer

This research was supported by the Advanced Research Projects Agency of the Department of Defense and was monitored by the Naval Training Equipment Center under Contract N61339-73-C-0066

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

SECTION I

INTRODUCTION AND SUMMARY

Performance measurement produces information needed for a specific purpose, such as the evaluation of trainee performance or the conduct of training. Performance measurement is therefore vital to improved training or improved evaluation. Typically, military man-machine system performance measurement involved the processing of large quantities of continuously varying information; consequently, such measurement is beyond the capabilities of manual processing and simple measurement devices, and thus must be automated.

Automation, however, places severe demands on exact definition of the conditions during which measurement takes place and a succinct definition of measures which have utility. The definition of useful measures itself has been a major technical challenge (e.g., Smode, 1971; Vreuls and Obermayer, 1971b). Where performance measurement has been used, it has been selected, commonly, on the basis of "common practice" or on the basis of an analysis of the skills, knowledges, task components and/or mission objectives. Several studies (cf., Vreuls and Obermayer, 1971a; Vreuls, et al., 1973; Knoop and Welde, 1973) have emphasized that analytic methods alone fail to satisfactorily define measurement.

Measurement defined only on the basis of common practice or analysis is likely to be overabundant, unwieldy and perhaps impossible to implement in an operational setting. The large quantities of information thus produced are likely to include (1) different ways to measure the same behavior and (2) measures of behavior and system performance which may prove to be unimportant. Although the measurement development process must start with a good analysis, it is necessary to seek empirical methods to reduce measurement to a small, efficient set.

The reduction of initially defined measures into a set which can be shown, mathematically, to have the desired properties is called the *measure selection* process. Previous research by the authors established and tested a descriptive structure for obtaining measurement in a man-machine training situation. The primary emphasis of the work reported herein was the design and development of measure selection techniques which were based on multivariate statistical models which consider the total set of measures, rather than consideration of individual measures without regard to what is happening to other measures at the same time.

SUMMARY OF METHOD

An empirical method was used to develop the measure selection techniques. Data were collected while human participants underwent 18, one-hour training sessions. Raw data were converted to analytically define performance measures through the use of measure producing programs which read the data tapes at the conclusion of training. The performance measures were then used as a data base for the development of multivariate statistical selection techniques.

Measure selection development was oriented to the use of measurement within automated, adaptive flight training systems. It was desired that the resultant measurement have the capability (1) of discriminating between different levels of proficiency and (2) to predict later performance based on measures of current performance.

RESULTS

As it was defined herein, the discriminant procedure worked for measure selection. In one of the test cases, 24 initial measures were reduced to seven which could discriminate between early and later training performance. The procedure also produced the weights for summing the measures into one composite score. Minor improvements were recommended.

The canonical correlation procedure to choose measures which predict later performance worked also; however, the data revealed the need for more complex criteria in the selection of predictive measures. More comprehensive algorithms were suggested.

IMPLICATIONS FOR FURTHER RESEARCH

It is possible to mathematically define an efficient set of measures which can significantly change during training of psychomotor skills for flight control. Thus, the discriminant technique should be applied to automated training, and flight training where instrumentation and support subsystems are available. Minor method improvement should be undertaken to fine-tune the discriminant procedure, as suggested herein.

Further design and testing of the predictive measure selection method is needed. It was felt that with additional data and with suggested program changes, that the next iteration with the predictive procedure should solve many of the presently encountered problems. However, the problems of assessing proper criteria for performance prediction should wait for data collection in training programs with a broader scope than considered in this study.

SECTION II

A METHOD FOR MEASUREMENT DEFINITION

ANALYSIS FOR MEASUREMENT

Detailed analyses of missions and human operator tasks are conventional ways to provide foundation information for the study of man-machine problems. These analyses provide a concise description of the various separate parts of the mission, that which is to be achieved during the mission, the various sequential and parallel activities taking place, the specific human operator tasks, and criteria for the performance of human operator functions and the accomplishment of the mission. For the purposes of measurement definition it is desirable to achieve an operational description of the mission and tasks, that is, a definition of the overt clearly identifiable operations taking place which are directly or indirectly affected by the human operator.

Analysis for comprehensive measurement begins with a complete decomposition of the mission into smaller parts for which activities and criteria are more easily defined. For example, the mission may be decomposed into separate maneuvers showing the normal and alternative sequences of maneuvers. The maneuvers may then be further divided into segments. Through this procedure the mission is divided into many parts and the total measurement problem is correspondingly divided into smaller problems.

One of the most difficult aspects of automated measurement, in practice, is to clearly identify these parts so that a computer can be programmed to recognize a segment or maneuver so that the appropriate measurement can be taken. One must be able to operationally define without equivocation when a segment starts, so that appropriate measurement calculations can begin, and when the segment ends, so that measurement stops. This is termed start/stop logic in this report (e.g., *if* specified conditions are met, *then* start measuring, and, *if* other conditions are met, *then*, stop measuring).

Within each segment, measurement is conceivably possible at a minimum of two levels: (1) measurement of the total man-machine system for comparison to expected mission goals, and (2) measurement of human operator activity in relation to design expectations. It is also possible to increase the number of hierarchical levels for measurement by also measuring the performance of the various subsystems including the human operator.

At any hierarchical level of measurement, the measures may be defined in terms of the system state variables, that is, those parameters (e.g., altitude, airspeed, angle of attack) which are totally sufficient for the description of system behavior. In fact, with system equations defined in terms of the state variables, one should be capable of the prediction of future system states. The remaining primary task of measurement is the definition of calculations, or transformations, to produce *measures* (or metrics) in terms of system *parameters* during the intervals defined by *start/stop logic*.

An analysis for measurement will reflect all activities occurring during a mission which may affect mission success. Unless one can somehow remove portions of the mission from consideration, a set of measures will be produced which will attempt to reflect everything going on. In effect, we implement the policy, "If it moves, measure it." To be practical, we should attempt to be efficient, and certainly should remove all irrelevant measurement. Analyses conducted for measurement should strive to simplify and remove irrelevant measurement. This will probably be accomplished only to the extent that (1) the analyst fully understands the tasks of the human operator and their relationships to system performance, and (2) research has sufficiently examined similar cases and alternative forms of measurement. Since these conditions are seldom met, the analyst is likely to be conservative and create an excessively large set of measures.

A STRUCTURE FOR MEASURE DEFINITION

Analyses suggest that most maneuvers can be thought of as collections of different segments for measurement purposes. A segment is any portion of a maneuver in which the desired behavior of a trainee or resulting system performance is relatively constant or follows a lawful relationship from beginning to end. Just as a primary task may continue while two subtasks proceed sequentially, measurement segments may overlap. Also, segments may repeat within a maneuver. Measurement sets within similar segments of any maneuver should be equivalent, although the desired value of some parameters might change.

The beginning and end of a measurement segment should be defined as a logical consequence of Boolean and relational expressions. Several relational expressions may be required to remove ambiguity. For example, one might define helicopter lift-off when:

{(altitude exceeds its initial value by more than one foot)

.OR.

(altitude rate exceeds 50-feet per minute)

.AND.

(collective control is greater than 20-degrees)

.AND.

(torque is greater than 50-percent)}

Specific Start/Stop functions and logical operators for combining these functions, used in the current study, are listed in table 1 and 2, respectively.

TABLE 1. GLOSSARY OF START/STOP FUNCTIONS[1]

MNEMONIC	FUNCTION	START/STOP WHEN:
B		Beginning of Record
E		End of Record
P		End, Best Fit Power of 2
G	PAR>DSR	Parameter Greater than Desired Value
L	PAR<DSR	Parameter Less than Desired Value
O	|PAR-DSR|>TOL	Absolute value of parameter minus desired value is greater than (outside of) tolerance
I	|PAR-DSR|<TOL	Absolute value of parameter minus desired value is less than (inside) tolerance
CO	|PAR-INIT|>TOL	Absolute value of parameter minus its initial value is greater than tolerance (or the change from initial is outside of tolerance)
CI	|PAR-INIT|<TOL	Absolute value of parameter minus its initial value is less than the tolerance

[1] These functional expressions were sufficient for the current development; they could be expanded as necessary.

TABLE 2. GLOSSARY OF LOGICAL OPERATORS FOR COMBINING START/STOP FUNCTIONS[1]

MNEMONIC	EACH PAIR OF FUNCTIONS (F) IS EVALUATED TRUE IF:
A	F_1 is True and F_2 is True
O	F_1 is True or F_2 is True
N	F_1 is True and F_2 is False
R	F_1 is False and F_2 is False

[1] These logical operations were sufficient for the current development; obviously, they could be expanded as necessary.

Thus, four observations for defining maneuver segmentation evolve from measurement analyses. First, maneuvers can be partitioned into any number of segments in which the determinants of performance can be mathematically defined and for which the conditions for starting and stopping measurement can be unambiguously defined. Secondly, within any maneuver an identical segment may repeat. Thirdly, different maneuvers may contain identical segments. Fourthly, segments for measurement purposes may overlap.

Having defined the conditions for measurement, a measure set can be constructed to represent all the trainee performance information which is desired for that segment. The set can contain an unlimited number of performance measures, each specified in terms of a *parameter*, a *sampling rate*, a *desired value* (if appropriate), a *transformation* and a *tolerance* if the transform requires one. A parameter is defined as a measure of (a) vehicle states in any internal or external reference plane such as pitch or roll attitude, (b) personnel physiological or positional states such as heart rate or eye movement, (c) control device states such as stick position, or (d) discrete events such as switch positions. The sampling rate is the frequency at which the parameter is sampled. Sometimes the value of the parameter means nothing unless it is compared to a desired value to derive an error score. Finally, a transformation is the mathematical treatment of the parameter such as a scalar value, a mean, a variance, a Fourier transform, etc. Common transforms used in manned vehicle research are shown in table 3. Specific transformations used in this study are presented in table 4.

An adequate description of multidimensional human operator performance will require many measures. Each measure of the set must be defined in terms of all of the following determinants:

a. Maneuver
b. Segment (when measurement starts and stops)
c. Parameter
d. Sampling Rate
e. Desired Value (if required)
f. Tolerance Value (if required)
g. Transformation.

