DOCUMENT RESUME

ED 226 711 IR 010 591

AUTHOR McDaniel, William C.; And Others
TITLE Evaluation of the Computer Aided Training Evaluation and Scheduling (CATES) Decision Model for Assessing Flight Task Proficiency.
INSTITUTION Naval Training Analysis and Evaluation Group, Orlando, Fla.
REPORT NO TAEG-TR-130
PUB DATE Sep 82
NOTE 59p.; For related document, see ED 201 316.
PUB TYPE Reports - Evaluative/Feasibility (142)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Competency Based Education; *Computer Managed Instruction; *Decision Making; Efficiency; Evaluation Criteria; *Evaluative Thinking; *Flight Training; *Mathematical Models; Postsecondary Education; Student Evaluation; Training Methods
IDENTIFIERS *Naval Training

ABSTRACT

The efficacy of the CATES system for making training decisions and determining student proficiency in Naval in-flight training proposed in an earlier study (Rankin and McDaniel, 1980) is compared with the present system of instructor judgments for performance assessment. The current study used 29 newly-designated naval aviators undergoing Fleet Replacement Pilot Training in the SH-3 aircraft. From an inventory of 190 tasks, 18 tasks were selected to evaluate the model. Standard training materials and equipment were used, and performance was graded as students proceeded through the training syllabus. The task performance information required to reach a decision and the level of student proficiency upon completion of the training program were then analyzed. Results indicate the CATES system requires less information to make a decision than the current human-judgment system and is reliably more accurate, suggesting greater consistency and accuracy of mathematical models to an actual training situation in a considerably more unstructured environment than previous research studies. This report also provides extensive details of the statistical decision model used by the CATES system, results of other evaluations using mathematical models versus human models in decision making, study definitions, research methods, results, and conclusions. Six appendices include related materials and 30 references. (LMM)

Technical Report 130

EVALUATION OF THE COMPUTER AIDED TRAINING EVALUATION AND SCHEDULING (CATES) DECISION MODEL FOR ASSESSING FLIGHT TASK PROFICIENCY

William C. McDaniel
Betty M. Pereyra
William C. Rankin
Paul G. Scott

Training Analysis and Evaluation Group

September 1982

Alfred F. Smode

ALFRED F. SMODE, Ph.D., Director
Training Analysis and Evaluation Group

W. L. MALOY, Ed.D.
Deputy Chief of Naval Education and Training for Educational Development and Research and Development

ACKNOWLEDGMENTS

Thanks are extended to Mr. Gary Hodak, Mr. Robert F. Browning, and Dr. Myron M. Zajkowski of the Training Analysis and Evaluation Group for their helpful suggestions during this evaluation. The constant support and cooperation of CAPT B. A. Spofford, Commander, Helicopter Antisubmarine Wing One; CDR L. Lewis, Commanding Officer, Helicopter Antisubmarine Squadron ONE (HS-1); and LCDR Jon Browning, HS-1 Training Officer, are expressly acknowledged. Without their interest, gathering the substantial amount of data required for analyses would have been impossible. The cooperation of the instructor pilots of HS-1 and the student replacement pilots undergoing flight training at HS-1 is gratefully acknowledged.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER: Technical Report 130
2. GOVT ACCESSION NO.:
3. RECIPIENT'S CATALOG NUMBER:

4. TITLE (and Subtitle): EVALUATION OF THE COMPUTER AIDED TRAINING EVALUATION AND SCHEDULING (CATES) DECISION MODEL FOR ASSESSING FLIGHT TASK PROFICIENCY

5. TYPE OF REPORT & PERIOD COVERED:

6. PERFORMING ORG. REPORT NUMBER:

7. AUTHOR(s): William C. McDaniel, Betty M. Pereyra, William C. Rankin, and Paul G. Scott

8. CONTRACT OR GRANT NUMBER(s):

9. PERFORMING ORGANIZATION NAME AND ADDRESS: Training Analysis and Evaluation Group
Department of the Navy
Orlando, FL 32813

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS:

11. CONTROLLING OFFICE NAME AND ADDRESS:

12. REPORT DATE: September 1982

13. NUMBER OF PAGES: 60

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 is 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)

Computer Managed Prescriptive Training
Computer Aided Training Evaluation and Scheduling (CATES) System
Scheduling
Pilot Proficiency
Decision Model
Proficiency Grading
Sequential Sampling Decision Model
Pilot Training
Performance Measurement

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

Determining student performance level and subsequent decisions to either continue or stop training has posed a perplexing problem for instructors and training managers who provide pilot training. In-flight pilot training involves both highly skilled human resources as well as sophisticated equipment. Therefore, training continued beyond established training objectives is costly. However, terminating training before the student pilot achieves the required skills is highly undesirable. (continued on reverse)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE
S/N 0102-LF-014-6601

20. ABSTRACT (continued)

A previous study (TAEG Report No. 94) proposed a Computer Aided Training Evaluation and Scheduling (CATES) system to improve proficiency judgments during in-flight training. This present study compared the efficacy of the CATES system with the present system of "human judgment" for assessing performance in flight training with regard to efficiency in reaching decisions and quality of decisions. The study also demonstrated that the CATES system can be used with some advantage in an actual flight training program.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION I

INTRODUCTION

Determination of student performance level and, subsequently, decisions to either continue or stop training have posed a perplexing problem for instructors and training managers. This problem is especially troublesome for instructors and training managers providing pilot training. In-flight training for pilots requires considerable resource expenditures involving both highly skilled human resources as well as sophisticated equipment. Training is generally accomplished by a one-on-one instructor-student relationship. Thus, training continued beyond established training objectives is costly. However, termination of training before the student pilot achieves the skills required of him in the precise aviation environment is also highly undesirable.

Rankin and McDaniel (1980) proposed a Computer Aided Training Evaluation and Scheduling (CATES) system for achieving improvements in the precision of proficiency judgments and in determining student proficiency during in-flight training. This method provides a computer managed, prescriptive training program based on individual student performance. The CATES system uses a proficiency grading system developed by Browning, Ryan, Scott, and Smode (1977). These grades are then evaluated as they are awarded using a sequential sampling technique as a means for making statistical decisions with a minimum sample introduced by Wald (1947). According to Rankin and McDaniel (1980), the conceptual CATES decision model augurs well with the present system of instructor judgments. What remains is to assess the efficacy of the CATES decision model using actual data and to determine from this assessment if the CATES system offers some practical advantage.

PURPOSE

The objectives of this study are twofold. The first objective is to compare the efficacy of the CATES system with the present system of "human judgments" for performance assessment in flight training with regard to:

- efficiency in reaching decisions,
- quality of decisions.

Increased efficiency in reaching training decisions; e.g., reduced information requirements to determine when to stop training, could result in significant reductions in training costs. Increased quality of training decisions would produce a more effective utilization of training resources and reduce the risk of incorrect decisions; e.g., the decision is made to stop training when additional training is needed. The second objective is to demonstrate that the CATES system can be used with some advantage in an actual flight training program.

ORGANIZATION OF THE REPORT

In addition to this introduction, four sections and six appendices are presented. Section II presents the development of the statistical decision

model used by the CATES system and results of other evaluations using mathematical models versus human models in decision making. Section III presents the method used for comparing the CATES system decision model with the present system of decision making and the operational definitions used in this evaluation. Section IV presents the results and comparisons of efficient use of information in reaching decisions and the quality of the decisions as evidenced by performance on a final flight evaluation. Section V presents a discussion of the results and formulates conclusions based on the findings with recommendations for further applications of the CATES system.

Appendix A contains a description of the Wald Binomial Probability Ratio Test. Appendix B is a listing of the tasks and respective task parameters that were used in this evaluation. Appendix C contains the tasks used to evaluate decision efficiency as a function of difficulty. Appendix D contains a sample grade card used for data recording. Appendix E contains a copy of the Naval Air Training and Operating Procedures Standardization Program (NATOPS) Evaluation Worksheet. Appendix F contains the mathematical equation used for estimating trials to a training decision for the CATES decision model.

SECTION II

DEVELOPMENT OF CATES DECISION MODEL

NEED FOR ACCURATE PROFICIENCY ASSESSMENT

Simulator effectiveness evaluations and transfer of training studies have been faced with the problem of determining accurate student performance levels during or after training (Caro, Shelnutt, and Spears, 1981). For example, errors in performance assessment leading to overtraining results in lowered training effectiveness ratios (Holman, 1979). The need for accurate proficiency assessment was recognized by the TAEG while preparing an evaluation of the training effectiveness of a new state-of-the-art operational flight trainer (OFT), Device 2F64C, at the east coast SH-3 Fleet Replacement Squadron (FRS), Helicopter Antisubmarine Squadron ONE (HS-1).

In an earlier study to determine the effectiveness of Device 2F87F (P-3 Operational Flight Trainer) in the FRS, the inadequacies of current FRS grading procedures for simulator effectiveness evaluations were recognized Browning, Ryan, Scott, and Smode, 1977; Browning, Ryan, and Scott, 1978). To overcome these inadequacies, the TAEG instituted a "proficiency grading system." The proficiency grading system provided a simple procedure for performance assessment by flight instructors. Each time a task was performed, performance was graded on a dichotomous scale that provided a grade of "P" if performance met established standards or a grade of "1" if performance was substandard. These grades were recorded in the sequence of student attempts, thus providing a history or protocol of student performance. The grading system provided two important attributes for evaluating student performance: (1) a static or cross sectional grade of performance on a task attempt and (2) a dynamic or longitudinal record of performance over several attempts.

Determination of proficiency was accomplished by arbitrarily defining the point at which proficiency was attained by the following rule:

1. over 50 percent of the trials (for a given task) on any flight had to be "P" and

2. at least 50 percent of the trials were "P" on all subsequent flights (Browning, et al., 1978, p. 23).

