AUTOMATED SCORING OF INSTRUMENT FLIGHT CHECKS

H. KINGSLEY POVENMIRE and LCDR KENT M. BALLANTYNE
U. S. Coast Guard, Aviation Training Center

U. S. Coast Guard helicopter pilots receive annual instrument and emergency training in the Variable Cockpit Training System simulator. Many portions of initial aircraft transition training are also conducted in the simulator. Initial instrument ratings as well as annual instrument renewals are given on the basis of simulator checkrides, which are automatically scored by the computer. Scores are kept for as many as 12 parameters at a time. This paper discusses the first three years of experience with automated scoring. Moderate correlations were found between subjective instructor scores and automated scores. Although not useful for individual pass-fail decisions, normative comparisons of the automated scores are responsive to changes in the training program.

BACKGROUND

Coast Guard helicopter pilot training has been conducted at the Aviation Training Center in Mobile, Alabama, since 1967. A study was conducted in 1969 to determine future Coast Guard aviator training requirements (Hall *et al*, 1969). After a thorough investigation of operational pilot tasks, several training goals were identified which had general applicability and might best be taught in a central location. Each training goal was evaluated with respect to simulation technology available. Decisions were then made concerning which simulator capabilities might be implemented to enhance training of helicopter pilots (Caro *et al*, 1969). For example, very little saving of flight time was envisioned through use of a visual system because of its limited ability to simulate the water environment. On the other hand, many helicopter malfunctions would require all six degrees of freedom in motion simulation.

Based on these two papers, the entire training program was revised. Concurrently, a flight simulator was designed and built. Both efforts came together in 1973 as the Variable Cockpit Training System (VCTS) became operational. The VCTS utilizes a highly sophisticated, computer-based flight simulator built by Reflectone, Inc., which consists of two separate cockpits. Each cockpit duplicates one of the two different types of Coast Guard helicopters.

SYSTEMS APPROACH TO TRAINING

Implementation of the VCTS was a major shift away from the traditional concept of aviator training towards a program based on the systems approach to training. The major concept underlying this new training system has been in the area of aircraft systems knowledge. The present methods stress teaching the student how to operate the system as opposed to giving him detailed explanations about how the system operates.

INSTRUCTOR AS TRAINING MANAGER

The key to success in using this new approach has been the instructor's background and the role he plays. Instructors are carefully selected from a group of operational pilots who have expressed a desire to come to the training center. In addition to extensive operational experience, each instructor pilot receives instruction in managing the training progress of his students. While working with a group of students, the instructor becomes involved in all facets of the course of instruction from briefing, through the simulator, to the aircraft.

PEER TRAINING

One instructor and two students work together as a team. Each student spends half the time in the copilot seat and the other half in the pilot seat during simulator training periods. Simulator flights then become primarily practice and validation sessions where students must react as a team to situations covered in the briefing and reading assignments.

For those courses involving actual flights, an additional instructor pilot is also provided. He flies in another aircraft with the second student. This team stays together until the pilots have reached course standards in the basic helicopter maneuvers, instrument and emergency procedures and systems knowledge. The students then fly with other instructors for training in operational maneuvers and for aircraft and simulator check flights.

Briefing and debriefing sessions are conducted in a dialectic manner. Questions are used to lead students from one concept

to another. In many cases one student may be able to shed some light on a point of confusion. Since the program makes use of the proficiency based learning concept, the length of any given course will vary considerably depending on the student's past experiences and his ability to progress through the new material.

The methods now in use have resulted in overwhelming acceptance by the students. They invariably praise the concept of personalized self-paced learning.

TRAINING COURSES

Although the facilities and methods described have been used for a variety of purposes, the major emphasis is on recurrent training. Transition training and initial qualification training are also conducted on a regularly scheduled basis. Students in all courses are qualified pilots.

Pilots who are current in one or both of the helicopter types come to Mobile annually for recurrent training. This involves one week of intensive training in instrument and emergency procedures conducted completely in the simulator. Upon completion, the pilot's instrument rating is renewed for another year. Approximately 500 pilots a year go through this course in one cockpit or the other.

Transition training is offered in each specific type of helicopter. Pilots who have never flown the helicopter to which they are being assigned take from three to six weeks to fully qualify, depending largely on their rate of learning and previous experience. Approximately 30 pilots go through this course in each helicopter annually.