TABLE 3. COMMON MEASURE TRANSFORMATIONS

TIME HISTORY MEASURES

- Time on Target
- Time Out of Tolerance
- Maximum Value Out of Tolerance
- Response Time, Rise Time, Overshoot
- Frequency Domain Approximations
  - Count of Tolerance Band Crossings
  - Zero or Average Value Crossings
  - Derivative Sign Reversals
  - Damping Ratio

AMPLITUDE-DISTRIBUTION MEASURES

- Mean, Median, Mode
- Standard Deviation, Variance, Quartile Range
- Minimum/Maximum Value
- Root-Mean-Squared Error, Mean-Squared Error
- Absolute Average Error

FREQUENCY DOMAIN MEASURES

- Autocorrelation Function
- Power Spectral Density Function
  - Bandwidth
  - Peak Power
  - Low/High Frequency Power
- Bode Plots, Fourier Coefficients
  - Amplitude Ratio
  - Phase Shift
- Transfer Function Model Parameters
  - Quasi-Linear Describing Function
  - Cross-Over Model

TABLE 4. GLOSSARY OF TRANSFORMATIONS

MNEMONIC	TRANSFORMATION
INIT	Initial Scalar Value
FINL	Final Scalar Value
AINI	Absolute Initial Scalar Value
AFIN	Absolute Final Scalar Value
MIN	Minimum Value
MAX	Maximum Value
AVG	Average Value $\frac{1}{N} \sum_{n}^{1} x$
AAE	Average Absolute Value $\frac{1}{N} \sum_{1}^{n} \lvert x \rvert$
ERS	Error Squared Value $\frac{1}{N} \sum_{1}^{n} x^2$
VAR	Variance $\sum_{1}^{n} x^2 - \frac{1}{N} (\sum_{1}^{n} x)^2$
RMS	Root-Mean-Square $\frac{1}{N} (\sum_{1}^{n} x^2)^{\frac{1}{2}}$
SDV	Standard Deviation $\frac{1}{N-1} (\sum_{1}^{n} x^2 - \frac{1}{N} (\sum_{1}^{n} x)^2)^{\frac{1}{2}}$
TOT	Time Out of Tolerance in Seconds and Tenths
RNG	Range, Distance Between the Largest and Smallest value
ELT	Elapsed Time in Seconds and Tenths
ZRX	No. Zero Crossings per Second
AVX	No. Average Crossings per Second
AUTO	Auto Covariance Function

TABLE 4. GLOSSARY OF TRANSFORMATIONS (Cont)

MNEMONIC	TRANSFORMATION
PERD	Periodicity of Auto Covariance Function, the tau shift values and covariance at peaks.
MLTR	Multiple Regression of a Parameter x and its derivative ($\dot{x}$) on Parameter y (Cooley and Lohnes, 1962). This particular transform computes successive multiple regressions of x, $\dot{x}$ on later (tau) values of y, (as in an auto covariance function) until maximum multiple regression coefficient is found. It returns (1) Tau in seconds, (2) the coefficient of multiple regression (3) the Beta weights and (4) B-weights at the point of maximum multiple regression.
HARM	Harmonic Analysis using procedures outlined Blackman and Tukey (1959), Cooley and Tukey (1965) and Villasenor (1968) produced the power spectral density function for the requested bandwidth.
FLTR	Relative power between 2 and 6 Radians-per-second using a pair of low-pass second-order digital filters as described by Norman (1973).

A representation of the assumed structure for measurement is shown in figure 1. As can be seen, it is hierarchical in nature. Objective performance for any trainee on any training day can be represented by a collection of measures for each maneuver. Each maneuver can contain any number of segments. Identical segments may repeat within a maneuver. Similar segments may appear in different maneuvers. Maneuver segmentation defines when measurement starts and stops. Within a segment any number of single or multiple parameter transformations may be employed. An unlimited number of transformations may be computed on any parameter.

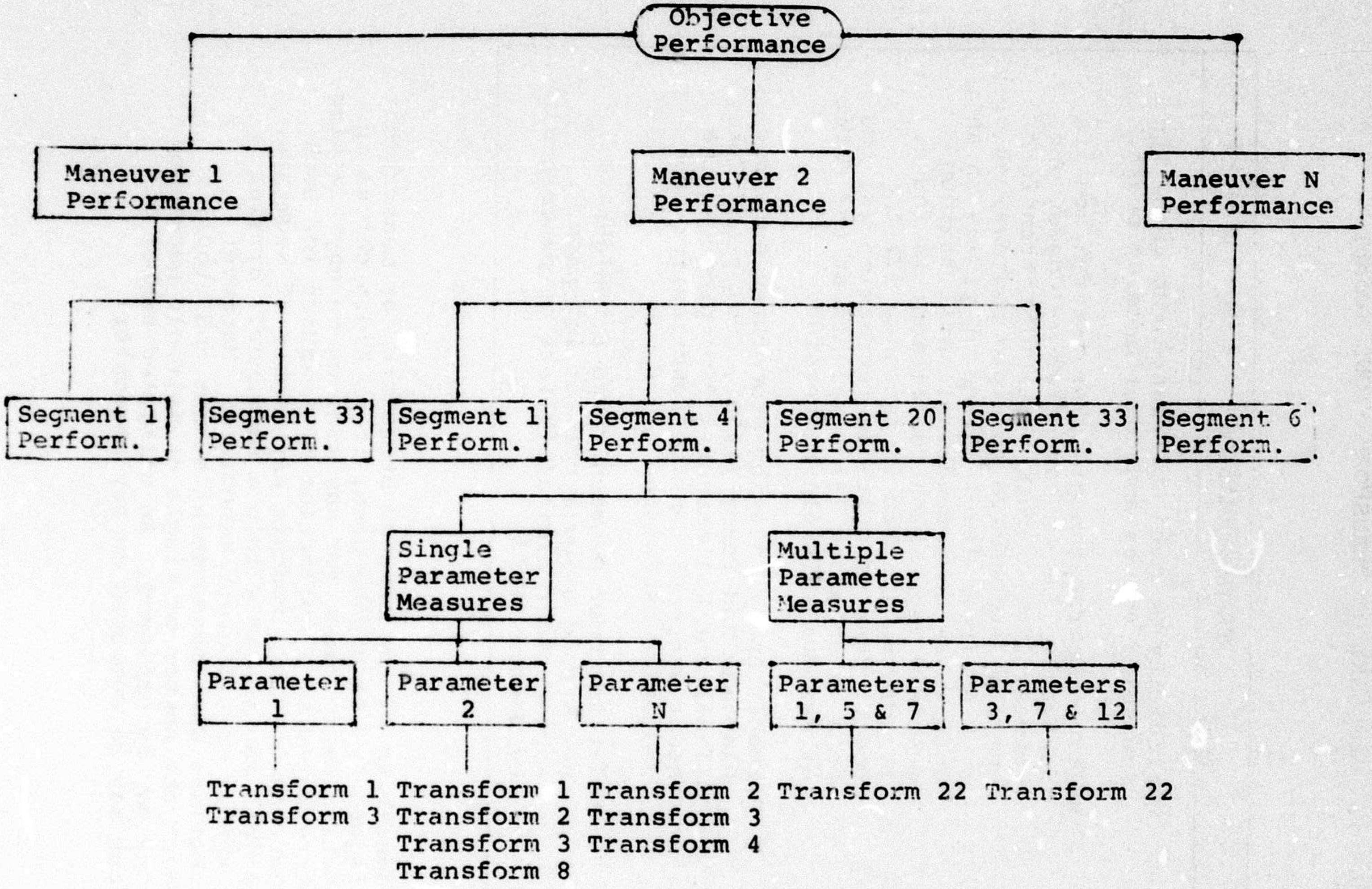


Figure 1. Representation of Measurement Model Components

THE CANDIDATE MEASURE SET

The measurement produced as a result of mission and task analysis, defined using the foregoing measurement structure, will be extensive for such applications as flight training. If the method is systematically applied, all human operator activity for which the analyst suspected a relation to mission performance will be measured. The selection of measurement also depends on the availability of knowledge of human performance and system models throughout the mission. The analyst may have some difficulty, therefore, in determining whether measurement of some segments is important, and, in determining which of several measurement alternatives are appropriate and best.

One procedure is to be very conservative: Measure if there is any reasonable doubt whether the measure can be excluded, and implement alternative forms of measurement if a clear-cut choice cannot be made. The result is a set of measures which is almost certainly redundant and too large. This set of measures is then used as an initial candidate set from which the final and more efficient measure set may be selected. Since presumably analytic resources have been exhausted, the candidate measure set can be reduced only through test with a modest number of human subjects performing tasks identical to, or related to, those involved in the mission. When additional data are empirically collected, further reduction of the candidate measure set should be possible.

A small set of measures is highly desirable from a number of points of view. Measurement while a student is performing is desirable for the purposes of computerized automated training since sufficient computing time is not available for large quantities of measures. Also, measurement may be intended for use with airborne instrumentation for which the capability for measurement is very restricted. Finally, large quantities and types of measurement make interpretation of results quite difficult whether the consumer of the information is a research scientist or an instructor.

MEASURE SELECTION CRITERIA. Reduction of the candidate measure set can be based on an analysis of data collected through a trial application of the measures, but, as a rather large number of measures is typical, and a number of subjects and trials will be required for an adequate statistical sample, more computer analysis is indicated. The criteria for selection must then be defined in quantitative operational form to permit machine selection. When the criteria are clearly stated, the type of computer programs required to mechanize the selection should also be apparent.

But, on what ground should a specific measure be excluded from further consideration? After consideration of the needs for measurement in training (Vreuls and Obermayer, 1971a, 1971b; Vreuls, et al., 1973), three general criteria have emerged:

(1) If two measures provide the same information for a given application, one member of the redundant pair may be discarded.

(2) Measures may be discarded if they are not sensitive to performance differences between individuals. The measurement retained should be able to discriminate between "good" and "bad" performers, students and instructors, and performance by a student early in training compared to that later in training.

