AD-783 256

EFFECTS OF GROUND-BASED AIRCRAFT SIMULATOR MOTION CONDITIONS UPON PREDICTION OF PILOT PROFICIENCY. PART I

Jefferson M. Koonce

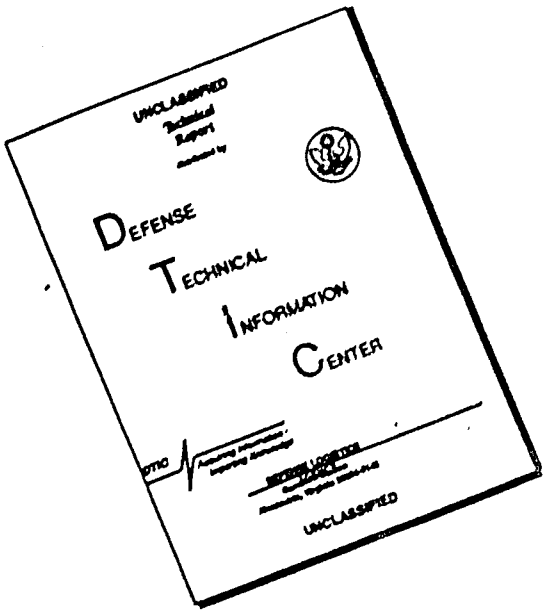
Illinois University

Prepared for:

Air Force Office of Scientific Research

April 1974

DISCLAIMER NOTICE



THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

AD 783 256

REPORT DOCUMENTATION PAGE		READ INSTRUCTIONS BEFORE COMPLETING FORM
1. REPORT NUMBER AFOSR - TR - 74 - 1292	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
4. TITLE (and Subtitle) EFFECTS OF GROUND-BASED AIRCRAFT SIMULATOR MOTION CONDITIONS UPON PREDICTION OF PILOT PROFICIENCY. Part I and [redacted]		5. TYPE OF REPORT & PERIOD COVERED Scientific Interim 6. PERFORMING ORG. REPORT NUMBER
7. AUTHOR(s) Jefferson M. Koonce		8. CONTRACT OR GRANT NUMBER(s) F44620-70-C-0105
9. PERFORMING ORGANIZATION NAME AND ADDRESS Aviation Research Laboratory Institute of Aviation University of Illinois, Savoy, Illinois 61874		10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS 9778 681313 61102F
11. CONTROLLING OFFICE NAME AND ADDRESS Air Force Office of Scientific Research 1400 Wilson Boulevard (NL) Arlington, Virginia 22209		12. REPORT DATE April 1974 13. NUMBER OF PAGES 213
14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)		15. SECURITY CLASS. (of this report) UNCLASSIFIED 15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Performance Prediction
Pilot Training
Flight Simulator Motion

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Three groups of thirty pilots with multi-engine and instrument ratings performed a simulated flight mission in a General Aviation Trainer - 2 (GAT-2) on each of two days. The experimental conditions for the graups differed in terms of GAT-2 mation (Graup I - no motion; Group II - sustained linear, scaled-down analog motion; Group III - washout motion). Each group af pilats then flew the same mission in a light twin-engine aircraft representative of the class of aircraft simulated by the GAT-2. The experimental design

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

was a two factor mixed design (groups by days) with repeated measures on one factor (groups).

The mission consisted of five maneuvers representative of those usually performed under instrument flight rules (IFR) without visual reference to the outside world and five maneuvers usually performed with outside visual contact under visual flight rules (VFR). In the simulator, all of the maneuvers were performed without outside visual reference.

Two trained observers, one of whom was also the safety pilot for the mission, recorded pilot performance on each mission in a specially designed booklet. The order of assignment of observers to the missions permitted recording of a pilot's performance on a single mission by two independent observers and also the recording of the pilot's performance on two successive missions by the same observer and two independent observers.

The results indicated that the proficiency of aircraft pilots can be predicted to a high degree from ground-based simulator performance measures. Of the three simulator motion conditions used greater prediction of operator performance from a simulator to flight can be obtained using sustained cockpit motion than by using washout motion or no motion. There was no significant difference between the predictive validities of performance with no motion and washout motion.

The experiment demonstrated that very high observer-observer reliabilities ($r = .771$ to .971) on the same mission can be obtained by recording performance on scales that are well defined and easy to follow, descriptive of the maneuver and behavior being recorded, and not too demanding upon the person doing the recording of performance. The performance measures taken in the simulator tended to be more reliable than those taken in the aircraft because of the elimination of degrading environmental factors and the reduction of safety oriented duties frequently imposed upon safety observers.

Simulator motion tends to increase subject acceptability of the device, lower performance error scores, and reduce the workload on the subjects and observers through the aiding effects of the motion onset cues. But the differential effects of motion on two performance trials in the simulator do not transfer to performance in flight. In the prediction of operator performance in flight the magnitude of the error scores resulting with the use of one motion system as opposed to another is not as important as the stability of the subjects' performances from one day to the next. Increasing the fidelity of the simulator motion system may bring much of the variability of flight into the simulated environment which was used to escape the variability of the operational environment.

The recorded pilot performance measures correlated very highly with the observers' overall subjective ratings of the missions ($r = .726$ to .878). The observers' overall ratings correlated slightly higher with performance on instrument flight maneuvers than with performance on visual flight maneuvers. Other possible indicies of pilot proficiency, such as the amount of multi-engine land, instrument or total flight time logged in the past six months, did not correlate very well with mission performance scores, in fact they correlated about as well as age.

AVIATION RESEARCH LABORATORY

INSTITUTE OF AVIATION

UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

WILLARD AIRPORT

SAVOY, ILLINOIS

61874

TECHNICAL REPORT

ARL-74-5

AFOSR-74-3

APRIL 1974

EFFECTS OF GROUND-BASED AIRCRAFT SIMULATOR MOTION CONDITIONS UPON PREDICTION OF PILOT PROFICIENCY

JEFFERSON M. KOONCE

PREPARED FOR

LIFE SCIENCES PROGRAM
OFFICE OF SCIENTIFIC RESEARCH
AIR FORCE SYSTEMS COMMAND

F44620-70-C-0105

10

EFFECTS OF GROUND-BASED AIRCRAFT SIMULATOR MOTION CONDITIONS UPON PREDICTION OF PILOT PROFICIENCY

BY

JEFFERSON MICHAEL KOONCE

B.S., Tulane University, 1959
M.S., Tulane University, 1961

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Psychology
in the Graduate College of the
University of Illinois at Urbana-Champaign, 1974

Urbana, Illinois

ACKNOWLEDGMENTS

The research reported in this thesis was supported by the Life Sciences Program, Air Force Office of Scientific Research under Contract No. F44620-70-C-0105. Dr. Charles E. Hutchinson was scientific monitor of the contract.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

The assessment of proficiency of an individual or a crew in an operational system generally is a difficult task. Blum and Naylor (1970), Gagné (1962), and Fleishman (1967) have delineated and analyzed the problems. When an individual or crew produces some tangible object, proficiency frequently is measured in terms of the quality and quantity of the product. In situations where there are no immedicte products to be judged or counted, proficiency becomes a more elusive quality to measure. Examples of such cases are the determination of proficiency at riding a bicycle, operating a boat, and driving a car. In these instances praficiency may be measured in terms of the degree of accuracy with which the individual controls the vehicle, the number of accidents he has during a specified period of time, the ability to recall and follow specific procedures, the timeliness with which he executes required tasks (steps), and the ability to handle properly the vehicle under emergency conditions, real or simulated (Fleishman, 1967; Chapanis, 1959).

A more difficult area for the determination of an individual's proficiency is in the control of a vehicle that operates in three dimensions, such as an aircraft, a submarine, or a space vehicle. Because of the widespread use and the dangers involved in operating aircraft, a vast amount of research has been devoted to the field af aviation in general and to the training and assessment of pilots in particular. Major research in this area started just prior to World War II and has more recently expanded into the field of space. Ericksen (1952b) presents a comprehensive review of the major research in pilot proficiency measurement up through the end of 1951 and Smode, Gruber, and Ely (1962) review the literature through 1961.

In all areas af human performance there may be needs for determining the proficiency levels of individuals (Glaser and Klaus, 1962; Vreuls and Obermayer, 1971). An instructar may wish to know how well his student is doing, a training director wauld like to knaw if his program has weaknesses and how severe they are, a supervisor would like ta know how well his personnel are maintaining the necessary skills, and a unit commander needs ta know the readiness status of his men ta

determine the level of task to which his men can be committed or what portion of them could be committed to a task.

For all skills of operational significance there is a continual need to know the degree of proficiency of the operator, and this proficiency generally is determined through measures of his performance on the necessary tasks. The measurement of task performance is dependent upon the definition of performance and the limits of the task. The limits of a task are to a degree determined by the purpose to be served. For example, in driving, if one wishes to measure performance for a daytime driver's permit, the measurement might be limited from driving on a rural road to driving in a densely populated area in the daytime. For a broader scope, one might wish to expand the limits to include the control of the car under all possible weather and road conditions. To certify a driver for delivering a special car somewhere across the country, the limits of the operator's task might be expanded to include vehicle control under all weather and road conditions as well as to include sufficient knowledge of the vehicle, laws, regulations, and minor maintenance to insure a relatively high degree of success at bringing the vehicle across the country.

Generally, the determination of the limits of the task(s) is not too difficult. But decisions regarding what measures of performance should be used do not follow a specific procedure. Ideally, the measures should be derived from a sound theoretical position, but such measures simply do not exist (Smode, Gruber, and Ely, 1962). As a result, sometimes everything that is measurable by whatever means available is measured and attempts are then made to determine their usefulness or, alternatively, systematic inspection and expert judgment are used to select a certain few critical measures which are thought to reflect proficiency at the specified task. The later method is the one most frequently used because of cost and time factors.

Methods of Measuring Performance

After the desired measures of the tasks have been identified, just how they can be measured must be determined. Frequently a specific behavior cannot be directly observed or measured so a measure of some different behavior which reflects

the desired behavior is taken. When the behavior or its resultant performance parameter can be observed or directly measured, the problem of which measurement method to use leads to much controversy.

First, from one point of view it is held that reliable performance data can be obtained only when the data are recorded by automatic or semiautomatic devices, such as cameras, video tape, electronic tape, strip-chart recorders, or telemetry equipment relaying measures to specific receiving stations (Vreuls and Obermoyer, 1971). This method is quite useful in research situations where few operational sites or vehicles are used and where precise information pays its way. But such automatic and semiautomatic recording devices are relatively expensive and frequently cannot be applied economically to the routine assessment of proficiency in a shop or in a large number of vehicles.

The numerical size and cost of the project might warrant automatic and semiautomatic measurement of performance especially when the individuals desiring the information cannot be present to collect the data personally, i.e., space flight. Also, this method of obtaining data about an operator's proficiency requires careful calibration and maintenance of the recording equipment, may require a rather long time for reduction of data and interpretation of the results, and may require an acceptance of arbitrarily small number of measures depending upon the number of channels of information that can be recorded. Automatic and semiautomatic recorders of performance frequently sample only a small portion of the behaviors that might be desirable for determining proficiency. Danneskiold (1955) discusses some of the difficulties encountered in using mechanical means of recording flight performance.

A second method of recording information about the performance of a task(s) is to use a human observer. Typically, human observers have been utilized for obtaining subjective ratings about how an individual performs his task. Within the past 20 years strides have been taken toward bringing more objectivity to the performance measures made by observers. These include standardization of tasks, training of observers, and the use of checklists as guides to the observation and

recording of performance. The use of human observers brings along the potential problems of bias and halo effects, inaccuracy of observation, failure to observe continuously, limitation of the number of factors that can be monitored and recorded at one time, and the sizable differences in standards between observers. On the other hand, human observers are more flexible, may be aware of more factors than a mechanical device, give useful subjective ratings of performance unmeasurable by mechanical devices, give more immediate feedback to the operator, and may also be one of the team members necessary to perform the task. The observers must be carefully trained as to the meanings of the various levels on the subjective scales and in the proper methods and techniques of gathering and recording relatively objective measures.

Another source of information about an individual's proficiency in performing a job is the individual himself. Information from the individual relative to his ability to perform the job may be obtained in the form of confidence ratings before the performance of the task.

Just which method of performance measurement should be used will be a function of what is to be measured, the space allotted in the vehicle for that which will do the measuring, the extent to which cost is a factor, and the degree of reliability required.

Frequently, performance measures are taken on a specific occasion and the results might not be indicative of typical performance of the individual because of anxieties about being checked or some external factors. Thus, it might be recommended that proficiency measures be composed of measures of performance from several occasions to give an average performance measure. Once such recommendation is from Gagné (1965): "The only good way to achieve dependable measures of human performance is to use repeated observations." In simple tasks that are not time consuming taking multiple measures for an average performance can often be easily accomplished, but when the task is complex, lengthy, and in some cases costly to measure, mutliple measures may not be feasible and proficiency must be estimated from a single trial. If the samples of behaviors taken in the measurement

of performance are sufficient in number then the consistency of the individual's performance within the measurement occasion can be checked.

The difficulties in obtaining measures of performance increase as the tasks become more complex. The measurement of performance for a punch press operator is easier than that of a train engineer which is easier to measure than the driver of an automobile. As we go from vehicle control in one dimension, to two dimensions, to the three dimensions of submarine and airborne vehicles the measurement of performance becomes more difficult, costly, and occasionally dangerous.

Simulators

Simulation of the job from which performance is to be measured is sometimes desirable to present a safe, controlled environment and standardized conditions of measurement. For the purpose of this paper simulators are considered to be dynamic physical representations of the vehicle or equipment used in performing the task that is being simulated. In general, simulators of tasks will be referred to as task-type simulators, such as driving simulators, flight simulators, space simulators, and train simulators. These simulators are not designed to represent a specific vehicle or tool but a class of vehicles or tools, and they have some means of displaying information to the operator that reflects the inputs he makes into the system. The display of information may be in the form of instruments and lights within the vehicle, representations of the visual field outside the vehicle, and/or the ability to see the operation of the tool being manipulated.

On the other hand, there are vehicle-specific simulators such as a lunar landing module simulator, a B-52 simulator, or a particular type of nuclear submarine simulator. These simulators are characterized by a high degree of fidelity in the representation of the specific vehicles they are intended to simulate, particularly with respect to the configuration and dynamics of controls and displays.

Simulators are used for the training of operators, for research generally in the areas of displays and control system evaluation, and for the maintenance and measurement of proficiency. Adams (1961), Gagné (1954), and Glaser and Klaus (1962) discuss the differences in simulators used for instruction and those used for proficiency measurement. They make a distinction between a simulator that

represents a class of vehicles (a training simulator), and a simulator that represents a specific vehicle (a simulator for maintenance of proficiency). In their approach, proficiency measurement should be conducted in the simulator of the vehicle that the operator generally uses. But one could use just as well a more general simulator and speak of the individual's proficiency at driving or flying, etc. Using the general task simulators for measurement of performance would considerably lessen the cost of the simulator because fidelity and cost go hand-in-hand, and general simulators could be used by a much larger proportion of the population than the vehicle specific simulators.

Simulators are useful in situations where it is not possible or practical for an observer to go along with the operator to make records of his performance, e. g., experimental or research aircraft and space vehicles. Frequently, automatic or semiautomatic recording or telemetry devices are placed on the vehicle to send back information about the operator's performance, but these devices are expensive, require special care, and may be limited in the number of things that can be recorded. Where an observer and automatic equipment cannot be used because of space, cost and power limitations, external measures of performance are frequently used as measures of proficiency.

External measures of performance may be such things as number of accidents, number of traffic citations, accuracy in hitting a target, economic handling of the vehicle, and others. Each is related in some way to some aspect of performance but is of limited value in assessing overall proficiency at driving, flying, boating, etc. Even when an observer or automated recording device is used in the real vehicle there are numerous problems in standardizing the measurement conditions. Traffic, weather, vehicle characteristics, and mission profile factors vary from one performance measurement occasion to another, and the effects of these varying factors add to the difficulty of obtaining reliable measures of performance. Using the actual working environment and equipment for performance measurement generally introduces certain safety problems, especially if the task is naturally dangerous or if performance in emergency-type situations is to be measured.

Simulators seem to offer solutions to most of these problem areas and, in addition, allow the measurement of many aspects of operator behavior not available

in the actual work situation. But simulation also has certain drawbacks. The fact that the operator realizes that he is not performing the task in the actual vehicle may cause him to respond differently to the simulator than he does to the vehicle being simulated. Frequently this behavioral response difference is not recognized or reported by the operator. When it is, the rationale generally given is that the simulator does not "feel" like the real thing or the stresses of the real environment are not present. Because of this, many manufacturers and users of simulators feel that increasing the fidelity of the control system "feel," the use of audio, visual, and motion systems, and adding more naturally functioning switches and displays will increase the operator's acceprance of the simulator. One thing that is certain is that such additions increase the cost and intricacy of the equipment.

In the simulation of some simpler tasks such as driving an automobile, the simulators presently available do a poor job of simulating the basic tasks, have questionable training value, and frequently are more costly than the vehicle they are intended to simulate. The transfer effectiveness of these simulators and the validity of the performance measures taken are questionable. In most cases one might just as well use the actual vehicle or equipment and if necessary simulate only the situation. Some of these limitations found in open-loop automobile driving simulators are described by Baron and Williges (1971).

The value of complex simulators, such as those designed for flight, submarine, and space operation, is accepted in terms of the cost savings, safety, standardization of tasks, and degree of built-in fidelity. For several years commerical airline companies have been using simulators successfully in the transition training and refresher training of pilots (Meyer, et al., 1967; TWA, 1969; and American Airlines, 1969). Both the trainees' and the evaluators' verbal and written reports support the belief that simulators can be used to evaluate pilot performance and such evaluations are indicative of performance in the aircraft. American Airlines has published the general statement that the rating success of their pilots in their optimized flight training program using simulators was 98% without prior practice in the airplane (American Airlines, 1969). Unfortunately, specific data bearing on the predictive validity of simulator performance of airline pilots have not been made public.

The apparent success the airlines have had in utilizing flight simulators does not obviate the need for a systematic controlled investigation of the relationships between ground simulator pilot performance and aircraft pilot performance. On the contrary, it is important to determine the degree of predictive validity of ground based simulator performance, and to identify the variables of which the predictive validity is a function.