In the third major course, fixed-wing pilots are given their initial helicopter qualification in the HH-52A single engine helicopter. Approximately 18 pilots per year go through this course. The end-of-course objectives for the Qualification Course are the same as for the HH-52A Transition Course. For planning purposes, an additional week is allowed to acquire the basic helicopter skills.

The switch in training philosophies and the implementation of the VCTS have resulted in a much higher quality of training at a reduced cost. Emergency procedures are practiced in the simulator that are not possible in the aircraft. The simulator operating costs are relatively stable at $65 per hour while aircraft operating costs are high and rising rapidly. Cost figures for Coast Guard helicopters are roughly ten times as much for the HH-52A and 15 times as much for the HH-3F.

OBJECTIVE SCORING TECHNIQUE

Recent interest in objective scoring of pilot performance has grown out of some longstanding uncertainties. In the one-on-one evaluation situation imposed by most training aircraft it is impossible to achieve totally objective performance scores. Training supervisors have had to make judgments concerning quality of training and student progress based on scores collected by a number of instructor pilots.

Many efforts have been made through the years to increase the objectivity of scoring by check pilots (Smith *et al*, 1952; Povenmire *et al*, 1973; Koonce, 1974). In every case it was found that reliability between observers increased when scores were based on instrument readings rather than outside references or quality judgments. With present computer technology,it is quite simple to automatically score instrument readings much more accurately than can be done even with an expert human observer.

The Coast Guard has been collecting data on a computer scored operational instrument checkride for over two years. We will herein report some preliminary results and point out some pitfalls to others presently designing similar systems.

Automated Checkride

The automated checkride incorporated in the VCTS simulator records frequency and time out of tolerance for 12 flight parameters. A simple computer language called SANSKRIT is supplied by Reflectone, Inc., to allow Training Division personnel to design checkrides.

Each maneuver is broken down into segments for scoring purposes. Errors are accumulated and stored for as many as nine segments of a maneuver. Each segment may be further divided into as many as 15 blocks to ensure proper sequencing through the segment. The checkride is automatically sequenced when the student reaches the "END CRITERIA" of each block.

Maneuvers are ordered in such a way as to simulate an operational instrument mission. The instructor acts as an FAA controller giving clearances and weather information which would be heard on an actual flight. He also

records information which will not show up on the recorded checkride such as cockpit procedures, planning and crew coordination. Prior to the checkflight, the student plans the flight as he would in the real world including all required paperwork.

A typical checkride commences with an instrument takeoff from Bates Field in Mobile destined for Gulfport 30 miles away. The student in the pilot seat is automatically graded on his takeoff, his airways tracking as he flies to Gulfport, and his holding procedures simulating delays for traffic or weather. He will make several approaches down to Minimum Descent Altitude. He is assisted by his flying partner acting as copilot. Malfunctions are automatically inserted along the way to allow the instructor to subjectively grade both crewmembers in their teamwork, judgment and trouble-shooting strategy. From an experimental standpoint, the copilot provides an unwanted source of variance. However, greater realism and training value are achieved by having two students act as a crew.

Establishing Criteria

Special attention must be paid to designing "end criteria" for each checkride block in order to assure proper sequencing. End criteria must be met not only by any acceptable procedures, but by any conceivable procedure. If poor performances do not produce a completed record, the lower end of the distribution will be eliminated.

"Scoring criteria" must be developed to allow any of the acceptable alternatives to be scored equally. Many operational instrument maneuvers, such as procedure turns, allow alternative procedures which are equally correct. This limits the number of items that can be scored during a given block.

The computer automatically prints out a record of error scores for each maneuver at the end of the checkride provided that all end criteria have been met. These records are retained in permanent storage for complex data analysis and reporting.

ANALYSIS AND DEVELOPMENT

Several questions must be answered before this data can be used. First, how many records must be collected before the data reaches an acceptable level of reliability? Second, do error scores discriminate between differences in pilot ability? Third, are these scores sensitive to changes in the training program?

Reliability

Each quarter we would analyze the data from all checkrides stored to date to evaluate the stability of the mean error scores. Oddly, these means kept going down rather than merely being unstable. We then ran separate analyses grouping students chronologically, each group of 20 students representing roughly eight to ten weeks. No student was scored twice during this period. There was a high correlation between time and group mean scores ($r = -.98$). The results of an analysis of variance indicate a high reliability of this effect as shown in Table 1. This phenomenon has held true in five of the six checkrides currently developed.