(3) Measures should also be retained if they lend themselves to early prediction of performance to be achieved by an individual; for example, the performance level to be achieved at termination of training, or, the prediction of deficiencies which may be remedied by an appropriate change of training.

If two measures correlate highly, then conceivably one of the pair may be removed from the candidate measure set. In fact, it may be quite necessary to remove such measures for the proper functioning of multivariate statistical analyses used for testing other measurement selection criteria. However, the investigator must also ensure that small differences between two imperfectly correlated measures are not important; for example, one subject of a larger group may be sufficiently different that the measures are definitely uncorrelated for him. Further, of course, the problem of specifying the magnitude of the correlation coefficient for which measures will be considered redundant measures remains to the judgment of the investigator.

A multivariate statistical technique, the multiple discriminant analysis, is available to derive a discriminant function composed of a weighted sum of the available measures which will discriminate best between two or more groups. The weighting computed indicates the relative amounts each of the measures contributes to the discrimination. If the investigator can establish test groups which are known to be different in ways which are of interest to him, and test data are collected, the multiple discriminant analysis can be used to find those measures of the candidate measure set which facilitate discrimination. The weightings, then, are the key to the definition of selection criterion: the criterion can be that the measures with the least weights are discarded. Of course, the threshold level for measure weights is also left to the investigators judgment.

Another multivariate statistical technique, the canonical correlation analysis, can be used to test the prediction qualities of a measure set. Measures are found through this technique which, as a whole, correlate when measured at one time (e.g., early in training) as compared to the same measures taken at another time (e.g., late in training). Again, weights are associated with the measures which may be used to define criterion for selection.

The specific details of the criteria used, and the mechanics of the selection techniques, are perhaps best presented in terms of the computer operations needed. A description of the computerized selection techniques is available in the following chapter.

SECTION III

COMPUTERIZED MEASURE SELECTION TECHNIQUES

SELECTION BY DISCRIMINANT ANALYSIS

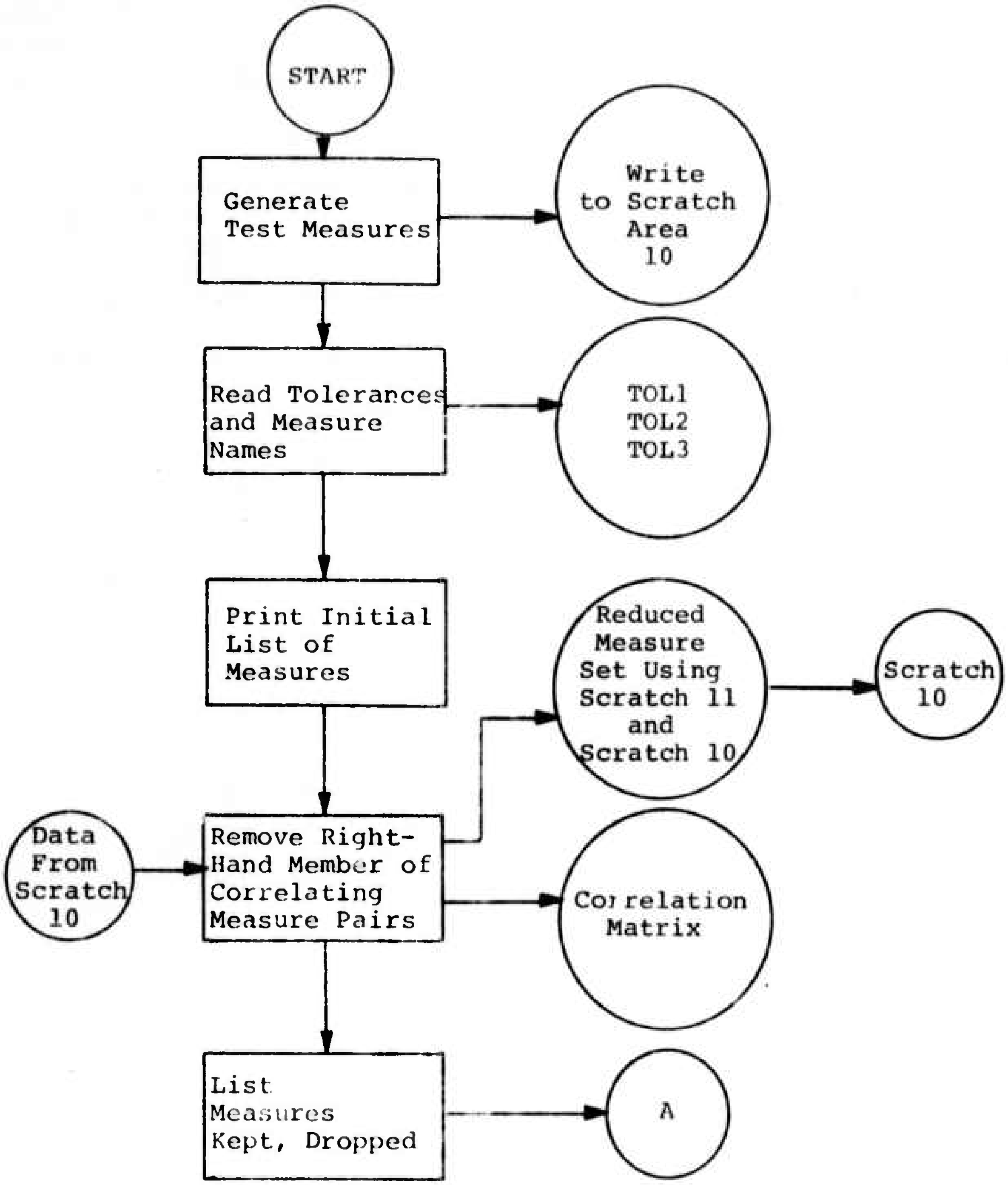
The programs generated to select measurement through discriminant analyses (Cooley and Lohnes, 1971) assume that a battery of measures have been taken for each of a number of groups of subjects. The primary purpose of these programs is to isolate the measures which best discriminate between the groups. For example, a pair of groups may consist of experienced and inexperienced subjects, respectively. The procedure adopted discards measurement which does not contribute to such discriminations.

The computer programs for selection based on discriminant analysis (DISCRM SELECT) iteratively discard measures until a minimum set of measures results. The iterative process stops when either one of two criteria is met: (1) the total number of remaining measures is less than the minimum number of factors determined through a principal components analysis (program PRINCO), or (2) discarding another measure will reduce discrimination to an unacceptable level.

The above procedure is satisfactory unless some of the measures are highly correlated, then the ability of the measure set to discriminate between groups cannot be clearly attributed to either of a pair of correlating measures. The procedure adopted, therefore, first performs a correlation analysis, and one of a pair of measures which correlate highly will be discarded. The right-hand measures of a pair of correlating measures in the correlation matrix is dropped in the programs developed.

FLOW DIAGRAM. The flow diagram for DISCRM SELECT is presented in figure 2. The output produced is listed in table 5. The version shown is a test program using a random number generator to produce data with known characteristics. Three tolerances must be specified: (TOL 1) the minimum percent of the original variance to be accounted for by any measure of the final reduced set of measures, (TOL 2) the minimum proportion of the variance of a specific measure extracted by all discriminant functions, and (TOL 3) the maximum correlation permitted between measures.

After test measures are generated, and tolerances and measure names are inputed, a correlation analysis is performed and the right-hand member of a pair of measures is discarded if the correlation coefficient exceeds TOL 3. A new list of measures is then printed indicating the measures which have been retained or discarded.


START
Generate Test Measures
Write to Scratch Area 10
Read Tolerances and Measure Names
TOL1
TOL2
TOL3
Print Initial List of Measures
Reduced Measure Set Using Scratch 11 and Scratch 10
Scratch 10
Data From Scratch 10
Remove Right-Hand Member of Correlating Measure Pairs
Correlation Matrix
List Measures Kept, Dropped
A

LEGEND: TOL1: Minimum Percent Variance to be Accounted for by any Measure of the Set