Motion for Simulators

Simulators vary considerably in fidelity, and their face validity increases with improvements in the fidelity of control-display relationships, motion systems, and visual systems. Several studies (Adams, 1961; Bergeron, 1970; Demaree, Norman, and Matheny, 1965; Federson, 1962; Mackie, 1972; Puig, 1970) have been performed on the degree of simulation and the fidelity of motion, and it appears that in many cases the degree of simulation and fidelity of motion, controls, displays, and visual and aural facets can be reduced considerably without adversely affecting the operator's performance on tracking tasks. The relationship between the degree and range of motion and the predictive validity of simulator performance measurement has not been empirically determined. It has been demonstrated that varying the fidelity of motion affects the operator's performance in the simulator (Rathert, Creer, and Sadoff, 1961) and it affects the relative merits of displays being evaluated (Matheny, Dougherty and Willis, 1963), but it has not been related to performance in the actual vehicle being simulated.

The simulation of vehicle motion might be useful in some situations and of no value in others. If the performance of maneuvers in the aircraft is accompanied by motion cues that the operator uses in his task or that make the task more difficult, then it might be valuable to simulate these motion cues in a simulator. It would add to the fidelity of the simulator and the pilot's acceptance. But if the task is such that the pilot should feel no motion cues or is taught not to pay attention to the motion cues that exist because they could cause some confusion (FAA AC 61-21), then to simulate those cues might degrade the performance in the simulator and would certainly add to the cost. An opposing view is that if the operator should

ignore the motion cues, then presenting them in the simulator would be more realistic and would aid the pilot in learning to ignore the motion cues. The absence of motion in the simulated vehicle would deprive the operator of cues that exist in the actual environment and it might affect his acceptance of the simulator and his performance. Intuitively, the increase in correspondence of the simulator to the real-world vehicle operation should increase the predictive validity of the measures of performance taken in the simulator.

Fraser (1966) held that simulator motion cues should be used in situations where they contribute to improved performance or interfere with satisfactory performance. Complete fidelity of simulated motion in all degrees of motion, acceleration and duration is impracticable or impossible. Therefore, if a requirement for motion cues is established, it then becomes necessary also to indicate the fidelity and the specific dimensions of motion. Motion bases for simulators range from the relatively inexpensive kinds used on general aviation trainers to very complex ones costing more than a million dollars, e.g., the NASA simulator for advanced aircraft.

It is believed generally that to move a simulator just because the vehicle being simulated moves is not enough, but that the simulated motion should be realistic. As an example (AFM 51-37), in an airplane entering a coordinated turn at a rate of three degrees per second (standard rate of turn), the pilot and occupants of the aircraft will initially detect the turn through the vestibular senses and after a few seconds they will sense that they are no longer in a bank but are upright. If the turn is entered very gradually, it may not even be detected by the passenger, but the pilot interpreting his instruments will know that he is in a turn. Now, if the simulator of an aircraft simulates a normal turn by tilting the cab, the pilot will receive an initial cue in the direction of the turn. If this position is maintained throughout the turn the pilot of the simulator would not have to interpret his instruments to determine his attitude because the inaccurate motion system would provide sufficient cues. Some simulated motion systems will remove the bank cues at a rate that is below the operator's threshold for the perception of angular acceleration and fail to give cues for banks that are entered below normal threshold, much as they occur in flight.

A summary of 25 research studies on man's perception of angular acceleration (Clark, 1967) revealed that the resulting thresholds vary from 0.035 to 8.2 degrees/second^2 with a median of 1.0 degree/second^2. It must be pointed out that most of the investigations were performed in laboratories, not in operational environments, the subjects' tasks were the detection of changes in position, and not the control of a complex vehicle, and of the 26 studies cited only five used more than five subjects. Discussing laboratory studies on acceleration thresholds, Rolfe (1968, p. 47) states that "a significantly higher rate of 'washout' can be employed in simulator motion systems where the attention of the operator is not directed solely to the perception of the motion." Unfortunately, because of the difficulty in defining and controlling all of the relevant variables affecting the detection of washout in operational complex task simulators, the limits of acceptable motion washout rates have not been determined systematically. The acceptable rate is most often empirically determined by test subjects performing the tasks for which the simulator is to be used. Just how the existence of and fidelity of simulator motion affects the predictive validity of performance measures taken in the simulator has not been determined.

Developing an Instrument for Measuring Performance

The performance measuring instrument to determine the predictive validity of complex simulators must be equally usable in both the simulators and the vehicles being simulated. Once the reliability and the predictive validity of the measures taken with the instrument have been determined, the instrument might be applicable in simulators of similar type vehicles. An example of this situation is where the instrument is validated in a multiple-seat vehicle and later used in simulators to determine the proficiency of the operators in single-seat vehicles.

Ericksen (1951), Flanagan and Gordon (1948), Glaser and Klaus (1962), Gordon (1948), and Smode, Gruber, and Ely (1962) make recommendations for development of a reliable instrument for measuring the performance of pilots. Some of the most important points are summarized in the following.

The overall task for which performance is to measured must be defined carefully and its limits clearly set. Frequently, it is useful to divide the overall

task into coherent subtasks or phases. However care must be taken to insure that in the process of defining subtasks, important overall aspects of the task are not lost.

If the developer of the measuring instrument is not intimately familiar with the task, he may need to discuss the task with operators and with individuals who currently evaluate operators' performance. He also should study the operating manuals, procedural guides, and applicable laws and regulations. Additional sources of information are accident reports and reports of critical incidents experienced by pilots. The design engineer is a valuable source of information about factors critical to successful operation of new systems.

Just how performance should be measured is an issue that brings much discussion. The various merits of machine, observer, and self-report techniques have already been discussed. Because of the cost, limited capability, and lack of easy widespread application, machine recording of performance is best suited for research needs. The use of observers and/or self-report techniques seems to have the most flexibility and capabilities in the field of performance measurement. Over the past 20 years many researchers have worked hard at increasing the objectivity of observers' measures of performance with the goal of giving observer measures high reliability. "Actually, methods of measurement align themselves at points along what we may call a subjectivity-objectivity continuum" (Smode, Gruber, and Ely, 1962, p. 96). Truly objective measures of performance by observers have not been attained, but as the performance measuring methods approach objectivity, they become more independent of the observer and there is more reliability in the measures taken (Miller, 1947).

Just how subjective a measuring instrument will be is determined in a large part by what it is designed to measure. If a measurement of a vehicle's speed at some point along a path is desired, an objective measure of it can be obtained by reading an instrument; but if one wishes to know how alert the operator of the vehicle is, subjective judgment from the observer may be necessary. To measure the performance of complex tasks, especially those that involve many procedures, purely objective measures may leave much of the task unassessed. What is needed is a

combination of objective and subjective measures that give a comprehensive representation of all of the factors that make up the task. Some fairly objective booklets used by observers to record performance of pilots contained criterion measures that were subjective in nature (Edgerton and Walker, 1945; Wilcoxon, Johnson, and Golan, 1952). The frequent use of such measures and the lack of trained observers resulted in many of the early attempts to measure flying performance objectively showing low observer-observer reliabilities. Increased reliability of the more subjective observations can be obtained through explicit anchoring of the scales and training of the observers (Greer, Smith, and Hatfield, 1962; Smith, Flexman, and Houston, 1952).

When using observers to record performance in complex tasks, the time and equipment required in the training of observers are important. One organization that used check pilots to record student performances on periodic check rides gave their observers extensive training in the use of their booklet, the Pilot Performance Description Record (PPDR): ten hours of ground classroom instruction plus discussion sessions, 20 hours of inflight practice with other observers acting as "student pilots," plus four months of actual use of the booklet on a preliminary training class before the experimental pilot training class began. The students' instructors who were to use the Daily Performance Record (DPR) received similar training except that they had only ten hours of inflight practice with fellow instructors serving as "student pilots" (Prophet and Jolley, 1969). In some cases, such practice of observers in the operational situation is too expensive or sometimes impossible. Simulators can be used for the training of observers where operational practice is impractical, and clearly defined scales and methods of recording performance can reduce training time.

The amount of work required to make the records of performance is a problem encountered when using a safety observer or crew member to record the operator's performance in a complex task. In aviation, for example (Ericksen, 1947; Wilcoxon, Johnson, and Golan, 1952), some objective scoring methods require the safety observer to read several instruments at 20- or 30-second intervals. Sometimes these booklets had as many as 20 criterion measures using up to six different types of scales on a single page. Because the safety pilot could not keep his attention directed to the

booklet and instrument panel, this frequently resulted in omitted data or a best guess on the part of the observer. Wilcoxon, et al., (1952) reported that 69 percent of the instructors using their booklets thought it was dangerous.

When measures of an operator's performance are taken by reading values from instruments in the vehicle, the time-sample, range, or limit method of scoring performance may be used. Ericksen (1947b) evaluated all three methods in the assessment of multiengine aircraft flying skills. He found that the time-sample method, taking instrument readings of the subject's performance at specific equally spaced periods during a maneuver, yielded higher observer-observer reliabilities than the other methods, but the ride-ride correlations were lower and this type of grading was difficult to accomplish in flight. Comparison of the other two methods for scoring the subject's total range of deviation on each measure of a maneuver showed that the range method was better than the limit method where the subject's maximum single deviation for each measure was recorded.

Because of variability in the operator's performance and the effects of outside factors, that which is measured in specific instances may not be entirely representative of the operator's level of proficiency. The representativeness of the operator's behavior is known to some degree to the operator. Garnder (1969), Rippey (1970), Shuford (1969), and Shuford and Gibson (1969) have proposed that additional information about the person's knowledge of a subject matter or ability to perform a task can be obtained through self-confidence ratings by the subject. Self-confidence ratings have been used extensively on written examinations using students, but very little work has been done on self-confidence ratings as sources of information about performance of psychomotor skills (Shuford and Gibson, 1969). Although the operator's performance at parking or landing, etc. may be poor on one occasion or good on another occasion, the operator has a fairly consistent level of confidence of his ability to perform the task based upon his prior experience on that task or similar tasks (Little, 1961; Jersild, 1929; Shuford and Gibson, 1969). Thus, the use of self-confidence ratings of ability to perform the various subtasks of the overall complex task might provide additional information about the operator's actual proficiency on the task. Self-confidence ratings on the subtasks are preferred to a single overall

confidence rating because the single roting might be unduly influenced by the operator s fear of foilure or very high confidence ossociated with one of the subtasks. When self-confidence ratings are used to obtain information obout an operator's level of proficiency, core must be taken not to make the situotion threatening for this might couse the operatar to give unreolistic ratings af his ability to perform the task (Klein and Schoenfield, 1941). Self-confidence ratings might be more useful in the routine assessment of proficiency than in periodic spot checks that might result in the revocation of the operator's license or suspension from his job.

PROBLEM

The research reported here was designed and performed to provide results that would have significance for several mojor problem areas that have been discussed in the Introduction. First, to what degree, if at all, can proficiency of aircroft pilot performance be predicted from measures of ground-based simulator pilot performance? Second, does the predictive validity of ground-based simulator pilot performance measures vary as a function of simulator motion conditions? Third, can o reliable pilot performance roting scale be developed that is useful and efficient for flight instructors and flight observers in on operational situation? Fourth, can a pilot performance rating scale be developed to produce results that correlote highiy with other indices of pilot performance? Fifth, is there a systematic, useful relationship between pilots' stated levels of confidence in their abilities and their measured performances?

The results of this research are also expected to provide information relevant to several subsidiary problems. For example, will predictive validity of pilot performance in a simulator vary as a function of contact (VFR) and instrument (IFR) flight conditions? What is the predictive validity of performance on specific simulated maneuvers? Is the predictive validity of performance on specific simulated maneuvers and on classes of flight (VFR and IFR) dependent upon the simulator motion condition? Whot is the effect upon reliability of pilot performance measures when the observer has additional duties such as Safety Pilot? What are the relationships between observer ratings of overall mission performances and observer ratings of individual maneuver performances? And, how well can pilot proficiency be predicted from various indices of flight experience and currency?

METHOD

Three groups of 30 pilots with multi-engine and instrument ratings performed a simulated flight mission in a Singer-Link General Aviation Trainer 2 (GAT-2) on each of two days. The experimental conditions for the groups differed in terms of GAT-2 motion (Group I - no motion; Group II - sustained linear, scaled-down analog motion; Group III - washout motion). Each group of pilots then flew the same mission in a light twin-engine aircraft, the Piper Aztec, which is representative of the class of aircraft simulated by the GAT-2. Thus the experimental design was a two-factor mixed design (groups by days) with repeated measures on one factor (groups). The repeated measures were on the 30 subjects per group over their days of participation. As the subjects were assigned to each of the three experimental groups, one-third of each group was selected to participate on a fourth day, giving them two aircraft missions. Thus, of the 30 subjects in each group that participated on Days 1, 2, and 3, only ten flew on Day 4.

The mission consisted of five maneuvers representative of those usually performed without visual reference to the outside world under instrument flight rules (IFR) and five maneuvers usually performed with outside visual contact under visual flight rules (VFR). In the simulator, all of the maneuvers were performed without outside visual references.

Two trained observers, one of whom was also the safety pilot for the mission, recorded pilot performance in a specially designed booklet. Observers were assigned to the safety observer (SO) and flight observer (FO) roles so that the FO on the first mission for a pilot was the SO on that pilot's second mission. On the second mission a new observer served as FO. Thus, the same observer was assigned to a given pilot only twice. This procedure permitted recording of a pilot's performance on the same mission by two independent observers and also recording of his performances on two successive missions by the same observer and by two independent observers.

Subjects

The subjects were recruited by means of a brief letter and questionnaire (Appendix A-1) sent to all multi-engine land (MEL) and instrument (I) rated pilots

TABLE 1. Order of Assignment of Subjects to Experimental Conditions

Subject	Experience Category					
	L/O	L/I	M/O	M/I	H/O	H/I
1	I +	II +	III+	I +	II +	III+
2	II +	III+	I +	II +	III+	I +
3	III+	I +	II +	III+	I +	II +
4	I	II	III	I	II	III
5	II	III	I	II	III	I
6	III	I	II	III	I	II
7	I	II	III	I	II	III
8	II	III	I	II	III	I
9	III	I	II	III	I	II
10	I +	II +	III+	I +	II +	III+
11	II +	III+	I +	II +	III+	I +
12	III+	I +	II +	III+	I +	II +
13	I	II	III	I	II	III
14	II	III	I	II	III	I
15	III	I	II	III	I	II
16	I	II	III	I	II	III
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
n						

I - No Motion II - Sustained Motion III - Washout Motion + - Two Aircraft Missions

TABLE 2. Distribution of All Subjects Among the Experimental Groups

	Multi-Engine Flight Time				Instrument Time (last 6 mos.)	
Group	L	M	H	Total	None	Some
I	14	4	12	30	6	24
II	13	6	11	30	5	25
III	13	6	11	30	7	23
Total	40	16	34	90	18	72
(%)	(44.4)	(17.8)	(37.8)		(20)	(80)

TABLE 3. Distribution of Subjects that Flew Two Aircraft Missions

	Multi-Engine Flight Time				Instrument Time (last 6 mos.)	
Group	L	M	H	Total	None	Some
I	6	1	3	10	2	8
II	2	3	5	10	3	7
III	4	2	4	10	3	7
Total	12	6	12	30	8	22
(%)	(40)	(20)	(40)		(26.7)	(73.3)

registered with the Illinois Department of Aeronautics who live in Champaign County and ten neighboring counties. Those who were interested in participating were asked to return a questionnaire (Appendix A-2) in a stamped pre-addressed envelope, both enclosed with the first letter. Two hundred and forty-seven letters were mailed and 133 questionnaires returned. Of the questionnaires received, 29 persons were not able to participate for some reason, and the remaining 114 potential subjects were sent a second letter (Appendix A-3) giving more detail of the mission to be flown. Appendix E-1 gives a breakdown of subject attrition.

The subjects were classified both in terms of their total multi-engine flying time (Low, < 100 hrs, Medium, 100 to 500 hrs, and High, > 500 hrs) and in terms of their instrument flying time logged in the past six months (0, none, 1 some). The subjects' questionnaires were sequentially numbered in the order in which they were received, and they were identified and placed in one of the six experience categories L/0, L/1, M/0, M/1, H/0, H/1). Table 1 shows how the subjects assigned to each category were assigned to the three experimental groups (I, II, and III) with one-third of the subjects having two aircraft missions (+).

If a subject assigned to one of the experimental conditions had to drop out, the next subject acquired in his experience category was used to replace him. The sequences within each experience category for assigning subjects to experimental groups were broken toward the end of the study to insure a balance of subjects between experimental groups based on both total multi-engine flight time and instrument flight time in the past six months. The resulting distributions of subjects per experimental group are given in Tables 2 and 3.

Observers

Ten multi-engine and instrument rated, certified flight instructors (CFI), plus one U. S. Air Force pilot with extensive multi-engine experience as a military flight instructor and flight examiner, served as observers. During the experiment three observers had to withdraw from participation, and they were replaced by three other CFIs with similar qualifications.

Of the 11 observers, three were paid hourly wages for their porticipation, seven were employed by the Aviation Reseorch Laboratory, University of Illinois, and the experimenter was serving on active duty with the U. S. Air Force.

Records were kept of the observers utilization so that their assignment to missions would be balanced, with opproximately equal numbers of sofety pilot and flight observer missions and the proper proportion of simulator missions to aircraft missions, 60% simulation to 40% aircraft (Appendix E-2).

Training of Observers

Each observer was given a copy of the Pilot Performance Record to study ond written descriptions of the FO and SO's duties (Appendix C-1, 2, and 3). Then the experimenter carefully briefed the observer on the mission to be flown in the study and the available alternatives to the mission profile. They discussed each type of measure in the record booklet, the methods of recording performance, and of each individual item, its meaning, interpretations, and just when in the mission it should be recorded. After careful review of the mission, the booklet, and the procedures to be followed, the observer flew at least one mission as pilot in the GAT-2 to get a better feel for the subject's task and the nature of the data acquisition task. The observer was permitted to fly on actual data collection missions as a third observer to practice recording data in the booklet and afterwards discuss areas of difficulty with the experimenter. There were no set criterio for the troining of the observers. When an observer felt that he was familior enough with the mission ond the performance items that he should have no problems in recording the desired information, he was put on the schedule as one of the observers. Most of the observers served as subjects before being trained for data collection. Thus, they were instructed in the proper methods of briefing and debriefing the subjects and the conduct of the mission.

