LAND
SEA
AIR
SPACE

Society of Automotive Engineers, Inc.
TWO PENNSYLVANIA PLAZA, NEW YORK, N.Y. 10001

AEROSPACE INFORMATION REPORT

AIR 1054

Issued 12-1-68
Revised

THE MEASUREMENT OF TRAINEE PERFORMANCE IN SIMULATORS AND PART-TASK TRAINERS

1. INTRODUCTION

1.1 Increasing attention is being given to measuring trainee performance in synthetic training devices, as evidenced by the recent Air Force, NTDC, and NASA-sponsored efforts in this area cited in the bibliography. A fair summary of well-informed opinions in this area would be:

a) Such measurement can be most useful (see par 2).

b) Subjective measures, such as the traditional instructor ratings,. do not meet the current requirements for reliable, valid, and practical measurement.

c) Synthetic training equipment, especially of the digital variety, makes feasible the achievement of needed measures more expeditiously than in operational equipment.

1.2 This report intends to acquaint the non-technical reader with the potentialities and limitations of measuring trainee performance in synthetic training devices, and outline the steps needed to achieve these measures. As an introduction, the report intends to be limited in depth; for those desiring to explore the subject in greater depth, publications cited in the Bibliography should be consulted. Much of the information cited herein is derived from the reports listed in the Bibliography (par. 9.); the report by Smode et al. was especially useful.

2. IMPORTANCE AND PURPOSES OF MEASUREMENT

Performance measurement can serve a wide variety of functions. The optimal kind of measurement depends on the functions to be served, but most measures should be relevant for a number of different functions. The specific purposes which can be served by measurement of trainee performance can be stated as follows:

a) Feedback for Training - It has been well established that furnishing trainees with knowledge of the results of their efforts enhances their learning. Feedback is of most value when it is prompt, accurate, and relevant.

b) Trainee Motivation - When objective standards of performance exist, each individual has an immediate and concrete goal to strive for, and also meaningful and healthy competition between individuals or crews is possible.

c) Prediction of Future Success - Measurement may involve the collection of data from which can be estimated (preferably with a specified success probability) how an individual or team will perform in some future context or universe of events. This prediction of future performance may be an evaluation of aptitude for training or an evaluation which uses measures of training achievement as the basis for prediction of subsequent operational performance.

d) Evaluation of Present Performance - Measurement may involve collecting data which can specify the knowledge, skill level, or performance level of an individual or team. The measures may reflect present performance level in both part-tasks and/or more comprehensive segments.

e) Evaluation of Learning Rate - Measurement data may be collected at several points in a training program in order to indicate the rate at which knowledge and skills are being acquired. Such measures provide a basis for judging an individual's or a crew's present stage of learning and readiness for the next phase in a training program.

f) Identification of Areas of Proficiency and Deficiency - Measurement data, particularly of a diagnostic nature, may be used to determine in what areas or tasks either individual crew members or an entire crew are proficient and in what areas they are deficient. Such measures pinpoint the need for and nature of further training and suggest task or environmental modifications to achieve a specified level of proficiency. In addition they can provide information to the trainee which will speed his learning.

g) Evaluation of Training Effectiveness - Measurement data may be used to determine the nature and extent of changes resulting from a training experience. Another important area subsumed here relates to training research and the evaluation of the differential effectiveness of alternative training methods, the contribution of component proficiencies to overall mission accomplishment, schedules, simulators and simulator features.

h) Selection and Placement of Individuals and Teams - Measurement data may assist in identifying individuals more likely to achieve a given level of proficiency: i.e., the identification of persons who either will require less training or will profit the most from a given amount of training. Included here are the assignment of individuals to crew positions and the assignment of crews to special missions.

SAE Technical Board rules provide that: "All technical reports, including standards approved and practices recommended, are advisory only. Their use by anyone engaged in industry or trade is entirely voluntary. There is no agreement to adhere to any SAE standard or recommended practice, and no commitment to conform to or be guided by any technical report. In formulating and approving technical reports, the Board and its Committees will not investigate or consider patents which may apply to the subject matter. Prospective users of the report are responsible for protecting themselves against liability for infringement of patents."

i) Refinement of Criterion Information - Measurement data may provide the basis for refinement of the criterion by helping to define further what constitutes successful or proficient performance.

j) Definition of Requirements - Measures of performance may permit statements of functional requirements for training equipment to be specified more precisely. Training standards can be made more precise and objective, and training equipment more effective.

k) Evaluation of Equipment and Procedures - This can include determination of whether given equipment or procedures permit attainment of required standards, as well as determination of the better of two equipments or procedures. For example, total training time to a proficiency criterion can be used as a measure of effectiveness of a specific equipment or procedure. Items to be evaluated include instruments, control feel, personal equipment (e.g., seating), and task sharing between captain and first officer.

l) Evaluation of Instructor Capability - The best measure of an instructor is the kind of student he turns out. Reliable and valid measures of trainee performance permit instructors to be evaluated directly by their product, rather than indirectly by their knowledge of the subject and teaching manner.

