IN PURSUIT OF THE FATEFUL FEW: A METHOD FOR DEVELOPING HUMAN PERFORMANCE MEASURES FOR TRAINING CONTROL

DONALD VREULS
Canyon Research Group, Inc.
Westlake Village, California

AND

IRA GOLDSTEIN
Human Factors Laboratory
Naval Training Equipment Center

Measurement produces information needed for assessment of trainee performance and subsequent control of the training process. Any device or system which is to control or evaluate the training process will be only as effective as its information source--measurement. Improvements in training efficiency and evaluation of training devices and methods are absolutely dependent on improved measurement.

The purpose of our work to date has been to *develop a method* for producing and selecting proper human training performance measures. Specific measures which resulted were of secondary importance. Therefore, this paper emphasizes those techniques which have resulted from method development studies sponsored by NAVTRAEQUIPCEN with partial support from the Naval Air Systems Command and the Advanced Research Projects Agency.

In order to measure many of the complex dimensions of man-machine system training performance, processing of large amounts of continuously varying information is required. Such measurement is beyond the capability of manual or simple measurement devices; it must be automated in order to produce information in time for efficient control of training.

Automated measurement places severe demands on the definition of (1) fool-proof algorithms for determining the conditions during which measurement is to occur, and (2) measure sets which produce only information necessary for effective use by the information receiving system. Too much information can overload the user system; not enough may reduce user effectiveness.

Historically, performance measures have been specified by analysis of skills, knowledges, tasks and mission requirements. Many NAVTRAEQUIPCEN sponsored studies (ref. 1, 2, 3, and 4) as well as others (ref. 5) have emphasized that analytic methods alone fail to produce satisfactory measurement. Analytically defined measures are likely to include (1) different measures of the same behavior and (2) measures which may prove to be unimportant. Although measurement development must start with a good analysis, empirical methods are required to reduce measures to a small, efficient set.

The reduction of analytically defined measures into a set which can be shown, mathmatically, to have the desired properties is called the *measure selection* process. Work to date has established and tested (1) a descriptive structure for obtaining measures in man-machine simulation training, and (2) a measure selection technique. The measure selection technique is based on multivariate statistical models which evaluate the total set of measures, taken together, and provide valuable scaling information.

MEASURE SELECTION PROCESS SUMMARY

The measure selection method was based on criteria that the final measure set should represent a comprehensive, yet minimum set of measures which (1) tended to eliminate redundant forms of information, (2) was sensitive to the skill change that occurred during training, and (3) had performance prediction qualities. The method that evolved from these studies contained a series of related analytical and empirical steps.

ANALYSIS FOR MEASUREMENT. The first step in the development of measures involves specification of performance measurement *candidates* (candidates for empirical selection analyses) which in the judgment of the investigators might contain information of importance for adaptive logics or other systems which are to control training. Examples from the literature can be used along with common task analytic techniques to perform analytic measure specification.

Measure Transformation - The analysis should identify the raw data parameters, such as (but not limited to) vehicular states and control inputs and the required sampling rates for each parameter. Typically, raw data are not in a form useful for automated measures; usually, error from some desired value or performance envelope contains more useful data. It is necessary, frequently, to transform error data into summary form, such as average error over a particular measurement interval in which the desired behavior of the trainee or system performance follows a lawful relationship.

Measure Start/Stop Logic - Conditions which define when measurement is to start and stop require detailed specification. It is emphasized that the specification of unambiguous rules to start and stop measurement can be underestimated. In practice, the construction of start/stop algorithms has been most challenging; iterative, empirical test of the algorithms has been mandatory.

When analytical means are exhausted, the measurement remaining is most often overabundant, unwieldly and perhaps impossible to completely implement in an operational setting. It is necessary, therefore, to seek other, empirical sources of information for further reduction of analytically defined candidate measures.

MEASURE SELECTION. The next step in the process requires collection of empirical data during training to provide a reliable sample for measure selection analyses. Computer measure selection techniques based on multivariate statistical models are used to reduce the candidate measures to a final set according to each of the aforementioned criteria The outcome of each analysis is evaluated by the investigators and merged into a recommended set for each maneuver and measurement interval.

The recommended measure sets are then processed to establish weighting coefficients for each measure. The weights reflect the importance of each measure to training, scale the measures on a common basis and permit summation of the many measures into one score.