This approach was not useful in evaluating proficiency for the assessment of Device 2F64C at HS-1. The number of flight tasks requiring training was considerably greater for HS-1 than those trained in the Browning, et al. (1977) study. This larger number of tasks presented a greater range of difficulty and precluded the training of all tasks during one flight or training session (Browning, McDaniel, and Scott, 1981). Further complicating the problem of proficiency determination by this arbitrary rule was the fact that many tasks were limited to one attempt or trial per flight or session. Therefore, in many instances the student would be declared "Proficient" or "Not proficient" on the basis of one trial if the cited rule was followed.

Other approaches for determining level of proficiency were investigated. One such approach was to arbitrarily assess proficiency as being reached after the student had demonstrated performance to standards on two, three, or four successive trials. Such an approach was used in the Initial Entry Rotary Wing Flight Training Program by the Army (USAAVNC Evaluation Team, 1979). The logic of such an approach was appealing; however, arbitrary selection of the number of proficient trials needed to demonstrate proficiency do not account for variability in student performance, task difficulty, and variability in instructor ratings (Rankin and McDaniel, 1980). Also, both the approach used by Browning, et al. (1977) and the USAAVNC Evaluation Team (1979) required training protocols that include initial and final levels of proficiency to make accurate performance determinations. Neither approach could accommodate situations where only a small number of training trials are given or where there are wide differences in learning rates of students. Further, instructor knowledge of arbitrary decision rules defined in these approaches may also bias performance ratings.

It appears that in actual practice, training decisions are more probabilistic than deterministic judgments. In other words, instructors and training managers infer a probability of a range of acceptable performance by the student in the future rather than making an absolute prediction of a specific level of performance. The CATES decision model provides a method for assessing flight task proficiency based on the probabilistic nature of decision making. Using this method, an analogy of the training program can be envisaged as a biasing process; students enter the training program with a low probability of performing the task to established standards. With successive trials, the probability of performing to established standards increases until it reaches the desired objective at which time training is terminated.

In summary, the CATES system promised to achieve two purposes. First, it appeared to offer a quantifiable method for the accurate quantification of student performance levels needed for simulator effectiveness evaluations. Second, and perhaps more important, the CATES system could provide training managers and instructors with a valuable tool to aid the decision-making process.

MATHEMATICAL DECISION MODEL USED IN CATES

A statistical decision model analogous to determining the probability that a student would perform a task to established standards is a sequential sampling method introduced by Wald (1947) and described in Rankin and McDaniel (1980). Appendix A provides a mathematical discussion of the Wald Binomial Probability Ratio Test used as the statistical decision model. The sequential sampling method differs from conventional sampling methods. Conventional sampling methods usually require a fixed number of items randomly drawn from a larger collection. The sampled items are examined and the decision is made to accept or reject the entire collection or lot based on this assessment. Sequential sampling does not use fixed sample sizes nor are the items drawn at random from the entire lot. Rather, the items are examined in the order they are produced. Thus, the sample size required to make a decision becomes variable and is dependent on four a priori parameters and the variability of the ordered sequence. The four a priori parameters are:

- minimum proportion of nondefectives at or below which the collection or lot is rejected or, conversely, the proportion of defectives above which the lot is rejected (P_1)
- desirable proportion of nondefectives at or above which the collection or lot is accepted (P_2)
- risk of making a TYPE I decisional error or declaring the lot acceptable when in fact it is not (Alpha (α))
- risk of making a TYPE II decisional error or declaring the lot unacceptable when in fact it is acceptable (Beta (β)).

The variability of the ordered sequence may either reduce or increase the sample size required to make a decision. For example, if the sequence contains items that are either consistently acceptable or consistently unacceptable, a decision may be reached with fewer items. If the sequence contains inconsistencies; i.e., both acceptable and unacceptable items, the sample size required to reach the appropriate decision will increase.

Originally, the sampling procedure was used to determine whether a collection of a manufactured product should be rejected because the proportion of defectives is too high or should be accepted because the proportion of defectives is below an acceptable level. In this industrial quality control setting, the inspector needs a chart similar to figure 1 to perform a sequential test to determine acceptable levels. As each item is observed, the inspector plots a point on the chart one unit to the right if it is not defective, one unit to the right and one unit up if the item is defective. If the plotted line crosses the upper parallel line or boundary, the inspector will reject the production lot. If the plotted line crosses the lower boundary, the lot will be accepted. If the plotted line remains between the two boundaries, another sample item will be drawn and observed/tested. Because sampling is expensive, a fixed limit on the number of items to be sampled may be set. If the limit is reached and the plotted line has not crossed either the upper or lower boundary, the inspector must then make a decision. Generally, the decision will be made to accept or reject based on the proximity of the last plotted point to the closest boundary (truncation). This decision model has been used in the educational and trainirg settings by Ferguson (1970) and Kalisch (1980). Previous use of the model in training was to evaluate performance after the learning period and to serve as an evaluation tool for computer-based instruction that conserved testing time by using a minimum sample of items.

The CATES system decision model uses sequential sampling during the learning period and eventually terminates it. Figure 2 illustrates the CATES decision model as proposed by Rankin and McDaniel (1980) for assessing flight task proficiency. This figure shows a trainee trial sequence of 11PPPPPP. Analyzing this sequence using the decision model on the second trial of the sequence, the plotted line would cross the lower boundary denoting the student is "Not Proficient." Thus, the student is remediated and the plot starts