3. CLASSIFICATION OF PERFORMANCE VARIABLES

3.1 There are seven basic classifications of measures useful in evaluating trainee performance:

1) Time: Measures dealing with time periods in production of performance.

2) Accuracy: Measures dealing with the correctness and adequacy of production of performance.

3) Frequency of occurrence: Measures dealing with the rate of repetition of behavior.

4) Amount achieved or accomplished: Measures dealing with the amount of output or accomplishment in performance.

5) Consumption or quantity used: Measures dealing with resources expended in performance in terms of standard references.

6) Behavior classification by observers: Measures dealing with classifying more complex behaviors into operationally defined subjective categories. Observations are placed into discrete classes on a continuum for the event observed.

7) Condition or state of the individual in relation to the task: Measures dealing directly with the state of the individual which describe behavior and/or results of acts that have occurred.

3.2 These classes of measures are graded on a quantative-qualitative continuum, with precise quantities (time, accuracy, frequency) at one end and more qualitative interpretations (categorization, descriptive reports) at the other. Each class or group includes a variety of subgroups and specific measures. These are listed in detail in Table 1.

3.3 As in other areas of psychometrics, the more objective, easier-to-obtain measures usually reflect but a single facet of the trainee's performance; more global measures tend to be unreliable, difficult to obtain, or both. In general, combinations of discrete scores are required; the task of combining separate scores into a useful overall score is often a demanding one. Since a given overall score can be achieved in a variety of ways, it is usually necessary to utilize sub-scores as well as the overall score in interpreting trainee performance.

Table 1

A Classification of Measures

TIME

1. Time to Initiate an Activity from the Onset of a Signal or Related Events
    - Time to perceive event
    - Reaction time
    - Time to initiate a correction
    - Time to initiate a subsequent activity (following completion of a prior activity)
    - Time to initiate a course of action
    - Time to detect trend of multiple related events
2. Time to Complete an Initiated Activity
    - Time to acquire, to lock-on, to identify
    - Time to complete single message
    - Time to complete a computational problem
    - Time to make an adjustment/manipulation/control positioning
    - Time to reach a criterion
3. Overall Time from Signal Onset to Activity Completion
    - Percent time-on-target
    - Time spent in an activity (communicating, repairing, computing, etc.)
    - Time to complete a sequence of activities
    - Build-up of time (cue length)
4. Distribution of Part Task Times in Completing an Activity
    - Time-sharing among events

ACCURACY

1. Correctness of Observation or Perception (Discrete/sequential)
    - Accuracy in identifying display readout
    - Accuracy in identifying extra-cockpit objects (environment, ground terrain, celestial navigation objects)
    - Accuracy in estimating distance, direction, speed

Table 1 (Cont'd)

- Time estimating accuracy
- Detection of a trend based on multiple related events
- Detection of change in presence of noise
- Correctness of observation sequence

2. Correctness of Response or Output
    - Accuracy in control positioning (pressures, direction, amplitude, rate, and duration)
    - Accuracy of in-flight maneuvers
    - Accuracy of retrofire maneuvers
    - Accuracy of intercept
    - Computing accuracy
    - Selection of action from among alternates
    - Correct symbol usage
    - Accuracy in spatial positioning (navigation)
    - Accuracy in weapon delivery
    - Accuracy in landing
3. Error Magnitude
    - Error amplitude measures
    - Error frequency measures
    - Error in bomb drop
4. Correctness of Response Sequence
    - Sequence of response
    - Sequential-manipulative accuracy (serial response, one activity; coordinated response with several controls)
5. Adequacy of Probability Estimation (Relative to an "Ideal Observer")
    - Accuracy in using unreliable information
    - Recognition of signal in noise
    - Recognition of out-of-tolerance condition

FREQUENCY OF OCCURRENCE

1. Number of Responses Per Activity or Interval
    - Number of actions made per unit
    - Number of communications per activity or interval
    - Number of adjustments to maintain in-tolerance (number of checks, replacements, problems solved)
    - Number of interactions with other members
    - Number of gross/significant errors per unit
2. Number of Defined Consequences of Performance Per Activity
    - Number of out-of-tolerance conditions
3. Number of Observing or Data-Gathering Responses
    - Number of requests for information
    - Number of interrogations/observations made
    - Number of discrete recordings/reportings made

AMOUNT ACHIEVED OR ACCOMPLISHED

1. Response Magnitude or Quantity Achieved
    - Degree or proportion of success (intercepts information collection, weapon delivery, rescue, landing, etc.)
    - Cumulative response output
    - Written test of knowledge (scores)
2. Man-Machine System Achievement
    - Attainment of training objectives
    - Assessment of "merit" in performance (influenced by man-machine interactions)

CONSUMPTION OR QUANTITY USED

1. Resources Consumed Per Activity
    - Fuel/energy conservation
    - Units consumed in activity accomplishment
2. Resources Consumed Per Time
    - Rate of consumption