Mission Profile

The missions flown in the simulator and in the aircraft consisted of the following ten maneuvers:

VFR		IFR	
Takeoff and Climbout	(T/O)	Cruise on a VOR airway	(CRU)
360-Degree Steep Turn	(360)	Holding at a VOR station	(HOL)
Chandelle	(CHN)	ADF Approach	(N-P)
Lazy Eight	(LZY)	ILS Approach	(PRE)
Landing	(LNG)	Missed Approach	(MIS)

These maneuvers are representative of those required by the Federal Aviation Administration for obtaining a multi-engine, commercial, or instrument pilot's certificate (Federal Aviation Regulation, Part 61, Subpart D, Section 61117). To save flight time the VOR approach was not included, because the ability to interpret VOR signals can be demonstrated during the cruise and holding phases, and because the procedures in the VOR instrument approach are very similar to those of the ADF approach.

The ADF approach is categorized as a non-precision instrument approach (N-P) along with VOR approaches. Both provide lateral guidance to and from stations on the ground with the range from the station being determined by distance measuring equipment (DME), timing, or intersections with bearings from other radio aids. The ILS approach has the additional feature of providing vertical guidance (glide slope) to a ground station and more precise course guidance, and it is therefore considered a precision approach (PRE). Instrument rated pilots should be capable of flying both types of approaches (PRE and N-P) to transition safely from the enroute flight structure to a position from which a visual landing can be made.

Despite the fact that most instrument approaches are terminated with the airfield in sight, lowering weather conditions or improperly executed approaches may result in the necessity of performing the published missed approach procedure, and instrument pilots should always be prepared to fly the missed approach.

Although the chandelle and lazy eight are maneuvers seldom performed in aircraft, except for training and flight checks, they were included because they require the pilot to demonstrate timing, coordination and planning while controlling the aircraft with pitch, bank, and airspeed continuously changing. Other patterns

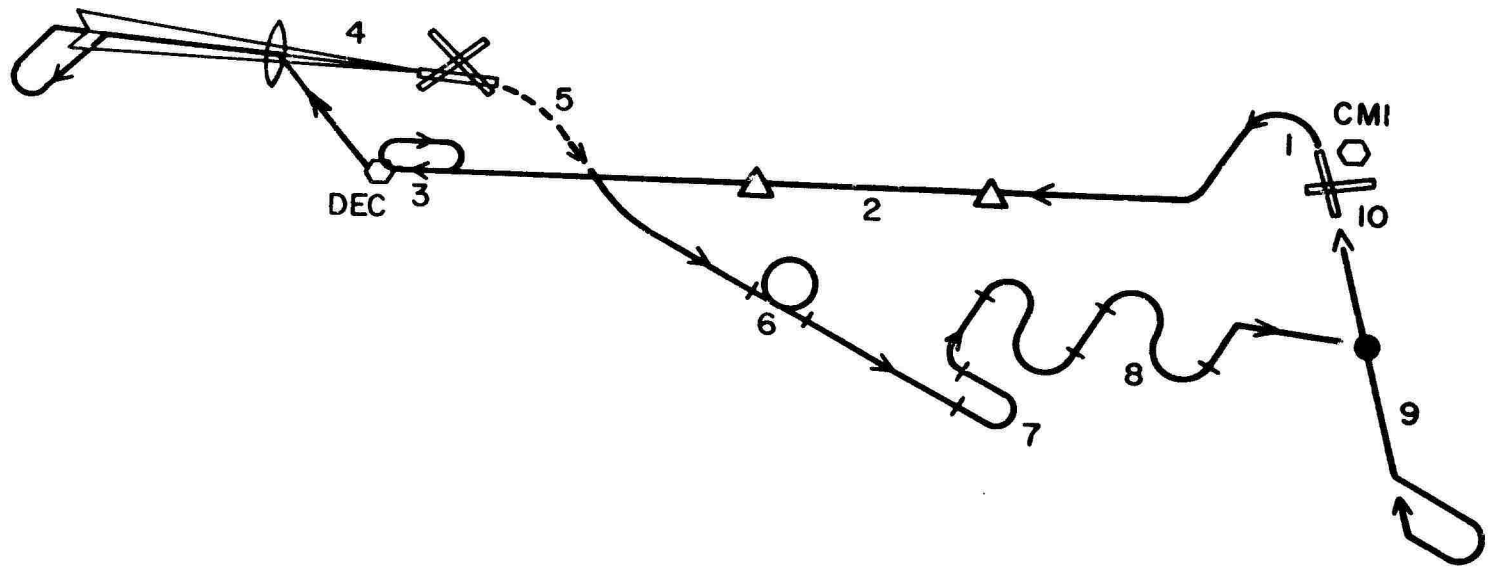
were not included because the pilot population sampled would not be as familiar with them as with the chandelle and lazy eight.

The maneuvers were put together to form the flight profile presented in Figure 1. This profile was flown in all of the simulator missions and in most of the aircraft missions: takeoff and climbout, cruise, holding, precision approach (ILS), missed approach, 360-degree steep turn, chandelle, lazy eight, non-precision approach, and landing. When the ILS approach at Decatur could not be made, because of strong tailwinds or failure of the ground station, the alternate flight profile (Figure 2) was flown wherein a VOR approach was made at Decatur and an ILS approach was made at Champaign. The IFR, VFR, and general flight rules as stated in the Federal Aviation Regulations (FARs), Part 91, Section B were applicable to the missions flown.

In the simulator, the visual maneuvers were flown in a different manner than in the aircraft. Directional control during the takeoff ground roll in the aircraft is primarily by reference to the outside world, mainly the runway centerline; while in the simulator, directional control was maintained by reference to the heading indication, a "gyro-stabilized" compass. In the aircraft, pilots were instructed to fly the 360-degree steep turn, chandelle, and lazy eight with primary reference outside the aircraft for attitude and directional control. These maneuvers had to be performed by reference to instruments in the simulator.

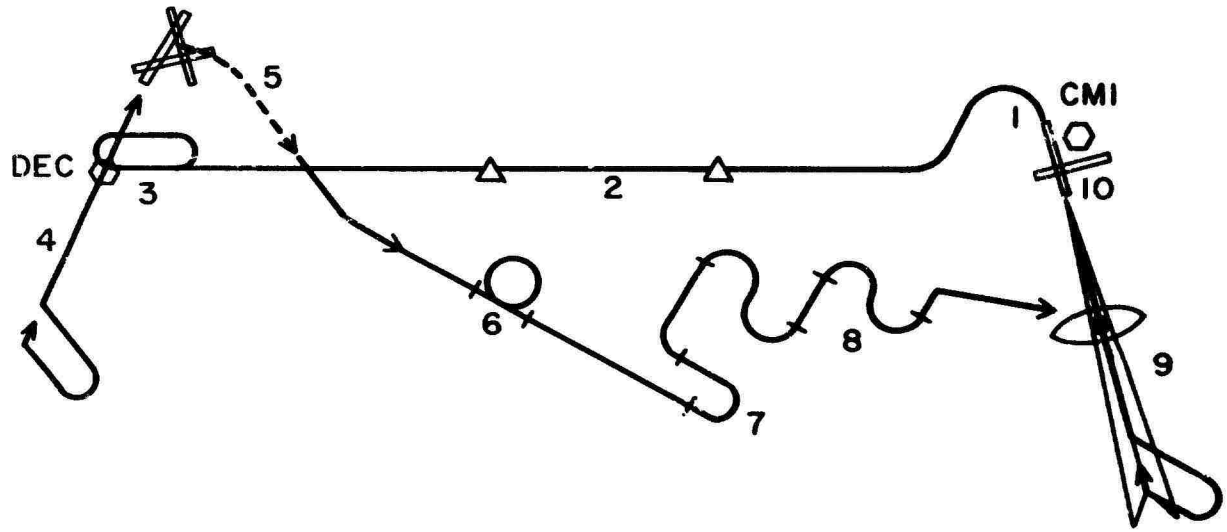
The landing in the simulator was performed as a straight-in approach from the missed approach point (MAP) of the ADF approach. From the MAP, the subject pilot was to maintain final approach airspeed and set landing configuration while maintaining a shallow descent. When 50 feet above the ground the pilot was to be at the threshold airspeed and begin his flare for landing. Throughout this straight-in approach, directional control was maintained by reference to the gyro compass. The performance recorded for landing in the aircraft included flight parameters and procedures in the traffice pattern as well as final approach and touchdown.

Generally the simulator missions required 1.2 to 1.5 hours to complete while the aircraft mission required 1.4 to 1.7 hours.



1. Take Off
2. Cruise
3. Holding
4. Precision ILS Approach
5. Missed Approach
6. 360° Steep Turn
7. Chandelle
8. Lazy Eight
9. Non-Precision ADF Approach
10. Landing

Figure 1 . Normal mission profile.



1. Take Off
2. Cruise
3. Holding
4. Non-Precision VOR Approach
5. Missed Approach
6. 360° Steep Turn
7. Chandelle
8. Lazy Eight
9. Precision ILS Approach
10. Landing

Figure 2 . Alternate mission profile.

Apparatus

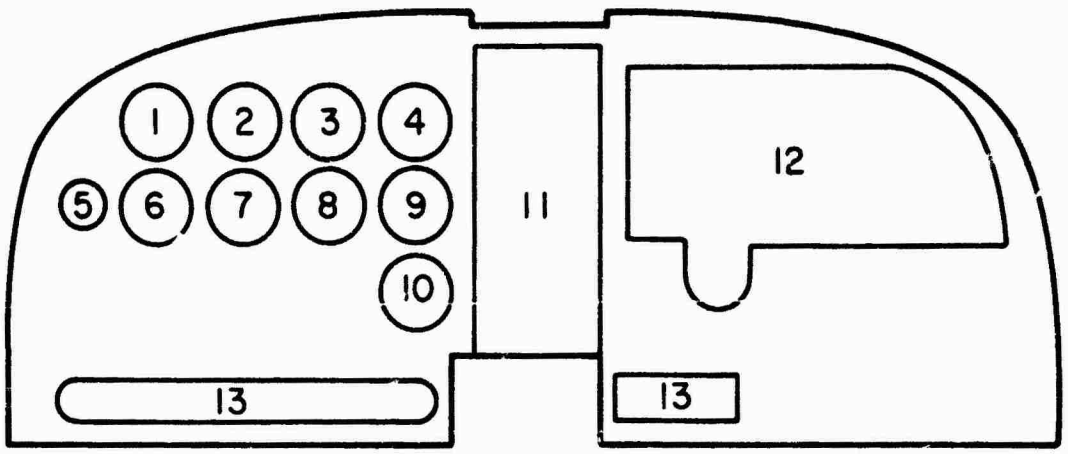
Simulator. The instruments for the display of flight and navigational information in the GAT-2 are similar to those in the aircraft that was used but the location of some of the instruments and some controls differed from simulator to aircraft (Figure 3). The windows in the GAT-2 were covered with frosted panels so that the subject had no visual horizon cues from outside the GAT-2.

The GAT-2 is equipped with a two-degree-of-freedom motion system to provide limited kinesthetic and vestibular sensations to the occupants. The motion system has been modified so that it may operate in either one of two modes, "sustained" or "washout." The sustained motion consists of a linear, scaled-down analog motion in pitch and roll (see Table 4 for limits). When operating with this mode, as the pilot enters a banked turn, the simulator cab is angularly displaced in the direction of turn and maintains that tilt until such time as the pilot brings the "aircraft" back to wings-level flight.

The washout modification of the system provides the same roll and pitch acceleration cues, but while steady-state flight attitudes are maintained, the simulator cab is returned to the neutral position at a rate that is below the pilot's vestibular and kinesthetic thresholds for acceleration. The rate at which the simulator cab is returned to the neutral position was determined by having several local instructor pilots fly the simulator with different rates of "washout," and the fastest rate which none of them reported as being noticeable was the rate that was used in this study.

The aileron and elevator control feel system of the GAT-2 was modified on the basis of data from spring tension studies in the Aztec aircraft throughout a wide range of airspeeds, altitudes, configurations, and flight attitudes. Because of the absence of sustained G forces in the GAT-2, some of the control system pressures had to be increased slightly to simulate the "feel" that the pilot would have in the aircraft.

There are three seats in the cabin of the GAT-2. The front seats, left and right, were for the subject and the safety observer, and the rear seat was for the flight observer (Figure 4).



1. Airspeed Indicator
2. Gyro Horizon
3. Altimeter
4. ADF Indicator
5. Clock
6. Turn and Slip Indicator
7. Directional Gyro Compass
8. Vertical Velocity
9. Omni Bearing Selector with Glide Slope
10. Omni Bearing Selector
11. Radio Controls and Tuning Heads
12. Engine Instruments
13. Systems Switches and Controls
14. Experimental Displays and Controls

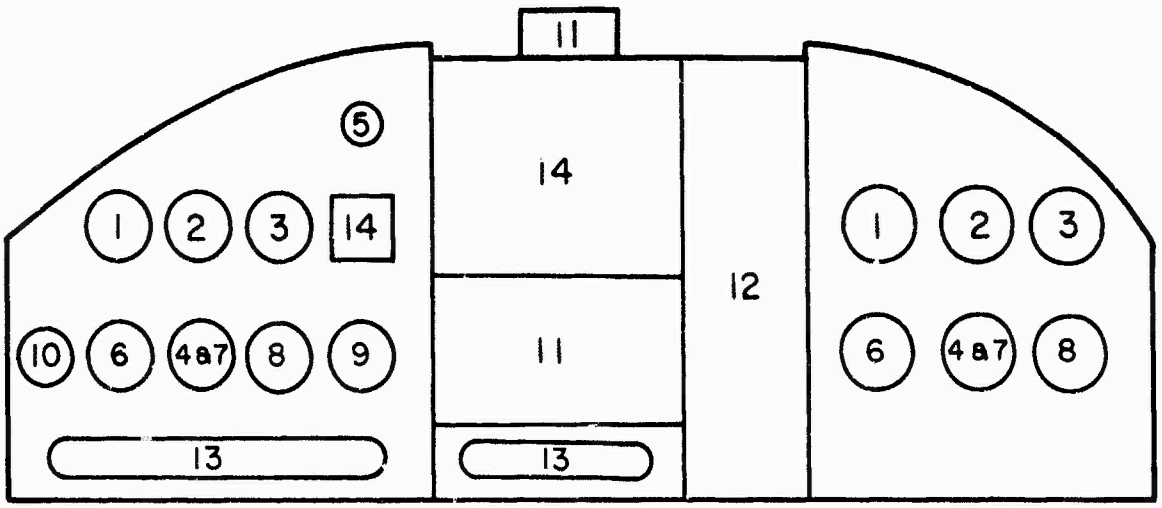
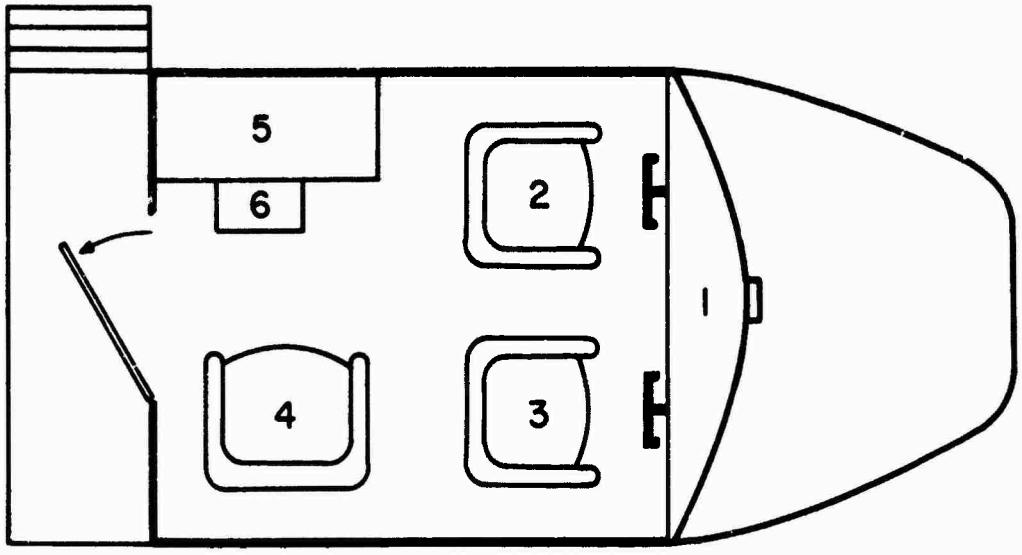


Figure 3. Aztec instrument panel (top) and GAT-2 instrument panel (bottom).



1. Instrument Panel
2. Pilot's Seat
3. Safety Observer's Seat
4. Flight Observer's Seat
5. Plotting Board
6. Systems Control Panel

Figure 4. Interior arrangement of GAT-2 simulator.

TABLE 4. GAT-2 Limits of Motion

Dimensions	Position	Velocity	Acceleration
Pitch	$+13^{\circ}$ to -8°	$\pm 13^{\circ}$/sec	$\pm 100^{\circ}/sec^2$
Roll	$\pm 13^{\circ}$	$\pm 25^{\circ}$/sec	$\pm 300^{\circ}/sec^2$
Vertical Translation at the Pilot's Position	+7.16 to -4.36 in.	±7.16 in/sec	±0.6g

Aircraft. The aircraft used in this study was a Piper Aztec-D, a light twin-engine propeller-driven aircraft that can carry six persons. The instrument panel of the Aztec displayed the same information as the GAT-2 but the location and type of presentation differed somewhat (Figure 3). There were also differences in the location of switches and controls.

Both the Aztec and the GAT-2 were equipped with two VOR receivers, one of which included glide slope for ILS, an ADF receiver, a marker beacon receiver, and flight instruments and communication radios required for operation under instrument flight rules (IFR), Federal Aviation Regulation, Part 91, Subpart A, Section 91.33, b, c, and d. The DME (distance measuring equipment) in the simulator was covered with masking tape, and in the aircraft it was not turned on to prevent the subject pilot from using its information as an additional aid during the instrument flight phase.

To simulate instrument flight conditions, the subject wore a translucent hood which tended to restrict his visual field to the instrument panel.

The aircraft was serviced before the flights so that its weight and flight characteristics would be about the same for each subject.

Both the simulator and the aircraft had placards (Appendix D-1) on the instrument panel which gave the appropriate power settings and airspeeds for the various phases of flight.