TOL2: Minimum Discrimination Communality

TOL3: Maximum Measure Intercorrelation Permitted

Figure 2. Flow Diagram of Discriminant Selection Process

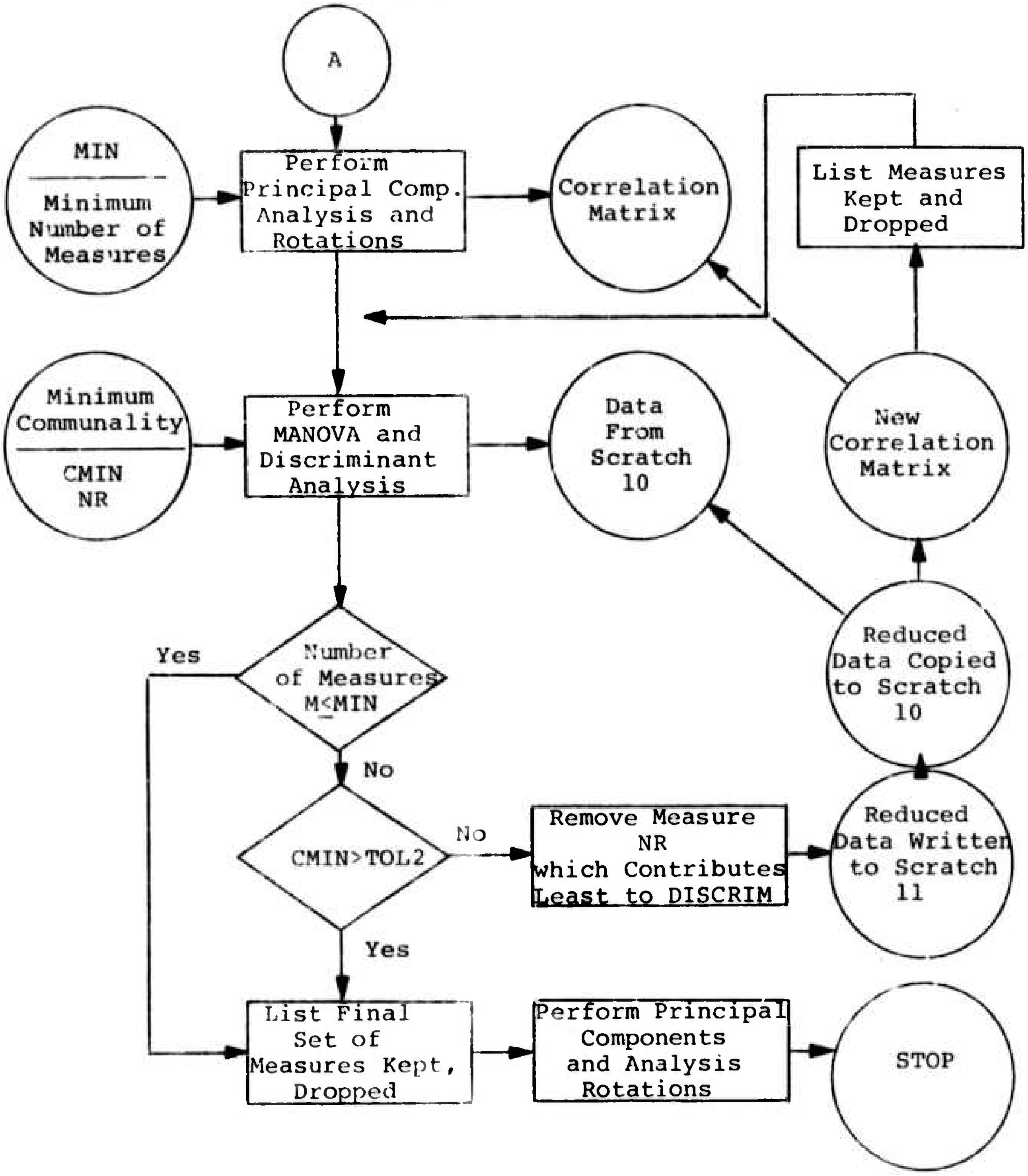

A
MIN
Minimum Number of Measures
Perform Principal Comp. Analysis and Rotations
Correlation Matrix
List Measures Kept and Dropped
Minimum Communality
CMIN NR
Perform MANOVA and Discriminant Analysis
Data From Scratch 10
New Correlation Matrix
Yes
Number of Measures M<MIN
Reduced Data Copied to Scratch 10
No
CMIN>TOL2
No
Remove Measure NR which Contributes Least to DISCRIM
Reduced Data Written to Scratch 11
Yes
List Final Set of Measures Kept, Dropped
Perform Principal Components and Analysis Rotations
STOP

Figure 2. Flow Diagram of Discriminant Selection Process (continued)

TABLE 5. DISCRM SELECT OUTPUT

1. Initial Output
   Criteria for Selection
   Correlation Matrix
   Measure Set Summary (after removing Correlating Measures)

2. Principal Components
   Correlation Matrix
   Sphericity Test
   Factors, % Trace, DF, CHI-SQUARE
   Factor Pattern
   Communality & Multiple R by Measure
   Factor Score Coefficients

3. Rotations
   VARIMAX
   QUARTIMAX

4. Multivariate Analysis of Variance
   Means & Standard Deviations by Group
   Test of Equality of Dispersions
   Univariate F-Ratios
   Multivariate Test - Wilks' LAMBDA & F-Ratios

5. Multiple Discriminant Analysis
   Multivariate Test - Wilks' LAMBDA & F-Ratios
   Chi-Square with Successive Roots Removed
   Row Coefficients Vectors
   Factor Pattern
   Communalities
   % Trace Accounted for by each Root
   Group Centroids

6. Measure Set Summary
   Measures Kept and Dropped

A principal components analysis is then performed on the reduced measure set. The full analysis is printed, along with VARIMAX and QUARTIMAX rotations of the factors. The variance of the factor analysis is compared to TOL 2 to determine the minimum number of factors required, and hence the minimum number of measures required (MIN).

Subsequently a Multivariate Analysis of Variance (MANOVA) and a Multiple Discriminant Analysis (DISCRM) is performed and all results are printed. The communality associated with each measure is computed and printed; this is the proportion of the variance associated with the specified measure which is extracted by all discriminant functions. The minimum communality (CMIN) is determined and the measures (NR) associated with CMIN is noted.

The computation will now stop with a final listing of measures kept and dropped if (1) the number of measures (M) is minimal (M<MIN), or (2) the minimum communality is greater than TOL 2, i.e., discarding another measure would significantly reduce the ability of the total measure set to discriminate between the experimental groups.

Otherwise, the computation iterates through the sequence again. However, the measure associated with CMIN is dropped, and a new correlation matrix for the reduced data base is computed.

SELECTION BY CANONICAL CORRELATION ANALYSIS

The programs called DISCRM SELECT were designed to aid in the selection of measures which are capable of discriminating between previously designated groups. Another series of programs, described in this section, were designed to select measures which relate performance exhibited at one time in training to that at another time. The basis of the method is a canonical correlation analysis (Cooley and Lohnes, 1971) which derives a linear combination of the measures and maximizes the correlation between the linear combination of one set of measures in relation to another set of measures. If the following linear combinations are formed:

$$y_1 = a_1x_1 + a_2x_2 + \ldots\ldots a_n\ x_n$$

$$y_2 = b_1z_1 + b_2z_2 + \ldots\ldots b_n\ z_n$$

Where x_i and z_i are the same measures collected at different points in the training sequence, then canonical correlation analysis determines the coefficients a_i and b_i so that y_1 and y_2 maximally correlate.

The quantities y_1 and y_2 are factors of their respective data groups. The computer programs generate the factor structure for each set of data which displays the correlation between each measure and factor. The factor which correlates between groups best is also indicated, and it is this factor which is used for measure selection. The measure which correlates least with this factor contributes least to inter-group correlation. It is this measure which is equated with the computer parameter RMIN in the CANON SELECT programs.

FLOW DIAGRAM. CANON SELECT iteratively reduces the measure set until the entire remaining measures contribute sufficiently to inter-group correlation. The flow diagram in figure 3 corresponds to a test version of the program which generates measures artificially; computer output categories are listed in table 6.

A canonical correlation analysis is performed (program CANON) and the measure with minimum weighting (RMIN) is selected. If this measure contributes less than a pre-specified amount to the canonical correlation (RMIN<TOL1) the measure is dropped from the data base and another canonical correlation is performed. These steps are performed iteratively, with a new list of measures printed at each step, until the minimum measure redundancy is equal to or greater than the pre-specified tolerance.