This can be attributed to two factors. Pilots completing the training would presumably discuss the sequence of events back at their home unit to the benefit of those pilots yet to come. This was partially counteracted by the freedom of instructors to substitute malfunctions of the same general type for those pre-programmed. Secondly, as instructors became efficient in the role of training manager and more familiar with the checkride sequence, they became better able to train students in the skills measured by the automated scoring system.

TABLE 1. ANALYSIS OF VARIANCE OF MEAN TOTAL ERROR SCORES FOR CHRONOLOGICAL GROUPINGS OF STUDENTS ON HH-52A CHECKRIDE 1.

	N	$\bar{X}$	SD
Feb 74-May 74	24	1698.2	711.2
Jun 74-Aug 74	21	1449.9	629.2
Sep 74-Jan 75	22	1141.3	569.9
Feb 74-Apr 75	20	1053.8	388.3
All	87	1350.1	638.6
	df 83 3	F=5.51	$p<.01$

Validity

A comparison was made between a group of nine instructor pilots, who are responsible for determining standard operating procedures, and a group of 23 operational pilots on their annual instrument checkride. This comparison showed significant differences favoring the instructors in total error score ($p<.001$) and in 38 of the 58

individual parameter scores ($p<.01$). Of the twenty remaining parameters scores where no significant differences were found, all but two favored the instructors (Povenmire, 1974).

The objective scores did seem to differentiate between two groups known to have different levels of ability but the question remains as to whether higher error scores indicate progressive poorer overall performance. There can be no doubt that the simulator computer can measure what it measures with unfailing accuracy. However, scoring parameters were restricted to those that a well-qualified instructor would consider critical if violated. As previously mentioned, the number of scoring criteria were further restricted to allow all correct procedures to be scored equally. These two factors severely limited the number of data points that could be scored.

This brings up an important concept. Totally objective evaluation is possible only when limited to those performance criteria that can be expressed in quantitative terms. Much to the dismay of those who write behavioral objectives, judgmental behaviors cannot be quantified. Such major areas as troubleshooting strategy, reordering of priorities and coordination of flight crew can only be evaluated subjectively.

It is generally assumed that some correlation exists between ability to accurately perform standardized maneuvers and judgmental maturity and sophistication. To test this assumption, we asked the instructors to give a single subjective grade using a four point scale for the entire checkride. This was done prior to looking at the computer printout of error scores.

The correlation between total error scores and instructor grades on each current checkride is shown in Table 2. A perfect correlation would be -1.0 due to the fact that higher subjective grades indicated better performance and higher computer generated error scores indicated poorer.

TABLE 2. CORRELATION BETWEEN TOTAL ERROR SCORE AND SUBJECTIVE GRADES.

	Checkride	r
HH-3F	3	-.46
	4	-.56
HH-52A	3	-.63
	4	-.70

These correlations are somewhat lower than might be expected between two instructors subjectively grading the same performance (Povenmire, 1970; Koonce, 1974). They are also lower than those found by Knoop (1973) using a complex computed function of instructor pilot performance. Higher correlations might be achieved with future refinement discussed below.

Distribution of checkride scores in each grade category shown in Figure 1 illustrates the wide areas of overlap between various categories.

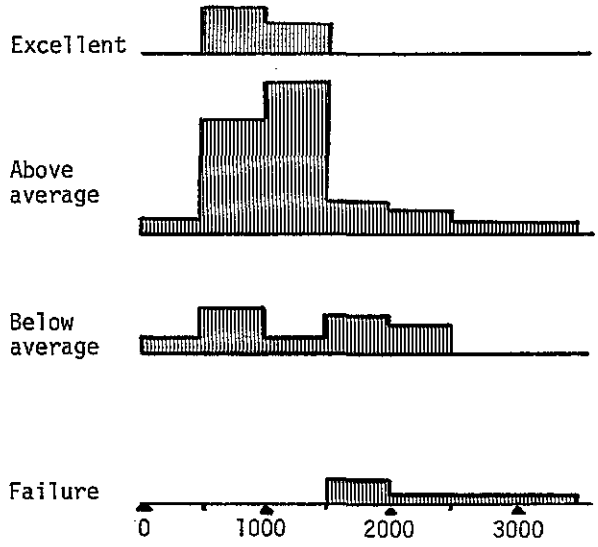

Figure 1. Distribution of checkride error scores in each subjective grade category.

Clearly, automated checkride scores cannot be the sole basis for passing or failing the final checkride. They can be used however to assist instructors in making pass-fail judgments. The greatest benefit of these objective scores involves their use in evaluating long-term stability of the training program.