Chart and approach plates. The instrument flight portion of the mission required an enroute chart for navigation from one airfield to another and approach plates, one for each type of instrument approach to be made. The low altitude enroute

chart, L-23 published by the National Ocean Survey, was used for the cruise segment from Champoign to Decatur, and approach plates by National Ocean Survey and by the Jeppesen Company were available for the instrument approaches (Appendix D-2 to D-9). Sufficient numbers of the charts and plates were available for the subject and the observers to have the type that each preferred available for each mission. Subjects who brought their own flight kits were permitted to use their own charts and approach plates.

Equipment. Each subject was offered a standard 9 x 12.5-inch clipboard with a pad of paper for holding the charts and plates and for copying clearances.

The translucent hood, used for simulating instrument flying conditions in the aircraft, was adjustable for head fit, and the width of the field of view was adjustable. Most subjects chose to open the hood a bit to prevent disorientation from the tunnel vision effect of the very narrow field of view with the sides closed in.

Checklists. The checklist booklet (Appendix D-10) was developed from information contained in the Piper Aztec-D owner's manual and the Institute of Aviation's Aztec Checklist. Individual checklists were developed for each of the following phases of flight: engine start, runup, before takeoff, after takeoff, cruise, descent, before landing, and after landing. These checklists and a list of appropriate power settings and airspeeds were typed on individual 4 inch x 6 inch cards and held together by a plastic ring binding to form a checklist booklet.

The checklist booklets were appropriate for both the GAT-2 and the Aztec. Each page was step indexed, making it easier to identify and select a desired checklist.

A sufficient number of checklists was made so that each observer could have a personal copy, and there were enough remaining for the subjects to use during the missions.

Pilot Performance Record

The Pilot Performance Record (Appendix B), independently developed for use in this study, was similar in principle to those used by Smith, Flexman, and Houston (1952) and Greer, Smith, and Hatfield (1962). Each of the maneuvers described in the section, Mission Profile, was carefully examined, broken down into phases, and

the factors that could be used to identify the properly performed maneuvers were listed. To reduce the total number of factors per maneuver, only those that could be clearly defined, readily observed, and fairly objectively recorded were retained. Those which were reflected in other potential measures and did not add any significant information to the description of the maneuver were deleted. The goal was to compile a comprehensive group of indices of desired performance describing each maneuver while avoiding factors that were ambiguous, redundant, or practically unmeasurable by a safety pilot.

Each maneuver was subjectively checked to see if the factors assigned to each maneuver adequately described the maneuver. These checks were made by several flight instructors until there was agreement that each maneuver was adequately described also was measurable by an observer.

Two general types of factors were identified: flight performance factors of airspeed, altitude, heading, course deviation, etc., and procedural factors such as reporting a fix, accomplishing a checklist, and lowering the gear at the proper time. Most of the performance factors referred to measures to be taken from cockpit instrument observations and were called scaled items because of the scales on which they would be recorded. Typically, the procedural factors could be defined in terms of responses such as yes-no, proper-improper, and o.k.-late, and they were called categorical items. These items were presented as two or three labeled blocks in a row; the one representing that which was observed was to be marked.

The scaled items involved the recording of two types of information. For some performance factors it was desirable to know the deviation from desired value at a given point in time or at a specific place in a pattern. This was recorded as a slash or arrow on a scale indicating its deviation from the desired value. Other scaled items required recording the range of performances over a period of time or between specified points in a pattern. Each of these was recorded by a horizontal line extending along the scale over the range of deviations from desired or as vertical slashes on the scale representing the amount of deviation from desired, each time the observer noted that flight performance parameter as a greater deviation than before.

Because the booklets were designed for use in many different types of aircraft, the scales were made for recording the deviation from the desired value on each variable rather than the absolute readings of the instruments. This required the observers to know what the desired value for each item should be.

The items, for each maneuver, were arranged on pages with a brief verbal description of the item on the left and the scale or categorical blocks to the right. For those maneuvers more difficult to record, the items were numbered in sequence. On the page facing the maneuver a sketch of the maneuver was presented with numbers corresponding to the items to be recorded, so that the observer could easily identify the point or phase for each item.

The maneuvers were arranged in a sequence, presenting the instrument flight portion first followed by VFR maneuvers and landing. For ease of selection of the page for the maneuver to be recorded, the pages were stepped on the bottom, and the names of the maneuvers written on the steps. The pages were bound at the top so that the observer could easily flip to the desired page.

Identifications of the subject, the observer, date, weather, and aircraft information were written on the first page. The back of the first page explained the subjective rating system, and a subjective rating scale was placed before each maneuver.

Following the last maneuver, landing, is a page for each observer to indicate his overall subjective rating of the flight just completed. The last page in the booklet provides spaces for recording information about the subject's total flight experience and his flight experience in the six months prior to participating in the study.

The booklet size, nine inches high by 6 1/2 inches wide, maintained convenient size and still was large enough to provide usable scales.

Procedures

Prior the data collection, the experimenter made personal visits to the tower chiefs at Champaign and Decatur, the airfields to which approaches would be made in this study. The purpose of the visits was to explain the study to be conducted and the importance of timely control of our aircraft to minimize the chances of the final phase

of the approaches being broken off because of other traffic. Letters (Appendix A-4 and A-5) were sent to the tower chiefs confirming what was discussed personally.

The flight schedule for the experiment offered five 2 1/2 hour block times per day, the first beginning at 0800 hours and the last ending at 2030 hours. The subjects were telephoned to arrange for their participation, and whenever possible all of the subject's missions were scheduled during that initial telephone contact. Whenever a subject or poor weather caused a mission to be cancelled, that mission and all of the remaining missions for that subject were rescheduled following the order of observer utilization established in the experimental design.

Each subject was met by the experimenter and briefed by the flight observer while the safety observer prepared the GAT-2 simulator or preflighted the Aztec aircraft. After the briefing and recording of the subject's confidence estimates about performing the maneuvers, the flight observer would escort the subject to the vehicle, introduce him to the safety observer, and they would proceed with the mission.

After each mission the flight observer collected the materials used and accompanied the subject to the briefing room for debriefing and confirmation of the subject's next scheduled mission. At the end of the subject's last mission, the flight observer obtained the data about the subject's flight experience and aircraft preference asked for on the last page of the booklet.

The observers turned in their Pilot Performance Records on the subject to the experimenter who checked them for legibility of entries and discussed any problems encountered. Frequently the experimenter went on the missions as a "fourth" to observe the briefings and conduct of the missions.

RESULTS

The results are presented in three phases: first, the reliability of the measures of performance; second, the effects of the experimental conditions upon performance levels in the simulator and aircraft; and third, the prediction of proficiency in piloting the aircraft using the relatively objective measures of performance, subjective confidence ratings, and other data about the subjects.

Summary data on the subjects are presented in Table 5. The results of analyses of variance indicated no significant differences among groups with regard to age or the various measures of flight time and flight currency.

Reliability of Performance Measures

The total number of measures that were taken for the ten maneuvers in a mission were 50 scale items, 79 categorical items, 10 confidence estimates by each subject, and 11 ratings of performance by each observer. For analysis purposes, the categorical items in the Pilot Performance Record were recorded as zero if correct and one if in error. The scale items were recorded as the maximum deviation and direction from desired if the item were observed over a period of time or as the amount and direction of the deviation from desired if observed only at one particular point or time in a maneuver. If an item were not noted by an observer, it was left blank. Most of the data analysis was performed on an IBM 360 computer using pre-programmed statistical procedures under the title SOUPAC.

The agreement between the safety observer and flight observer ratings of the subjects' performances on the various items were expressed in terms of correlation coefficients. Pearson product-moment correlations were used for the scale variables, and phi coefficients were used for the dichotomous variables. Table E-3.1 in the appendix gives the reliability coefficients for each recorded scale item for each group and day, except for the fourth day on which ten subjects from each of the three groups were systemtically selected. Most of the reliability coefficients were based upon thirty pairs of observations, but some had as few as 26 pairs because of mission data. The recording of the observers' ratings far each maneuver was begun after some

TABLE 5. Flight Experience and Age of Subjects

	Mean	Median	Range
Total			
Flight Time	2,658.3	1,675	240 - 12,596 hours
Multi-engine Time	1,144.7	250	10 - 10,000 hours
Instrument Time	321.5	150	16 - 2,200 hours
Simulator Time	67.4	29	0 - 600 hours
Past 6 Months			
Flight Time	120.4	60	0 - 999 hours
Multi-engine Time	28.6	10	0 - 250 hours
Instrument Time	11.7	8	0 - 78 hours
Age	37	36	22 - 59 hours

missions had already been flown; thus the correlations between observer ratings per maneuver and performance scores for the maneuvers were based upon 19, 20, and 18 pairs of observations for groups I, II, and III.

The calculation of correlations between observers for the categorical items and coordination sometimes resulted in no correlation coefficient at all because the standard deviation of one or both observers was zero, a result of no errors being recorded. The categorical items also had a high number of perfrect observer-observer reliability coefficients resulting from very few errors occurring on a particular item and both observers recording them. An alternative method of estimating the reliability between observers on the categorical items and coordination is the calculation of the percent agreement between the observers on each item (Appendix E-3.2). Most of the measures of agreement on these items were very high except for the judgment of pitch and coordination at the 90-degree points throughout the lazy eight maneuvers. Agreement was somewhat higher on coordination at these points because the vehicle used was fairly stable, and there were few noted deviations from coordinated flight. Pitch reference when passing through the ninety-degree points was difficult to judge because it was continuously changing and there was no precise reference pointer to read or refer to.

Table 6 presents the frequency and cumulative percent distributions of observer-observer reliability coefficients by type of item, vehicle where recorded, and type of measure (maximum deviation, check point, observer rating). In the calculation of the percent of reliability coefficients above a given level, zero correlations resulting from no deviations being recorded were omitted because they are a function of whether or not the subject erred, not the reliability of the observers' records. Inspection of Table 6 reveals that for the individual items recorded on each mission there is a tendency for the observer-observer reliability coefficients to be greater in the simulator than in the aircraft and on measures of maximum deviation over a period of time than on check-point measures.

Applying the typical formula for standard scores to the deviation scores might result in a distorted interpretation of the results. For example, a person who makes an error gets a standard score of zero while a subject who performs perfectly on the same

TABLE 6. Frequency and Cumulative Percent Distributions of Observer-Observer Reliability Coefficients for Each of Eleven Measures

	Maximum Deviation Scores															
	Airspeed				Altitude				Heading				Bank (Deg)			
	Simulator		Aircraft		Simulator		Aircraft		Simulator		Aircraft		Simulator		Aircraft	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %
.95-.99	18	30.0			16	38.10	9	32.14	12	25.00	7	21.88				
.90-.94	18	60.0	10	25.0	17	78.57	14	82.14	18	62.50	6	40.63				
.85-.89	11	78.33	7	42.5	6	92.86	2	89.29	9	81.25	6	59.38	4	33.33		
.80-.84	8	91.67	7	60.0	0		0		5	91.67	6	78.13	2	50.00		
.75-.79	1	93.33	9	82.5	2	97.62	1	92.86	1	93.75	0		2	66.67	1	12.50
.70-.74	0		3	90.0	1	100.00	1	96.43	0		2	84.38	0		2	37.50
.65-.69	1	95.00	3	97.5			0		0		2	90.63	0		0	
.60-.64	2	98.33	0				0		1	95.83	2	96.88	0		1	50.00
.55-.59	0		0				1	100.00	0		0		0		1	62.50
.50-.54	1	100.00	0						0		0		0		0	
.45-.49			1	100.0					0		1	100.00	1	75.0	0	
.40-.44									0				0		1	75.00
.35-.39									0				1	83.33	0	
.30-.34									1	97.92			1	91.67	0	
.25-.29									1	100.00			0		2	100.00
.20-.24													0			
.15-.19													0			
.10-.14													1	100.00		
.05-.09																
<.04																
N=	60		40		42		28		48		32		12		8	

TABLE 6 (continued)

Check Scores												
	Airspeed				Altitude				Heading			
	Simulator		Aircraft		Simulator		Aircraft		Simulator		Aircraft	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %
.95-.99	1	1.67			7	12.96	4	11.11	10	23.81	1	3.57
.90-.94	10	18.33	2	5.00	9	29.63	9	36.11	4	33.33	0	
.85-.89	12	38.33	1	7.50	11	50.00	5	50.00	3	40.48	1	7.14
.80-.84	9	53.33	7	25.00	8	64.87	7	69.44	4	50.00	3	17.86
.75-.79	7	65.00	9	47.50	5	74.07	3	77.77	3	57.14	3	28.57
.70-.74	3	70.00	2	52.50	5	83.33	2	83.33	2	61.90	1	32.14
.65-.69	6	80.00	5	65.00	1	85.19	0		2	66.67	1	35.71
.60-.64	5	88.33	4	75.00	2	88.88	1	86.11	0		1	39.29
.55-.59	0		4	85.00	0		3	94.44	3	73.81	3	50.00
.50-.54	2	91.66	0		2	92.59	0		1	76.19	4	64.29
.45-.49	2	95.00	2	90.00	4	100.00	1	97.22	2	80.95	1	67.86
.40-.44	0		2	95.00			1	100.00	1	83.33	0	
.35-.39	1	96.67	0						1	85.71	0	
.30-.34	1	98.33	1	97.50					2	90.48	1	71.43
.25-.29	1	100.00	0						1	92.86	0	
.20-.24			0						1	95.24	0	
.15-.19			0						0		3	82.14
.10-.14			1	100.00					0		0	
.05-.09									0		0	
<.04									2	100.00	5	100.00
N=	60		40		54		36		42		28	

TABLE 6 (continued)

	Check Score				Maximum Deviation Scores											
	Time				Coordination				Rate of Turn				Observer Ratings on Maneuvers			
	Simulator		Aircraft		Simulator		Aircraft		Simulator		Aircraft		Simulator		Aircraft	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %
1.00					2	4.76	1	3.57								
.95-.99	5	27.78	2	16.67	0		0								1	2.50
.90-.94	4	50.00	2	33.33	1	7.14	0						3	5.00	0	
.85-.89	3	66.67	3	58.33	0		0						6	15.00	3	10.00
.80-.84	2	77.78	0		0		3	14.28	2	11.11			14	38.33	1	12.50
.75-.79	1	83.33	1	66.67	0		0		2	22.22			8	51.67	6	27.50
.70-.74	1	88.89	0		0		0		1	27.78	3	25.00	6	61.67	2	32.50
.65-.69	1	94.44	0		3	14.29	0		3	44.44	0		10	78.33	9	55.00
.60-.64	1	100.00	2	83.33	2	19.05	0		1	50.00	3	50.00	6	83.33	3	62.50
.55-.59			0		0		2	21.43	2	61.11	0		1	90.00	6	77.50
.50-.54			0		2	23.81	1	25.00	3	77.78	1	58.33	2	93.33	1	80.00
.45-.49			0		2	28.57	0		1	83.33	1	66.67	1	95.00	2	85.00
.40-.44			0		2	33.33	0		2	94.44	0		0		5	97.50
.35-.39			1	91.67	4	42.86	1	28.57	0		2	83.33	0		1	100.00
.30-.34			0		0		1	32.14	1	100.00	2	100.00	2	98.33		
.25-.29			0		0		1	35.71					0			
.20-.24			0		1	45.24	0						1	100.00		
.15-.19			0		1	47.62	3	46.43								
.10-.14			0		1	50.00	0									
.05-.09			0		2	54.76	2	53.57								
<.04			1	100.00	19	100.00	13	100.00								
N=	18		12		42		28		18		12		60		40	

item may get a standard score other than zero. Because the error scores may not be distributed equally about zero error, the following method of calculating standardized error scores was used so that a measure of "goodness" of performance would not be lost. A standardized error score for each item for each maneuver was developed by dividing the error score for the item by the standard deviation of all error scores on that item over all observers, groups, and days. Using the absolute values of the standardized

$$\text{Standard Score} = \frac{X - \overline{X}}{\sigma_x} \qquad \text{Standardized error score} = \frac{Xe - O}{\sigma_e}$$

error scores, the average standardized error score per maneuver was calculated for each observer for every mission. The absolute values of the standardized error scores were used so that negative errors would not cancel out positive errors. Because some observers might have failed to record all of the items on each maneuver, the average of those items recorded was used. The average standardized error scores (maneuver performance scores) by each observer were added over the five contact maneuvers for a composite VFR performance score, over the five instrument maneuvers for a composite IFR performance score, and over all ten maneuvers to form a total mission score for each subject by each observer.

One subject received no scores by either observer for the precision approach on Day 1. He had such difficulty with the approach that the safety observer talked the subject through the approach, and neither observer recorded data. This subject was assigned the greatest error scores of the subjects in his group for the items of that maneuver.

The observer-observer correlations over two days in the simulator and one day in the aircraft for each group on each of the ten maneuvers are given in Appendix E-4, and are summarized in Table 7. With N = 30, a correlation of at least 0.31 is significantly greater than zero at $p = .05$. The observer-observer correlations for the steep turn, chandelle, and lazy eight on days 3 and 4 in the aircraft seem to stand out as a block of correlations somewhat lower than the rest. The observer-observer correlations for the total mission scores, instrument maneuver scores, and contact maneuver scores over the three days for the three groups are

TABLE 7. Observer-Observer Correlations for Total, IFR and VFR Performance Scores and for Each Maneuver (N = 30 pairs per cell)

					Individual Maneuvers								
		TOT	IFR	VFR	T/O	CRU	HOL	N-P	PRE	MIS	360	CHN	LZY	LNG
Day 1	GI	.893	.968	.788	.826	.873	.906	.910	.924	.889	.803	.526	.503	.829
	GII	.967	.985	.927	.929	.950	.949	.958	.983	.883	.921	.767	.941	.832
	GIII	.943	.965	.872	.783	.934	.904	.944	.933	.918	.912	.542	.535	.596
Day 2	GI	.937	.957	.882	.840	.888	.861	.945	.897	.918	.858	.875	.799	.863
	GII	.898	.932	.848	.878	.945	.699	.864	.940	.813	.787	.439	.823	.751
	GIII	.971	.965	.958	.850	.891	.833	.954	.940	.969	.898	.889	.862	.742
Day 3	GI	.771	.925	.527	.837	.701	.854	.752	.881	.916	.231	.544	.485	.732
	GII	.860	.953	.665	.840	.969	.928	.899	.906	.870	.696	.587	.379	.713
	GIII	.905	.886	.485	.739	.886	.896	.810	.742	.779	.333	.425	.385	.803
Day 4	All Ss	.835	.912	.773	.862	.583	.899	.790	.892	.937	.559	.531	.688	.717

presented in Tables 8 through 16. They show the relationships between two observers' scores for the same performance by two independent observers (SO and FO, same day), the same observer's records of the subject's performances from one day to the next (FO one day to SO next day), two different observers' records of the subject's performances on two successive days (SO one day to FO next day), and the effects of position in the cockpit and concurrent duties upon performance records by independent observers (SO one day to SO next day vs. FO one day to FO next day). These tables permit comparisons between groups on the total scores (Tables 8 - 10), IFR scores (Tables 11 - 13) and VFR scores (Tables 14 - 16).