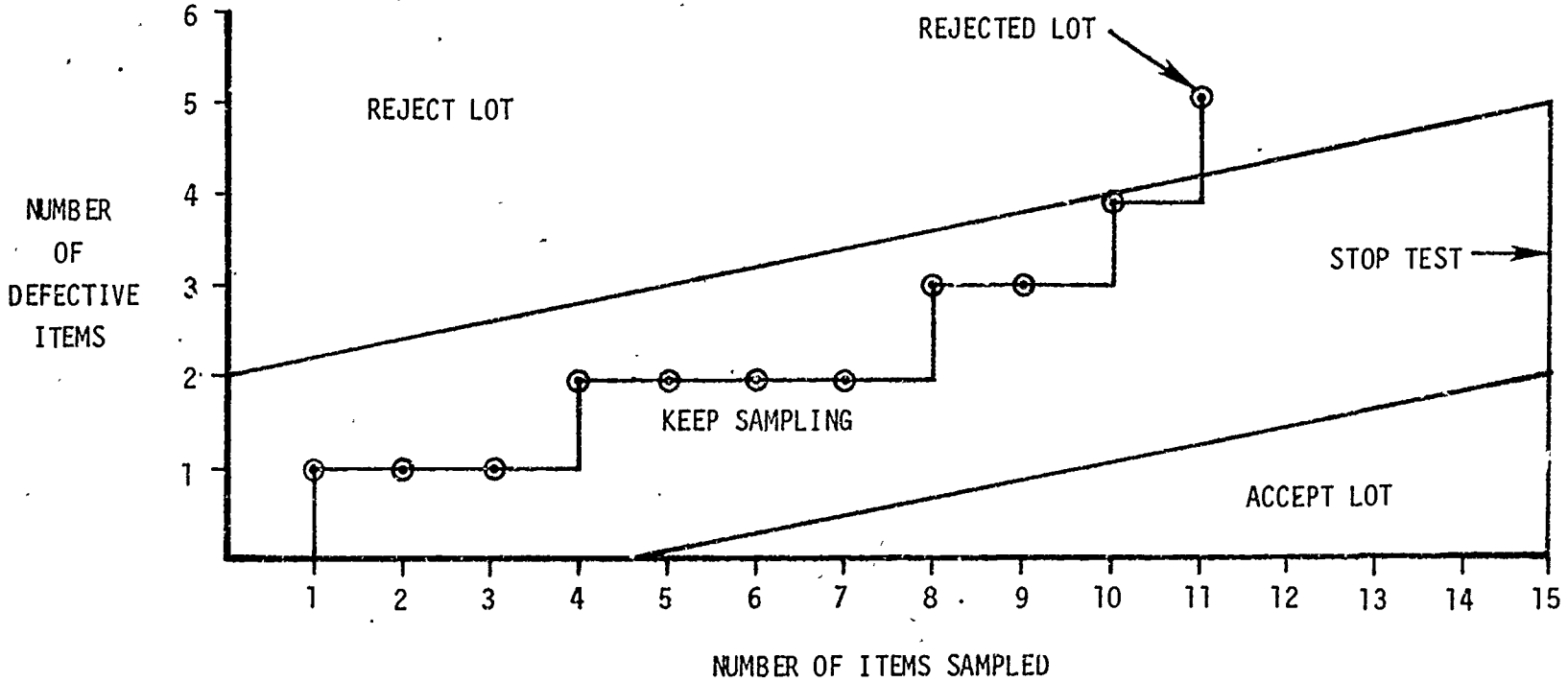

Figure 1. Hypothetical Sequential Sampling Chart

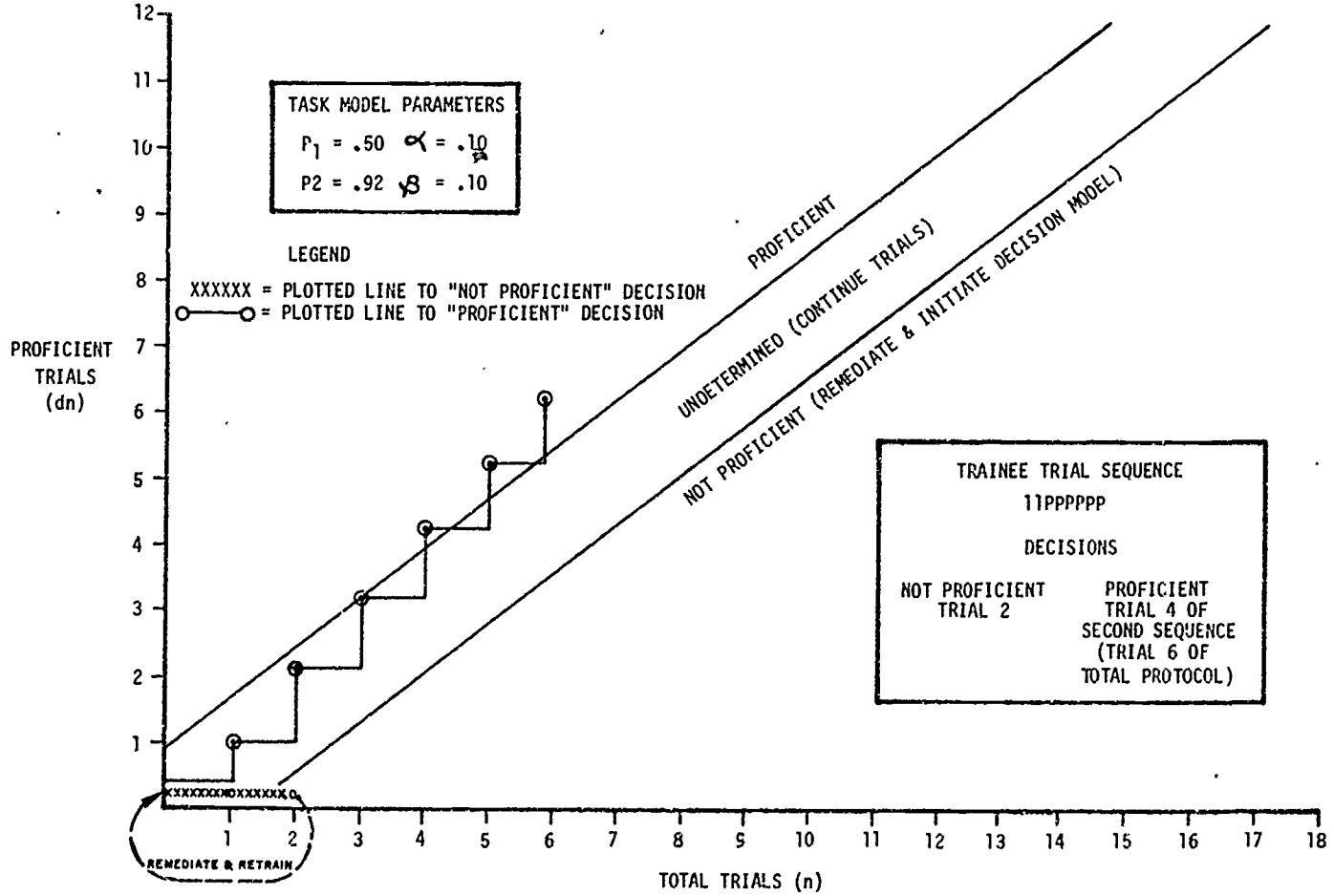

Figure 2. Sequential Sampling Decision Model for Running Takeoff Task

with the next trial. In this particular sequence, that trial is the first P trial in the sequence. On the sixth trial in the overall sequence (fourth trial in the new sequence), the plotted line crosses the upper boundary denoting the student is "Proficient" and training may cease. In this example, the student received two additional training trials after the CATES decision of "Proficient, Stop Training."

SIMILARITY OF CATES DECISION MODEL AND CURRENT DECISION METHOD

The mathematical algorithm used in the CATES system closely parallels the current decision method used by the training manager or instructor to determine when to terminate training. Like CATES, the human judgment method bases decisions on varying numbers of practice trials on the task rather than requiring a fixed number of practice trials. Consistency of student performance on training tasks is also considered in determining the appropriate amount of trials. Students that perform consistently well on a task are considered proficient with less task performance information than those students that perform inconsistently. Instructors and training managers also appear to consider the risks involved in making an inappropriate decision.

The advantage of the CATES decision model appears to be the quantification of acceptable (proficient) performance, unacceptable (not proficient) performance, and the risks (alpha and beta) involved in making an inappropriate decision. The problem then is to assess the advantages offered by the mathematical algorithm in increasing the effectiveness of training decisions. The quantifying of performance and risk gained through the use of the mathematical algorithm is an obvious advantage in training effectiveness evaluations. Other practical advantages involve a better means to aggregate inconclusive information concerning student performance and a decision accuracy greater than the current method.

ADVANTAGES OF MATHEMATICAL DECISION MODELS

Considerable investigation has been conducted on human decision behavior and the cognitive processes humans employ to make choices and solve decision-related problems. Comprehensive reviews of the experimental literature are available: Imhoff and Levine (1981), Lee (1971), Nickerson and Feehrer (1975), Rapoport and Wallsten (1972), Slovic, Fischhoff, and Lichtenstein (1977), and Slovic and Lichtenstein (1971). Some relevant areas of study include: statistical decision theory (Fishburne, 1964), game theory (Luce and Raiffa, 1957), and probabilistic information systems (Edwards, 1962).

It is generally found that decisions reached by mathematical models are considerably more consistent and accurate than decisions based on human judgment (Dawes and Corrigan, 1974; Meehl, 1954; Sawyer, 1966). It appears that human judgment decisions require more data than mathematical models as a result of poorly defined parameters and biases in the processing of information for decisions (Slovic, 1976; Tversky and Kahneman, 1974). Dawes (1979) proposes that mathematical models are especially good at aggregating information resulting in the more efficient use of available information. Dawes further suggests that humans have expertise in perceiving and sorting information that cannot be matched by a mathematical model.

Given that human judgment excels in perceiving and sorting information and that mathematical models are especially good at combining or aggregating information, it appears that a combination of these models should considerably enhance the decision-making process. It follows that a combination of people assessing trial performance and a mathematical model determining the integration and quantity of these assessments should substantially increase the validity and reliability of training decisions. The potential value of a CATES decision model has been recognized for aviation management. Mixon (1981) recommended the decision model be used to assess proficiency of naval flight officers undergoing training at the A-6 aircraft Fleet Replacement Squadrons.

Although previous research has indicated mathematical models may provide a potentially valuable decision making tool, results have generally been limited to laboratory studies and experiments. Evidence is needed to support the practical use of a mathematical decision model in a considerably more unstructured environment. To satisfy this need, this evaluation was conducted to extend the knowledge of the mathematical decision model to a direct application in training.

APPLICATION OF THE CATES SYSTEM DECISION MODEL

To examine the practicality of the CATES system decision model in a realistic training situation, an evaluation was conducted "in-situ" at HS-1, Naval Air Station, Jacksonville, Florida. Concurrent with this study, the TAEG was evaluating Device 2F64C (Browning, McDaniel, and Scott, 1981; Browning, McDaniel, Scott, and Smode, 1982). The test plan for this evaluation required instructors to use the proficiency grading system to record task trial performance of students undergoing flight training. As discussed previously in this section, the recording of task trial data forms an integral, necessary component of the CATES system decision model. In addition to the current method of making training decisions, data were recorded in a manner usable by the CATES system. Although the proficiency grading system posed an additional requirement for the instructors, it does not appear to overburden them in accomplishing their duties. Further, most instructors seem to have accepted the proficiency grading system as a more useful method than current grading practices.

Rankin and McDaniel (1980) envisaged that full implementation of the CATES system would require computer support. Although computer support is available to HS-1 through the Aviation Training Support System (ATSS), the TAEG and HS-1 agreed that before using the ATSS for computer support, the efficacy of the CATES system should be evaluated to determine if advantages could be realized. If advantages using CATES were realized, full implementation could be initiated.

Full implementation would require data input to the ATSS. Although this may appear to be an additional requirement, the CATES system may provide a more efficient method of management control than the present system of maintaining "hard copy" records.

In summary, the CATES system appears to place little additional burden on the training manager than current methods used and may actually relieve

certain requirements. This is contingent upon how well the CATES system "works" in the actual training environment. The method used to determine how well the CATES system "works" is presented in the next section.

SECTION III

METHOD

STUDENTS

The student sample consisted of 29 newly designated naval aviators undergoing Fleet Replacement Pilot Training in the SH-3 aircraft at HS-1. The students were recent graduates of Undergraduate Pilot Training at Pensacola, Florida, and had no prior flight experience in the SH-3 aircraft.

TASKS

The student was required to master approximately 190 flight tasks during Fleet Replacement Pilot Training to become qualified to fly the SH-3 aircraft. From the task inventory of 190 tasks, 18 tasks (appendix B) were selected to evaluate the CATES decision model proposed by Rankin and McDaniel (1980). These 18 tasks were representative of the range of difficulty for tasks in the inventory as well as tasks introduced in early and later stages of training.

Task difficulty was determined by a task sort into categories of "easy," "medium," and "difficult" and rank ordering of the 18 tasks by subject matter experts (HS-1 instructor pilots). From this pool of 18 tasks, 9 tasks were selected with 3 from each category to assess the efficient use of information needed to reach a decision. These nine tasks and categories are presented in appendix C.

INSTRUCTORS

Flight task training was provided by the 28 regular HS-1 flight instructors. All instructors had completed at least one tour in an operational assignment and the training course for flight instructors at HS-1. All instructors were briefed on the grading procedures currently in use as well as the proficiency grading system.

MATERIALS AND EQUIPMENT

Standard training materials and equipment were used by students and instructors at HS-1. No additional equipment or materials were required to obtain data and/or information necessary for this study. The primary data collection instruments were the standard syllabus grade card (appendix D) and the Naval Air Training and Operating Procedures Standardization Program (NATOPS) Flight Evaluation Worksheet (appendix E).

