SIMPLIFYING THE MEASUREMENT OF COMPLEX SKILLS IN A TRAINING SIMULATOR

BRIAN D. SHIPLEY, JR.
US Army Research Institute

WILLIAM V. HAGIN AND VERNON S. GERLACH
Arizona State University

INTRODUCTION

The Evaluation Problem

Student pilots must learn to operate a complex system in an unstable, frequently dangerous environment. These operational factors make stringent demands on an instructor pilot (IP) as he evaluates student pilot (SP) performance in an aircraft. The IP must monitor SP behaviors, unsafe performance conditions, and dangers in the airspace. Observations in all relevant areas of performance must be remembered or recorded and then used to arrive at a meaningful evaluation of performance skill.

Clearly, the IP must process large quantities of information. As training tasks become more complex or dangerous, demands on the IP tend to increase and IPs tend to be less and less able to adequately process all the essential information. As the information processing load increases, the integrity of safety procedures, training effectiveness, and evaluation methods may be compromised.

Evaluation Methods in Pilot Training

After more than 30 years, rating scales remain the basic IP evaluation method. Rating scales are still used because they best meet the needs of training management and the operational user - the IP - without intruding seriously into the IP's operational tasks (Koonce, 1974).

As currently used, however, rating scale methods are marginally adequate for evaluations in training research and development because most observations using them lack sufficient discrimination; i.e., the observations tend to accumulate on two, or at best three points in the rating scale (Knoop & Welde, 1973). To make IP observations more discriminating in training research and development, extensive observer training and quality control programs must be developed and carried out (Horner, Radinsky, and Fitzpatrick, 1970; Koonce, 1974). This is a continuing and cumbersome process during the design and conduct of a study. Furthermore, no practical alternative to IP ratings of airborne performance has been readily available.

The increasing sophistication of flight simulators allows greater opportunity to utilize objective data recording and processing methods. When such simulators are used in both training and research, it is important that the objective measures be reliable, readily accessible, simply computed and, most importantly, meaningful to each user - IP, student pilot, and researcher.

There are a variety of objective measures that can be used with a flight simulator. In this study, objective measures were divided into two broad categories: exact and summary indicators. An exact indicator was a measure easily observed at one or a few points in the sequence of events which comprise the profile of a training task. Most exact indicators used in training research are not well suited for general applications because they are not usually derived from theoretical or general training requirements. Instead, they are selected or developed to reveal potential differences in performances which will answer some specialized research question. When that specialized need has been satisfied, all further interests in the measurements are dropped (Koonce, 1974).

By contrast, summary indicators are very popular in most training research applications. Selected to reveal general trends or overall differences, a summary indicator was a global measure computed continuously or from a large number of discrete observations made throughout a single performance. Summary indicators have been criticized because they are usually insensitive to the effects of particular events or skills in a performance (Knoop and Welde, 1973), and because they are usually based on massive quantities of data with no definitely specified relationship to the evaluation problem (Christensen and Mills, 1967; Koonce, 1974).

The purpose in this study was to investigate exact indicators of particular skills or events that could be used with a performance state evaluation model to simplify objective

Based on the senior author's dissertation, the research covered in this report was completed at Arizona State University under the provisions of Air Force Office of Scientific Research Grant No. AFOSR 75-2900 with Vernon S. Gerlach as principal investigator.

measurement procedures in pilot performances; in short, to reduce excessive detail while not being uniformatively general (Youtz and Ericksen, 1947). A step in the solution of this detail versus generality dilemma would be achieved if a general set of exact indicators were derived from a theoretical and/or training requirements analysis, and if such measures were empirically shown to carry sufficient information to evaluate performances in a training research effort. The research hypothesis was that a set of exact indicators would successfully replace a global summary indicator typically used in pilot training research.