MEASUREMENT MODEL

Our early work concentrated upon development of a descriptive framework for relating system performance and human behavior to segments of maneuvers constituting a training mission. The descriptive model (fig. 1) that emerged permitted measurement of a variety of tasks and performance dimensions in order to describe unique as well as common aspects of maneuvers. To accomplish this, the model defines *each measure* in terms of the following six determinants:

(1) A maneuver segment
(2) A parameter
(3) A sampling rate
(4) A desired value (if required)
(5) A tolerance value (if required)
(6) A transformation

A *segment* is any portion of a maneuver for which desired student behavior or system performance follows a lawful relationship from beginning to end, *and* for which the beginning and end can be defined unambiguously. Measurement start and stop conditions define a segment. Conditional tests of ongoing performance, such as the following, are required to start and stop measurement:

IF (1) ALTITUDE is GREATER THAN .OR. LESS THAN 1000-feet from the initial value,
.AND.
(2) HEADING is GREATER THAN 90° .AND. LESS THAN 270°,
THEN start measuring.

A *parameter* is any quantitative index of (1) vehicle states in any reference plane, (2) personnel physiological states, (3) control device states, or (4) discrete events. A *sampling rate* is the temporal frequency at which the parameter is examined. Often, parameters have no utility unless compared to a *desired value* or a *tolerance* to derive an error score. Finally, a *transformation* is any mathematical treatment of the parameter, to include measures of central tendency, variability, scalar values, Fourier transforms, pilot/system transfer functions, etc. Common measure transforms are shown in table 1.

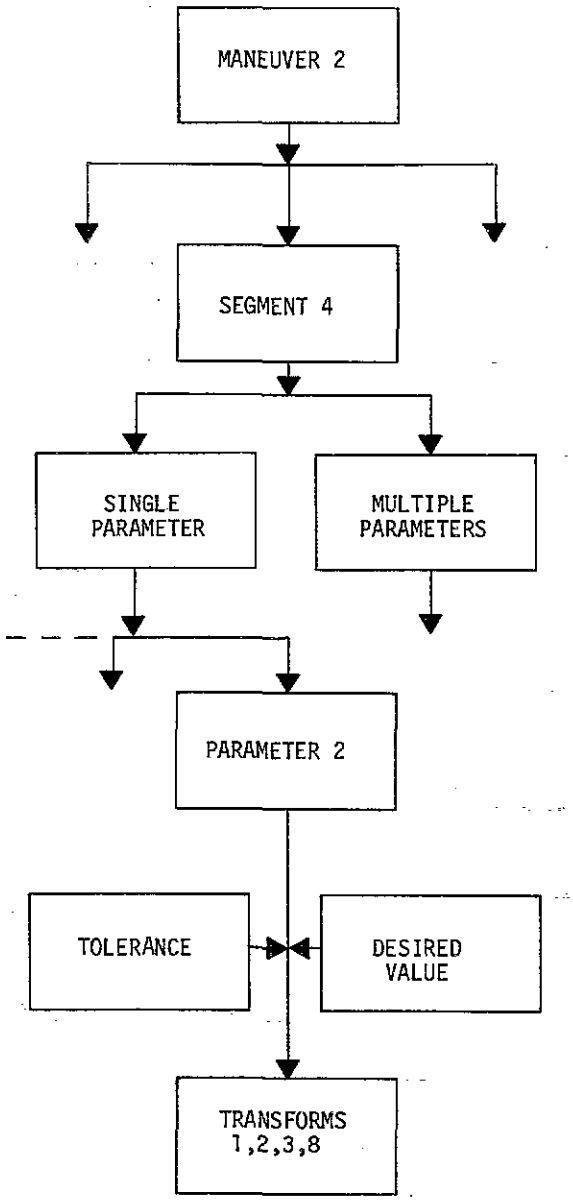

Figure 1. Measurement Model Components

TABLE 1. COMMON MEASURE TRANSFORMS

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, Range
- Minimum/Maximum Value
- Root-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

Computer programs have been written in FORTRAN IV for a PDP-9 and a Sigma-7 computer to generate measures from time history data tapes (recorded in real-time) in accordance with the measurement model. These programs operate in batch processing modes. They provide the flexibility of specifying each measure in terms of the aforementioned six determinants, using simple input lists. Thus, different measures can be created without reprogramming.