To facilitate retrieval of task trial information and calculate CATES system decisions, data from the grade cards were entered on a WANG 2200 MVP computer at the TAEG.

PROCEDURE

As students proceeded through the training syllabus, performance was graded on the Syllabus Grade Card using both current procedures; i.e., NATOPS,

and the proficiency grading system procedure. The NATOPS procedure grades task performance in three categories or classifications: "Q" or Qualified (performance meets or surpasses NATOPS standards, "CQ" or Conditionally Qualified (performance not to established standards, but does not exhibit safety violations), "U" or Unqualified (performance not to standards and safety violations are exhibited). The NATOPS grade for each task is a summary of all task trials; i.e., there is only one NATOPS grade for each task for each flight or session. In addition to the NATOPS grading procedure, the grades for each practice trial on each task were recorded in the sequence the trial was attempted (Proficiency Grading System). Syllabus grade cards were collected after each training flight or session. Data from each grade card were then entered into the WANG 2200 MVP.

Upon completion of the training syllabus and at the discretion of the instructor pilot/training manager, each student was scheduled for a final NATOPS flight evaluation. The instructor pilots/training manager were not apprised of any decisions made by the CATES system decision model.

The NATOPS flight evaluation for each student was made by one of eight designated instructor pilots. Flight evaluation grades were recorded on the NATOPS Flight Evaluation Worksheet. It should be noted that the worksheet does not specify discrete tasks in the same manner as the syllabus grade cards. However, if the student performance is below standards set by NATOPS, the evaluator is required to specify the task and explain why the task was not performed to standards. Thus, performance of specific tasks on the flight evaluation could be obtained. Upon completion, the NATOPS evaluation flight worksheets were collected. These worksheets were reviewed and a determination was made concerning the evaluation grade for each task and each student; i.e., Qualified, Conditionally Qualified, or Unqualified.

DEPENDENT MEASURES FOR THE CURRENT DECISION METHOD

Two dependent measures were extracted from the data collected: (1) task performance information required to reach a decision and (2) the level of student proficiency upon completion of the training program.

Task performance information required to reach a training decision was determined as the total number of practice trials the student attempted in the flight training program. Each practice trial was envisaged as a "bit" of information the instructors acquired concerning student performance.

The level of student proficiency was determined by the NATOPS grade awarded for each task on the last evaluation of training. Grades awarded on this basis would be more likely to use the same standards as required by the NATOPS flight evaluation. A grade of Qualified would indicate the instructor was confident the student was proficient and would perform the task to standards on the NATOPS evaluation. A grade of Conditionally Qualified would indicate the instructor was less confident the student would perform to standards and could benefit from additional training.

CATES SYSTEM PARAMETERS

The CATES system decision model requires four values be established: (1) the lowest acceptable proportion of proficient trials at or below which the student is considered "Not Proficient" (P_1), (2) the proportion of proficient trials at or above which represents proficient performance (P_2), (3) the probability of a TYPE I decision error (Alpha or (α)), (4) the probability of a TYPE II decision error (Beta or (β)).

The P_1 parameter values were determined from an examination of first trial performance data from a group of 17 students undergoing training at HS-1. The proportion of acceptable trial performances to the total of first performances was used to set the P_1 value for each task.

The P_2 parameter values were determined from the performance of 50 naval aviators on the NATOPS flight evaluation. The proportion of Qualified grades to the total grades awarded was calculated. The P_2 values were then established at one-half standard deviation units below the mean proportion.

In the present study, parameter values for (α) and (β) were arbitrarily selected as .10. The parameters for the representative sample of 18 tasks used in this study are shown in appendix B.

DEPENDENT MEASURES FOR THE CATES SYSTEM DECISION MODEL

As in the current decision method, two dependent measures were extracted from the data collected. These were: (1) task performance information required to reach a decision and (2) the level of student proficiency.

Task performance information required was determined to be the total number of practice trials attempted before a CATES system decision was reached. It is important to note that because training and task practice terminated at the discretion of the instructor or training manager, there is a possibility the CATES system decision model would not have sufficient task trial information to reach a decision. If the task protocol had not resulted in crossing the upper boundary of the decision model (indicating student proficiency was "Undetermined"), an estimate was made of the number of additional trials required to make a decision. This estimate is based on a mathematical equation developed by Hoel (1971) and is shown in appendix F. Parameter values used in this equation were the same values set for each task. The estimated trials to a decision were then added to the total number of trials actually attempted. Using this procedure, it was possible for the CATES system decision model to require either less, equal, or more trial information to reach a decision than the current decision method for each task or student.

A proficient level of performance by the student was determined if the CATES system decision model reached a "Proficient, Stop Training" decision based on actual trials. Thus, a "Proficient, Stop Training" decision was considered equivalent to the current decision method of awarding a "Qualified" grade for the task on the last training flight/session.

CRITERION FOR EVALUATION OF DECISIONS

The criterion used to evaluate the accuracy of the training decisions made by the current decision method and the CATES system decision model was the student's graded task performance on the NATOPS flight evaluation. If a decision concerning proficient level of performance was reached and subsequent task performance on the NATOPS flight evaluation was graded as Qualified, the decisions were considered correct. If the grade on the NATOPS flight evaluation was either Conditionally Qualified or Unqualified, the decision was considered incorrect.

The information requirements and accuracy for the current decision method and the CATES system decision model were compared. Results of that comparison are described in the next section.

SECTION IV

RESULTS

MODEL INFORMATION REQUIREMENTS

The first analysis dealt with the amount of information required by the two decision methods to reach a decision as a function of task difficulty. Results of the analysis of variance (ANOVA) are shown in table 1.

TABLE 1. SOURCE TABLE FOR ANOVA OF INFORMATION REQUIREMENTS OF TWO DECISION METHODS AND THREE TASK DIFFICULTY LEVELS

Source	Sum of Squares	df	MS	F
A (Decision Method)	1250.702	1	1250.702	109.81*
Error	318.922	28	11.390	
B (Task Difficulty)	73.215	1 (Adj).	36.607	3.22
Error	637.075	43 (Adj)	11.376	
AB (Method x Task Difficulty)	452.332	2	226.166	43.22*
Error	293.058	56	5.233	

*$P < .05$

Because the ANOVA was a repeated measures design, it was suspected that certain assumptions or requirements of the ANOVA may have been violated; i.e., lack of homogeneity, additivity. A conservative F-test with reduced degrees of freedom was conducted using the procedures recommended by Myer (1979). This conservative F-test still revealed significant differences for the A main effect (Decision Method) and the AB interaction effect (Method x Task Difficulty). However, for the B main effect (Task Difficulty), the test of significance failed to reach the critical level of .05. An epsilon factor (.7693) was determined from the variance-covariance matrix as recommended by Greenhouse-Geisser. With this adjustment to the degrees of freedom, the B main effect (Task Difficulty) did not reach the .05 level of significance.

To determine significant differences within the interaction effect, the Tukey's Wholly Significant Difference (WSD) was computed. Any differences in the means greater than 2.224 may be considered significant at the .05 level. Figure 3 graphically shows the relationship between decision method and task difficulty as a function of average trials required to reach a "stop training" decision. The figure shows that the CATES decision model required less information to make a "stop training" decision across all levels of task difficulty. The information requirements for the CATES decision model become greater as task difficulty increases. Reliable differences were found between information requirements for easy ($\bar{x}$ = 4.8