There appears to be a trend for the day-to-day correlations by the same observer (FO-SO) to be slightly higher than the day-to-day correlations of different observers (SO-FO). Also, the opposite diagonals of these indicate that the correlations between the safety observers from one day to the next tend to be lower than those of the flight observers from one day to the next. A chi square test for safety observer correlations being lower versus higher than flight observer correlations or for FO-SO compared to SO-FO is not appropriate because of the number of cells with expected frequencies less than five.

Effects of Simulator Motion upon Performance

The maneuver scores of the safety observer and of the flight observer were also added to obtain a single performance score (SO + FO) for each subject on each maneuver. These subject scores per maneuver were added to determine the subject's composite scores for IFR maneuvers, VFR maneuvers, and for the total mission in the same manner as previously described for the individual observers.

The mean performance scores of the groups for two days in the simulator and the third day in the aircraft for total mission, instrument, and contact scores are shown in Figures 5 - 7. The Day 4 performances are represented by a single point for the ten subjects from each group.

The performance scores (SO + FO) for the total mission, instrument maneuvers, and contact maneuvers were first analyzed over three days, two days of simulator missions, one day aircraft. The F ratios for between days and for groups by days

TABLE 8. NO MOTION, TOTAL MISSION: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group 1 (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	SO	FO
Day 1	SO		.893	.632	.730	.405	.603
	FO			.770	.812	.430	.548
Day 2	SO				.937	.432	.428
	FO					.479	.444
Day 3	SO						.771
	FO						

TABLE 9. SUSTAINED MOTION, TOTAL MISSION: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group II (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
Day 1	SO		.967	.800	.787	.651	.703
	FO			.811	.778	.633	.673
Day 2	SO				.898	.558	.742
	$\overline{FO}$					.651	.767
Day 3	$\overline{SO}$						.860
	FO						

TABLE 10. WASHOUT MOTION, TOTAL MISSION: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group III (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
Day 1	SO		.943	.600	.617	.394	.458
	FO			.666	.673	.503	.530
Day 2	SO				.971	.386	.443
	$\overline{FO}$					.478	.511
Day 3	$\overline{SO}$						.905
	FO						

TABLE 11. NO MOTION, INSTRUMENT MANEUVERS: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group 1 (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
	SO		.968	.715	.742	.597	.646
Day 1							
	FO			.774	.790	.556	.582
	SO				.957	.536	.512
Day 2							
	$\overline{FO}$					.602	.593
	$\overline{SO}$						.925
Day 3							
	FO						

TABLE 12. SUSTAINED MOTION, INSTRUMENT MANEUVERS: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group II (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
Day 1	SO		.985	.774	.719	.728	.763
	FO			.797	.723	.697	.739
Day 2	SO				.932	.743	.767
	$\overline{FO}$					.816	.822
Day 3	$\overline{SO}$						.953
	FO						

TABLE 13. WASHOUT MOTION, INSTRUMENT MANEUVERS: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group III (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
Day 1	SO		.695	.627	.659	.549	.535
	FO			.670	.700	.589	.545
Day 2	SO				.965	.500	.448
	$\overline{FO}$					.549	.480
Day 3	$\overline{SO}$						.886
	FO						

TABLE 14. NO MOTION, CONTACT MANEUVERS: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group I (N = 30 pairs each)

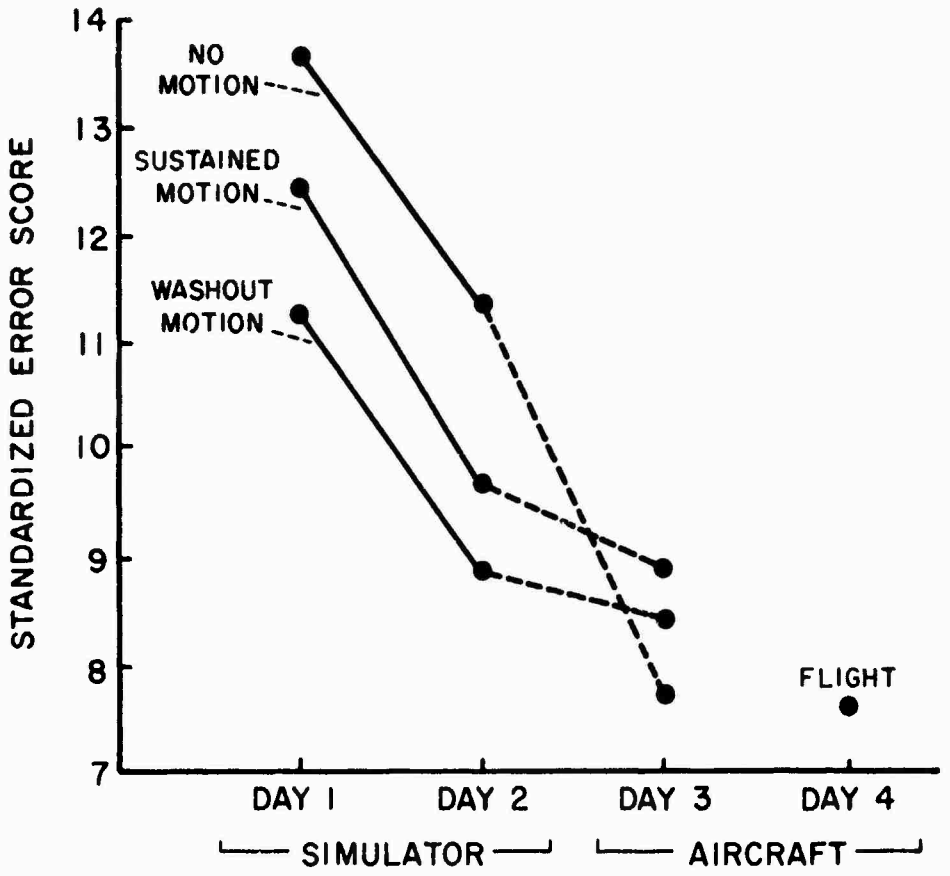
		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
	SO		.788	.500	.594	.209	.525
Day 1							
	FO			.680	.702	.233	.420
	↓						
	SO				.882	.251	.271
Day 2							
	$\overline{FO}$					.312	.263
	↓						
	$\overline{SO}$						.527
Day 3							
	FO						

TABLE 15. SUSTAINED MOTION, CONTACT MANEUVERS: Correlation Coefficients between Performance Scores Recorded by Same and Different Observers on Same and Different Days, Group II (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
Day 1	SO		.927	.771	.736	.380	.577
	FO			.792	.744	.446	.583
Day 2	SO				.848	.254	.624
	$\overline{FO}$					.297	.549
Day 3	$\overline{SO}$						.665
	FO						

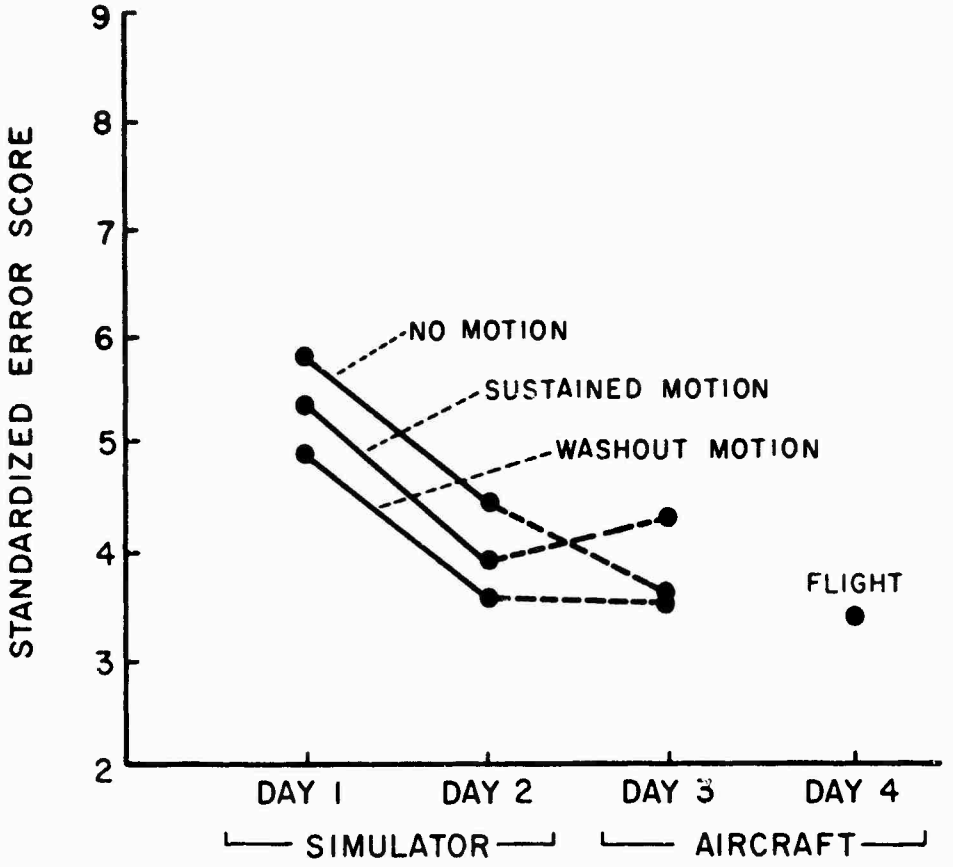
TABLE 16. WASHOUT MOTION, CONTACT MANEUVERS: Correlation Coefficients between Performance Measures Recorded by Same and Different Observers on Same and Different Days, Group III (N = 30 pairs each)

		Day 1		Day 2		Day 3	
		SO	FO	SO	FO	$\overline{SO}$	FO
	SO		.872	.510	.508	.228	.342
Day 1							
	FO			.564	.530	.319	.446
	↓						
	SO				.958	.202	.377
Day 2							
	$\overline{FO}$					.298	.431
	↓						
	$\overline{SO}$						.485
Day 3							
	FO						



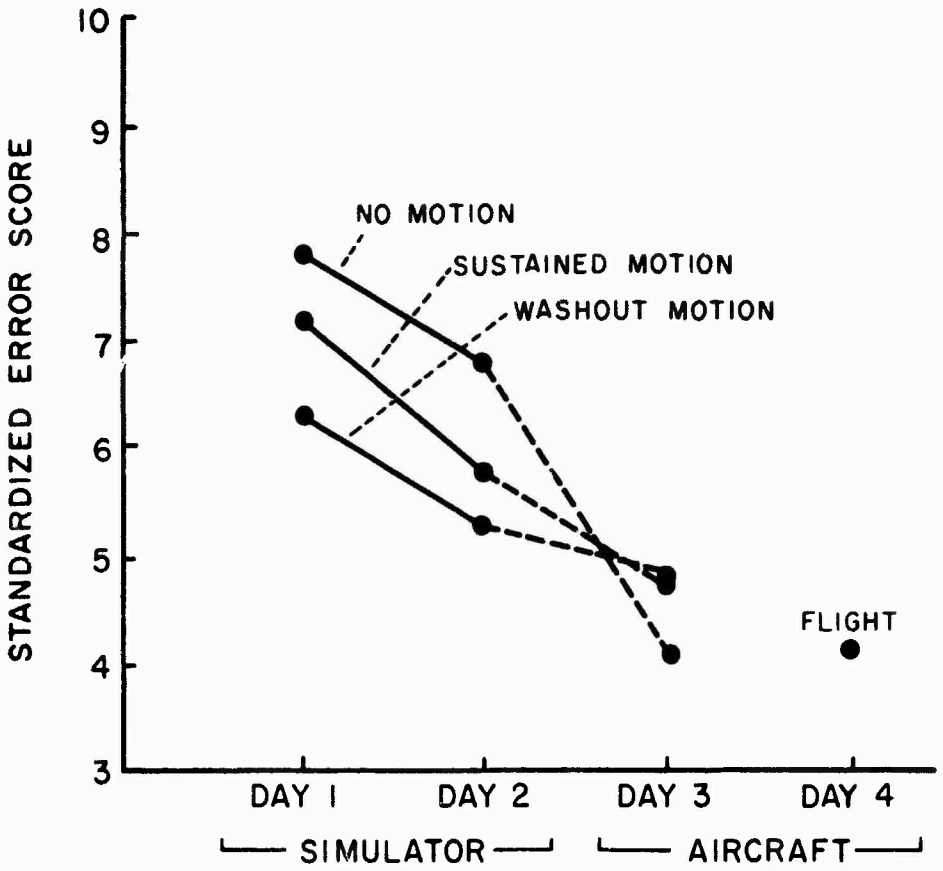
GROUP	DAY 1	DAY 2	DAY 3	DAY 4
I	13.71	11.27	7.67	7.54
II	12.39	9.62	8.92	
III	11.21	8.81	8.40	

Figure 5. TOTAL MISSION: Composite performance scores for each group on two days in the simulator and two days in the aircraft (N = 30 per Group, Day 1-3; N = 30, ten from each group, Day 4).



GROUP	DAY 1	DAY 2	DAY 3	DAY 4
I	5.80	4.44	3.58	
II	5.28	3.90	4.20	3.43
III	4.97	3.56	3.56	

Figure 6. INSTRUMENT MANEUVERS: Composite performance scores for each group on two days in the simulator and two days in the aircraft (N = 30 per Group, Day 1-3; N = 30, ten from each group, Day 4).



GROUP	DAY 1	DAY 2	DAY 3	DAY 4
I	7.90	6 83	4.09	4.11
II	7. 11	5.73	4.72	
III	6.24	5.25	4.84	

Figure 7. CONTACT MANEUVERS: Composite performance scores for each group on two days in the simulator and two days in the aircraft (N = 30 per Group, Day 1-3; N = 30, ten from each group, Day 4).

interaction were significant ($p < .05$) for all three types of measures (Tables 17 - 19). For total mission, instrument, and contact scores, Group I (no motion) had higher mean error scores for the two simulator missions than the groups that had motion in the simulator. Of the two motion groups, the sustained motion group had higher mean error scores than the washout motion group.

A separate analysis of variance was performed on the performance scores for the two days of simulator missions to see if there were a significant difference among groups in the simulator that might have been masked by the strong interaction that occurred when the Day 3 aircraft mission was included. The analysis, Tables 20 - 22, indicated o significant difference between days for total mission, instrument, and contact scores and a significant difference among motion conditions for the contact maneuvers only. The differences among motion conditions in total mission scores barely missed arbitrary statistical reliability, $p = .0568$.

Analysis of variance of the scores for the three groups in the aircraft, Days 3 and 4, showed no significant difference among groups or between days and no significant interaction within total mission scores, instrument maneuver scores, or contact maneuver scores (Appendix E-5).

Analysis of the performance from the simulator to the aircraft, Day 2 - Day 3, (see Appendix E-6) showed a significant decrease in contact error scores between the days (p, remote), but not a significant decrease in instrument error scores ($p = .2291$). When the contact and instrument error scores are combined to obtain total mission error scores, there is an overall significant difference in scores between Days 2 and 3 (p, remote). The Newman-Keuls *a posteriori* test for significant differences between means on Day 2 in the simulator and on the first day in the aircraft, Day 3, indicated that Group I, no motion, had a significantly higher ($p < .05$) total mission error score in the simulator than Group III, washout motion, but in the aircraft there was no significant difference among the groups on Day 3 ($p > .05$). For the contact maneuver scores, the only significant difference among groups was between Groups I and III on Day 2 in the simulator, ($p < .05$).

TABLE 17. TOTAL MISSION: Analysis of Variance of Composite Performance Scores for Days 1, 2, and 3

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	45.0676	1.374	.259
Subjects (S/G)	87	32.8041		
Within Subjects				
Days (D)	2	386.2327	66.916	remote
G x D	4	30.2495	5.241	.001
D x S/G	174	5.7719		

TABLE 18. INSTRUMENT MANEUVERS: Analysis of Variance of Composite Performance Scores for Days 1, 2, and 3

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	8.1098	1.026	.363
Subjects (S/G)	87	7.9006		
Within Subjects				
Days (D)	2	66.2406	55.628	remote
G x D	4	3.5374	2.971	.021
D x S/G	174	1.1908		

TABLE 19. CONTACT MANEUVERS: Analysis of Variance of Composite Performance Scores for Days 1, 2, and 3

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	15.4797	1.463	.237
Subjects (S/G)	87	10.5830		
Within Subjects				
Days (D)	2	145.2100	54.296	remote
G x D	4	14.8878	5.567	remote
D x S/G	174	2.6744		

TABLE 20. TOTAL MISSION, SIMULATOR: Analysis of Variance of Composite Performance Scores for Days 1 and 2

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	93.1715	2.9644	.0568
Subjects (S/G)	87	31.4301		
Within Subjects				
Days (D)	1	289.0309	58.5949	remote
G x D	2	0.6332	0.1284	.8797
D x S/G	87	4.9327		

TABLE 21. INSTRUMENT MANEUVERS, SIMULATOR: Analysis of Variance of Composite Performance Scores for Days 1 and 2

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	11.2198	1.6899	.1905
Subjects (S/G)	87	6.6395		
Within Subjects				
Days (D)	1	86.4260	76.9212	remote
G x D	2	.0063	.0056	.9944
D x S/G	87	1.1236		

TABLE 22. CONTACT MANEUVERS, SIMULATOR: Analysis of Variance of Composite Performance Scores for Days 1 and 2

Source	df	MS	F	p
Between Subjects				
Groups (G)	2	39.7463	3.4622	.0351
Subjects (S/G)	87	11.4140		
Within Subjects				
Days (D)	1	59.3570	25.6475	remote
G x D	2	.6442	.2784	.7577
D x S/G	87	2.3143		

A discriminant analysis (Appendix E-7) was applied to the scores for the three groups on Day 2 in the simulator for both the total mission scores and for contact maneuvers scores since these were the only ones on which the groups differed significantly on Day 2. From each of the analyses, only the first discriminant function was statistically significant, and the group means for each of these discriminant functions were closely related to the mean error scores on Day 2 plotted in Figures 5 and 7. No meaningful interpretation of the standardized discriminant weights could be made in the light of how the significant discriminant functions separated the three groups.