Having developed the measurement model and programs for producing measures, emphasis was shifted to measurement selection methods.

MEASURE SELECTION TECHNIQUES

Empirical studies were conducted in the context of visual reference flight and instrument flight to provide data for measure selection method development. These studies collected data while students underwent training for as long as 18 one-hour sessions. With these data it was possible to examine the changes in performance from the beginning to the end of training.

The purpose of measure selection method development was to create a set of techniques that would identify those measures which were redundant, changed as a function of training and predicted future relationships. The following four techniques were explored:

UNIVARIATE SELECTION. Considering each measure independent of all other measures, the average value of each measure on a given training day was compared to the average value of that measure on the criterion day, at the end of training. Statistically significant differences were found by t-tests of the means. Those measures which significantly changed were felt to be sensitive to the performance change as a function of training.

Univariate selection, however, gave no consideration to the relationships *between* measures, did not provide weighting or scaling information, and did not consider redundant measures.

REDUNDANT INFORMATION ELIMINATION. Many of the exploratory measures were different transforms of parameters which were closely related; also, many of the parameters were related due to airframe cross-coupling effects.

Highly correlated measures were omitted in order to reduce redundant information forms and to prevent possible computational problems in the following multivariate analyses. The correlations were computed only for a given maneuver segment on a given training day. If a measure was eliminated on one day of a comparison between two training days, it was dropped, also, from the measure set of the remaining day of the particular comparison.

DISCRIM SELECT. Computer programs were written to select measures using a special application of the multiple discriminant analysis (ref. 6). The analysis assumed that a battery of measures have been taken for each of a number of groups of participants. The primary purpose of DISCRIM SELECT was to isolate those measures that best discriminated between groups. For example, a pair of groups might have consisted of experienced and inexperienced people. The procedure discarded those measures that did not contribute to such discriminations when all measures were considered together as a set.

DISCRIM SELECT was based on an algorithm that iteratively discarded measures until a minimum set of measures resulted. The process stopped when either of two criteria was met, (1) the total number of remaining measures was less than the minimum number of factors required to describe the variance or (2) discarding another measure would have reduced the overall discrimination to an unacceptable level.

Two tolerances associated with the above

criteria had to be specified by the investigators: (1) the minimum percent variance to be accounted for by any factor and (2) the minimum measure communality. The percent variance of each factor in the original data was determined by a Principal Components analysis and compared to the tolerance specified. Measure communality was the loading of a measure onto a particular discriminant function, and could be thought of as the amount of variance (or discriminating "power") of a particular measure, given the discriminant function.

Trial analyses revealed that between 90 and 95 percent of the original variance was retained when the minimum variance for any factor was set at 7 percent; this tolerance set the minimum possible measure set size equal to the *minimum number* of "significant" factors. Trial analyses also revealed that in most cases measures which exhibited communalities less than .20 were not significant contributors to the discriminant function.

The DISCRIM SELECT computer programs were designed, therefore, to aid in the selection of measures which were capable of *discriminating* between previously designated groups, such as between instructors and students, or between early and later performance by the same trainees.

CANON SELECT. Another series of computer programs were designed to aid in the selection of measures which *related* performance measured at one time in training to that measured at another time. The basis of the method was a canonical correlation analysis (ref. 6) which derived linear combinations of the measures and maximized the correlation between a linear combination of one set of measures and a linear combination of another set of measures taken at another time.

If linear combinations were formed

$$y_1 = a_1x_1 + a_2x_2 + \ldots + a_nx_n$$
$$y_2 = b_1z_1 + b_2z_2 + \ldots + b_nz_n$$

where x and z were the same measures collected at different points in the training sequence, the canonical correlation analysis determined coefficients a and b so that y_1 and y_2 maximally correlated. The relation

$$a_1x_1 + a_2x_2 + \ldots + a_nx_n = b_1z_1 + b_2z_2 + \ldots + b_nz_n$$

is called the canonical correlation equation. For N measures, there were N such equations, the first of which maximally correlated, the second of which correlated next best and the Nth correlated least. A test of significance was available. For a given set of measures, it was possible to establish none, or many significant canonical correlation relationships.