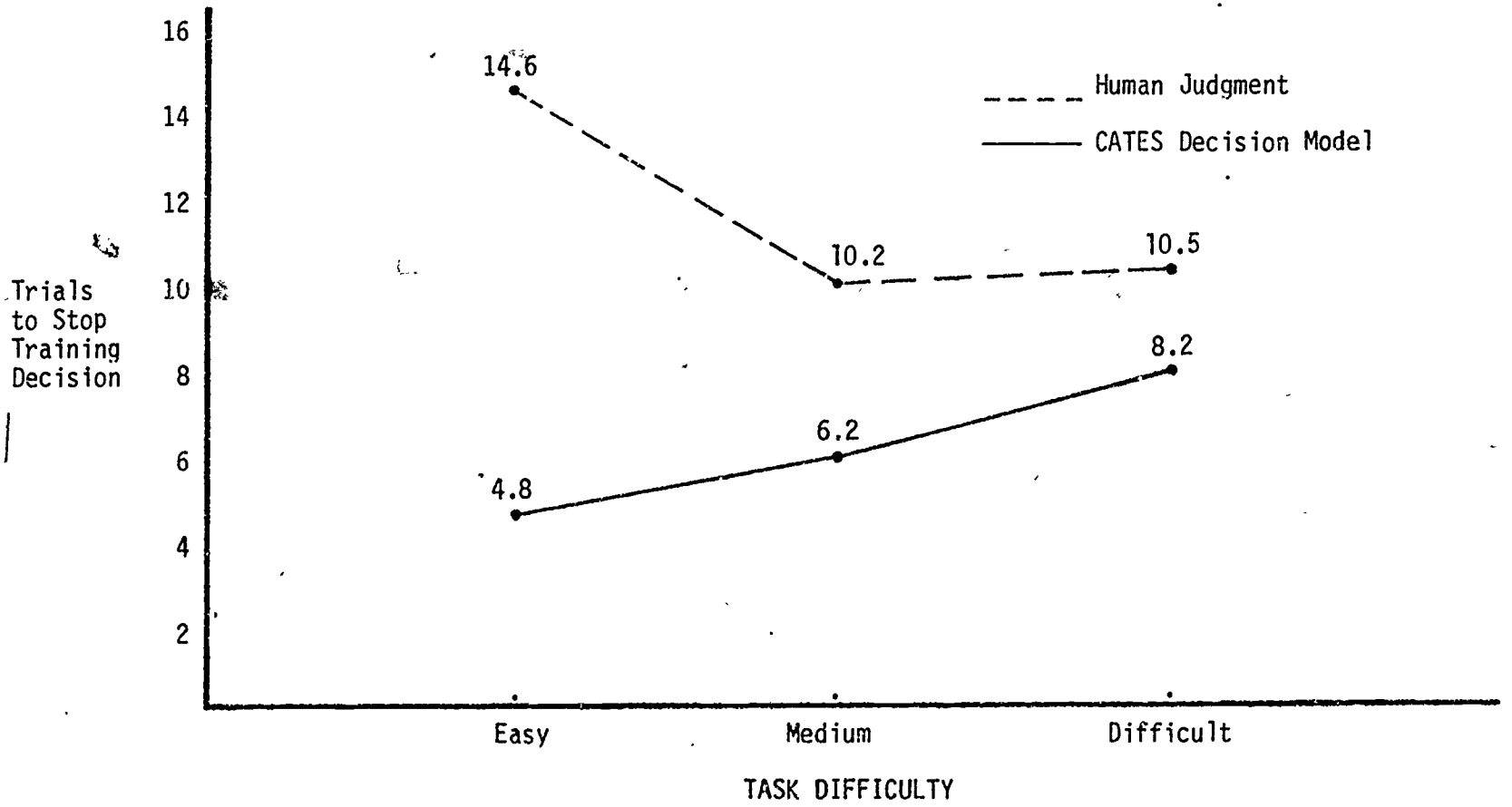

Figure 3. Trials Required to Reach a "Stop Training" Decision for Two Decision Methods Across Three Levels of Task Difficulty

trials) and difficult tasks ($\bar{x}$ = 8.2 trials) assessed by the CATES model. For the human judgment procedure, it appears the reverse is true. More information was collected on the easy tasks ($\bar{x}$ = 14.6 trials) than on the medium ($\bar{x}$ = 10.2 trials) or difficult tasks ($\bar{x}$ = 10.5 trials). Differences in information requirements for medium and difficult tasks reached by human judgment were not reliable. The data indicate the CATES model requires less information to reach a decision than human judgment and the information requirements for CATES appears to trend in a logical manner; i.e., more difficult tasks require more trial information.

ACCURACY OF DECISION METHODS

To determine the degree to which the two decision methods were able to "predict" the student's performance on the final NATOPS flight evaluation and compare the judgments made by each method across the three levels of task difficulty, the following analysis was done. The judgment made for each method was the proportion of "Qualified" or "Proficient, Stop Training" decisions to the overall possible decisions that could be made. There were 87 possible decisions (3 tasks X 29 students) for each level of task difficulty. From these 87 possible decisions, the last instructor grade awarded was determined. If the final grade was a "Q" or Qualified, it was counted as a Qualified judgment made. If it was a "CQ" or Conditionally Qualified judgment, it was not considered to be a Qualified judgment. For the CATES decision model, only those student task protocols that crossed the upper boundary resulting in a "Proficient, Stop Training" judgment were considered as a "Qualified" judgment. Each of these judgments from both methods were then matched against the task-student evaluation made on the final NATOPS flight evaluation. A Qualified judgment made was considered correct if a Qualified grade for that task was awarded on the NATOPS flight evaluation.

Table 2 shows the results of this examination of the proportion of Qualified judgments made and the proportion of correct judgments for the nine tasks. A test for proportions revealed no significant differences on the proportion of qualified judgments made between decision methods. There were no significant differences found in the proportions of correct judgments made between methods.

TABLE 2. PROPORTION OF QUALIFIED JUDGMENTS AND PROPORTION OF CORRECT DECISIONS MADE BY EACH DECISION METHOD ACROSS THREE LEVELS OF TASK DIFFICULTY

METHOD	TASK DIFFICULTY					
	Easy (N=87)		Medium (N=87)		Difficult (N=87)	
	Qualified Judgments	Correct Judgments	Qualified Judgments	Correct Judgments	Qualified Judgments	Correct Judgments
CATES	.9885	.9884	.6092	.8302	.6092	.8112
Instructor	.9885	.9884	.7816	.7794	.7126	.7097

Although no significant or reliable differences were found, it was noted that the proportion of Qualified judgments made decreased as task difficulty increased. This supports the intuitive judgment that the more difficult or complex tasks are somewhat more difficult to evaluate with confidence. The CATES decision model appeared to be more conservative or less willing to make a judgment as task difficulty increased. However, once a decision had been made, the CATES decision method tended to be more correct than the instructor method.

Considering this trend toward increased accuracy or correctness of judgments made, the entire sample of 18 tasks was assessed for Qualified judgments made and the accuracy of the judgments. Results indicated that for 12 of the 18 tasks, CATES was more correct in the judgments made. Proportions of correct decisions were equal for the instructor and CATES method on 2 of the 18. Instructor judgments appeared to be more correct on 4 of the 18 tasks. A sign test revealed that CATES was reliably more correct in judgments than the instructors beyond the .05 level of significance. This finding would support a conclusion that if CATES decisions were used to determine proficiency across the training syllabus, a more accurate assessment would be made concerning student proficiency than the present method of instructor judgments.

SECTION V

DISCUSSION, CONCLUSIONS, AND RECOMMENDATIONS

Results of this evaluation indicate the CATES system decision model, using the parameter values established in the present study, requires less information to make a decision than the current system of human judgment. Decisions reached by the CATES system reflected a higher proportion of correct decisions in reference to the NATOPS flight evaluation. Across a representative sample of 18 tasks, the CATES system was reliably more accurate than the current method of human judgment. The finding that the CATES system requires less trial information to reach a decision of greater precision strongly supports the CATES system decision model's superior efficiency when compared to the current method of making training judgments. Results of this study extend previous research results suggesting greater consistency and accuracy of mathematical models to an actual training situation in a considerably more unstructured environment.

The proportion of judgments concerning student proficiency for easy tasks was high and equal for both methods. As task difficulty increased, however, the CATES system model made a lower proportion of decisions than the current method of instructor judgments. The conservatism or riskiness of the CATES system model is established through parameter values, specifically alpha (α) and beta (β). Since these parameter values were held constant across all tasks and levels of task difficulty, it is reasonable to conclude the instructors were willing to take more risks in decisions made on medium and difficult tasks. This willingness to take greater risks may result in the lowered proportions of correct decisions made by the instructors. Results of this study are similar to a study of human decision making behavior in a sequential testing situation reported by Becker (1958). According to Becker, subjects appeared to operate more like Wald's sequential sampling model when the problem was difficult than when the problem was easy. Typically, subjects required relatively more samples or information on easy problems and relatively less information on the difficult, as if they set alpha (α) and beta (β) lower for the easy problems.

The reasons why instructors obtained considerably more information than the CATES system required for easy tasks remains unclear. This may have resulted from:

- easy tasks being introduced earlier in the training program allowing more time for practice
- easy tasks being prerequisite to the performance of the more difficult tasks; e.g., normal starting of the engines were required to accomplish more difficult flight tasks
- instructors allowing students to perform easy tasks so that successful performance would motivate the student to perform better on the more difficult tasks
- instructors being reinforced by the student's demonstrated high levels of performance on the easy tasks thus increasing the

probability the instructor will request the student to perform the task again

- instructors obtaining information about student performance of easy tasks is done at a lower "cost." Easier tasks are probably less complex to evaluate and do not require as high a degree of actual physical risk to the student and instructor than more difficult complex tasks present.

Whether there is single or multiple causes, it would appear that easy tasks are "overlearned" as a result of significantly more practice allowed. This overlearning probably results in considerable performance consistency by the student resulting in high agreement between the current decision methods, the CATES system decision model, and the NATOPS flight evaluation. Considering the greater consistency of performance and the high agreement between decision and evaluation method, it appears that overlearning is highly desirable.

The more salient issue from the training managers point of view concerns the cost of "overtraining." The results indicated the amount of training provided for easy tasks in excess of that required to make a CATES system decision; however, it was not within the scope of this study to determine the economic or training costs incurred by training beyond acceptable proficiency levels. If such an evaluation were conducted in the future, it would be necessary to consider several possible causes of "overtraining" rather than simply the amount and cost of providing training beyond required levels.

Of considerable interest to the training manager is the issue of "undertraining" the medium and difficult tasks. Neither the CATES system nor the current human judgment method were able to render qualified or proficient judgments in 20 to 40 percent of the proficiency decisions. A paradox seems to exist in the data. While the CATES system decision model appeared to be more conservative in making a judgment than the current decision method, the amount of trial information needed to reach a decision was reliably less for the CATES decision model. It would appear logical that a relatively conservative method would require more data or task performance information. Training trial sequences were individually examined to determine reasons for this apparent paradox. The observation was made that students demonstrating consistent proficient performance continued to perform training trials well after the CATES decision (overlearning). Conversely, students with more variable task protocols were not afforded the opportunity to practice the task with a sufficient number of trials needed to reach a CATES decision. It would appear that the paradox of the more conservative model requiring less information to make a decision could be attributed to under and overtraining in the medium and difficult tasks.

An important methodological restriction was placed on this evaluation. Students proceeded through the training program at the discretion of the instructor/training manager under the current decisional method. In the event the CATES system reached a decision, training may have continued.