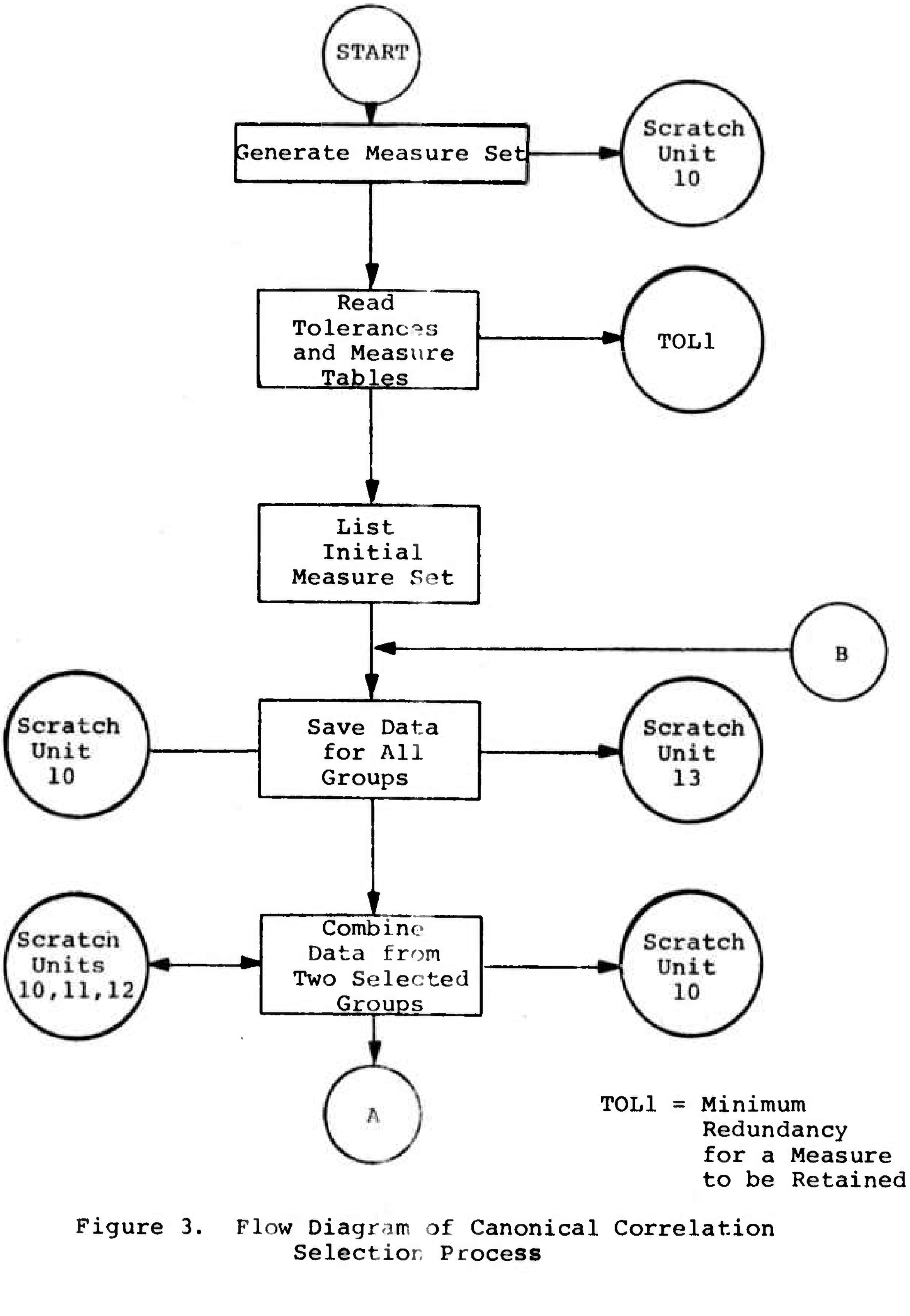


Figure 3. Flow Diagram of Canonical Correlation Selection Process

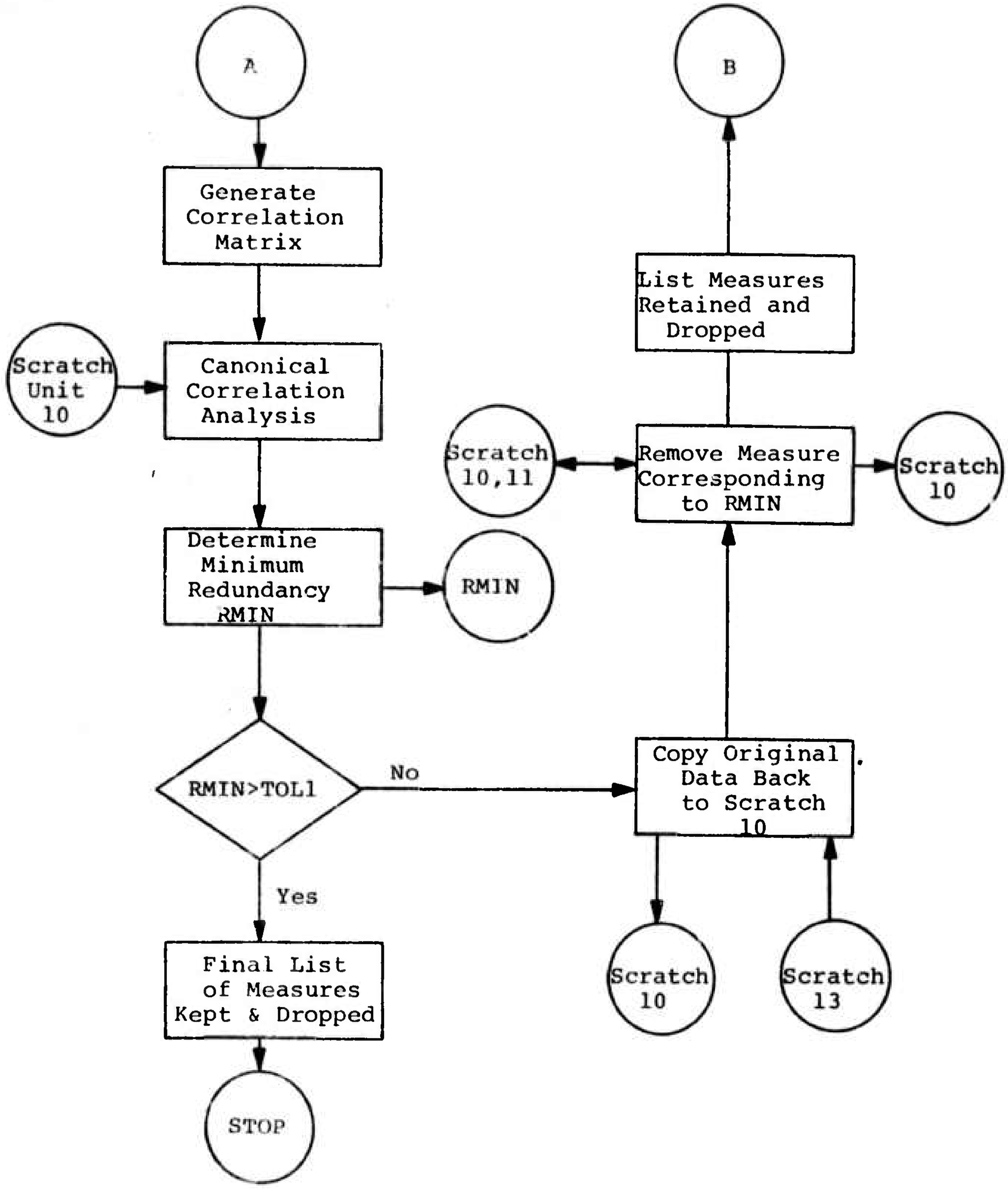

A
Generate Correlation Matrix
Scratch Unit 10
Canonical Correlation Analysis
Determine Minimum Redundancy RMIN
RMIN
RMIN>TOL1
No
Yes
Final List of Measures Kept & Dropped
STOP
Copy Original Data Back to Scratch 10
Scratch 10
Scratch 13
Remove Measure Corresponding to RMIN
Scratch 10,11
Scratch 10
List Measures Retained and Dropped
B

Figure 3. Final Diagram of Canonical Correlation Selection Process (continued)

TABLE 6. CANON SELECT OUTPUT

1. Measure Set Summary (Measures Kept & Dropped)
2. Correlation Matrix
3. Canonical Weights (Left and Right Set)
4. Factor Structure (Left and Right Set)
5. Variance Extracted, Redundancy (Left and Right Set)
6. Total Variance, Redundancy (Left and Right Set)
7. Total Set: Wilks' LAMBDA, CHI SQUARE, Degrees of Freedom
8. CHI SQUARE Tests with Successive Roots Removed

SECTION IV

DEVELOPMENT OF MEASURE SELECTION TECHNIQUES

Measure selection techniques were developed within a computer-controlled training environment. The environment was the automated instrument flight maneuvers (IFM) training system developed by Johnson (1972) on the Training Device Computer System (TRADEC) located at the Naval Training Equipment Center. IFM automatically sequenced the trainee through a series of maneuvers and simulated flight conditions as a function measured trainee performance on the previous and antecedent trials. The performance measures (and weighting coefficients for summing the various components of error into one composite score) were derived during IFM system design from task analytic data; the measures were never formally tested.

In order to produce data for empirical measure selection studies, the IFM system was modified to control a measure selection experiment and to produce raw data for subsequent (non-real-time) conversion into candidate measures and further measure selection analyses.

DATA COLLECTION METHOD

A data base for preliminary measure selection analyses was created by conducting a study in which trainees flew each of four principal maneuvers of IFM until their performance was assumed to be very good by virtue of having flown the simulator for 14-18 hours. Measure selection methods were developed using the preliminary data base so produced.

PARTICIPANTS. Four participants were used. They were low-time private pilots who were unskilled at instrument flight at the onset of data collection. All were light plane pilots; none were familiar with jet fighter dynamic response.

APPARATUS. The test equipment was the TRADEC, which was configured as a fixed-wing aircraft (F-4E). TRADEC hardware included an XDS Sigma-7 computer and associated peripherals, an aircraft cockpit mounted on top of a four-degree-of-freedom motion platform (pitch, roll, yaw and heave), and a host of related equipment. A digital computer program provided the basic flight simulation (cf., Kapsis, et al., 1969; Erickson, et al., 1969). The basic flight program was converted into a computer-controlled training device by the automated IFM program.

IFM was modified from an automated training configuration to an automated data collection configuration. The computer-controlled speech synthesizer (COGNITRONICS) was used to brief participants on the task requirements for each trial, and issue corrective commentary when various vehicle states were out of tolerance. The task scheduler was used to set the experimental conditions for the next trial as prescribed by the experimental design.

EXPERIMENTAL DESIGN. Each of the four participants (see table 7) were trained on four basic instrument flight maneuvers for 18, one-hour sessions over a period of seven weeks. The four maneuvers were (1) straight and level flight, (2) standard rate climbs and descents, (3) level turns, and (4) climbing and descending turns. Six trials of each maneuver were flown during each training session. Each successive odd and even numbered training session was pooled into one unit called a training "day"; thus, sessions 1 and 2 became Day 1, sessions 3 and 4 became Day 2, etc. This pooling resulted in 48 observations (4 participants by 6 trials by 2 sessions) for each maneuver for each day.

Two task stressors were used, turbulent air and aircraft weight and center of gravity. The turbulent air was generated in the flight program from a random number generator. When used, its intensity was set to a "light turbulence" level as defined by the TFM program. The aircraft weight was either light or heavy. The light aircraft carried 2,500 pounds of fuel, had a gross weight of 33,600 pounds and a center of gravity at 29.0 percent mean aerodynamic chord. The heavy aircraft carried 12,896 pounds of fuel, had a gross weight of 43,996 pounds and a center of gravity at 30.2 percent mean aerodynamic chord. The weight increases and aft center of gravity shift reduced the longitudinal axis short-period damping coefficient, which decreased the simulator pitch axis stability, making it more difficult to control. Task stressors were not changed during a trial.

Each participant received exactly the same order of experimental trials each day. Thus, maneuver one always was flown first and maneuver four always was flown last. This fixed order permitted the study of measures for each maneuver under identical antecedent conditions (and subsequent order effects) across training days.

Performance data from Days 1, 3, 5 and 7 were primary units for measure selection analyses. It was assumed that after 14, one-hour training sessions (the conclusion of Day 7), the participants would be relatively proficient on the basic maneuvers. Data were collected during Day 2, 4, 6, 8 and 9 for further examination of the effects of the task stressors on the measure set in a later study (beyond the scope of the current effort).

MEASUREMENT. Eighteen (18) pilot/system performance parameters shown in table 8 were collected on magnetic tape at a rate of five times-per-second from the beginning to the end of training. Only the raw data from the straight and level maneuver trials were transformed into candidate measure sets for the purpose of

TABLE 7. EXPERIMENTAL DESIGN

		G1T1				G2T1			G1T2	G2T2
		DAY1	DAY3	DAY5	DAY7	DAY2	DAY4	DAY6	DAY8	DAY9
M1	P1	*	*	*	*	*	*	*	*	*
	P2	*	*	*						
	P3	*	*							
	P4	*								
M2	P1	*								
	P2	*								
	P3	*								
	P4	*								
M3	P1	*								
	P2	*								
	P3	*								
	P4	*								
M4	P1	*								
	P2	*								
	P3	*								
	P4	*								

Legend:	M=Maneuvers:	M1 = Straight and Level
		M2 = Standard Rate Climbs and Descents
		M3 = Level Turns
		M4 = Climbing and Descending Turns
	P=Participants	
	G=Center of Gravity:	G1 = Light Aircraft, Fore cg.
		G2 = Heavy Aircraft, Aft cg.
	T=Turbulence:	T1 = Smooth Air
		T2 = Light Turbulence

DAY=Two successive one-hour training sessions.