The quantities y_1 and y_2 were factors of their respective groups. Such factors were not correlated with other factors. The computer programs generated the factor structure which displayed the correlation (or loading) of each measure on each factor. The factor structure was useful for quantifying the degree to which measures contributed to the relationship of performance at one time in training to that at another time. Measures which contained low loadings on factors, and factors which accounted for little variance in the data were prime candidates for omission.

CANON SELECT iteratively produced a list of measures ranked on the basis of their importance to relating performance at one time in training to that at another time. During each iteration, the measure which contributed least to the canonical correlation was removed, and the full analysis was performed again with the remaining measures. The process continued until only three measures remained.

Several attempts to develop a suitable algorithm for CANON SELECT were conducted. The most satisfactory procedure was found to be the following:

1. The variance extracted by each canonical factor was multiplied by the canonical correlation coefficient to determine the redundancy.

2. The loading of each measure onto each factor was multiplied by the redundancy. In this way a number was derived which reflected both the importance of each factor and the importance of each measure to the factor.

3. The maximum weighting, as found in (2) above, was determined for each measure.

4. Finally, all of the maximum weights in (3) above were examined. The measure which produced the smallest maximum weight was designated as the least-needed measure, and was subsequently dropped from the next iterative analysis.

5. The process repeated until only three measures remained.

Interpretation of the above data was necessary to form a recommended set which related performance at one time to performance at another time. The interpretation process started with the measures which were ranked from most important to least important. The ranked list was examined and only significant factors were retained. Thus, measures were omitted

which were loaded only on non-significant factors.

Measures thus produced were recommended for inclusion in the final measure sets. It is noted that the above procedure is not necessarily the only procedure that can be adopted. It was, however, the most satisfactory procedure that has been found to date.

FORM RECOMMENDED MEASURES. The objectives of the previous measure selection techniques were to define the set of measures which best discriminated and predicted performance for each maneuver segment and task variation. The results of DISCRIM SELECT and CANON SELECT were brought together to find the commonalities and differences in order to recommend measurement which would contain both discriminant and predictive information. Also, it was desirable to determine the weighting coefficients for each measure to reflect the training importance of each measure and scale the measures on a common basis for possible summation into a single score.

Final measure set specification was a manual, judgmental procedure based on data. The minimum measure set which resulted from DISCRIM SELECT was entered into a table for each maneuver segment and task variation, such as aircraft weight change or addition of turbulent air. To each table was added any measures produced by CANON SELECT.

The resulting measure sets were examined to insure that all vehicular outer loops which represented task instructions (such as hold heading, airspeed and altitude) were present. Measures which appeared infrequently (once or twice for any maneuver segment) were eliminated if the information they contained was represented by a high correlation ($.80<r<.90$) with another measure that was contained in the set, and/or they had been added to the table by CANON SELECT and they were weakly related (loadings less than .20) to the canonical factor.

Finally, DISCRIM SELECT was modified to perform an analysis on the recommended measure set. The analysis assured that a significant discriminant function was retained. Also, the weights assigned to each measure of the final set were computed; the weights provided scaling information for each measure relative to other measures in the set and coefficients for the linear summation of the measures into a single score, the discriminant function.

EXAMPLE MEASUREMENT STUDY

A brief summary of one measure development study is presented. The purpose of this particular study was to further refine measure selection techniques in the context of automated instrument flight training. Note that the previously explained techniques were a *result* of this study.

DEVICE. The NAVTRAEQUIPCEN Training Device Computer was used. It was configured as an F-4E, contained an experimental automated Instrument Flight Maneuvers (IFM) training program (ref. 7) and had a four degree-of-freedom motion platform.

The IFM system was modified to automatically control a measurement study. It set the initial conditions for each experimental trial and issued instruction for the trainees through a computer controlled voice synthesizer. The program produced time history records of 18 pilot system performance parameters in real-time for subsequent batch processing for measure production and measure selection analyses.

SUMMARY OF PROCEDURES. Twelve low-time student and private pilots were used as trainees. They averaged 55 hours of flight time, 3.7 hours of prior instrument time and had a median age of 24 years. The trainees represented the anticipated experience level for initial simulator instrument flight training.

Each trainee flew four basic instrument flight maneuvers, (1) straight and level flight, (2) standard rate climbs and descents, (3) level turns and (4) climbing and diving turns. Aircraft weight and resultant center-of-gravity shift, and turbulence were varied.