Although in a strict sense, the CATES system would consider the additional task training and trial information unnecessary to reach a decision, in a practical sense this additional training may have been an important factor in the final NATOPS flight evaluation. Certainly no implication should be made that training trials beyond a CATES system decision of proficiency is unnecessary overtraining. The next logical step in eliminating this methodological flaw would be to further evaluate the CATES system with methodology similar to that used in the present study, with the exception that the procedure should provide for additional training beyond the current decision method if additional information is required to reach a CATES system decision. Thus, the proportion of proficient judgments made by the CATES system would increase. If results were found similar to this study, strong evidence would be available for employing the CATES system in an important role for training decisions.

It should be noted that the criterion measure for both the current decision method and the CATES system decision model was performance on the NATOPS flight evaluation. Although the NATOPS flight evaluations are conducted using specially selected, experienced naval aviators, no measures of validity or reliability have been determined for that procedure. Essentially, performance on the NATOPS flight evaluation is determined in the same manner as used by the current decision method on training flights. The fact that NATOPS evaluators are specially selected, experienced, and trained may very well result in a greater reliability for the NATOPS evaluation of flight performance. Nevertheless, it is still subject to the problems of human variability; i.e., biases, varying standards, personal interaction. Determination of the validity and reliability of criterion measures is a difficult and elusive task. However, if naval aviation continues to use the NATOPS flight evaluation as a yardstick to measure flight performance, it is desirable that this task be undertaken.

CONCLUSIONS AND RECOMMENDATIONS

Based on the results of this study, the following conclusions and recommendations are made.

CONCLUSION	RECOMMENDATION
The CATES system is particularly useful to manage the training syllabus at the lowest element of the syllabus (flight task). Changes to the syllabus resulting from addition/deletion/modification of the flight tasks can be made quickly and efficiently.	HS-1 should consider extending the current ATSS of managing the syllabus at the event level to tasks trained within each event.
The determination of student performance at the task level rather than event/session/flight level provides a more well defined picture of student performance. It allows the instructor/training manager to determine student strengths and weaknesses in a more timely manner.	HS-1 should focus on student performance of individual tasks in the training syllabus. Capabilities of the ATSS to record student performance on events/sessions/flights should be extended to record student task performance within an event/session/flight.
The Proficiency Grading System provides student task performance information with better definition and specificity than the currently used NATOPS grading procedures.	HS-1 should continue to use the proficiency grading procedures for Category I replacement pilots. The proficiency grading procedure should be extended to include all categories of replacement pilots.
The CATES system decision model appears to be more efficient and accurate than the current method of determining student task proficiency and making training judgments.	Based on positive results of future evaluations, HS-1 should consider incorporating the CATES system decision model to augment the current method for making training decisions.
Method used for establishing CATES system model parameters; i.e., P_1, P_2, (α), and (β), appears to be reasonable and in general agreement with the present system of making training decisions.	Continue using this method to establish parameters for all tasks to be trained in the replacement pilot training syllabus.
Data from this study indicate considerable variability in instructor judgments. Levels of	HS-1 should continually train and standardize instructors to reduce variability in grading student

risk appear to vary with task difficulty and instructors when using the NATOPS grading procedures. This variability and instructor bias may affect the reliability of the NATOPS grading procedure to a considerable degree.	performance, thus increasing the reliability of the grading.
The CATES system decision model is useful to preclude the "under-training" of tasks.	Student task performance should be evaluated using the CATES decision model. If proficiency level cannot be determined by the CATES system within parameters used by the system, training should be continued until a decision is reached.
The CATES system decision model may be useful to determine excessive task training (overtraining).	Tasks that are trained beyond levels required by the CATES system decision model should be carefully monitored to ensure the additional training is desirable for improving student performance across the overall flight syllabus.
The CATES system could be adapted to other FRS flight training programs.	If subsequent evaluations reveal the CATES system continues to result in greater efficiency and higher accuracy in reaching training decisions, other FRSs may consider incorporating the CATES system into their training programs.
The CATES system may provide a more efficient and accurate method of determining student performance in Undergraduate Pilot Training.	If subsequent evaluations reveal the CATES system continues to result in greater efficiency and higher accuracy in reaching training decisions, the Chief of Naval Air Training should consider evaluating the CATES system for possible inclusion in Undergraduate Pilot Training.
The validity or reliability of the NATOPS flight evaluation has not been determined. The evaluation is subject to the same vagaries and variability noted in evaluating student performance in the training program.	Naval Air Systems Command should consider initiating a program to determine the validity and reliability of the NATOPS flight evaluation program.

POST NOTE

This study provides evidence that a mathematical decision model, specifically the CATES system decision model, can powerfully augment present training decision methods for replacement pilots undergoing training at the FRS. It is worthy to note that in addition to achieving more accurate and precise training decisions, the CATES system also provides a useful tool for the management of a curriculum. The CATES system provides documentation as well as student performance measures at the lowest level or element in the curriculum; i.e., the flight task. This documentation and recordkeeping, combined with the apparent effective tool for making training decisions, makes the CATES system especially amenable as a computer-based or computer-managed instructional system.

As a result of this conceptual logic and the findings in this study, HS-1 is aggressively pursuing the incorporation of the CATES system into the ATSS to aid in increasing the efficiency of training management. Upon completion of this effort, it is envisaged that the procedures used in incorporating the CATES system into the ATSS accompanied by a user's manual will be published in a future TAEG report.

In addition, further evaluation of the CATES system decision model is being planned at HS-1 to provide additional training required to reach a CATES system decision. Such an evaluation will extend the findings in this study by providing actual rather than estimated information required to reach a decision. This planned evaluation will also determine if the additional training will impact on the NATOPS flight evaluation in terms of accuracy and precision of decisions similar to the findings in this study. Results of this study will also be published as a TAEG report.

REFERENCES

Becker, G. "Sequential decision making: Wald's model and estimates of parameters." Journal of Experimental Psychology, 1958, 5, pp. 628-636.

Browning, R. F., McDaniel, W. C., and Scott, P. G. Preparation and Design for a Training Effectiveness Evaluation of Device 2F64C for Replacement Pilot Training. TAEG Report 108. August 1981. Training Analysis and Evaluation Group, Orlando, FL 32813.

Browning, R. F., McDaniel, W. C., Scott, P. G., and Smode, A. F. An Assessment of the Training Effectiveness of Device 2F64C for Training Helicopter Replacement Pilots. Technical Report 127. July 1982. Training Analysis and Evaluation Group, Orlando, FL 32813.

Browning, R. F., Ryan, L. E., and Scott, P. G. Utilization of Device 2F87F OFT to Achieve Flight Hour Reductions in P-3 Fleet Replacement Pilot Training. TAEG Report 54. April 1978. Training Analysis and Evaluation Group, Orlando, FL 32813 (AD A053650).

Browning, R. F., Ryan, L. E., Scott, P. G., and Smode, A. F. Training Effectiveness Evaluation of Device 2F87F, P-3C Operational Flight Trainer. TAEG Report 42. January 1977. Training Analysis and Evaluation Group, Orlando, FL 32813 (AD A035771).

Caro, P. W., Shelnutt, J. B., and Spears, W. D. Aircrew Training Device Utilization. AFHRL-TR-80-35. January 1981. Air Force Human Resources Laboratory, Brooks AFB, TX 78235.

Dawes, R. M. "The robust beauty of improper linear models in decision making." American Psychologist, 1979, 34, pp. 571-582.

Dawes, R. M. and Corrigan, B. "Linear models in decision making." Psychological Bulletin, 1974, 81, pp. 95-106.

Edwards, W. "Dynamic decision theory and probabilistic information processing." Human Factors, 1962, 4, pp. 59-73.

Ferguson, R. "A model for computer-assisted criterion-referenced measurement." Education, 1970, 91, pp. 25-31.

Fishburne, P. C. Decision and value theory. New York: John Wiley & Sons, Inc., 1964.

Hoel, P. G. Introduction to mathematical statistics. New York: John Wiley & Sons, Inc., 1971.

Holman, G. L. Training Effectiveness of the CH-47 Flight Simulator. Research Report 1209. May 1979. Army Research Institute, Fort Rucker, AL 36362.

REFERENCES (continued)

Imhoff, D. L. and Levine, J. M. Perceptual-motor and cognitive performance task battery for pilot selection. AFHRL-TR-80-27. January 1981. Advanced Research Resources Organization, Washington, DC.

Kalisch, S. J. Computerized Instructional Adaptive Testing Model: Formulation and Validation. AFHRL-TR-79-33. February 1980. Air Force Human Resources Laboratory, Brooks AFB, TX 78235.

Lee, W. Decision theory and human behavior. New York: John Wiley & Sons, 1971.

Luce, R. D. and Raiffa, H. Games and decisions. New York: John Wiley & Sons, 1957.

Meehl, P. E. Clinical versus statistical prediction: A theoretical analyses and a review of the evidence. Minneapolis: University of Minnesota Press, 1954.

Mixon, T. R. "A model to measure bombardier/navigator performance during radar navigation in device 2F114, A-6E weapon system." Proceedings of the Human Factors Society. 12-16 October 1981. Rochester, NY.

Myers, J. L. Fundamentals of experimental design. Third Edition. Boston: Allyn and Bacon, Inc., 1979.

Nickerson, R. S. and Feehrer, C. E. Decision making and training: A review of theoretical and empirical studies of decision making and their implications for the training of decision makers. Technical Report NAVTRAEQUIPCEN 73-C-0128. July 1975. Bolt, Beranek, and Newman, Cambridge, MA.

Rankin, W. C. and McDaniel, W. C. Computer Aided Training Evaluation Scheduling (CATES) System: Assessing Flight Task Proficiency. TAEG Report 94. December 1980. Training Analysis and Evaluation Group, Orlando, FL 32813 (AD A095007).

Rapoport, A. and Wallsten, T. S. "Individual decision behavior." Annual Review of Psychology, 1972, pp. 131-176.