* Twelve trials were administered on each maneuver, each day.

TABLE 8. RAW DATA PARAMETERS

PARAMETER	UNITS	ABBREVIATION
1. SYSTEM CLOCK COUNT		CLOK
2. ELEVATOR STICK FORCE	POUNDS	ELVF
3. ELEVATOR STICK DISPLACEMENT	INCHES	ELVS
4. ANGLE OF ATTACK	UNITS	ALPH
5. PITCH ATTITUDE	DEGREES	PTCH
6. CLIMB/DESCENT RATE	FEET PER MINUTE	HDOT
7. ALTITUDE	FEET	ALT
8. RIGHT THROTTLE DISPLACEMENT	DEGREES	THRR
9. AIRSPEED	KNOTS	A/S
10. AILERON STICK FORCE	POUNDS	AILF
11. AILERON STICK DISPLACEMENT	INCHES	AILS
12. ROLL ATTITUDE	DEGREES	ROLL
13. TURN RATE	DEGREES PER SECOND	TURN
14. HEADING	DEGREES	HEAD
15. RUDDER PEDAL FORCE	POUNDS	RUDF
16. RUDDER PEDAL DISPLACEMENT	INCHES	PED
17. SIDESLIP	DEGREES	BETA
18. TURBULENT AIR INTENSITY	ARBITRARY UNITS	RUFF

preliminary measurement selection method development. The transforms available in the measure producing programs are shown in table 4 in Section II. Specific maneuver one-candidate measures are shown in table 9.

RESULTS

Measure selection analyses were conducted three ways by using (1) t-tests, (2) multiple discriminant analyses, and (3) canonical correlation analyses. The purpose of the analyses was to develop and test selection methods. Only a small sample of the data are presented.

t-TESTS. The t-tests considered each measure independent of all other measures. No consideration of measure correlation was given. As a result, 17-measures were found to be significantly different between Day 1 and Day 7, as shown in table 10.

DISCRIMINANT SELECTION. The 24-candidate measures were reduced to seven measures which could significantly discriminate between Day 1 and Day 7 performance as shown in table 11. The greatest reduction in the candidate measure set occurred during the initial correlation analysis. All measures were intercorrelated. The right-hand measure of a pair was eliminated if the correlation exceeded .69. This criteria reduced the candidate set from 24-measures to 9-measures.

The discriminant selection procedure further reduced the set from nine to seven measures shown in table 11, based on two criteria, (1) any measure of the final set must account for more than seven percent of the total variance, and (2) the minimum measure communality is .200. Communality can be thought of as the proportion of variance associated with each measure which is extracted by all discriminant functions. The discriminant vectors shown in table 11 reflected the weighting coefficients for the summation of measures into a discriminant function.

The composition of the discriminating set was of interest. Three measures represented outer-loop vehicle states--heading, altitude and airspeed. Four measures represented control input states--elevator stick range and crossover power, aileron stick crossover power, and rudder pedal range. Thus, over half of the measures which discriminated between early and late performance were control input measures.

It was of interest also to examine the change in the descriptive capability of the resulting measure set. The factor loadings from the principal components analysis are shown in table 12. It was apparent that the loadings on factor I were higher on Day 7 than on Day 1, and that the amount of variance accounted for by that factor was 21 percent higher on Day 7.

TABLE 9. CANDIDATE MEASURES FOR MANEUVER ONE MEASURE SELECTION METHOD DEVELOPMENT

MEAS. NO.	PARA-METER	DESIRED VALUE	TRANS-FORM	ABBREVIA-TION IN ANALYSIS	GLOSSARY	
1	ELVS	0	RNG	ELRG	ELEVATOR STICK	RANGE
2			FLTR	ELF1		CROSSOVER POWER
3			AAE	ELF2		AVERAGE DISPLACEMENT
4	AILS	0	RNG	AIRG	AILERON STICK	RANGE
5			FLTR	AIF1		CROSSOVER POWER
6			AAE	AIF2		AVERAGE DISPLACEMENT
7	PED	0	RNG	PDRG	RUDDER PEDAL	RANGE
8			FLTR	PDF1		CROSSOVER POWER
9			AAE	PDF2		AVERAGE DISPLACEMENT
10	ALPH	0	RNG	ALRG	ANGLE OF ATTACK	RANGE
11			SDEV	ALSD		STANDARD DEVIATION
12	PTCH	0	RMS	PTRM	PITCH ATTITUDE	ROOT-MEAN-SQUARED ERROR
13			SDEV	PTSD		STANDARD DEVIATION
14			RNG	PTRG		RANGE
15	ROLL	0	RMS	RORM	ROLL ATTITUDE	ROOT-MEAN-SQUARED ERROR
16			RNG	RORG		RANGE
17	HEAD	360	RMS	PSRM	HEADING	ROOT-MEAN-SQUARED ERROR
18			RNG	PSRG		RANGE
19	ALT	25000	AAE	HAA	ALTITUDE	AVERAGE ABSOLUTE ERROR
20			RNG	HRG		RANGE
21	HDOT	0	AAE	HDAA	ALTITUDE RATE	AVERAGE ABSOLUTE ERROR
22			RNG	HDRG		RANGE
23	A/S	350/280*	AAE	ASAA	AIRSPEED	AVERAGE ABSOLUTE ERROR
24			RNG	ASRG		RANGE

*One-half of the trials were at 350-knots IAS, the other half at 280-knots

TABLE 10. AVERAGE MANEUVER ONE MEASURES

MEASURE	DAY1	DAY3	DAY5	DAY7
ELRG	1.707*	1.184	1.059	1.086
ELF1	.033*	.019	.018	.017
ELF2	.687	.713	.700	.708
AIRG	1.064*	.787	.692	.641
AIF1	.056*	.025	.021	.021
AIF2	.253	.274	.256	.269
PDRG	.186*	.119*	.066	.079
PDF1	.014	.011	.010*	.015
PDF2	.149	.161	.150	.151
ALRG	2.378*	1.664	1.491	1.607
ALSD	.446*	.307	.269	.274
PTRM	2.611	2.606	2.591	2.591
PTSD	.837*	.495	.436	.414
PTRG	3.766*	2.383	2.101	2.142
RORM	2.616	2.324	2.616	2.331
RORG	12.198	9.305	10.066	8.705
PSRM	2.982*	2.219*	2.058*	1.733
PSRG	3.868*	2.934	2.893	2.756
HAA	.052*	.028*	.021	.017
HRG	.173*	.093*	.074	.066
HAAA	.456*	.242	.217	.190
HDRG	2.354*	1.338	1.151	1.125
ASAA	8.290*	4.114*	3.341	3.015
ASRG	16.916*	10.818*	8.444	8.492

*Measure is significantly different than Day 7, P<.05 based on t-test; 48 observations per number.

In general, the performance dimensions expressed by the factor structures appeared to be more integrated on Day 7 than on Day 1. Four factors accounted for 88 percent of the variance on Day 7, whereas, five factors accounted for only 86 percent of the variance on Day 1. Also note the integration of Pedal Range into the first and second factors by Day 7.

TABLE 11. MEASURES SELECTED BY DISCRIMINANT ANALYSIS*

MEASURE	P<**	DISCRM VECTOR	COMMUN-ALITY	MEANS DAY1	DAY7
ELRG	.01	0.142	.4088	1.71	1.09
ELF1	.01	-1.758	.2409	.03	.02
AIF1	.01	6.548	.3475	.06	.02
PDRG	.01	2.358	.3506	.19	.08
PSRM	.01	-0.005	.2496	2.98	1.73
HAA	.01	9.067	.4573	.05	.02
ASRG	.01	0.052	.5895	16.92	8.49

*The overall discrimination is significant, P<.01 for an F-ratio approximation of 9.18 with 7/88 df.

**The probability that the differences between the means were due to chance based on univariate F-ratios.

TABLE 12. FACTOR LOADINGS FOR FINAL MEASURES

DAY	MEASURE	FACTORS I	II	III	IV	V	VI	VII
DAY 1	ELRG	.63*				-.62		
	ELF1	.54	-.56	.35			-.39	
	AIF1	.44	-.61	-.32	.37		.31	
	PDRG			.90				
	PSRM	.72	.40					.40
	HAA	.63	.55					-.37
	ASRG	.54	-.42		-.63			
	%Variance**	.30	.20	.16	.11	.09	.07	.07
DAY 7	ELRG	.81			-.42			
	ELF1	.61	.58		.36			
	AIF1	.62		.65				
	PDRG	.84	.35					-.32
	PSRM	.65		-.64				
	HAA	.75	-.47			-.34		
	ASRG	.65	-.59		.32			
	%Variance	.51	.16	.14	.07	.06	.03	.03

*Factor loadings less than .30 are generally considered insignificant, and were omitted from the table.

**Percent variance accounted for by each factor.

Further rotation of the factor loading (table 13) suggested that each measure of the final set essentially represented an independent performance dimension on Day 1. As training progressed to Day 7, measure tended to double-up on three factors; this was interpreted to indicate an increase of control integration and coordination.

CANONICAL CORRELATION SELECTION. The output from CANON SELECT computer programs satisfied the initial requirements for selecting predictive measures. For the sample data shown in Tables 14, 15 and 16, the 24-candidate measures were reduced to a predictive set of 18-measures. The criterion for measure rejection was a correlation of less than .25 with the first canonical factor. This criterion was set low, deliberately, for the initial tests.