BEHAVIOR CATEGORIZATION BY OBSERVERS

1. Classifying Activities or Handling of Events
    - Impromptu response invention (improvising)
    - Communication effectiveness
    - Redundant communications
    - Emotional content of communication
    - Priority assignment to an activity or among activities
2. Overall Judgments of Performance
    - Coordination of effort/movement
    - Procedural synchronization of action
    - Relevance of response
    - Substantive content of communication
    - Intelligibility of voice report
    - Use made of available references, job information, test equipment
    - Visual-perceptual orientation
    - Crew cohesiveness
    - Quality of checks (fault location)
    - Use made of performance information available from symptons/checks/errors
    - Adequacy/goodness of behavior (gross rating of a complex performance)
    - Adherence to safety procedures (handling of equipment)

CONDITION OR STATE OF THE INDIVIDUAL IN RELATION TO THE TASK

1. Description of Behavior at Prescribed Times
    - Response perseveration
    - Anticipation of probable events
    - Alertness to events
2. Description of Condition
    - Behavioral intactness of individuals/crew
    - Physiological condition of individual/crew (life support) (by means of attachment on body surface or equipment near the body: electrocardiogram, electroencephalogram, temperature, galvanic skin response, sound at ear drum, etc.)

Table 1 (Cont'd)

3. Self Report of Experience
   Report of illusory phenomena (apparent movements; quality and duration of illusory movements)
   Protocols of experience

4. METHODS OF DATA HANDLING AND DISPLAY

Data on trainee performance can vary in form, timeliness, and permanence.

4.1 Form - The two broad categories here are analog and digital. Analog-type data are provided by most "repeater" instruments, CRT's, and plotters, both XY and XT. Digital-type data are provided by indicator lights, digital readout, counters, and printouts.

4.2 Permanence - Data can be displayed in either a transient or permanent manner. Permanent or hard copy data comprise numeric or alpha-numeric printouts, and charts, both XY and XT. Transient data are displayed (usually at the instructor's console) with a variety of devices: indicator lights, digital readouts, "repeater" instruments, meters, and CRT's.

Typical ways of providing data relevant to the assessment of trainee proficiency in current flight and mission simulators include the following:

1) In simulators where the instructor can watch the trainee and his instruments directly, cockpit instruments and indicators furnish appropriate data.

2) In simulators where the instructor's station is not in the cockpit, "repeater" instruments, "repeaters" of switch positions, and other status displays are provided.

3) XT recorders are used to provide a time history of relevant parameters (e.g., altitude and heading during a bombing run on a F-4C Weapons System Training Set).

4) XY recorders are used in two ways:

a) with the X axis representing East-West and the Y axis North-South, the device is a cross-country or approach flight path recorder;

b) with the X axis representing distance from the end of a runway and the Y axis height above that runway (or above the glide slope), the device shows vertical position data.

5) Alpha-numeric readouts of trainee performance are provided in real time on a teletypewriter (F-111A Mission Simulator).

A hard copy of 3, 4, and 5 above may be used by students after a problem.

4.3 Timeliness - Data can be provided in real time, or stored for later use, either by the instructor (experimenter) or by the trainee. Data presented in other than real time usually is in the form of hard copy (e.g., charts, printouts) but can be in a transient form, such as a CRT display.

5. ESTABLISHING PERFORMANCE CRITERIA

5.1 Four steps are needed in the selection of useful and relevant criteria:

1. Define the activity. Specify, to the extent possible, the activity in which it is desired to determine successful and proficient performance.

2. Analyze the activity. Consider the activity in terms of the purpose or goals, the types of behaviors and skills that seem to be involved, the relative importance of the various skills involved, and the standards of performance which are expected.

3. Define proficient and successful performance. Identify the elements that make for successful performance and weight these elements in terms of their relative importance.

4. Develop sub-criteria to measure each element of success. As appropriate, develop a combined measure of successful performance which includes each element weighted in accordance with its relative importance.

5.2 With automated scoring, it is sometimes possible for the trainee to "beat the system," i.e., achieve a high score without performing in the desired manner. For this reason, it is necessary that the measures taken actually reflect the performance criteria, and not merely correlate with them. "Beating the system" is not a problem when the instructor does the scoring because the instructor is usually aware of trainee deviations from correct procedures.

5.3 Instructor scoring is limited, however, by the limited attention span of a human, compared with that of a computer, and the unreliability of his performance. A simulator computer can utilize a wide variety of data sources within a short time interval to derive a trainee score; the number of displays an instructor can scan is severely limited. The computer, given identical trainee performance, will produce the same score time after time; with instructor scoring, variations will occur from one instructor to the next, and with a given instructor from one session to the next.

6. VALIDATION

"Validation" means determining the extent to which the developed performance measures actually measure what they purport to measure. Two important aspects of validity here are content validity and predictive validity.

6.1 Content validity is the extent to which the performance measures reflect the tasks included in the curriculum. Every major element in the curriculum should be covered by measurement data.

6.2 Predictive validity is the extent of the relationship of the scores obtained to meaningful external criteria. For example, if licensed transport pilots did not score higher than beginning students on an item purporting to measure smoothness of approach, the item would be suspect.