APPROACH

The Performance State Evaluation Model

Movements of an aircraft through space and time (i.e., the flight path) can be described with values from several variables such as altitude, heading, airspeed, pitch, and power. To accomplish his mission, a pilot must control the movements of the aircraft by using information from some of these variables to make changes on others. Any set of values on these variables becomes a system "state." A state, then, is the complete set of variables needed to describe the entire performance of the pilot and the aircraft at any instant in time (Etkin, 1972). To evaluate a pilot's performance, values are observed in one or more states and if they meet certain requirements, the performance is assigned an appropriate score or mark to indicate the level of achievement.

A review of measures used in previous pilot training research produced three possibilities which might be effectively used as exact indicators. These three measures were the range, the maximum deviation from the standard, and performance time. The range and maximum deviation were established in World War II as objective measures of instrument flying skills (Hagin, 1947). Since that time, many researchers have validated the utility of these two measures.

In parallel research during World War II, performance time was also studied (Miller, 1947). Performance time appeared promising as a simplifying metric inasmuch as it was also found to be a good indicator in two recent studies. In a training experiment, Brecke, Gerlach, and Shipley (1974) found that differences in means and variances on performance time discriminated between treatment groups. In another study, Knoop and Welde (1973) obtained 87 different objective measures from aircraft data as well as ratings from pilots and student pilots on the same performances. Of the 87 different measures, 7 or 8% were time measures. When correlated with the ratings, 11 or 14% of the 77 significant correlations (.05) involved the time measures.

Performance Time Algorithm

To empirically study time and deviation measures, a situation was selected in which students in undergraduate pilot training were learning a complex instrument flight task, the Vertical S-A, in a simulator (USAF ATC Syllabus, 1975). Performance state criteria were analytically determined for seven "states" of the Vertical S-A by Brecke and Gerlach (1972). Shipley, Gerlach, and Brecke (1974) derived flight path standards and performance times from these criteria. In the present study, an algorithmic, performance state evaluation model was developed for the maneuver and performance times and deviations from the standard flight path values were then studied as potential indicators of skill in three empirical investigations.

Since it is impossible for a performance to deviate significantly from standard times and still satisfy all other performance criteria throughout an entire flight, performance times can be used with deviations from flight path standards to simplify measurement and evaluation procedures for many maneuvers. In the algorithmic model, if total time is not within given tolerances, some state or states must contain errors; these errors will be reflected as deviations from the standard. The state or states which contain the errors can be quickly located by examining the performance time for each state. For a state that fails the time test there are three possible outcomes: (a) if the entry conditions to that state deviate significantly from the standard, at least some source of the error will be in the preceding state or states; or, (b) if the entry conditions for a state are within the given tolerances, the sources of the error must be located in that state; and, (c) if the end point deviations are out of tolerance, the following state should also contain information about the pilot's performance.

Data for the three investigations were obtained from an earlier experimental study by Brecke (1975) in which measures of time, airspeed, heading, vertical rate, pitch, power, and altitude were automatically obtained, at a sampling rate of one second, from performances of the Vertical S-A maneuver in a flight simulator. (Readers interested in more detail about the experiment or the data should consult AFOSR Technical Reports #50430 and #60229 by Brecke (1975) and Shipley (1976) respectively.)

INVESTIGATION I

The first investigation was carried out to compare "standard" time values for the Vertical S-A states with observed performances from two experienced pilots. The purpose was to discover whether or not the performance state evaluation model would reveal any significant differences among the performances of these two pilots. If the model detected significant differences among performances of experienced pilots, it should also make similar discriminations among performances of student pilots. Such discriminations are essential to effective evaluation in regular training and in experimental research on the effects of training methods.

Method

Two experienced pilots performed a sequence of 12 trials, 6 each, on the Vertical S-A in a flight simulator according to procedures described by Brecke (1975). Two trials of one pilot's performances were not included in this investigation; his first trial was a descending rather than a climbing Vertical S-A and data from his last trial was unusable because of a data recorder malfunction.

Procedures

Performance time and maximum altitude data were obtained from computer printouts of these performances and the means and variances were then computed for each pilot's data on total time, times for the seven performance states, and maximum altitude. Deviations of observed from corresponding standard values were also computed and used as the basis of tolerance limits in subsequent analyses of student performances.