The Prediction of Pilot Proficiency

The day-to-day (ride-ride) correlations among the composite performance scores of the subjects in all groups combined for the total mission, instrument maneuvers, and contact maneuvers on days 1, 2, and 3 are given in Tables 23 - 25. The Day 2 - Day 3 correlation on instrument maneuver scores for the sustained motion was significantly greater than that for washout motion ($p = .05$).

The Day 3 - Day 4 correlation over groups for the total mission was .647, for instrument maneuvers .679, and for contact maneuvers .455. The observer-observer reliability coefficients for the total, IFR, and VFR scores on Day 4 were 0.835, 0.912, and 0.773, respectively.

All of the day-to-day correlations of the total mission scores, IFR scores and the VFR scores in Tables 23 - 25 and the Day 3 - Day 4 correlations were significantly greater than zero ($p = .05$). The day-to-day correlations of performance on each maneuver for each of the three groups are in Appendix E-8.

In the prediction of performance from one day to another for the total, instrument and contact scores, the appropriate individual maneuver scores were merely summed to obtain the scores used in the correlations. Since assigning different weights to the individual maneuvers in the simulator might give a more accurate estimate of the pilot's performance in the aircraft, multiple correlations from the ten maneuvers in the simulator to a measure of performance in the aircraft were computed. Two criteria were readily available, the sum of the ten maneuvers in the aircraft (TOT) and the observer's total rating for the mission (OTR). The former was desirable because of its relative

TABLE 23. TOTAL MISSION: Day-to-Day Correlation Coefficients between Composite Performance Scores for Groups I, II, and III Combined

	No Motion		Sustained Motion		Washout Motion	
	Day 2	Day 3	Day 2	Day 3	Day 2	Day 3
Day 1	.771	.542	.822	.696	.652	.510
Day 2		.482		.724		.492

TABLE 24. INSTRUMENT MANEUVERS: Day-to-Day Correlation Coefficients between Composite Performance Scores for Groups I, II, and III Combined

	No Motion		Sustained Motion		Washout Motion	
	Day 2	Day 3	Day 2	Day 3	Day 2	Day 3
Day 1	.770	.611	.770	.743	.667	.576
Day 2		.577		.809		.513

TABLE 25. CONTACT MANEUVERS: Day-to-Day Correlation Coefficients between Composite Performance Scores for Groups I, II, and III Combined

	No Motion		Sustained Motion		Washout Motion	
	Day 2	Day 3	Day 2	Day 3	Day 2	Day 3
Day 1	.675	.419	.809	.560	.549	.394
Day 2		.324		.499		.384

objectivity, but it has the shortcoming of weighting each of the maneuvers equally. The item discrimination coefficients in Appendix E-9 indicate that there is a tendency for the instrument approaches, holding, and the steep turn maneuvers to correlate higher with the total booklet score (TOT) than the rest of the maneuvers. The observer's total rating, on the other hand, was based upon some unknown subjective weighting system, probably different from observer to observer, but perhaps a better "grade" of the subject's overall performance than TOT. Figures 8 - 10 show that the relationships between the observers' ratings of the individual maneuvers to their overall rating of the mission tend to be greater for instrument maneuvers than contact maneuvers.

Table 26 presents the results of both sets of multiple correlations along with product-moment correlations on TOT and OTR. The multiple correlation for the three groups from ten moneuvers on Day 2 in the simulator to TOT on Day 3 were higher than the Day 2 TOT to Day 3 TOT correlations. When using the OTR as criteria there was a slight but even further increase in the correlations. The correlations between OTR and TOT on Day 3 and the correlations of the ten maneuvers and the equally weighted sum of the ten maneuvers to the observer's total ratings on Day 3 serve as a check of agreement of the measures on the criterion construct. These correlations are all significantly greater than zero at $p = .01$.

The observers in each group were split into two samples so that each sample had one observer from each mission. The division of observers was ordered so that an equal number of SOs and FOs were in each sample. Multiple correlations were computed for one half of the data and the weights determined from that data were applied to the remaining half of the data for validation across independent observations of the same performances. The weights were based on the proportions of explained criterion variance attributed to each of the ten predictor variables. This was determined by noting the amount that the squared multiple correlation would drop when recalculated on only nine predictor variables, each variable was replaced as the successive variable was dropped (Darlington, 1968; Bale, et al., 1973). The resulting correlations and the cross-observer validation correlations are presented in Table 27.

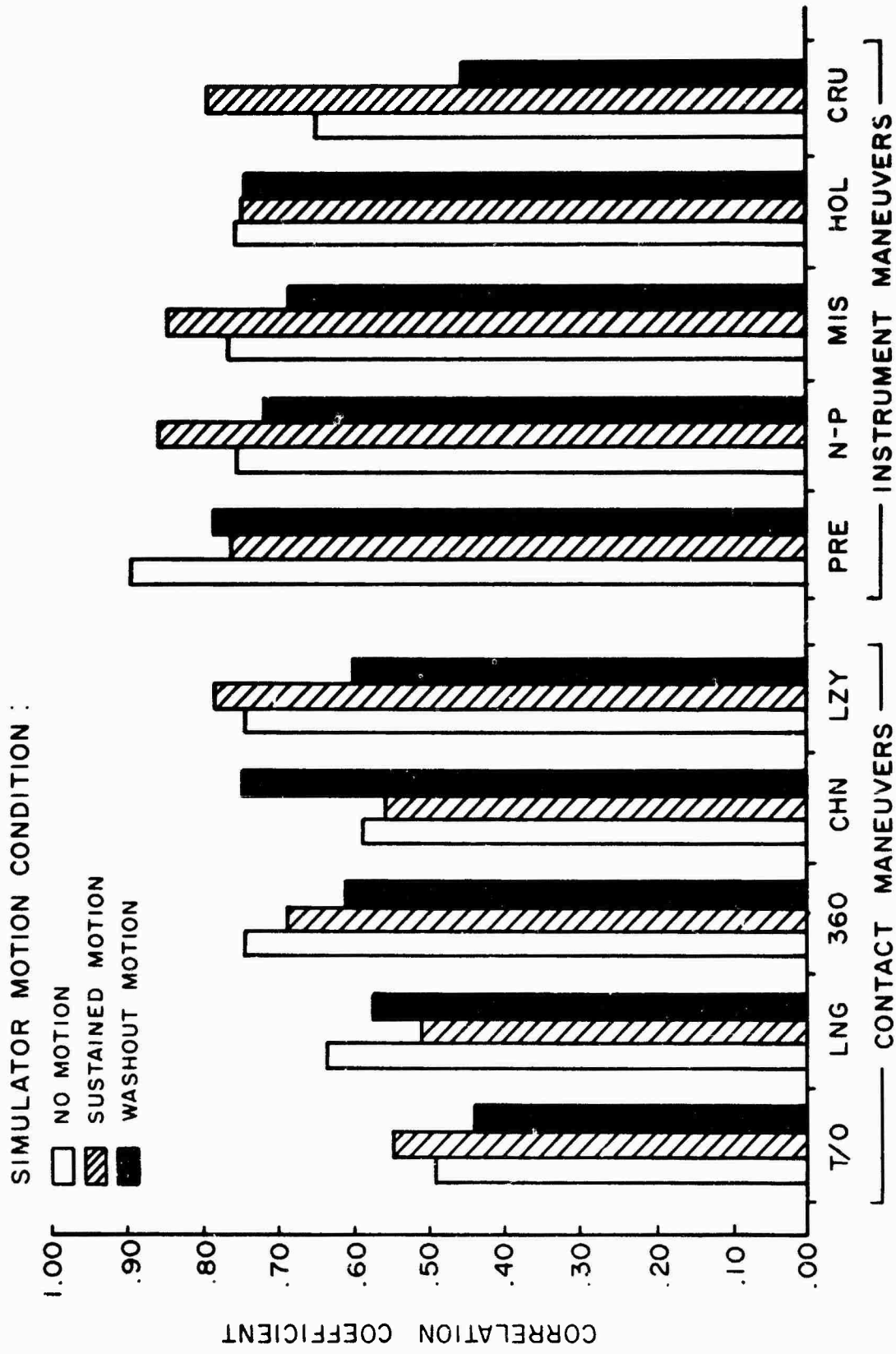


Figure 8. DAY I, SIMULATOR: Correlations between observer ratings on each maneuver with observer ratings for the total mission ($N_{GI} = 39$, $N_{GII} = 40$, $N_{GIII} = 36$).

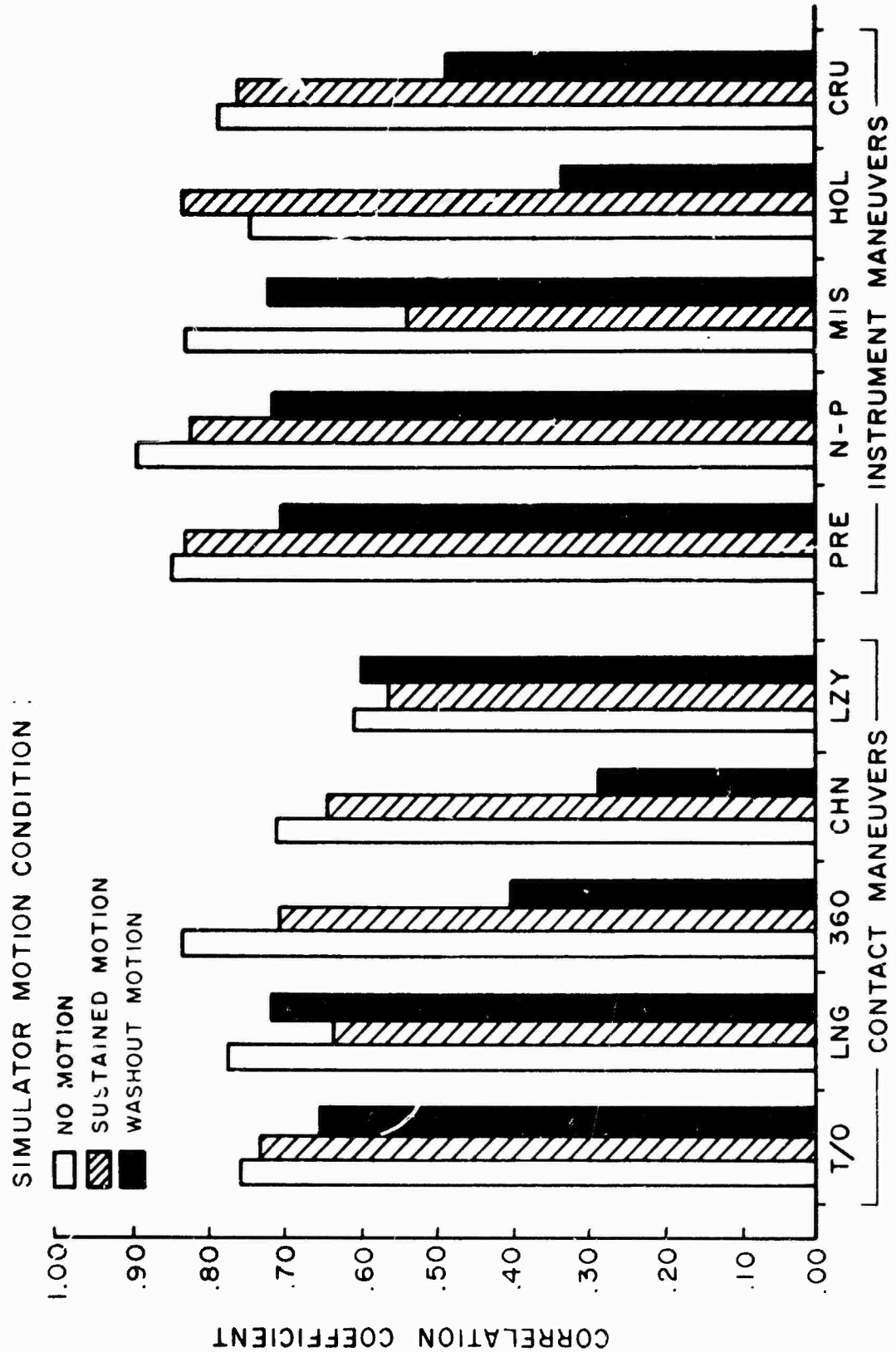


Figure 9. DAY II, SIMULATOR: Correlations between observer ratings on each maneuver with observer ratings for the total mission ($N_{GI} = 39$, $N_{GII} = 38$, $N_{GIII} = 40$).

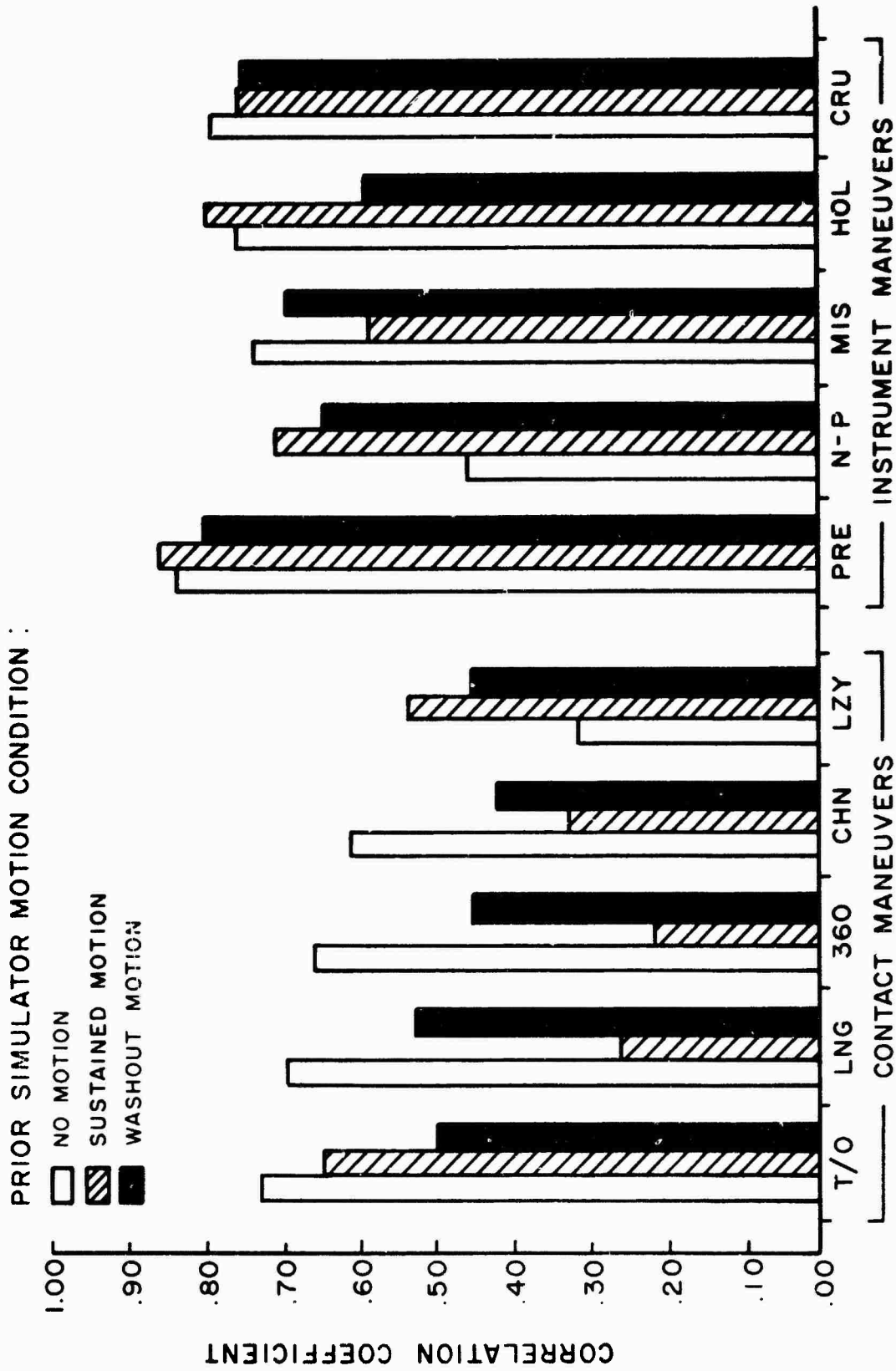


Figure 10. DAY III, AIRCRAFT: Correlation between observer ratings on each maneuver with observer ratings for the total mission ($N_{GI} = 41$, $N_{GII} = 42$, $N_{GIII} = 41$).

TABLE 26. Prediction of Pilot Performance in Aircraft on Day 3 from Performance Measures Taken in Simulator on Day 2 as a Function of Simulator Motion Condition and Correlations between Criterion Measures in Flight (N = 30 per correlation)

PREDICTORS	CRITERIA	
	Day 3 Performance Scores, Total Mission	Day 3 Observers' Ratings, Total Mission
	No Motion	
Day 2		
Performance Scores, Total Mission	.482**	.529**
Performance Scores, Ten Maneuvers	.722	.763*
Day 3		
Performance Scores, Total Mission		.878**
Performance Scores, Ten Maneuvers		.929**
	Sustained Motion	
Day 2		
Performance Scores, Total Mission	.724**	.709**
Performance Scores, Ten Maneuvers	.874**	.911**
Day 3		
Performance Scores, Total Mission		.852**
Performance Scores, Ten Maneuvers		.943**

TABLE 26 (continued)

PREDICTORS	CRITERIA	
	Day 3 Performance Scores, Total Mission	Day 3 Observers' Ratings, Total Mission
	Washout Motion	
Day 2		
Performance Scores, Total Mission	.492**	.417**
Performance Scores, Ten Maneuvers	.647	.651
Day 3		
Performance Scores, Total Mission		.726**
Performance Scores, Ten Maneuvers		.906**

*$p < .05$

**$p < .01$

TABLE 27. Cross-Observer Validation for Multiple Correlations between Composite Scores for Ten Maneuvers Performed in the Simulator on Day 2 with Composite Scores for the Same Maneuvers Performed in Flight on Day 3 (N = 30 per correlation)

	10 MAN D_2 = TOT D_3		10 MAN D_2 = OTR D_3	
	Prediction Half	Validation Half	Prediction Half	Validation Half
Group I, No Motion	.727	.612*	.785*	.577*
Group II, Sustained Motion	.883**	.615*	.876**	.716*
Group III, Washout Motion	.661	.506*	.685	.329*

*$\underline{p} < .05$

**$\underline{p} < .01$

Note: The validation half coefficients of correlation computed by using weights derived from the data of the other half of the observers are Pearson product-moment coefficients of correlation.