The trainees each flew six trials of each of the four basic maneuvers for 18 one-hour sessions. Adjacent one-hour training sessions were pooled together to form a training "day" for analysis purposes. There were nine training days. On each training day there were formed 12 trials for each of 12 trainees, or a total of 144 opportunities for measurement for each maneuver and segment. A total of 19 weeks was required to collect all of the data.

Candidate measure sets were constructed for each maneuver and segment. For example, the candidate measures for maneuver 4, final climbing and diving turns are shown in table 2. The whole maneuver consisted of three segments: (1) an initial climbing or diving turn through 1000-feet of altitude change and a 90-degree heading change, (2) a reversal from a climb to a dive (or a dive to a climb) along with a turn reversal, and (3) a final climbing or diving turn back to the initial heading and altitude. The candidate measures shown are for the third segment.

The segmentation rules were to start measuring when altitude change was greater than 1000-feet from the initial altitude and heading was more than 90-degrees from the initial value. Measurement stopped when altitude returned to its initial value. Several sets of segmentation rules were tested; these rules were the most satisfactory.

TABLE 2. EXAMPLE CANDIDATE MEASURES FOR MANEUVER 4, SEGMENT 3

MEASURE	ABBREVIATION
ELEVATOR STICK CROSSOVER POWER[1]	ELF1
ELEVATOR STICK AVERAGE DISPLACEMENT	ELF2
ANGLE-OF-ATTACK RANGE	ALRG
ALTITUDE RATE AVERAGE ABSOLUTE ERROR[2]	HDAA
RIGHT THROTTLE RANGE	THRG
AIRSPEED AVERAGE ABSOLUTE ERROR[3]	ASAA
AILERON STICK CROSSOVER POWER	AIF1
AILERON STICK AVERAGE DISPLACEMENT	AIF2
SIDESLIP ROOT-MEAN-SQUARED ERROR	BERM
ROLL ATTITUDE AVERAGE ABSOLUTE ERROR[4]	ROAA
RUDDER PEDAL CROSSOVER POWER	PDF1
RUDDER PEDAL AVERAGE DISPLACEMENT	PDF2
HEADING ABSOLUTE FINAL VALUE	PSAF
ELAPSED TIME	TIME

[1]The relative power between two and six radians-per-second generated by the pilot (ref. 8).
[2]Error from 1000 feet-per-minute climb or dive.
[3]Error from a desired 280-knots IAS.
[4]Error from a left or right 30-degree bank.

RESULTS. Candidate measure sets were created differently for each of the maneuver segments to reflect different dimensions of control and different criteria of performance. The data were used to refine measure selection techniques and recommend measurement for use by IFM. General results are discussed along with some specific results obtained for the example maneuver 4, segment 3.

t-Tests - As many as 23 and as few as 11 candidate measures (average of 16 measures) were constructed for each maneuver segment. Univariate statistical tests of the mean differences between initial and final training revealed that as many as 21 and as few as three measures significantly changed. The average number of measures selected by t-tests were 10-measures for each maneuver and segment.

Redundant Forms - Many of the candidate measures were redundant according to the intercorrelations on a given training day. Measures with intercorrelations greater than r=.90 were considered to be equivalent; one measure of each pair was dropped. As many as 10-measures (as few as zero) were found to be equivalent for each maneuver segment. An average of four measures were omitted per maneuver segment.

It was possible for a given measure to correlate with more than one other measure. In many cases, chains of intercorrelating measures were formed. After removing redundant measures, an average of 12-measures remained for multivariate measure selection analyses for each maneuver segment.

Discriminant Analyses - After elimination of redundant measures, the multiple discriminant analysis further reduced the candidate measures to an average of six measures per maneuver segment. Table 3 shows the measures selected for the example maneuver; the discrim vector was the weight of each measure and the communality was related to variance. The CHI SQUARE (χ^2) test of significance is shown.