Sawyer, J. "Measurement and prediction, clinical and statistical." Psychological Bulletin, 1966, 66, pp. 178-200.

Slovic, P. "Towards an understanding and improving decision." Science, Technology, and the Modern Navy, Thirtieth Anniversary 1946-1976. 1976. Office of Naval Research, Arlington, VA.

Slovic, P., Fischhoff, B., and Lichtenstein, S. "Behavioral decision theory." Annual Review of Psychology, 1977, 28, pp. 1-39.

REFERENCES (continued)

Slovic, P. and Lichtenstein, S. "Comparison of Bayesian and regression approaches to the study of information processing in judgment." Organizational Behavior and Human Performance, 1971, 6, pp. 649-744.

Tversky, A. and Kahneman, D. "Judgment under uncertainty: Heuristics and biases." Science, 1974, 185, pp. 1124-1131.

United States Army Aviation Center Evaluation Team. Evaluation of the 175/40 Initial Entry Rotary Wing Flight Training Program. TR-79-02. May 1979. U.S. Army Aviation Center, Fort Rucker, AL 36362.

Wald, A. Sequential analysis. New York: John Wiley & Sons, Inc., 1947. (Reprinted by Dover Publications, 1973).

APPENDIX A

WALD BINOMIAL PROBABILITY RATIO TEST

WALD BINOMIAL PROBABILITY RATIO TEST

The Wald binomial probability ratio test was developed by Wald (1947) as a means of making statistical decisions using as limited a sample as possible. The procedure involves the consideration of two hypotheses:

$$H_0: \; P \leq P_1$$

and $H_1: \; P \geq P_2$ where

P is the proportion of nondefectives in the collection under consideration, P_1 is the minimum proportion of nondefectives at or below which the collection is rejected, and P2 is the desired proportion of nondefectives, at or above which the collection is accepted. Since a simple hypothesis is being tested against a simple alternative, the basis for deciding between H_0 and H_1 may be tested using the likelihood ratio:

$$\frac{P_{2n}}{P_{1n}} = \frac{(P_2)^{dn} \; (1 - P_2)^{n-dn}}{(P_1)^{dn} \; (1 - P_1)^{n-dn}}$$

Where: P_1 = Minimum proportion of nondefectives at or below which the collection is rejected.

P_2 = Desirable proportion of nondefectives at or above which the collection is accepted.

n = Total items in collection.

dn = Total nondefectives in collection.

The sequential testing procedure provides for a postponement region based on prescribed values of alpha (α) and beta (β) that approximate the two types of errors found in the statistical decision process. To test the hypothesis $H_0: P = P_1$, calculate the likelihood ratio and proceed as follows:

1. if $\frac{P_{2n}}{P_{1n}} \leq \frac{\beta}{1-\alpha}$, accept H_0

2. if $\frac{P_{2n}}{P_{1n}} \geq \frac{1-\beta}{\alpha}$, accept H_1

3. if $\frac{\beta}{1-\alpha} < \frac{P_{2n}}{P_{1n}} < \frac{1-\beta}{\alpha}$, take an additional observation.

These three decisions relate well to the task proficiency problem. We may use the following rules:

1. Accept the hypothesis that the grade of P is accumulated in lower proportions than acceptable performance would indicate.

2. Reject the hypothesis that the grade of P is accumulated in lower proportions than acceptable performance would indicate. By rejecting this hypothesis, an alternative hypothesis is accepted that the grade of P is accumulated in proportions equal to or greater than desired performance.

3. Continue training by taking an additional trial(s); a decision cannot be made with specified confidence.

The following equations are used to calculate the decision regions of the sequential sampling decision model.

$$dn \leq \frac{\log \frac{\beta}{1-\alpha}}{\log \frac{P_2}{P_1} + \log \frac{1-P_1}{1-P_2}} + n \frac{\log \frac{1-P_1}{1-P_2}}{\log \frac{P_2}{P_1} + \log \frac{1-P_1}{1-P_2}}$$

$$dn \geq \frac{\log \frac{1-\beta}{\alpha}}{\log \frac{P_2}{P_1} + \log \frac{1-P_1}{1-P_2}} + n \frac{\log \frac{1-P_1}{1-P_2}}{\log \frac{P_2}{P_1} + \log \frac{1-P_1}{1-P_2}}$$

Where: dn = Accumulation of trials graded as "P" in the sequence

n = Total trials presented in the sequence

P_1 = Lowest acceptable proportion of proficient trials (P) required to pass the NATOPS flight evaluation with a grade of "Qualified."

P_2 = Proportion of proficient trials (P) that represent desirable performance on the NATOPS flight evaluation.

Alpha(α) = The probability of making a type I error (deciding a student is proficient when in fact he is not proficient).

Beta(β) = The probability of making a type II error (deciding a student is not proficient when in fact he is proficient).

The first term of the two equations will determine the intercepts of the two linear equations. The width between these intercepts is determined largely by values selected for alpha (α) and beta (β). The width between the intercepts translates into a region of uncertainty; thus as lower values of alpha (α) and beta (β) are selected this region of uncertainty increases.

The second term of the equations determines the slopes of the linear equation. Since the second term is the same for both equations, the result will be slopes with parallel lines. Values of P_1 and P_2 as well as differences between P_1 and P_2 affect the slope of the lines. This is easily translated into task difficulty. As P_2 values increase, indicating easier tasks, the slope becomes more steep. This in turn results in fewer trials required in the sample to reach a decision.

As differences in P_1 and P_2 increase, the slope also becomes steeper and the uncertainty region decreases. This is consonant with rational decision making. When the difference between the lower level of proficiency and upper level of proficiency is great, it is easier to determine at which proficiency level the pilot trainee is performing. The concept of differences in P_1 and P_2 is analogous to the concept of effect size in statistically testing the difference between the means of two groups. In such statistical testing, when alpha (α) and beta (β) remain constant, the number of observations required to detect a significant difference may be reduced as the anticipated effect size increases (Kalisch, 1980).

APPENDIX B

TASKS AND PARAMETER VALUES USED IN EVALUATION

TASKS AND TASK PARAMETERS USED IN EVALUATION

Task Description	Parameters			
	Alpha(α)	Beta(β)	P_1	P_2
1. Normal Landing	.10	.10	.18	.78
2. Normal Approach	.10	.10	.18	.78
3. Free Stream Recovery	.10	.10	.12	.55
4. Single Engine Approach	.10	.10	.18	.75
5. Single Engine Landing	.10	.10	.53	.75
6. Single Engine Malfunction Analysis	.10	.10	.06	.51
7. ASE Off Landing	.10	.10	.30	.86
8. Alternate Approach Pilot Procedures	.10	.10	.18	.69
9. Windline SAR Pilot Procedures	.10	.10	.38	.80
10. Normal Start	.10	.10	.12	.65
11. Rotor Engagement	.10	.10	.47	.75
12. Single Engine Malfunction Takeoff Abort	.10	.10	.41	.75
13. Automatic Approach Pilot Procedures	.10	.10	.24	.90
14. Servo Malfunction	.10	.10	.25	.62
15. Manual Throttle	.10	.10	.35	.51
16. ASE Malfunction	.10	.10	.35	.62
17. SAR Manual Approach	.10	.10	.25	.80
18. Shutdown Checklist	.10	.10	.29	.91

APPENDIX C

TASKS AND LEVEL OF DIFFICULTY USED TO EVALUATE EFFICIENCY

TASKS AND LEVEL OF DIFFICULTY USED TO EVALUATE EFFICIENCY

Level of Difficulty	Tasks
Easy	Normal Start Shutdown Checklist Normal Landing
Medium	SAR Manual Approach Alternate Approach Pilot Procedures Single Engine Malfunction Takeoff Abort
Difficult	Windline Search and Rescue Pilot Procedure ASE Off Landing Freestream Recovery

APPENDIX D

SAMPLE GRADE CARD FOR DATA RECORDING

HS 1-(TAEG) TRAINING FORM REV. 2 (11 DEC 80) BSF-3

FRP _____ COMP _____

INST _____ INCOMP _____

DATE _____ PILOT TIME _____ COPILOT TIME _____

COPILOT NAME _____

TASK CODE		COMD QUALIFIED	UNQUALIFIED	DISCUSS, INTRO, NOT OBS	NUMBER TRIALS/ PROFICIENCY
DA100	TAC NAV CHECK				
DA200	COUPLER DOPPLER CHECK				
B6500	NITE LIGHTING PROCEDURE				
BE300	INSTRUMENT TAKEOFF				
BB100	INSTRUMENT DEPARTURE				
DA300	PRE-DIP CHECKLIST				
DB100	AUTO APPROACH PILOT PROCEDURES				
DC100	ALTERNATE APPROACH PILOT PROCEDURES (INTRO)				
DC200	ALTERNATE APPROACH COPILOT PROCEDURES				
DB300	HOVER DEPARTURE PROCEDURES				
DA500	SONAR DEPLOYMENT VOICE PROCEDURES				
DF100	USE OF CABLE ALTITUDE (INTRO)				
DE100	FREESTREAM RECOVERY				
EB100	IFR SAR SCENARIO DEMO				
BE402	TACAN APPROACH				
BE409	MISSED APPROACH				
BE403	GCA APPROACH				
CE300	MANUAL THROTTLE				
BA500	CHECKLISTS				
CE500	SINGLE ENGINE MALFUNCTION ANALYSIS				
	MALFUNCTIONS/EMERGENCIES (GRADE IF GIVEN)				
FA756	ELECTRICAL FIRE				
DE912	BEEPER TRIM FAILURE				
FD845/846	FUEL CONTROL CONTAMINATION				
FB878	ASE MALFUNCTION (.879 TO .890)				
DE938	RADAR ALTIMETER FAILURE				
FD835/836	COMPRESSOR STALL				
FD803/804	LUBE PUMP SHAFT FAILURE				
FD843/844	P-3 SIGNAL LOSS				
FA751	GENERATOR FAIL (.751/752)				
DE200	SONAR RAISE MALFUNCTIONS				
DE400	BOTTOMED DOME				
DE500	HUNG DOME				