To recapitulate, the convention was used in the procedure to designate the predictor measurement as the "left" side (of the intercorrelation matrix) and the criterion as the "right" side, although the canonical correlation model was completely symmetrical. Since we were interested in the possibilities of using the measurement for prediction, our attention was focused on the left side of measures. The following output was produced:

a. Canonical factors -- the coefficients were produced which defined factors for each measure set, so that the factors of the two sets have the highest correlation.

b. Factor structure -- the correlation of each measure with each canonical factor.

c. The proportion of shared variance (R_c^2) between the corresponding canonical factors.

d. Redundancy -- the product of the proportion of shared variance and the proportion of the variance extracted by a canonical factor (i.e., the proportion of the variance of one set accounted for, or "explained", by a specific canonical factor of the other set.

e. Bartlett's test for significance of canonical correlation.

A sample case was extracted from the data to illustrate the method. The data shown in tables 14, 15 and 16 were derived from a test of the ability of pooled Day 1 and Day 3 data to predict pooled Day 5 and Day 7 data. Since the data were pooled, each measure had 96 observations; the left set represented Day 1 and Day 3 and the right set represented Day 5 and Day 7.

TABLE 13. ROTATED FACTOR LOADINGS FOR FINAL MEASURES*

		FACTORS						
DAY	MEASURE	I	II	III	IV	V	VI	VII
DAY 1	ELRG					-.97		
	ELF1		.96					
	AIF1						.97	
	PDRG			.99				
	PSRM							.94
	HAA	.95						
	ASRG				.98			
DAY 7	ELRG					-.86		
	ELF1	.94						
	AIF1			.94				
	PDRG	.35		.38				-.76
	PSRM				.92			
	HAA		.35				-.86	
	ASRG		.92					

*Factor loadings less than .30 are generally considered insignificant, and were omitted from the table.

The canonical factors were ordered (see table 14) on the canonical correlation coefficient (R_c); the first canonical factor of the left set, and the first canonical factor of the right set had the highest correlation. The redundancy for each factor is listed, indicating for each factor of the left set, the proportion of variance of the right set accounted for.

In the sample case shown in table 14, the first canonical factor of the left set extracted 14.7 percent of the variance of that set and explained 11.2 percent of the variance of the right set. The first canonical factor of the right set accounted for only four percent of the variance of the right set. It can be seen that although the first canonical factor had the highest canonical correlation, it accounted for only a small portion of the total variance. The contributions of the remaining factors were evident.

The output data for the test of significance are shown in table 15. The roots are related to factors; removal of a root is equivalent to dropping a factor from the canonical correlation. Table 15 reveals that nearly all of the factors were needed to adequately account for the shared variance between the left and right set data in this particular case.

TABLE 14. SAMPLE CANONICAL CORRELATION OUTPUT

FACTOR	LEFT SET VARIANCE EXTRACTED	LEFT SET REDUNDANCY	RIGHT SET VARIANCE EXTRACTED
1	.147	.112	.040
2	.064	.045	.051
3	.049	.031	.082
4	.064	.038	.052
5	.033	.018	.030
6	.033	.017	.044
7	.024	.011	.067
8	.019	.007	.020
9	.063	.018	.074
10	.022	.006	.047
11	.014	.003	.018
12	.024	.005	.020
13	.010	.002	.045
14	.038	.007	.017
15	.034	.005	.052
16	.038	.004	.022
17	.015	.001	.047
18	.054	.003	.038
19	.047	.002	.039
20	.026	.001	.032
21	.078	.002	.062
22	.020	.000	.018
23	.019	.000	.061
24	.062	.000	.024
Total:	.997	.338	1.000

A more detailed examination of the example case factor structures is shown in table 16, which presents the correlation of each measure with each factor. Only three factors are shown. The measures which contributed least to a specific predictive factors where, therefore, identified.

As an initial test, only the factor associated with maximum correlation was considered. Measures which correlated least were successively removed from the measure set until all remaining measures met a priori criteria (exceeding correlation of .25). However, it was apparent from the resulting data that a number of factors contribute to the canonical correlation. Therefore, the simple criterion based on the first canonical factor was insufficient. Alternative criteria are presented in the following discussion sections.

TABLE 15. EXAMPLE CHI SQUARE TESTS WITH SUCCESSIVE ROOTS REMOVED

ROOTS REMOVED	CANONICAL R	R^2	CHI^2	DF	LAMBDA PRIME
0	.87	.76	650	576	.0001
1	.84	.70	550	529	.0004
2	.80	.64	465	484	.0014
3	.77	.60	393	441	.0038
4	.75	.56	329	400	.0094
5	.72	.52	272	361	.0212
6	.68	.47	221	324	.0438
7	.61	.38	176	289	.0818
8	.54	.29	143	256	.1315
9	.51	.26	119	225	.1854
10	.47	.22	98	196	.2500
11	.45	.20	80	169	.3217
12	.43	.18	64	144	.4023
13	.42	.18	50	121	.4922
14	.38	.15	36	100	.5991
15	.33	.15	25	81	.7018
16	.27	.07	17	64	.7893
17	.25	.06	11	49	.8520
18	.19	.04	7	36	.9080
19	.16	.03	4	25	.9431
20	.14	.02	2	16	.9692
21	.10	.01	-	9	.9893
22	.02	.00	-	4	.9993
23	.01	.00	-	1	.9998

TABLE 16. PARTIAL CANON SAMPLE FACTOR STRUCTURES

	LEFT SET FACTORS			RIGHT SET FACTORS		
MEASURE	I	II	III	I	II	III
1	-.50	*				.42
2		-.49				.34
3		.50			.59	
4	-.47					.37
5			-.34			.36
6						
7				-.38		
8		-.30				
9	-.31			.31		
10	-.43		-.37			.34
11	-.47		-.35			
12		.54			.58	
13	-.57					.31
14	-.55					.34
15						
16	-.36					.31
17	-.33			-.36	-.36	
18	-.33					
19	-.63	-.30		-.43		.41
20	-.55					.31
21	-.56					.31
22	-.54					.35
23						
24		-.30	-.35			
Variance Extracted	.15	.06	.05	.08	.05	.08

*Factor loadings less than .30 are generally considered insignificant, and were omitted.

SECTION V

DISCUSSION

CANONICAL CORRELATION SELECTION

The criteria used for the current test were based on the degree of correlation between each measure and the first canonical factor; when all measures correlated at a specified level, no further reduction of the measure set was attempted. However, it was apparent from the data collected that a number of canonical factors significantly contributed to canonical correlation (or prediction). Thus, the criteria for measure selection must be expanded.

MULTIPLE SIGNIFICANT CANONICAL FACTORS. When a number of canonical factors are significant, the first measure to be removed from the measure set should be the one which correlates the least with the group of significant factors. But, the measure which correlates least with factor I may correlate best with factor II (as shown in table 16). The correlation across a group of significant factors must be assessed in some manner. The following steps are suggested as a partial solution to this problem:

a. Determine the significant factors. This can be done using the statistical test presented in the output along with a rule of thumb for discarding trivial factors. Cooley and Lohnes (1971, pg. 176) state, "As a rule, the authors frequently treat canonical correlations of .30 or less as trivial."

b. Multiply the columns of the factor structure by the redundancy of the respective factor to weight measure correlations with the proportion of variance accounted for in the criterion measures.

c. Using the weights computed in (b) above, find the greatest weight for each measure.

d. The measure which is a candidate for removal is the measure corresponding to the least of the numbers computed in (c) above.

PREDICTIVE AND CRITERION SET COMPOSITION. It should be noted in the preceding that the right side, or criterion, measures were not considered during measure selection. In the current application, however, corresponding right and left measures were the same. If a measure is to be removed, one should consider whether or not it is to be removed from just one side or both. There are several possibilities that have yet to be explored.

Prediction of the Full Original Set of Measures. If we assume that the original and complete set of measures is better than any subset for describing performance, and our goal is to predict total performance, then measures should be removed only from the left set. Removal of a given measure from both sides simultaneously may take away a measure which contributes least on the left side; however, it is quite possible that the removed measure might be important to the right side, or criterion side. In application it might be feasible to have an expanded criterion set for measure development, while the operational measure set might be reduced for practical reasons.

Prediction of the Reduced Set of Measures. Practical considerations might dictate that the same set of measures is to be used for predicting as well as for measuring that which is to be predicted. The algorithm for developing the reduced set must search iteratively for that measure which contributes little to both sides of the canonical correlation model and simultaneously remove the measure from both sides. The steps similar to those suggested (in a-d) above, applied for each measure across both right and left side factors, represent a feasible method. However, it must be noted that the composition of the predictor and criterion sets might be somewhat different; thus, the utilization of this method might create a larger predictor (left side) set than would result with consideration of only the left set alone.

Prediction of Specific Performance. If only specific performance characteristics are to be predicted, then the factors which relate to this performance must be located. Specific measures which are of major importance to the desired performance may be used to identify the pertinent factors, then the measures which load least on these factors may be discarded.

Multiple Predictive and Criterion Sets. The discussion of predictive and criterion set composition is concluded (but not exhausted) by noting that it is possible that multiple sets might be required in order to predict specific terminal behaviors. It would be unwieldly, and perhaps unwise, to expect the development of just one, all-encompassing predictive and/or criterion set. Since skill shifts during training, we can anticipate that specific set composition will be a function of the time and place during training that the prediction is to be made, as well as the specific behavior that is to be predicted.

DISCRIMINANT SELECTION

The discriminant analysis procedures appeared to work well to strip-down the candidate measures to a very small subset which could discriminate between early and late performance. Perhaps too much so. Initial measure rejection on the basis of measure intercorrelations appeared to be quite drastic. In a few analyses the wrong measure of a correlated pair was dropped relative to outside criteria, such as ease of implementation. In a few cases there was apparent conflict between the criterion of discrimination and the criterion of adequate performance description. These issues are discussed briefly in the following:

CORRELATION CRITERION. The candidate measure set intercorrelations were higher than expected. The Day 1 vs. Day 7 data shown previously (Table 11) were typical. In those data, by dropping the right-hand member of a pair which correlated better than r=.69, a substantial reduction in measures (from 24 to 13) was seen on basis of Day 1 data alone. Further measure reduction (from 13 to 9) occurred when the remaining Day 7 measures were correlated. The final set reduction (from 9 to 7) occurred during the discriminant analysis.