TABLE 3. MEASURES SELECTED BY DISCRIM FOR MANEUVER 4, SEGMENT 3

	LIGHT AIRCRAFT		HEAVY AIRCRAFT	
MEAS	DISCRM VECTOR	COMMUN-ALITY	DISCRM VECTOR	COMMUN-ALITY
ELF1				
ELF2				
AIF1				
AIF2				
ALRG	.31	.60	.52	.35
HDAA	.75	.77	1.20	.79
ROAA	.04	.17	.02	.25
PSAF	.01	.19	.01	.22
ASAA	.04	.52	.05	.67
χ^2	110		44	
df	5		5	
p<	.01		.01	

Table 3 shows the measures selected which were sensitive to *training* during two different aircraft weight configurations. Other comparisons were made to find the measure sets which would be sensitive to the effects of task stressors (turbulence alone and in combination with a heavy aircraft) at the completion of training.

Over all maneuver segments, the composition of the minimum discriminant set changed as a function of training and task stressors, as illustrated below:

Measure Type	Training	Stressors
Control Input	25%	56%
System Performance	72%	37%
Elapsed Time	3%	7%

Control input measures (stick, pedal and throttle) represented 25-percent of the minimum measures during training and 56-percent of the minimum measures which described performance changes due to task stressor changes.

Canonical Correlation Analysis - Following elimination of redundant measures, CANON

SELECT iteratively reduced the measures, forming a ranking of measures from most desirable to least desirable in terms of relating performance at one time to that at another time in training. Overall, CANON SELECT produced an average of seven measures for each maneuver segment. The results of the example case are shown in table 4. Measures which showed predictive relationships often were different than those that permitted discrimination.

TABLE 4. SUMMARY OF MEASURES SELECTED FOR MANEUVER 4, SEGMENT 3

MEAS	LIGHT AIRCRAFT DISCRM	LIGHT AIRCRAFT CANON	HEAVY AIRCRAFT DISCRM	HEAVY AIRCRAFT CANON
ELF1		X		X
ELF2		X		X
AIF1		X		
AIF2		X		X
ALRG	X	X	X	
HDAA	X		X	
ROAA	X	X	X	
PSAF	X	X	X	
ASAA	X	X	X	

Over all maneuver segments, the composition of the significant canonical relationships did *not* change materially as a function of task stressor, and was very similar to the composition of the stressor discriminant set:

Measure Type	Training	Stressors
Control Input	59%	56%
System Performance	37%	40%
Elapsed Time	4%	4%

Notice, however, that control input measures accounted for more than half of the measures that demonstrated significant canonical relationships overall. Again, control inputs were critical for a complete description of performance.

Final Recommended Measures - Having established the minimum discriminant measure sets and the minimum canonical sets which reflected predictive qualities, the results of both analyses were brought together. DISCRIM SELECT was modified to perform an analysis on the resulting measure sets to produce weights and communalities for each measure.

The results are shown for the example case in table 5. Even though measures that did not contribute to the discrimination were added (because they contained significant canonical relationships), the resulting discriminant functions were statistically significant. Note that the measure set changes slightly in that AIF1 drops out as a measure for training with a heavy aircraft. More importantly, notice that the weights for any particular measure may change substantially as a function of heavy or light aircraft training.

TABLE 5. FINAL MEASURES AND WEIGHTS FOR MANEUVER 4, SEGMENT 3

MEAS	LIGHT AIRCRAFT DISCRM WEIGHT	LIGHT AIRCRAFT COMMUN-ALITY	HEAVY AIRCRAFT DISCRM WEIGHT	HEAVY AIRCRAFT COMMUN-ALITY
ELF1	-67.90	.02	1.14	.11
ELF2	-1.41	.00	.26	.03
AIF1	-5.16	.07		
AIF2	.10	.04	.09	.00
ALRG	.41	.48	.03	.35
HDAA	.89	.61	1.23	.78
ROAA	.15	.13	.01	.25
PSAF	.01	.13	.13	.22
ASAA	.01	.41	.05	.67
χ^2	145		44	
df	9		8	
p<	.01		.01	

Over all maneuver segments, an average of 9.5-measures were recommended. Weights were produced for summation of these measures into one score for each maneuver segment and task condition.

DISCUSSION

MEASURE SET COMPOSITION. The measure selection analysis data have made two critical points which have an enormous impact on the design of performance measurement systems for automated simulator flight training.

First, control input measures contained a significant amount of information about training and the effects of two task stressors. Typically, control input measures are not found in many training device measurement systems. Even advanced systems may not evaluate control inputs, primarily because without empirical data, such as those contained herein, it has been difficult to assess the implication of control inputs. The discriminant analysis removes some of these difficulties by (1) selecting measures and (2) assigning weights for the utilization of measures.