Results

The means and variances of the performance times revealed three significant differences. There were two significant differences (.05) between means: a difference of 14.35 seconds on total time and a difference of 10.08 seconds on the time for the climb state. Finally, the variances were not homogeneous for time in the third transition state.

INVESTIGATION II

The second investigation was carried out to determine whether total performance time and/or maximum altitude discriminated between performances of treatment groups in a training experiment. An *a priori* prediction, that differences among maximum altitude variances would discriminate among treatment group performances, was based on a training requirements analysis (Brecke and Gerlach, 1972; Gerlach, Brecke, Reiser and Shipley, 1972) and on the outcomes of a prior experiment (Brecke, Gerlach and Shipley, 1974).

Method

Thirty-nine student pilots were randomly assigned to one of five groups. Members of four experimental groups studied the objectives and different preflight instructions on how to perform the Vertical S-A; members of a control group studied only the objectives. After each subject had studied his assigned materials, he performed a sequence of six trials on the Vertical S-A according to procedures described by Brecke (1975).

Procedures

Total performance time and maximum altitude values were obtained from the Brecke data. Total performance time values were analyzed with a two between- one within-subjects analysis of variance; repetition of performance trials on the Vertical S-A maneuver was the within subjects variable. Dunnet's method (Myers, 1966) was used to compare the means of the treatment groups with the means of the control group.

It was hypothesized that subjects in groups given experimental instructions would exhibit less variability among their performances at maximum altitude than subjects in either the control group or the groups given current instruction. To test this hypothesis, *F*-ratio tests for differences between group variances were used rather than tests for differences in means, i.e., ANOVA or *t*-tests (Winer, 1972).

Results

Total time. In the ANOVA on total time, there were significant main effects for type of instruction, number of response items in the instructional material, and performance trials. A significant interaction was also found between number of response items and performance trials. There were no significant differences on Dunnett's tests between any treatment group and the control group.

Maximum altitude. As predicted, the comparisons among the treatment groups' variances were significant on the maximum altitude measure. At maximum altitude, the variances of the two groups that received experimental instructions were significantly smaller (.05) than the variances of the control group and the two groups receiving current instructions. The control group variance was not significantly different at maximum altitude than the variances of either group receiving current instruction.

Summary

To summarize the results of the first two investigations, performance times and maximum altitude discriminated among the performances of both individuals and groups. To the extent that the observed variations in performance time and maximum altitude reveal variability in the pilots' pitch and power control skills, these findings provide strong support for the hypothesis that exact indicators could be used as alternatives to summary indicators. In the third and final investigation, this hypothesis was examined by comparing the relationship between values on the two types of indicators.

INVESTIGATION III

In this last investigation, the hypothesis was that a set of exact indicator values would predict values of a summary indicator. If scores from a summary indicator in the Brecke (1975) data could be predicted with scores from a set of exact indicators, evidence would be obtained to support a recommendation to use the exact indicators in place of the summary indicator. A multiple regression analysis was used to test this hypothesis.

Method

To evaluate the results of his experiment, Brecke (1975) used a summary indicator of variability in pilot performances called error amplitude. Error amplitude scores in Brecke's study were computed with procedures described by Shipley, et al. (1974). Conceptually, error amplitude is a normalized root mean square measure obtained with the deviations of each observation from the corresponding standard and a set of tolerance limits for each parameter. In the present investigation, error amplitude scores from Brecke's data were used as the criterion variable.

The exact indicators used as variates in the regression analysis were deviations from the standard on the performance state times, total time, and maximum altitude. Three combinations of performance state times were also included as variates in the regression design; these combinations were (a) the sum of the first three performance state times; (b) the sum of the last three performance state times; and (c) the sum of the times for all seven performance states. These sums were included in the design to test whether gross measures of time would be better predictors than more specific measures. Prior experience measures included as variates in the design were total hours as a pilot, flying time in undergraduate pilot training, total hours in the T-4 simulator, and number of minutes to complete the instructional materials in the Brecke (1975) experiment.