Sensitivity

Beginning in January 1976, the instrument renewal course was changed by giving the automated checkride on the first flight

with no warm-up instead of at the end of the course. If performance was satisfactory, the emphasis shifted to emergency training. If the initial check showed weaknesses in instrument flying the instructor-manager continued to emphasize IFR procedures until the student performance met criteria. In either case, student performance on instruments and certain mandatory malfunctions met end-of-course criteria as before.

This change in the training sequence was immediately reflected in the mean error scores for both checkrides in both cockpits as shown in Figure 2.

DISCUSSION

Although limited in scope to mechanical piloting ability this automated performance scoring system was responsive to this major change in the training system. After each such change, however, a period of time is required to collect enough data to evaluate the effect. Data shown in the 1976 portion of Figure 2 were collected over an 18 week period.

Giving the recorded checkride first has given us a better indication of general operational readiness of Coast Guard pilots. It has also given the instructor-manager more information on student capabilities and therefore more flexibility in the conduct of training. Students and instructors alike have expressed nearly unanimous approval for this sequence.

On the other hand we have eliminated the ability to judge the effectiveness of the training program itself. Course requirements will not permit giving two fully automated checkrides during this one-week course. We are currently developing a very short checkride to be given at the end of the week, consisting of three maneuvers in which aircraft control is weighted very heavily as opposed to judgmental behaviors.

Refinements

Several areas for further development may provide more accurate descriptions of total performance. There is presently no differential weighting applied to the frequency and time out of tolerance on various parameters except by reducing tolerances at critical points. A student receives the same number of error seconds if he exceeds his cruising altitude by 100 feet as he does by going 200 feet below minimums on an ILS approach. Conversion to modified standard scores using standard deviations from an optimum (Koonce, 1974) seems promising as it

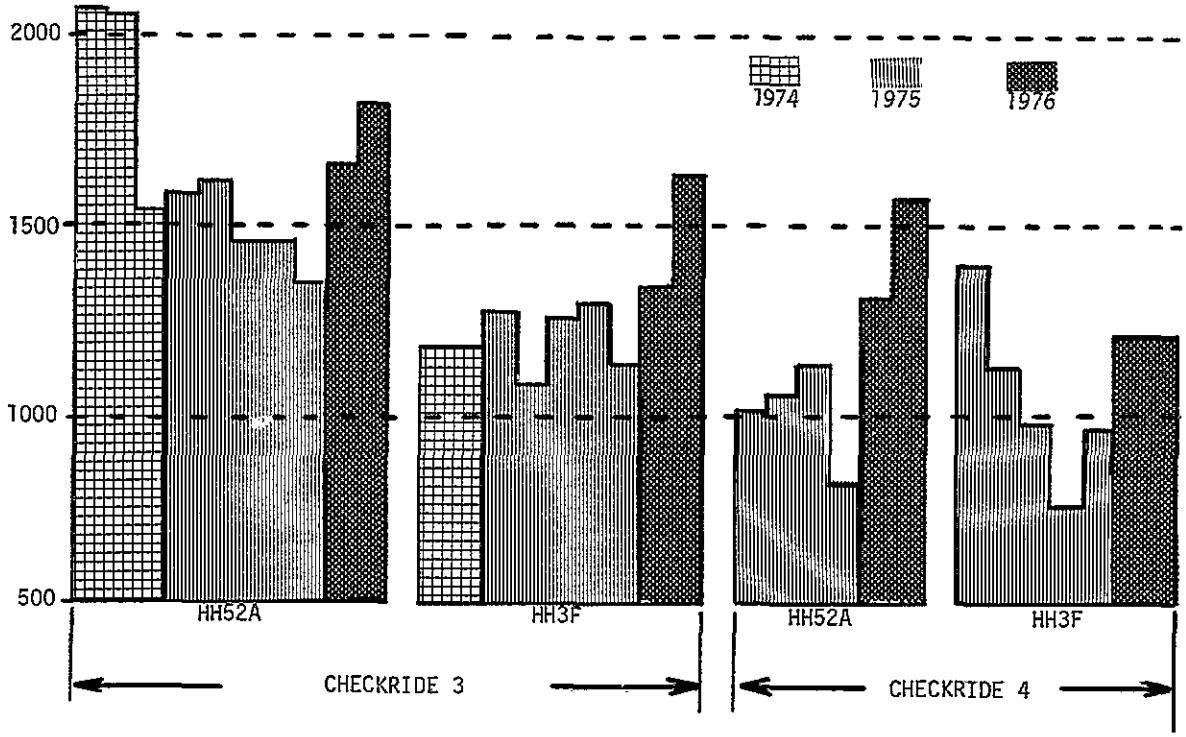

Figure 2. Mean scores for each ten classes in sequence for each current checkride.

tends to weight each parameter according to how much emphasis is placed on it by the sample of operational pilots. For this purpose, error scores would be more meaningful in terms of maximum deviation and root square error rather than time and frequency out of tolerance.