HS 1 (TAEG) TRAINING FORM REV. 2 (11 DEC 80) BSF-3
SIDE 2

TASK CODE		UNQUALIFIED	COND QUALIFIED	QUALIFIED	NUMBER TRIALS/ PROFICIENCY
COCKPIT PROCEDURE					
PREPARATION					
HEADWORK					
DISCUSS	AUTO AND ALTERNATE APPROACHES				
	HOVER DEPARTURE PROCEDURES. MANUAL CLIMBOUT				
	SWIMMER DEPLOYMENT				
	PROCEDURES (40 FOOT HOVER, 15 FOOT HOVER AND 10 FOOT				
	10 KNOT APPROACH)				
SYSTEMS KNOWLEDGE:					
COUPLER, LIGHTING					

TASK CODE	TASK COMMENTS

INSTRUCTOR SIGNATURE

TRAINING OFFICER REVIEW

SIGNATURE

APPENDIX E

NATOPS WORKSHEET

H-3 PILOT NATOPS EVALUATION WORKSHEET
(Rev. 9-79)

PILOT _____

FLIGHT DATE _____ GRADE _____

FLIGHT DURATION _____ BUNO _____

SIDE NUMBER _____

OPEN BOOK EXAM DATE _____ GRADE _____

CLOSED BOOK EXAM DATE _____ GRADE _____

ORAL EXAM DATE _____ GRADE _____

OVERALL FINAL GRADE _____

EVALUATOR _____

NOTES:
1. A grade of unqualified in any critical area/sub area will result in an overall grade of unqualified for the flight.
2. A grade of conditionally qualified in a critical area will result in an overall grade of conditionally qualified for the flight.
3. Only the numbers 0, 2, or 4 will be assigned to sub areas. No interpolation is allowed.

Unqualified 0.0

Conditionally Qualified 2.0

Qualified 4.0

50

GRADE

PILOTS ORAL EMERGENCY WORKSHEET

1. Electrical Malfunctions.
   a. Generators
   b. Electrical fire
2. ASE Malfunctions
   a. Pitch
   b. Roll
   c. Collective
   d. Yaw
3. Transmission Malfunctions
   *a. Chip detected
   *b. Pressure loss
   c. Tail takeoff
   d. Torque system
4. Engine Malfunctions
   *a. Engine fire
   *b. Flex shaft
   c. Oil pressure
   d. Oil temperature
   e. Hot start
   f. Post shutdown fire
   g. PMS
5. Rotary Rudder Malfunctions
   *a. Tail Rotor control/drive loss
   *b. TGB/IGB chip light
6. Fuel System Malfunctions
   a. Fuel filter bypass
   b. Fuel boost pump
7. Hydraulic Malfunctions
   a Primary
   b Auxiliary
   c. Utility
   d. Sensing unit
8. Water Operations
   *a Water landing
   b Water takeoff
   c Fuel dumping
9. Rotor Brake Malfunctions
   a. Inflight
   b. Shutdown

PILOTS ORAL EMERGENCY WORKSHEET GRADE

10. Discussion Items
    - *a. Power settling _____
    - *b. Blade stall _____
    - *c. Dynamic rollover _____
    - d. Sonar hoist _____
    - e. MAD reeling machine _____
    - f. AKT 22 antenna _____

GENERAL COMMENTS OVERALL GRADE _____

PILOT EVALUATIONS WORKSHEET

Area I. Ground Operations _____

a. Brief/debrief/flight gear _____
b. Records check *_____
c. Preflight/postflight *_____
d. Checklist procedures/systems check _____
e. Start/engagement _____
f. Taxi/lookout *_____
g. Disengagement/shutdown _____
h. General _____

Area I. Ground Operations

CONDITIONALLY QUALIFIED. Did not fully instruct or debrief the crew. Flight equipment improperly worn or in marginal condition. Did not fully examine flight records. Minor omissions or errors on preflight or postflight. Improper or incomplete use of checklists. Non standard procedures. Inattention or misinterpretation of visual signal. Rough or erratic start, engagement, disengagement or shutdown.

UNQUALIFIED. Did not conduct brief or debrief. Flight equipment missing, not worn or in an unsafe condition. Failed to sign for aircraft or accepted aircraft with grounding discrepancy. Failed to note or record downing discrepancy after flight. Any omission or error on preflight or postflight which would affect safety of flight. Exceeded published limitations during start, engagement, disengagement or shutdown. Did not utilize checklist or perform required systems checks. Marginal control of helicopter while taxiing. Ignored visual signal. Did not use pre-takeoff checklist.

Area II. Normal Flight Operations

a. Checklist procedures
b. Transition/climb
c. Cruise flight
d. Systems knowledge/usage
e. Normal landings/takeoffs
f. Hover/low work
g. General

Area II. Normal Flight Operations

CONDITIONALLY QUALIFIED. Incomplete use of takeoff, post-takeoff, or landing checklist. Application of power erratic but did not exceed limitations. Unable to maintain altitude within ±50 feet of assigned altitude. Maintained airspeed within ±10 knots. Heading control varied ±5 degrees between final approach and landing. Hover altitude 15 feet ±5 feet. Unable to fully explain aircraft systems or limitations.

UNQUALIFIED. Did not use checklist. Did not check instruments prior to leaving hover. Failed to use sufficient power or exceeded aircraft or engine limitations. Safety precautions not observed. Leveled off in excess of 50 feet from assigned altitude. Airspeed tolerance ±10 knots exceeded. Hover in excess 15 ±5 feet, excessive nose attitude or lateral drift on touchdown. Running landings/takeoffs in excess of 40 knots, yaw in excess of 10 degrees or lateral drift on touchdown/takeoff. Unsatisfactory knowledge of aircraft systems or limitations.

Area III Emergency Operations _____

a. Autorotations _____
b. Single engine landings waveoffs _____
c. AUX off landings _____
d. ASE off landings takeoffs _____
e. Emergency procedures _____
f. General _____

Area III Emergency Operations

CONDITIONALLY QUALIFIED Did not pre-brief co-pilot on autorotations. Airspeed, Nr and heading control erratic Groundspeed exceeded 15 knots or slight drift at recovery Did not establish and maintain minimum safe single engine speed on landings or waveoffs Minor difficulty in controlling Nr during single engine Power, heading and altitude control erratic during AUX or ASE off flight Did not fully comply with emergency procedures bud did not jeopardize aircraft or crew

UNQUALIFIED During autorotation did not call for full power. Airspeed, Nr and heading control beyond safe limits Implemented techniques that would have jeopardized the successful completion and recovery of the autorotation Failed to call for full power PMS off during single engines. Failed to note or correct low unsafe Nr conditions during single engine. Exceeded rate of descent limits during single engine approach or engine limits ASE off and AUX off flight unsafe or excessive lateral drift rate of descent on touchdown Failed to comply with established emergency procedures which resulted in jeopardizing aircraft/crew or exceeded engine airframe limitations

Area IV Coupler Sonar Operations (Hooded) _____

a. Checklist voice procedures _____
b. Automatic approach _____
c. Alternate approach _____
d. Climb out _____
e. Coupler sonar emergencies _____
f. Systems knowledge grade _____
g. General _____

Area IV Coupler Sonar Operations (Hooded)

CONDITIONALLY QUALIFIED Minor deviations from established checklist and voice procedures Erratic control of aircraft during automatic, alternate approach and climb out Erratic altitude control of 150 feet ±20 feet Reacted slowly to emergencies Unable to fully explain systems or limitations

UNQUALIFIED Checklist not used or unsate/improper procedures utilized Allowed aircraft to descend through 30 feet in hover without attempting to correct Made omissions or errors in emergency procedures that could jeopardize aircraft or crew Attempted to hover downwind without correcting Unsatisfactory knowledge of systems or procedures Unable to consistently maintain 150 ±30 feet while hooded

Area V Search and Rescue Operations	*
a. Navigation	
b IFR procedures (Hooded)	*
c VFR procedures	*
d. Crew/cockpit coordination	
e General	

Area V Search and Rescue Operations (Hooded)

CONDITIONALLY QUALIFIED. No coordination of visual lookout doctrine. Used nonstandard voice approach pattern or hoist procedures but none which would seriously affect the mission. Did not fully properly utilize copilot crew and systems in accomplishing rescue

UNQUALIFIED. Could not follow wind line rescue pattern. Hovered downwind without correcting. Unable to consistently maintain 110 ± 30 feet hooded. Allowed aircraft to descend below 80 feet during approach hover without correcting. Exceeded aircraft limitations or procedures that would have jeopardized aircraft or crew

APPENDIX F

MATHEMATICAL EQUATION FOR ESTIMATING TRIALS TO REACH STOP TRAINING DECISION FOR THE CATES DECISION MODEL

MATHEMATICAL EQUATION USED TO ESTIMATE TRIALS TO A "PROFICIENT, STOP TRAINING" DECISION

$$\text{Additional Estimated Trials to } P_2 = \frac{\beta \log \frac{\beta}{1-\alpha} + 1-\beta \quad \log \frac{1-\beta}{\alpha}}{1 - P_2 \log \frac{1-P_2}{1-P_1} + P_2 \log \frac{P_2}{P_1}}$$

Overall Estimated Trials = Trials Performed by Student + Additional Estimated Trials to P_2 (estimated trials required to cross the upper boundary)