The second critical point has been seen in every measurement study conducted to date: Different measure sets are required when the task changes, even with the simple addition of light turbulence. Measure set composition changes alter both (1) the specific measures selected for each task and (2) the weighting coefficients for these

measures if the data are to be summed into a single score.

Measures which are not useful for one condition, but which are "carried along" to cover a second condition, might degrade the power of the set to describe the first condition. Thus, one must be cautious in the application of universal measure sets to cover a variety of task situations. To guard against degrading power of measurement, only empirical studies offer an avenue to assure proper measure selection and compatibility at this time.

MEASURE SEGMENT START/STOP LOGIC. In spite of the broad capacity of existing programs to define when measurement starts and stops, considerable testing was required to derive a satisfactory set of logical conditions for starting maneuver 4, segment 3. In retrospect, additional logic would have eased the problem.

The point made by experience to date is that one can never be certain that the final solution has been achieved. Trainees will enter maneuvering flight, and performances will unfold in ways that are beyond the wildest dreams of the measurement analyst. This will occur in spite of standard, straightforward geometric requirements of even the most simple maneuvers. Allowances must be made for algorithm modification in the development of automated measurement systems.

REDUNDANT MEASURES ANALYSIS. The specification of initial candidate measures is a direct function of the skill of the analyst. Two kinds of errors have a high probability of occurrence. The most probable error is overmeasurement. In the face of uncertainty caused by sparse evidence, the tendency is to adopt the philosophy, "If it moves, measure it." The second kind of error is to omit an important information form, such as control input.

These two kinds of errors represent a dilemma for the measurement analyst. If the candidate measure sets are terse, the risk of missing important information is high. Yet, if the sets are abundant, the risk and cost of overmeasurement can be so enormous that data collection becomes impractical.

The use of correlation analyses to reduce redundant forms of information appears to be a useful tool to ease the dilemma. It serves as a first step check on the analyst. Also, it permits the analyst a little latitude to experiment with measures in selected areas of uncertainty. However, heavy dependence on the redundant measures analysis should be avoided.

MINIMUM STATISTICAL SAMPLE. The measure selection techniques which have developed appear to have been effective. However, considering the amount of data collected, one wonders if a smaller statistical sample could have sufficed.

A relatively small number of trainees have been used. The adequacy of this sample depends on the population to which one wishes to extrapolate. On the other hand, a large amount of data have been collected and numerous observations of each trainee have been obtained. Techniques developed in the work to date would be more easily applied in the future if the amount of data collection could be reduced. Consequently, it is appropriate to ask if sufficient statistical power would be maintained with fewer observations.

The data needed to make this examination did not exist for planning the measurement studies to date. Now that data are on hand, it is possible to modify computer programs and empirically find the minimum allowable sample. Such re-analysis should permit future applications of the methods which have been developed with increased efficiency.

PERFORMANCE PREDICTION. As it has been developed, the canonical correlation analysis allows prediction based on an equation that related a linear combination of measures under one set of conditions with a linear combination of the same measures under another set of conditions. The technique worked quite well.

The canonical correlation analysis generally produced a number of variables on both sides of the canonical equation. There were, however, cases where (1) measures which were more important to the relationships were different on each side of the equation and (2) particular combinations of measures early in training predicted a singular measure later in training. Such relationships may be both meaningful and useful.

Although beyond the results of the work to date, it appears most desirable to have the capability to predict (1) specific measures, (2) diagnostic combinations of measures, and (3) discriminant scores. Specific measures may be meaningful and important to control training. Measures may be combined to permit prediction of the need for remedial training or a specific training unit. Discriminant scores may be predicted to allow determination of future class standing for a specific individual (since the discriminant score

is the best combination of measures for discriminating between alternative levels of performance).

CONCLUSIONS AND RECOMMENDATIONS

We have developed a set of techniques that produce a reasonably minimum set of performance measures. The measures have the ability to discriminate performance levels and have prediction qualities for each maneuver segment and task variation tested. Weighting coefficients for the measures resulted; therefore, the relative importance of each measure for training or task variation was scaled. The weights permitted summation of measures into one score, the discriminant function, for use by an adaptive logic which requires one score for performance assessment.