Procedures

A set of 30 correlation matrices formed the basis of the regression analysis: one matrix for each of six performance trials across the five groups in Brecke's experimental design. The correlation coefficients were averaged over the five groups, after a Fisher's r to Z transform (Hays, 1963), and the means were then converted back to equivalent r values. The result was a set of six matrices of average correlations, one matrix for each trial. Finally, each of these matrices were submitted to a sequence of two stepwise regression analyses. The first analysis was based on the average correlations from the exact indicators and the sums; the second was based on the average correlations from the exact indicators, the sums, and the prior experience measures.

Results

In the results of the regression analyses, support would be obtained for the hypothesis if two criteria were satisfied: (a) if a small set (4 to 6) of the 12 exact indicators were consistently selected in the regression equations; and (b) if equations using these consistently selected indicators would account for a substantial proportion of the variance (50% or more) in the criterion.

The outcomes of the regression analyses confirmed the hypothesis. It was found that, of 16 possible indicators and experience measures, a set of 9 were frequently selected in the regression analyses. Various combinations from this set of nine were selected in equations of just five variables for each of the trials and these five variable equations accounted for 62% or more of the variance in the summary indicator, error amplitude.

The type of variable selected in these equations with five variables changed across the trials. For the first two trials, prior experience and combined times were predominant in the equations. By the last two trials, the exact indicators were predominant.

As measures of learning over trials, the summary indicator and the exact indicators were compared by making a plot of the means of each variable across the trials. It was found that deviations from maximum altitude exhibited virtually the same learning trend over trials as error amplitude (r = .98).

These results provided convincing evidence to support the hypothesis that in the Vertical S-A, exact indicators could be used to replace a summary indicator.

DISCUSSION

In these analyses, exact indicators were found more sensitive to the effects of different experimental treatments than a summary indicator. In addition, when used alone, a set of exact indicators were found to account for moderate (34%) to large (82%) proportions of the variance in a summary indicator. These findings are provocative because they suggest that a few well chosen observations, i.e., fewer data points, can be used to obtain evaluations superior to those obtained from typical summary indicators based on many data points.

The results of these investigations also support the use of a performance state evaluation model. Performance times and deviations from maximum altitude were found to discriminate precisely between both individual and group performances and to identify the sources of errors within a given individual performance. It is interesting that performance time for the transition from climb to maximum altitude was consistently found among the predominant predictors in the regression analyses. In previous research, Gerlach, et al. (1972) predicted, on the basis of a theoretical and a training requirements analysis, that this location was the most likely source of student performance errors in the Vertical S-A. Coupled with the demonstrated sensitivity of deviations from maximum altitude, it is clear that the evaluation model could be effectively utilized for performance of the Vertical S-A in a flight simulator.

For the operational user, these findings suggest that exact indicators can be used to replace summary indicators and also global ratings. Maximum altitude, an example of an exact indicator in the Vertical S-A, was found to carry nearly all the information given by error amplitude, a summary indicator. Since maximum altitude in the Vertical S-A occurs at a well defined point, IPs could easily observe it and use its value to replace a global rating. The dilemma of excessive detail versus uninformative generality in pilot training measurement and evaluation can be given a workable solution to the extent that similar values and their locations can be derived for other pilot training maneuvers.

REFERENCES

Brecke, F. H. Cues and practice in flying training (Tech. Rep. No. 50430). Tempe: Arizona State University, College of Education, April, 1975.

Brecke, F. H., & Gerlach, V. S. Model and procedures for an objective maneuver analysis (Tech. Rep. No. 21201). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Brecke, F. H., Gerlach, V. S., & Shipley, B. D. Effects of instructional cues on complex skill learning (Tech. Rep. No. 40829). Tempe: Arizona State University, College of Education, August, 1974.

Christensen, J. M., & Mills, R. G. What does the operator do in complex systems? Human Factors, 1967, 9, 329-340.

Etkin, B., Dynamics of atmospheric flight. New York: John Wiley, 1972.