Greater reliability might be achieved in a more controlled flight profile than is possible here. In certain phases of pilot training judgmental behavior need not be evaluated. End-of-phase objectives can be fully described by performance on a series of standardized maneuvers where aircraft control parameters are most important.

CONCLUSIONS

Although we have relinquished more precise control of the test situation in favor of more realism in the test environment, results show that automated objective scoring is possible. As more data are collected adding increased stability,we should be able to evaluate instrument proficiency of Coast Guard aviators on a long-term basis. With the addition of a five minute checkride at the end of the course,we will regain an objective means for monitoring the training program itself.

REFERENCES

Caro, Paul W., Jr.; Hall, Eugene R.; and Brown, CMDR Gilbert E., Jr. Design and procurement bases for Coast Guard aircraft simulators. Ft. Rucker, HumRRO Technical Report 69-103, Dec. 1969.

Hall, Eugene R.; Caro, Paul W., Jr.; Jolley, Oran B.; and Brown, Gilbert E., Jr. A study of U.S. Coast Guard aviator training requirements. Ft. Rucker, HumRRO Technical Report 69-102, Dec. 1969.

Knoop, Patricia A. Advanced instructional provisions and automated performance measurement. *Human Factors*, 1973, *15*, 583-597.

Koonce, Jefferson M. Effects of ground based aircraft simulator motion conditions upon prediction of pilot proficiency. Savoy, IL: Technical Report ARL 74-4, University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, April 1974.

Povenmire, H. Kingsley, Alvares, Kenneth M., and Damos, Diane L. Observer-observer flight check reliability. Savoy, IL: Technical Report LF-70-2, University of Illinois at Urbana-Champaign, Institute of Aviation, Aviation Research Laboratory, Oct. 1970.

Povenmire, H. Kingsley and Roscoe, Stanley N. Incremental transfer effectiveness of a ground based flight trainer. *Human Factors*, 1973, *15*, 534-542.

Povenmire, H. Kingsley. Automated performance measurement used in quality control of the U.S. Coast Guard aviator training system. Oklahoma City, *Proceedings*, 16th Annual Conference of the Military Testing Association, 243-256, October 1974.

Smith, James F., Flexman, Ralph E., and Houston, Robert C. Development of an objective method of recording flight performance. Lackland AFB, TX, Technical Report 52-15, Human Resources Research Office, Dec. 1972

ABOUT THE AUTHORS

MR. H. KINGSLEY POVENMIRE is Chief of the Training Analysis and Support Branch of the U.S. Coast Guard Aviation Training Center. He serves as a technical advisor to the Chief of Aviation Training on matters affecting the quality of training. He is responsible for developing a quality control program for Coast Guard aviation training. He was formerly a research associate and head of the flight operations group at the Aviation Research Laboratory of the Institute of Aviation, University of Illinois. Prior to joining the research staff, he was a flight instructor at the University of Illinois for three years. In September 1968, he was appointed supervisor of the basic flight instruction group. He received his B.A. from San Diego State College in 1960, and his M.S. in Education from the University of Illinois in 1972.

LCDR KENT M. BALLANTYNE, U.S. Coast Guard, has been assigned to the Coast Guard Aviation Training Center in Mobile, Alabama since 1971 serving as an HH52A helicopter flight instructor. Also, since August of 1975, he has been Chief of the Synthetic Training Branch which involves supervising the administration and maintenance of the Variable Cockpit Training System. During previous tours of duty, he was a research and rescue pilot at Coast Guard Air Stations in Salem, Massachusetts and Annette, Alaska. Prior to flight training, he was a deck watch officer aboard the Coast Guard Cutter, USS CAMPBELL, for two years. He holds a B.S. degree from the Coast Guard Academy in New London, Connecticut and an M.B.A. degree from the University of South Alabama in Mobile.