Because of the strong intercorrelations of the predictor variables and the relatively small sample size, meaningful interpretations of the weights assigned to the various maneuvers could not be made. Each mission was treated as a set of ten variables (maneuvers) and canonical correlational analysis was run between Day 2 and Day 3 for each group. The goal was to create linear combinations of the ten predictor variables (Day 2) that are highly correlated with linear combinations of the criterion variables (Day 3) and to develop some meaningful interpretations for the weights involved with each canonical variate. The computed canonical correlations were significant, but because the ten variables on each day were so highly intercorrelated, useful interpretations of the linear combinations making the canonical variates could not be made even when structure matrices were computed to aid the interpretation.

The day to day correlations of the subjects' confidence in their ability to perform the various maneuvers are in Appendix E-10. A frequency distribution with cumulative percentages for Day 2 - Day 3 correlations of subject confidence ratings on the ten maneuvers over the three groups is given in Table 28. All but one of the correlations are significantly greater than zero, and 80% of all correlations are greater than 0.70.

The correlations between the subjects' estimates of ability to perform the maneuvers and the recorded maneuver scores (SR to (SO + FO)) and the correlations of each observer's record of each maneuver with his rating of the performance on the maneuver (SO - SOR) (FO - FOR) are presented in Appendix E-11. Table 29 shows the frequencies and cumulative percent of correlatic of the subjects' ratings (SR) with the subjects performance (SO + FO). Only one-third of these correlations was significantly greater than zero and none is higher than -.533. The average correlation of subjective confidence ratings to subsequent performance was 0.262 in the simulator and 0.224 in the aircraft. The average of these correlations for instrument and contact maneuvers for each group on Day 2 and Day 3 are in Table 30.

Tables 31 - 33 indicate the relationships between the total mission score, instrument score, contact score, and the ratings of the safety observer and flight observer for Days 1, 2, and 3. Although they correlate highly with the total score, the instrument and contact scores correlate lower with each other, and they are

TABLE 28. Frequency Distribution and Cumulative Percent Correlations between Subject Confidence Ratings and Their Performance Scores for Each of Ten Maneuvers Performed on Days 2 and 3 (N = 30 pairs per correlation)

	Group I No Motion		Group II No Motion		Group III No Motion		Total	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %
.95-.99								
.90-.94			1	10	1	10	2	6.67
.85-.89	3	30	1	20	1	20	5	23.33
.80-.84	4	70	1	30	4	60	9	53.33
.75-.79	1	80	1	40	2	80	4	66.67
.70-.74	2	100			2	100	4	80.00
.65-.69			3	70			3	90.00
.60-.64			1	80			1	93.33
.55-.59			1	90			1	96.67
.50-.54								
.45-.49								
.40-.44								
.35-.39								
.30-.34								
.25-.29			1	100			1	100.00

TABLE 29. Frequency Distribution and Cumulative Percent of Correlations between Subject Confidence Ratings and Subject Performances on Ten Maneuvers (N = 30 per correlation)

	Day 2				Day 3				Day 2		Day 3		TOT		TOT	
	VFR		IFR		VFR		IFR		Sim		Acft		IFR		VFR	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %
.55 .59			1	6.67					1	3.33			1	3.33		
.50 .54			3	26.67			1	6.67	3	13.33	1	3.33	4			
.45 .49			2	40.00			0		2	20.00	0		2	23.33		
.40 .44	2	13.33	0				3	26.67	2	26.67	3	13.33	3	33.33	2	6.67
.35 .39	0		2	53.33	1	6.67	2	40.00	2	33.33	3	23.33	4	46.67	1	10.00
.30 .34	1	20.00	0		2	20.00	0		1	36.33	2	30.00	0		3	20.00
.25 .29	2	33.33	1	60.00	1	26.66	1	46.67	3	46.67	2	36.33	2	53.33	3	30.00
.20 .24	0		2	73.33	3	46.67	2	60.00	2	53.33	5	53.33	4	66.67	3	40.00
.15 .19	4	60.00	0		2	60.00	3	80.00	4	66.67	5	70.00	3	76.67	6	60.00
.10 .14	3	80.00	1	80.00	1	66.67	2	93.33	4	80.00	3	80.00	3	86.67	4	73.33
.05 .09	2	93.33	1	86.66	3	86.66	0		3	90.00	3	90.00	1	90.00	5	90.00
.00 .04	0		2	100.00	2	100.00	0		2	96.67	2	96.67	2	96.67	2	96.67
.05 .00	1	100.00					1	100.00	1	100.00	1	100.00	1	100.00	1	100.00

TABLE 29 (continued)

	Day 2		Day 2		Day 2		Day 3		Day 3		Day 3	
	GI		GII		GIII		GI		GII		GIII	
r	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %	f	Cum %
.55 .59	1	10										
.50 .54	2	30	1	10					1	10		
.45 .49	1	40	1	20					0			
.40 .44	1	50	0		1	10	2	20	0		1	10
.35 .39	1	60	1	30	0		2	40	1	20	0	
.30 .34	1	70	0		0		0		1	30	1	20
.25 .29	1	80	2	50	0		1	50	0		1	30
.20 .24	0		0		2	30	2	70	1	40	2	50
.15 .19	1	90	2	70	1	40	0		2	60	3	80
.10 .14	0		1	80	3	70	1	80	2	80	0	
.05 .09	1	100	2	100	0		2	100	0		1	90
.00 .04					2	90			2	100	0	
.05 .00					1	100					1	100

TABLE 30. Means of Correlations between Subject Confidence Ratings and Total Mission Performance Scores on Day 2 and Day 3
(N = 5 correlations per cell, 30 pairs per correlation)

	Instrument		Contact		Total	
Group	Day 2	Day 3	Day 2	Day 3	Day 2	Day 3
No Motion	0.492	0.361	0.269	0.155	0.386	0.260
Sustoined Motion	0.330	0.235	0.160	0.180	0.247	0.207
Woshout Motion	0.130	0.207	0.175	0.200	0.154	0.204
TOTAL	0.325	0.269	0.201	0.177	0.263	0.223

TABLE 31. Correlations between Total Mission, Instrument Maneuver and Contact Maneuver Scores and Observer Ratings of Total Missions for Day 1 (N = 30 pairs per correlation)

		IFR	VFR	SO TR	FO TR
Total	GI	.876	.936	.764	.708
	GII	.882	.948	.856	.881
	GIII	.955	.956	.715	.802
IFR	GI		.649	.807	.661
	GII		.687	.791	.841
	GIII		.832	.753	.827
VFR	GI			.614	.633
	GII			.787	.792
	GIII			.620	.710
SO TR	GI				.556
	GII				.895
	GIII				.895

TABLE 32. Correlations between Total Mission, Instrument Maneuvers and Contact Maneuver Scores and Observer Ratings of Total Missions for Day 2 (N = 30 pairs per correlation)

		IFR	VFR	SO TR	FO TR
Total	GI	.904	.947	.836	.768
	GII	.914	.954	.804	.867
	GIIi	.932	.968	.756	.795
IFR	GI		.719	.892	.851
	GII		.749	.877	.886
	GIII		.812	.794	.771
VFR	GI			.689	.610
	GII			.661	.757
	GIII			.670	.749
SO TR	GI				.890
	GII				.881
	GIII				.794

TABLE 33. Correlations between Total Mission, Instrument Maneuver and Contact Maneuver Scores and Observer Ratings of Total Missions for Day 3 (N = 30 pairs per correlation)

		IFR	VFR	SO TR	FO TR
Total	GI	.927	.896	.913	.831
	GII	.927	.921	.829	.872
	GIII	.922	.920	.770	.674
IFR	GI		.665	.889	.791
	GII		.708	.874	.885
	GIII		.696	.652	.643
VFR	GI			.769	.721
	GII			.655	.724
	GIII			.768	.598
SO TR	GI				.827
	GII				.886
	GIII				.692

significantly greater than zero. The observers' overall ratings on the completed mission tend to correlate higher with the instrument scores than the contact scores.

Other factors that might be related to performance scores or the ratings of observers are the data about the subject's age and flying time. The tables of correlations between performance measures and observer's total ratings of the missions with the data on the subject (Appendix E-12) show few significant relationships. The subject's age (year of birth) tends to relate with performance scores and ratings for Days 1 and 2 but not too well in the aircraft, Day 3, for Groups I and II. But, Group III has lower correlations in the simulator and higher in the aircraft performance and rating scores than Groups I and II with respect to age.

Flight time in the past six months seems to correlate well with aircraft performance for all three groups while the amount of multi-engine flight time and instrument flight time logged in the past six months show rather low relationships to performance and ratings in both the simulator and aircraft. Measures of various types of total time logged do not have any significant relationship with the performance measures and observer ratings taken in this study.

DISCUSSION

The discussion of the results follows the same sequence used in presenting the results. First, the stability of pilot subjects' performances and the reliability of the performance measures are considered. This is followed by a discussion of the effects of the experimental conditions upon performance in the simulator and in the aircraft. Then the implications of the results for prediction of aircraft pilot proficiency are discussed. Finally, there are discussions of the relationships between various measures and indices of flight performance and flight experience.

Reliability of Performance Measures

The low reliability of the traditional subjective rating of operator performance has given impetus to the development of automatic performance recording devices whose application has been hindered by the lack of flexibility for general use, the need of costly electronic recording and deciphering equipment, and considerable difficulty in developing meaningful interpretations of the data collected. In general, the measurement of performance on complex tasks by human observers can be accomplished with a reasonably high degree of reliability if relatively objective measuring instruments are used.

The reliability coefficients of the measures of performance in this study were based upon observer-observer reliabilities, two observers recording the same performance at the same time. Most reliability coefficients in aviation studies are inferred from split-half reliability coefficients or from observer-camera correlations. The observers in this study were positioned so that they could not see each other's records of performance, and none reported obtaining any recording cues from the other observer. A screen to preclude the possibility of cues from one observer to the other was ruled out as a potential flight hazard after a few practice missions had been flown.

Several factors were found to affect the observer-observer reliability coefficients for the items. First, there tends to be higher observer-observer

reliabilities for items referenced to instruments in the vehicle itself than for items referenced to points in the world outside the vehicle. For example, the reference of the pitch of the vehicle above or below the horizon or the alignment of the vehicle parallel to roads several thousand feet below the vehicle are difficult for two observers located in different positions in the vehicle to score congruently. One outside point of reference item that showed high observer-observer reliability was the point of touchdown on landing. The measure of this item was well anchored by counting the number of runway lights between the point of touchdown and the point of intended touchdown, and the units of measurement for touchdown were 100 feet, half the distance between lights. Also, the painted stripes on the runway could be used if landing were made to one side of centerline.

Second, the instruments themselves vary in the accuracy to which they can be read. Altitude, airspeed, and heading have a scale with a prominent pointer and they can be more easily read than bank, coordination, and rate of turn. The later three instruments have thick blunt pointers with few scale indices. Because of the positions of the instruments relative to the observers' line of vision, lack of visibility and parallax errors also affect the observer-observer reliabilities. There is greater parallax error for the safety observer reading the subject's instruments than for the flight observer. Also, the flight observer can move about to get a better view of difficult-to-see instruments.

The lack of understanding of how to record the performance information is a source of observer error and was predominant in the recording of heading error on rollout in the steep turn, chandelle, and lazy eight maneuvers. For example, if the subject were to roll out of a left turn on a heading of 270 degrees, but he overshot to 250 degrees, some observers recorded the error as +20 degrees because he overshot, others recorded the error as -20 degrees because 250 is arithmetically 20 degrees less than 270. Errors of this type can be prevented by more explicit instructions to observers and more detailed training. One problem that is difficult to solve, even through extensive training, is the development of high agreement between observers as to just when the rollout has stopped. This is especially true with subjects who,

knowing that their heading or rollout is to be recorded, will attempt to rudder the aircraft heading closer to the desired after the turn has stopped.

Observer-observer reliability coefficients also differ as a function of the type of information recorded from the instruments. Comparing the correlations on maximum deviations as opposed to check scores for airspeed, altitude, heading and bank (Table 8), it is clear that fewer check scores have observer-observer correlations above .70 than maximum deviation scores. These check scores were frequently referenced to procedural points (i.e., when power was retarded, and when leveloff or missed approach was executed) and to horizon or heading referenced points, and therefore the observers may have differed as to exactly when the check scores were taken. Erickser's (1947) time sample check scores resulted in higher observer-observer correlations than the range or limits methods, but his observers' check scores were anchored to specific times on the same clock (every 20 seconds), and the check scores were averaged for each observer before calculation of the correlation coefficients.

Procedural items, such as reporting arrival at a fix, performing a particular checklist, etc., were sometimes difficult for the flight observer to record because he could not hear what the subject said to the safety observer or he could not see when the subject moved a switch to the proper position. "Hotmike" communications between the cockpit personnel could aid the flight observer in recording some of the procedural items.

Environmental factors of noise masking communications, turbulence making pointers move erratically and extra vehicular references difficult, and unusual g-forces causing observers to become nauseated contributed to less accurate measures of performance and lower observer-observer reliabilities. Fortunately, most of these can be eliminated or reduced when performance measures are taken in vehicle simulators. This is supported by the data in Table 8 comparing the frequency of high observer-observer correlations from simulator data with those from aircraft data.

Although the results (Appendix E-1 and Table 9) do not indicate that motion and type of motion in the simulator affect the observer-observer reliabilities, verbal reports from the observers indicated that the motion tended to give them cues to changes in flight attitude. Some observers claimed that being flight observer with the washout motion system tended to nauseate them. This could be attributed to the fact that the flight observer, who is not too deeply involved with the task except for the recording of data, tends to be aware of the washout and the conflict between those cues and the instrument displays becomes uncomfortable. Both pilots in the front seats are involved with the mission to a greater extent and do not notice the washout of the motion onset cues.

When the observers were tired they found it difficult to keep up with the mission when there was no motion at all. This is attributable to the lack of cues to alert them to what was going on.

Two factors that seemed to affect the day-to-day correlations of observers were knowledge of the subject and the duties of the observer. The correlations between scores by an observer on one day (FO) and his scores for the same subject on the next day (SO) tended to be higher than for different observers on the two days (SO-FO). This finding is in agreement with that of Caro (1968) who noted that prior knowledge of the subject may tend to affect the results of checkrides.

The particular duties the observer is responsible for during the mission may also affect the day-to-day correlation. The results indicate that, for observers assigned the additional duties of safety pilot, making all radio transmissions outside the aircraft, and manager of fuel, ventilation, and other systems, the day-to-day correlation will tend to be lower than those between flight observers who primarily gather data.

The Pilot Performance Record used in this study also served as an aid to the observers in making their subjective grades of the subject's performances resulting in high observer-observer correlations on the subjective ratings. Greer, Smith, and Hatfield (1962) showed that observer's subjective grades made with the use of

relatively objective instruments, such as their performance record booklet (PPDR), are significantly more reliable than subjective grades without reference to such a device.

Despite the numerous factors affecting observer-observer reliability, the use of a carefully constructed, relatively objective booklet for the recording of pilot performance tends to result in high observer-observer reliability coefficients. If the booklet does not demand too much of the observer's time and if the proper use of the scales is fairly obvious, extensive and costly training sessions for observers may not be necessary.

Performance Levels of Subjects

The time between missions in this study varied between one day and three weeks, with a median just over two and one-half days. Due to work schedules of the subjects, weather, and availability of the observers, the time between missions could not be fixed. Attempts were made to schedule the subjects on three of four successive days to prevent considerable learning or forgetting from intervening. Some flight instructors and persons who used their aircraft regularly in their businesses did fly between the missions of this study. But, these pilots were probably at a fairly stable level of performance to begin with, and their flight activities between missions would not affect their performance greatly. The pilots with the most potential for relearning, those who had not flown in a long time, had the lowest probability of flying between sessions because they were not current in any aircraft, and they would have to pay to become requalified. Many stated that they hadn't been flying because of the cost, or they just couldn't afford the time to maintain their currency. A few pilots who had not flown for several years did study the instrument approach plates for Champaign and Decatur between the sessions because they felt so inadequate with their first-mission performances. Some of the subjects had never flown a lazy eight, chandelle, ADF instrument approach, or ILS instrument approach before.

The performance of subjects in simulators tends to differ as a function of the presence of motion and the type of motion used. In this study the group of subjects with no motion consistently had higher error scores than the groups with motion for total instrument maneuver scores, total contact maneuver scores, and total mission scores. The group with no motion had significantly higher total error scores and contact maneuver scores on Day 2 than the group with washout motion. The observed differences among groups on instrument maneuver error scores were not statistically reliable. The group with sustained motion had greater error scores than those with washout motion and lesser error scores than those with no motion, but their scores were not significantly different from those of the other groups.

While controlling an aircraft in three dimensions, operators tend to rely upon kinesthetic and vestibular cues for indication changes in attitude. With no motion in the simulator the operator must rely fully upon the instrument displays for attitude information, and if the simulated vehicle does not have an automatic attitude controller, the operator cannot permit his attention to wander from the attitude indicator for very long. The operator of a simulator with an operative motion base is aided by the attitude of the simulator and the onset cues as the simulator changes position. Thus he tends to notice changes in attitude more rapidly and correct back to desired attitude sooner, resulting in lower error scores. The beneficial effects of the motion cues seem to be greater when the subjects are performing contact maneuvers in which one normally relies upon motion cues than when they are performing the instrument maneuvers in which pilots are generally taught to disregard vestibular and kinesthetic cues.

Because instrument maneuvers can be performed solely by reference to the instrument, the error scores for pilots performing instrument maneuvers tend to be lower than their error scores for contact maneuver normally flown with external visual reference when using flight simulators with no external visual reference. In fact, many subjects found it difficult to envision the performance of contact maneuvers with reference to instruments alone, and two of the maneuvers, chandelle and lazy eight, are not frequently performed by general aviation pilots. The

significant difference between days on the contact maneuvers was indicative of the extensive learning that took place in the performance of the infrequently practiced contact maneuvers by reference to instruments. In changing to the aircraft, there was a significant decrease in contact maneuver error scores, and there was no longer a significant difference among groups.