Gerlach, V. S., Brecke, F. H., Reiser, R., & Shipley, B. D. The generation of cues based on a maneuver analysis (Tech. Rep. No. 21202). Tempe: Arizona State University, Instructional Resources Laboratory, December, 1972.

Hagin, W. V. Objective measures of single-engine instrument flying skill at the basic level of flying training. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D. C.: Army Air Force Aviation Psychology Program, Research Report No. 8, 1947.

Hays, W. L. Statistics for psychologists. New York: Holt, Rinehart & Winston, 1963.

Horner, W. R., Radinsky, R. L., & Fitzpatrick, R. The development, test, and evaluation of three pilot performance reference scales (Tech. Rep. No. 70-22). Air Force Systems Command, Brooks Air Force Base, Texas: Air Force Human Resources Laboratory, August, 1970.

Knoop, P. A., & Welde, W. L. Automated pilot performance assessment in the T-37: A feasibility study (Tech. Rep. No. 76-6). Air Force Systems Command, Wright-Patterson Air Force Base, Ohio: Air Force Human Resources Laboratory, April, 1973.

Koonce, J. M. Effects of ground based simulator motion conditions upon prediction of aircraft pilot proficiency. Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, 1974.

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.

Myers, J. L. Fundamentals of experimental design. Boston: Allyn & Bacon, 1966.

Shipley, B. D. An automated measurement technique for evaluating pilot skill

(Tech. Rep. No. 60229). Tempe: Arizona State University, College of Education, February, 1976.

Shipley, B. D., Gerlach, V. S., & Brecke, F. H. Measurement of flight performance in a flight simulator (Tech. Rep. No. 40830). Tempe: Arizona State University, College of Education, August, 1974.

Winer, B. J. Statistical principles in experimental design. New York: McGraw-Hill, 1971.

Youtz, R. P., & Ericksen, S. C. Analysis of the pilot's task. In N. E. Miller (Ed.), Psychological research on pilot training. Washington, D.C.: Army Air Forces Aviation Psychology Program, Research Report No. 8, 1947.

ABOUT THE AUTHORS

DR. BRIAN D. SHIPLEY, JR. is a Research Psychologist with the US Army Research Institute for the Behavioral and Social Sciences (ARI). He is assigned to the Fort Rucker, Alabama Field Unit of ARI where he is primarily responsible for research on the development of a simulator based testing system to improve methods for the selection of students into the Army's rotary wing pilot training course. Prior experience includes teaching in the public schools of Colorado, Arizona, and Oregon and doing research work on a program of pilot training. He holds an A.B. degree in education from the University of Northern Colorado, an M.A. degree in education from Arizona State University, an M.S. degree in psychology from the University of Oregon, and the Ph.D. in education from Arizona State University.

DR. WILLIAM V. HAGIN is a professor at Arizona State University. Prior to that, he was Technical Director of the Flying Training Division of the Air Force Human Resources Laboratory at Brooks Air Force Base, Texas. He was also a Staff Scientist and Department Manager for Philco-Ford Company at Palo Alto, California. Formerly, he was a Colonel in the U.S. Air Force where in his last position, he was Laboratory Director at AFSC Electronics Systems Division, L.G. Hanscom Field. Bedford, Massachusetts. He holds a B.S. degree in chemistry and secondary education from the University of Denver and the M.S. degree and Ph.D. in experimental psychology from the University of Texas. He is a member of the Human Factors Society and the Scientific Research Society of America.

DR. VERNON S. GERLACH is a Professor of Education at Arizona State University. Former positions in education include: Professor of Education at University of Minnesota; Director of Audiovisual and Instructional Materials Center at Washington District Schools, Phoenix; Elementary School Principal at Immanuel School, Mankato, Minnesota; Professor of Education and Director of Student Teaching at Bethany College, Mankato, Minnesota; and others. He holds the M.A. degree in elementary education from the University of Minnesota and the Ed.D. from Arizona State University in elementary education. He is a member of the American Association for the Advancement of Science and the American Educational Research Association.