The specific results of the measurement work to date are being incorporated in the NAVTRAEQUIPCEN Automated Instrument Flight Maneuvers training system for real or near-real time use by that system. An evaluation is envisioned to examine the use of the discriminant function with an adaptive logic that requires one performance score for future application to this class of adaptive logics.

It is not recommended, however, that future adaptive logics be tailored to use only one score simply because a method for obtaining a single score is available (except in cases where the use of one score is clearly the most reasonable design trade-off). Our reason for this position is that measures which were sensitive to many dimensions of task performance have been seen in all measurement studies to date. Future adaptive logics should be designed to fully use the available information, if at all possible. More efficient training might result if adaptation occurs in areas of task deficiency and proficiency, rather than on the basis of a single score of overall task performance alone.

Work on statistical sample size should be undertaken to examine the possibility of reducing the sampling requirements and increasing the efficiency of future measurement studies. The existing data base from our measurement studies can be used for certain kinds of empirical sampling studies.

More work on performance prediction should be done to explore predictive relationships in the present data and to clarify the role of prediction in future, more complex adaptive training systems. It is clear from current NAVTRAEQUIPCEN work on self-organizing adaptive training systems that the success of such systems will be predicated on the ability to predict.

The measure selection process, as it has been defined herein, is one way to approach measurement. It is not the only way. Many variations on the theme are possible, and may be useful for specific situations; however, the basic method appears to work. It is recommended that the method be applied to the development of meaningful measurement for existing and future training systems that have measurement capability.

REFERENCES

1. 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 1971.

2. 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 1971.

3. 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.

4. Vreuls, D., Obermayer, R.W., and Goldstein, I. Trainee Performance Measurement Development Using Multivariate Measure Selection Techniques. NAVTRAEQUIPCEN 73-C-0066-1, In Press.

5. 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, Ohio, April 1973.

6. Vreuls, D., Obermayer, R.W., Goldstein, I. and Norman, D.A. Trainee Performance Measurement in Four Instrument Flight Maneuvers. NAVTRAEQUIPCEN 74-C-0063-1, In Press.

7. Charles, J.P., Johnson, R.M., and Swink, J.R. Automated Flight Training (AFT) Instrument Flight Maneuvers. NAVTRAEQUIPCEN 71-C-0205-1, 1973.

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

ACKNOWLEDGEMENTS

The work reported herein has been successful because of the collective efforts of many Contractor and NAVTRAEQUIPCEN personnel, too numerous to list here. The contributions of two individuals, however, cannot go unrecognized: Mr. Richard W. Obermayer of Manned

Systems Sciences, Inc., has made significant technical contributions to the measure selection process development. Mr. James S. Duva, Acting Head, Human Factors Laboratory, NAVTRAEQUIPCEN provided support for our work and shared our belief even in the early days, that performance measurement in the present context could be made to work.

ABOUT THE AUTHORS

MR. DONALD VREULS, Senior Staff Scientist at Canyon Research Group, Inc., has executed, directed, or managed 20 research and development contracts in basic and applied problems related to quantifying, optimizing or predicting human performance in industrial, education, military training, environmental and man-machine systems. He has applied behavioral engineering analytic methods to analysis and design of training systems and has originated and developed computer software concepts for human performance measurement and evaluation. As a consultant and member of several design teams, Mr. Vreuls has influenced design based on human learning, physiological, perceptual motor and cognitive capabilities. With research and applied experience in commercial and military aviation, he has originated hardware/software avionic systems design concepts. Mr. Vreuls has also authored or co-authored 29 technical reports, journal articles, and professional presentations. Prior to establishing Canyon Research Group, Inc., he was a Program Manager and Principal Investigator at Manned Systems Sciences, Inc. (1968-1973). From 1961 to 1968 he conducted human factors research for The Bunker-Ramo Corp.

MR. IRA GOLDSTEIN is a Research Psychologist in the Human Factors Laboratory at the Naval Training Equipment Center. His professional activity over the past dozen years has focused on the role of computers in behavioral science research and their employment to improve training. He has contributed a number of technical papers on computer-controlled measurement of human performance in perceptual-motor and decision-making tasks. Before joining the Center in 1969, Mr. Goldstein was employed in private industry for 4 years. Previously, he had been with the United States Air Force's Decision Sciences Laboratory from 1958 to 1965.

PAPERS PUBLISHED, BUT NOT PRESENTED