It was possible, as a result, that the criterion for selection (the ability of the set to discriminate) did not influence the final measure set as much as the investigators would have liked. The resulting "very clean" rotations of the factor structures suggested the possibility of performance dimension oversimplification. As a consequence, it is suggested that further work with the DISCRIM SELECT procedure examine a slightly larger correlation tolerance in the range of r=.74 to r=.82.

MEASURE PRIORITIES FOR SELECTION. The arbitrary decision rule was established that the right-hand member of a correlated pair was to be dropped. At first it was thought that the data could be arranged generally from left to right to reflect external criteria, such as those measures which are easier, faster or less expensive to implement. However, this simple, linear scheme did not always produce the desired result.

A priority of measures scheme should be added to the DISCRIM SELECT procedure. It should cause the rejection of a lower priority of any pair of measures. Also, it might be necessary for reasons other than discrimination to retain a particular measure at all costs. The priority scheme might require addressing such complexities as the following: If A, B, AND C are dropped, keep D.

DESCRIPTION VS. DISCRIMINATION. The reduction of measures into a set which significantly discriminates might result in a final set which has weakened power to describe all important dimensions of performance. For example, holding roll attitude might be a very important part of straight-and-level flight performance; however, if there is no substantial change in the variance due to roll attitude holding during training, the measure might fail to emerge from a discriminant analysis. This problem was attended to in the early design of the procedure.

The DISCRIM SELECT procedure was organized so that the ability to describe performance should have been retained by the final measure set. Following ejection of highly-correlated measures, a principal components analysis produced a list of factors, ordered according to the amount of variance each contributed. On the basis of investigator specified criteria (the minimum percent variance to be accounted for by any factor in the final, reduced set--in this case it was 7 percent), the minimum measure set size was defined. After initial tolerance testing, the procedure worked well, most of the time.

Statistically, we could expect the results to go awry once in awhile. They did. In at least one case it was judged that the discriminant analysis stopped too soon because it hit the minimum measure set size; a low communality in the bottom measure which would have been dropped in the next iteration, suggested that an additional iteration would have produced a significant discrimination. In a second case the procedure went too far; although the second-to-the last iteration produced a significant discrimination, the last iteration resulted in an insignificant overall discrimination.

The first case above (stopping too soon) has to be accepted if we continue to insist that the discriminating set should have sufficient description power. However, further testing of the percent variance tolerance appears necessary. Seven percent might have been too high; preliminary testing suggested that five percent was too low. Trials in the range of 5.5 percent to 6.5 percent appear warranted.

The second case above (going too far) can be corrected by adding the capability to test the statistical significance of the overall discrimination in the program. A subroutine to compute the exact probability of the F-ratio should be added. The logic should be changed to test for a significant F. Once achieved, the program should continue to iterate normally unless F becomes insignificant. If that should happen, the previous iteration should be the result. Note that F must first become significant before an insignificant F can cause a stop.

STATISTICAL ASSUMPTIONS. Multivariate techniques were explored as a part of a method to reduce the number of measures which could be used to describe significant aspects of performance changes during training (rather than the more traditional application in personnel selection). Although limited in terms of the number of subjects, the study involved the collection and processing of more than 20-million numbers. Because of practical constraints, it was necessary to make the assumption that the number of observations (participants x replications) could be used to replace the number of participants found in more conventional use of the multivariate technique. While the use of observations in this sense remains a researchable issue, it is emphasized that the work reported herein is being continued in order to establish a larger data base.

SECTION VI

CONCLUSIONS

METHODS ARE AVAILABLE

Engineering hardware and behavioral research methods are available to provide pilot-system performance measurement for many operational and training tasks. The major constraints appear to be related primarily to the amount of time and effort required to define and test measurement. In order to minimize these costs of obtaining performance information, and to maximize the utility of that information, method improvement should be undertaken.

METHOD IMPROVEMENT

The initial methodology for reducing candidate performance measures which were developed during this study requires further elaboration and refinement. Reduction of measures to the set which yields information concerning performance prediction will require further tests of criteria for rejecting measures; rejection on the basis of simple correlations appears erroneous in some cases. The discriminant procedure also requires refinement in the area of elimination of correlated measures; a priority elimination scheme appears warranted in some cases. Also, exercise of the selection techniques with a larger data base is mandatory.

The measurement development method requires the combined use of analytical and empirical techniques; however, the dependence on empirical data collection is more than desired. Empirical methods are costly and time consuming, partly because multivariate statistical procedures require such large samples for maximum effectiveness (cf., Lane, 1971). Often in practical settings sufficient time is just not available for the full use of this method. It is hoped that means can be found to permit heavier emphasis on analysis.

Over time, empirical data collection for measurement development may be reduced if (1) attempts are made to collect empirical results which are generalizable, and (2) measurement-relevant information is catalogued for used by others. If some attempt is made to preserve measurement development information, conceivably future data collection efforts may be reduced.

The work reported here is based on simulation research. Similar work involving inflight performance measurement will require expensive inflight and ground measurement equipment installations. As considerable expense is involved, justification of the expense is required in terms of the benefits accruing from the availability of performance information;

however, such a tradeoff analysis for justification also requires a measurement system for the generation of data. Perhaps small scale test systems should be developed for the purpose of exploring potential payoff.

The current methods using ground-based computer equipment can be improved by (1) new measures suggested by the empirical tests, (2) better computer algorithms for definition of measures, (3) implementation so that measurement can be computed and used as a simulated flight is performed, and (4) selection techniques which include diagnostic as well as discriminating and predicting measurement properties. Further, if test and evaluation efforts can be initiated which focus on measurement and operational information needs, measurement development efforts should benefit from the feedback provided.

SECTION VII

RECOMMENDATIONS

It is recommended that:

a. The discriminant selection method improvement be undertaken by further work with selection criteria along with the addition of a priority scheme to control measure rejection during initial correlation analysis.

b. Canonical correlation prediction method improvement be undertaken by implementing new algorithms which will consider measures which load onto more than one factor and will consider measures on both sides of the prediction equation.

c. More data be collected with the same experimental design, data collection and measure producing software to permit the acquisition of more observations and participants for the above method improvement.

d. After the desired measure sets and the conditions which control measurement are defined from the above work, a real-time programming effort be taken to modify the Instrument Flight Maneuvers program accordingly.

e. Following Instrument Flight Maneuvers program modification, conduct an evaluation of the measurement subsystem during automated training.

f. The performance measurement methods reported and referenced herein be considered for application to simulation and instructional aircraft environments. As a supporting comment, sufficient work has been done to-date to justify the conclusion that statistical and rational methods can be applied to the sensible specification of performance measures in manned-vehicle training. Training commands appear to have specific needs for improved measurement. Since measurement of the kind addressed herein may take some investment and lead-time, it is suggested that fine-tuning the selection methods need not hold back the process of obtaining measurement capability. Ultimately, measurement studies, or at least verification of measurement, must be conducted in operational training settings to insure the best utilization.

SECTION VIII

REFERENCES

Blackman, R.W. and Tukey, J. The Measurement of Power Spectra. New York, Dover, 1959.

Cooley, N.W. and Tukey, J. An Algorithm for the Machine Calculation of Complex Fourier Series, Mathematics of Computation, Vol. 19, No. 90, April 1965, pp 297-301.

Cooley, N.W. and and Lohnes, P.R. Multivariate Data Analysis. New York: John Wiley, 1971.

Erickson, E.S., Kapsis, P.B., Ciolkosz, M.D. Software Documentation for the Research Tool Digital Computer System Volume II Program Report. U.S. Navy, NAVTRADEVCEN 67-C-0196-7, September 1969.

Erickson, E.S., Kapsis, P.B., Ciolkosz, M.D. Software Documentation for the Research Tool Digital Computer System Volume IIA Detailed Program Description. U.S. Navy, NAVTRADEVCEN 67-C-0196-7, September 1969.

Johnson, R.M. AFT Program Description. Contract N61339-71-C-0205, Report SDR-111(AFT)PD, Logicon, San Diego, May 1972.

Kapsis, P.A., et al. Software Documentation for the Research Tool Digital Computer System Volume I Math Model Report. U.S. Navy, NAVTRADEVCEN 67-C-0196-7, September 1969.

Knoop, P.A. and Welde, W.E. Automated Pilot Performance Assessment in the T-37: A Feasibility Study. AFHRL-TR-72-6, Air Force Human Resources Laboratory, Wright-Patterson Air Force Base, April 1973.

Lane, N.E. The Influence of Selected Factors on Shrinkage and Overfit in Multiple Correlation. Naval Aerospace Medical Research Laboratory, Naval Aerospace Medical Institute, Pensacola, Florida, September 1971.

Norman, D.A. Personal Communication on the Implementation of a Second-Order, Low-Pass Digital Filter Program. 1973.

Obermayer, R.W. and Vreuls, D. Measurement for Flight Training Research. Proceedings of the 16th Annual Meeting of the Human Factors Society, Beverly Hills, California, October 1972.

Smode, A.F. Human Factors Inputs to the Training Device Design Process. NAVTRADEVCEN 69-C-0298-1. September 1971. U.S. Naval Training Device Center, Orlando, Florida. 327-378.

Vreuls, D. and Obermayer, R.W. Study of Crew Performance Measurement for High-Performance Aircraft Weapon System Training Air-to-Air Intercept. NAVTRADEVCEN 70-C-0059-1, February 1971a, U.S. Naval Training Device Center, Orlando, Florida.

Vreuls, D. and Obermayer, R.W. Emerging Developments in Flight Training Performance Measurement. U.S. Naval Training Device Center 25th Anniversary Commemorative Technical Journal, November 1971b. 199-210.

Vreuls, D., Obermayer, R.W., Goldstein, I., and Lauber, J.K. Measurement of Trainee Performance in a Captive Rotary-Wing Device. NAVTRAEQUIPCEN 71-C-0194-1. July 1973. U.S. Naval Training Equipment Center, Orlando, Florida.