The very steep drop in contact maneuver scores for all groups could be attributed to the exceptional difficulty the subjects had with those maneuvers in the simulator, especially without the benefit of motion cues, and the possibility of further learning. Also, when flying the aircraft no precautions were taken to prevent the subjects from using their instruments, and a rapid crosscheck of instruments developed in the simulator might have been beneficial in the aircraft.

For the instrument maneuvers, group performances did not differ significantly in the simulator or in the aircraft, and the transfer from the simulator to aircraft did not result in a significant decrease in error scores. Because the instrument maneuvers were performed as they should have been, without external visual references and because instrument pilots should be capable of performing the instrument maneuvers in the simulator as well as in the aircraft, there was little learning exhibited in the simulator and no significant changes in error scores when changing to the aircraft. Yet the no-motion group in the simulator showed the greatest drop in error scores going from the simulator to the aircraft.

With respect to total mission scores in the simulator, it can be said that no motion results in the greatest error scores while either type of motion results in the smaller error scores. But when transferring to the aircraft differences among the groups are not significant. Possibly because the sustained motion cues are different from the washout motion cues that the pilots are accustomed to experiencing, there are slightly greater error scores for the sustained motion than for washout motion. It is possible that this difference, though not statistically significant, might become even less if the subjects were to have a greater opportunity to adapt to the sustained motion system.

Several subjective impressions concerning the simulator motion were stated by the subjects and observers or observed by the experimenter. First, with respect to no motion, subjects who were assigned to this group and had prior knowledge of the simulator's motion base occasionally asked why the motion wasn't turned on. Observers noted a greater amount of subject fatique towards the end of the one and one-half hour no-motion missions presumably due to greater workload. After numerous missions, observers tend to find it more difficult to keep up with the mission when motion is not used.

Although motion aided the subjects and perhaps the observers too, it has some drawbacks. Just as a pilot can experience vertigo in an aircraft, so it can happen in the simulator, and some observers became nauseated with the conflict between sensory cues and the simulator instrument displays. Several flight observers reported noticing the washout of bank cues, and it bothered them. Some subjects reported that the sustained motion, while unlike the aircraft in a turn, did help keep them aware of direction of bank in the absence of the visual cues used in the aircraft.

The Prediction of Aircraft Pilot Proficiency

The subjects' estimates of their ability to perform the various maneuvers correlated very highly from one day to the next in the same vehicle and between vehicles, which demonstrated the stability of their confidence estimates. But these measures correlated very poorly with the subject's performance score on each of the maneuvers indicating low criterion-related validity. The main difficulty with using subjects' estimates of their abilities to predict performance in complex tasks, such as flying, is that there aren't any clearly defined anchors for good or poor performances. One can more accurately estimate his approximate score in marksmanship using a certain number of shots because the index of desired performance is clear, and the stability of one's performance is greater from trial to trial.

It appears that the prediction of vehicle performance scores from simulator performance scores might not rest simply upon the fidelity of the motion cues from simulator to aircraft, but upon what these cues do to the variability in operator performance from day to day.

Despite the teachings about vertigo, pilots still suffer spatial disorientation when operating IFR. This vertigo results in the operator's feeling that his attitude is one way while the instruments tell him that he is in a different attitude. This phenomenon is experienced naturally in flight and is induced in flight simulators through the washout motion system. Operators differ in their susceptibility to spatial disorientation from day to day for many reasons known and unknown, and there seems to be a good bit of variation between individuals.

In day-to-day performance in simulators with washout motion system, the pilots' performance scores may vary from one day to the other in different ways because of some degree of disorientation induced by the washout motion, while with sustained motion or no motion there is not so much day-to-day variation in performance. These differences in day-to-day performances as a function of motion and type of motion are not too noticeable in the instrument maneuvers in which pilots have always been instructed not to depend upon motion cues for they may induce spatial disorientation. But, visual and motion cues are more important in the performance of contact maneuvers, and with the motion cues of the "washout" type, there tends to be greater variability in subjects' day-to-day performances. This is because of a confusion between the attitudes that the body perceives and what is displayed on the instruments in the simulator. With the sustained motion in the simulator, not only are the conflicts about attitude reduced, but the motion cues tend to substitute for visual cues which do not exist in the simulator. For example, when the operator is in a right hand bank, the simulator will lean to the right, and the subject will feel that his body is tilted to the right which is in agreement with the visual cues that he would see if visual cues were provided. Thus, the cockpit instrument, the simulator cab, and the operator's feeling of a right bank would all be in agreement.

When predicting aircraft performance from simulator performance, which motion system produces the greatest errors is not as important as under which motion system is there greater stability of performance. No motion causes some day-to-day simulator variability in performance because of its lack of cues, and the washout motion causes even more variability because of its conflicting cues. Because of this, predictability of performance in the aircraft from performance in no-motion or washout-motion simulators is understandably lower than the prediction from the more stable measures of performance using the sustained motion system. The correlation coefficients from simulator to aircraft for the no motion and washout groups are significantly greater than zero and they are greater than most predictions of aviation performance found in the literature. But much better prediction of performance can be obtained using a sustained motion system.

Attempts to account for performance in the simulator or aircraft by correlating those scores to measures of the subject's age and flight experience yielded few significant correlations and only two of trends. It appears that younger pilots tend to have lower performance error scores and higher, more favorable, observer total mission ratings. Two factors may be influencing this trend. First, younger pilots tended to have more recent experience, and many of the older pilots had retired from flying months or even years before this study. Second, the observers themselves were all relatively young instructor pilots and knew many of the younger subjects. This might have had some biasing effect on the total ratings, but still, the total ratings did correlate very highly with the performance scores.

Another trend noted in the correlation of subject data with performance scores was that persons with greater amounts of flight time logged in the past six months tended to have lower error scores and higher observer ratings, particularly on Day 1 in the simulator and Day 3 in the aircraft. The skills and confidence that recent experience with the flight procedures bring was reflected on Day 1 in the simulator. However, the learning of the non-current subjects from Day 1 to Day 2 tended to reduce the significance of this factor on the second day in the simulator, but it did not affect the Day 3 aircraft correlations because the non-current subjects still had not flown an aircraft recently.

In situations where the subject pilot population is limited to a select proficient group of pilots, as with the airlines or military organizations, one may find that there is very little variation between pilots on the maneuvers used in the Pilot Performance Record, and thus the predictive correlations will be lowered. Further development of the booklet to include performance under emergency or degraded situations and systems knowledge may result in a greater spread of overall performance scores, but such expansion must be undertaken with care to insure that the items are clearly defined, relatively objective, and adequately describe the subjects' performances.

Appendix F is a selected bibliography of articles related to pilot proficiency measurement, the prediction of pilot proficiency, flight simulators, and the effects of simulator motion upon simulator performance.

CONCLUSIONS

The results of this study indicate that proficiency of aircraft pilots with instrument and multi-engine land ratings can be predicted to a high degree from ground-based simulator performance measures. Of the three simulator motion conditions used, the highest predictive validity for pilot performance in the simulator was obtained with the use of a sustained motion system. There was no significant difference between the predictive validities of performance with no motion and a washout-type motion system.

The predictive validity of simulated instrument flight performance is higher than the predictive validity of contact flight performance in a simulator with no visual reference outside the cockpit. Simulator performance on certain maneuvers, such as holding, precision approach, and non-precision approach, have significantly higher predictive validities than other maneuvers such as steep turns, chandelle, and lazy eight. Simulator motion conditions have an effect upon the predictive validities of the individual maneuvers, in that the sustained motion resulted in higher predictive validities than the no motion or washout-type motion systems; and there were no significant differences between these later two motion systems in the predictive validities of the individual maneuvers. The differences in predictive validity of instrument flying and contact flying as a function of the type of motion system used are the same as for the individual maneuvers, the sustained motion condition in the simulator results in the greatest predictive validities to the aircraft.

Simulator motion tends to increase the subject's acceptance of the device, lower performance error scores, and reduce workload on the subjects and the observers through the aiding effects of the motion onset cues. But the differential effects of motion on the simulator performance does not transfer to the performance in the aircraft. In predicting the pilot's performance in the aircraft, the magnitude of his error scores in the simulator as a function of the type of motion

used is not as important as the stability of the performance from one mission to the next. Increasing the fidelity of the simulator motion system may bring much of the variability in performance found in the aircraft into the simulated environment which was used to escape the variability of the operational environment.

One key to the high predictive validities obtained is the high reliabilities of the measures used. This experiment demonstrated that very high observer-observer reliabilities (r = .771 to .971) on the same mission can be obtained by recording performance on scales that are well defined, easy to follow, descriptive of the maneuver and behavior being recorded, and are not too demanding upon the person doing the recording of performance. Such a device for recording pilot performance can be used in operational situations with minimum specific training of the observer whether he be a flight instructor, check pilot or just a flight observer. Performance measures taken in simulators tend to be more reliable than those taken in aircraft because of the elimination of degrading environmental factors and the reduction of safety orientated duties frequently imposed upon the observer.

The recorded pilot performance measures correlated very highly with the observers' overall subjective ratings of the missions (r = .726 to .878). The observers' overall ratings correlated slightly higher with performance on instrument flight maneuvers than with performance on visual flight. Within the class of instrument flight maneuvers the approaches and the missed approach performance correlated highest with the observers' overall rating of the missions. Of the contact maneuvers, the takeoff and landing correlated highest with the observers' overall ratings.

Other possible indicies of pilot proficiency, such as the amount of multi-engine land, instrument or total time logged in the past six months, do not correlate very well with mission performance scores; in fact they correlate about as well as age.

REFERENCES

Adams, J. A. Some considerations in the design and use of dynamic flight simulators. In H. W. Sinaiko (Ed.) Selected papers on human factors in the design and use of control systems. New York: Dover, 1961.

American Airlines. Optimized flight crew training: a step toward safer operations. Fort Worth, Texas: American Airlines Flight Training Academy, April 1969.

Bale, R. M., Rickus, G. M. Jr. and Ambler, R. K. Prediction of advanced level aviation performance criteria from early training and selection variables. Journal of Applied Psychology, 1973, 58, 347-350.

Baron, M. L. and Williges, R. C. Transfer of training assessment by means of response surface methodology. Savoy, Ill.: University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Technical Report ARL-71-24/AFOSR-71-9, October 1971.

Bergeron, H. P. Investigation of motion requirement in compensatory control tasks. IEEE Transactions on Man-Machine Systems, 1970, MMS-11, (2).

Blum, M. L. and Naylor, J. C. Industrial psychology. New York: Harper and Row, 1968.

Caro, P. W., Jr. Flight evaluation procedures and quality control of training. Fort Rucker, Ala.: HumRRO Division No. 6 (Aviation), Human Resources Research Office, Technical Report 68-3, March 1968.

Chapanis, A. Research techniques in human engineering. Baltimore: Johns Hopkins Press, 1959.

Clark, B. Thresholds for the perception of angular acceleration in man. Aerospace medicine, 1967, 38, 443-450.

Computing Services Office. CSO Volume 9 (USER): Statistical Systems. Book 2: SOUPAC Program Descriptions. Champaign, Illinois: University of Illinois at Urbana-Champaign, October 1973.

Danneskiold, R. D. Objective scoring procedure for operational flight trainer performance. Port Washington, N. Y.: Office of Naval Research, Special Devices Center, Technical Report SPECDEVCEN 999-2-4, February 1955.

Darlington, R. B. Multiple regression in psychological research and practice. Psychological Bulletin, 1968, 69, 161-182.

Demaree, R. G., Norman, D. A., and Matheny, W. G. An experimental program for relating transfer of training to pilot performance and degree of simulation. Port Washington, N. Y.: U. S. Naval Training Device Center, Technical Report NAVTRADEVCEN 1388-1, June 1965.

Edgerton, H. A. and Walker, R. Y. History and development of the Ohio State Flight Inventory, Part I: Early versions and basic research. Washington, D. C.: Civil Aeronautics Administration, Division of Research, Report No. 47, July 1945.

Ericksen, S. C. Measures of two-engine flying skill (contact). In N. E. Miller (Ed.) Psychological research on pilot training. Washington, D. C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Ericksen, S. C. Objective measures of multi-engine instrument flying skill. In N. E. Miller (Ed.) Psychological research on pilot training. Washington, D. C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Ericksen, S. C. Development of an objective proficiency check for private pilot certification. Washington, D. C.: Civil Aeronautics Administration, Program Planning Staff, Report No. 95, May 1951.

Ericksen, S. C. A review of the literature on methods of measuring pilot proficiency. Lackland AFB, Texas: Human Resources Research Center, Research Bulletin 52-25, August 1952.

Fedderson, W. E. The role of motion information and its contribution to simulation validity. Fort Worth, Texas: Bell Helicopter Company, Army-Navy Instrumentation Program, Report No. D 228-429-001, April 1962. (AD 281 855)

Federal Aviation Administration. Federal Aviation regulations for pilots. North Hollywood, Calif.: Pan American Navigation Service, November 1970.

Federal Aviation Administration. Flight training handbook. Washington, D. C.: FAA, Advisory Circular AC 61-21, 1965.

Flanagan, J. C. and Gordon, T. A. Objective flight-check techniques: Scientific methods for use in the observation of flight crew requirements. Woods Hole, Mass.: Flight Safety Foundation, 1948.

Fleishman, E. A. Studies in personnel and industrial psychology. Homewood, Ill.: Dorsey Press, 1967.

Fraser, T. M. Philosophy of simulation in a man-machine space mission system. Washington, D. C.: National Space and Aeronautics Administration, 1966.

Gagné, R. M. Training devices and simulators: Some research issues. American Psychologist, 1954, 9, 95-107.

Gagné, R. M. Simulators. In R. Glaser (Ed.) Training research and evaluation. Pittsburgh, Pa. and New York: University of Pittsburgh Press, 1962 and Wiley, 1965.

Gardner, W. C. The use of confidence testing in the academic instructor course. Proceedings of the 11th Annual Conference of the Military Testing Association. New York: Military Testing Association, 1969.

Glaser, R. and Klaus, D. U. Proficiency measurement assessing human performance. In R. M. Gagne (Ed.) Psychological principles in system development. New York: Holt, Rinehart, and Winston, 1962.

Gordon, T. The development of a standard flight-check for the airline transport rating based on the critical requirements of the airline pilot's job. Washington, D. C.: Civil Aeronautics Administration, Division of Research, Report No. 85, April 1949.

Greer, G. D., Jr., Smith, W. D., and Hatfield, J. L. Improving flight proficiency evaluation in Army helicopter pilot training. Ft. Rucker, Ala. and Washington, D. C.: U.S. Army Aviation Human Research Unit and Human Resources Research Office, The George Washington University, TR 77, May 1962.

Jersild, A. Determinants of confidence. American Journal of Psychology, 1929, 41, 640-642.

Klein, G. S. and Schoenfield, N. The influence of ego-involvement on confidence. Journal of Abnormal Social Psychology, 1941, 36, 239-258.

Little, K. B. Confidence and reliability. Educational Psychology Measurement, 1961, 21, 95-101.

Mackie, R. R. Factors leading to the acceptance or rejection of training devices. Orlando, Fla.: Naval Training Equipment Center, Report NAVTRAEQUIPCEN 70-C-0276-1, August 1972.

Matheny, W. G., Dougherty, D. J., and Willis, J. M. Relative motion of elements in instrument displays. Aerospace Medicine, 1963, 34, 1041-1046.

Meyer, D. E., Flexman, R. E., VanGundy, E. A., Killian, D. C., and Lanahan, C. J. A study of simulator capabilities in an operational training program. Wright-Patterson AFB, Ohio: Aerospace Medical Research Laboratory, Technical Report, AMRL-TR-67-14, May 1967. (AD 656 308)

Miller, N. E. The problem of measuring flying proficiency. In N. E. Miller (Ed.) Psychological research on pilot training. Washington, D. C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

Prophet, W. W. and Jolley, O. B. Evaluation of the integrated contact-instrument concept for Army fixed wing flight instruction. Alexandria, Va.: HumRRO Division No. 6 (Aviation), Human Resources Research Organization, Technical Report 69-26, December 1969.

Puig, J. A. Motion in flight training: A human factors view. Orlando, Fla.: United States Naval Training Device Center, Technical Report NAVTRADEVCEN IH-177, October 1970. (AD 880 445)

Rathert, G. A., Creer, B. Y., and Sadoff, M. The use of piloted simulators in general research. Paris: NATO Advisory Group for Aeronautical Research and Development, Report 365, April 1961. (AD 404 196)

Rippey, R. M. A comparison of five different scoring functions for confidence tests. Journal of Educational Measurement, 1970, 7, 165-170.

Rolfe, J. M. Vehicle simulation for training and research. Farnborough, England: RAF Institute of Aviation Medicine, IAM-R-442, March 1968.

Shuford, E. H., Jr. Confidence testing: A new tool for measurement. Proceedings of the 11th Annual Conference of the Military Testing Association. New York: Military Testing Association, 1969.

Shuford, E. H. and Gibson, D. L. A new method for predicting performance. Lexington, Mass.: The Shuford-Massengill Corp., 1969.

Smith, J. F., Fleaman, P. E., and Houston, R. C. Development of an objective method of recording flight performance. Lackland AFB, Texas: Human Resources Research Center, TR 52-15, December 1952.

Smode, A. F., Gruber, A., and Ely, J. H. The measurement of advanced flight vehicle crew proficiency in synthetic ground environments. Wright-Patterson AFB, Ohio: Aeromedical Research Laboratory, Technical Documentary Report No. MRL-TDR-62-2, February 1962. (AD 432 038)

Trans World Airlines. Flight simulator evaluation. Kansas City, Mo.: Trans World Airlines, Flight Operations Training Department, June 1969.

U. S. Air Force, Air Training Command. Instrument flying. Washington, D. C.: Superintendent of Documents, U. S. Government Printing Office, Air Force Manual 51-37, 1 November 1971.

Veruls, D. and Obermayer, R. W. Emerging developments in flight training performance measurement. Naval Training Device Center's 25th Anniversary Commemorative Technical Journal, November 1971, 199-210.

Wilcoxon, H. C., Johnson, W., and Golan, D. L. The development and tryout of objective check flights in pre-solo and basic instrument stages of naval air training. Pensacola, Fla. and N. Y.: U. S. Naval School of Aviation Medicine and the Psychological Corporation, Joint Project Report No. NM 001 058.24.01, July 1952.