DOCUMENT RESUME

ED 201 316 IR 009 270

AUTHOR Rankin, William C.; McDaniel, William C.
TITLE Computer Aided Training Evaluation and Scheduling (CATES) System: Assessing Flight Task Proficiency. TAEG Report No. 94.
INSTITUTION Naval Training Equipment Center, Orlando, Fla. Training Analysis and Evaluation Group.
PUB DATE Dec 80
NOTE 30p.

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Competency Based Education; *Computer Managed Instruction; *Diagnostic Teaching; *Flight Training; Flow Charts; Grading; Performance Tests; Standards; *Student Evaluation

ABSTRACT

This report proposes a method for achieving improvements in the precision of determining Fleet Replacement Squadron (FRS) student aviator proficiency. The proposed method, called the Computer Aided Training Evaluation and Scheduling (CATES) system, provides a computer managed, prescriptive training program based on individual student performance. Designed to formalize and quantify the parameters of the decision process used to determine proficiency, the CATES system (1) clearly defines the level of skills required of the FRS graduate; (2) adds precision to instructor pilot judgments by providing a more clearly defined comparison standard; (3) increases reliability of instructor pilot judgments by grading each task execution rather than using the instructor's subjective judgment; (4) lists tasks for individual students depending upon whether proficiency has been attained, has not been attained, or has not been determined; and (5) provides an acceptable and workable performance assessment schema for use in a computer-managed instruction system. Twelve references are listed, and a mathematical discussion of the Wald Binomial Probability Ratio Test is appended. (Author/LLS)

TAEG Report No. 94

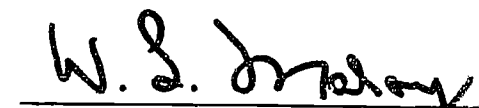
COMPUTER AIDED TRAINING EVALUATION AND SCHEDULING (CATES) SYSTEM: ASSESSING FLIGHT TASK PROFICIENCY

William Rankin
William McDaniel

Training Analysis and Evaluation Group

December 1980

ALFRED F. SMODE, PhD Director
Training Analysis and Evaluation Group



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

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

SECTION I

INTRODUCTION

A major problem in the Fleet Replacement Squadron (RAG) is determining the appropriate amount of in-flight training that should be given a pilot trainee to meet the objectives of the FRS in serving the needs of fleet squadrons. The difficulty centers on determining performance requirements and assessing skill levels appropriate for the FRS graduate. This inability to achieve precise determination of pilot performance at all times commensurate with squadron goals has hampered the management of training, particularly the scheduling of replacement pilot training in terms of both effectiveness and efficiency).

These obstacles are of most concern in the training of first-tour pilots (recent graduates of Undergraduate Pilot Training [UPT] entering their first operational aircraft type). These students must acquire many skills and learn to organize much information as minimal in the very short time period available in order to graduate to a fleet assignment. Extending training beyond assured proficiency is expensive in resource use; training to less than required proficiency incurs significant costs.

Improving precision in judging pilot proficiency and enhancing management ability in prescribing training sensitive to individual differences in student performance and instructor evaluation continues to be a prime requirement in military flight training.

This report proposes a method for achieving improvements in the precision of proficiency judgments and in determining student proficiency. This proposed solution, identified as the Computer Aided Training Evaluation and Scheduling (CATES) system, provides a computer managed prescriptive training program based on individual student performance. In essence, the CATES system emphasizes the following:

- clearly defines the level of skills required of the FRS graduate
- adds precision to instructor pilot judgments by providing a more clearly defined comparison standard
- increases reliability of instructor pilot judgments by grading each task execution rather than using the instructor pilot's "subjective average" of all task executions
- lists tasks for individual students indicating one of the following decisions:
  - .. desired proficiency attained
  - .. proficiency below acceptable limits, or
  - .. proficiency undetermined, continue training/practice.
- provides an acceptable and workable performance assessment schema for use in a Computer Managed Instruction (CMI) system.

The report describes the problems encountered in attempts to determine sufficient task performance of students and the conceptual development of the CATES system as a method that may be used in making proficiency determinations. An effort is in progress to test the operational feasibility of the CATES system and to evaluate the validity of proficiency determinations made by the system. Results of this effort will be presented in a future report.

ORGANIZATION OF THE REPORT

In addition to this introduction, three sections and one appendix are presented. Section II presents the method designed to strengthen grading criteria. Although the criteria continue to be based on subjective judgments of instructor pilots, by clarifying basics to be measured and by providing a standard or which to base subjective judgments the criteria should reflect a greater precision.

Section III presents the method for formalizing and quantifying the parameters of the proficiency determination process.

Section IV presents preimplementation considerations of the CATES system and its applicability at a specific FRS. Issues to be tested as well as future implications of the CATES system are discussed.

The appendix provides a mathematical discussion of the Wald Binomial Probability Ratio Test.

SECTION II

IMPROVEMENT OF GRADING PROCEDURES

CURRENT PRACTICE

Determination of the proficient performance of aircraft flying tasks continues to be a subjective judgment made by instructor pilots. Current practice in training squadrons consists of "flights" during which a subset of tasks from the training syllabus are performed a varying number of times by the pilot trainee at the discretion of the instructor pilots. During or shortly after each flight, the instructor pilot "grades" the pilot trainee on the tasks performed using a standard scale but also employing his own personal criteria. While instructors differ in their personal rating bias (hard-easy), they attempt to grade in terms of "average performance at this stage of training." It is usual for the pilot trainee to be exposed to several different instructor pilots. After a specified minimum number of flights, and a recommendation by an instructor pilot, the pilot trainee is scheduled for a final "check flight." His performance on selected tasks is graded by an instructor pilot acting in the independent role of "check pilot." Should the pilot trainee not perform the flight consonant with the standards of performance expected of him by the "check pilot," he is rescheduled for additional "check flights" until he is deemed proficient.

Student exposure to training tasks can be variable due to instructor differences and varying performance standards. In addition, each individual pilot trainee exhibits variability in successive performances on complex procedural and psychomotor tasks. This variability of skilled task performance has been well documented (Fitts and Posner, 1968). Further compounding this problem of inconsistent performance, the pilot trainee is transitioning from a level of performance well below the required level to a required standard of performance. This transition reflects different learning rates by the individual pilot trainees. Learning rates are also highly variable within and between individuals (Sidman, 1960). It is quite obvious that determination of asymptotic performance commensurate with desired performance standards is difficult to ascertain using the current practice.

PROFICIENCY GRADING SYSTEM

In a series of studies conducted by the Training Analysis and Evaluation Group (TAEG) to determine the effectiveness of Device 2F87F (P-3 Operational Flight Trainer) in the FRS, the inadequacies of current grading procedures 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 system provided a clearer picture of the trainee's flight task performance in both simulator and aircraft training. The proficiency grading system still required a subjective judgment by instructor and check pilots. However, the instructors graded task performance against a precise standard: "P was defined as performance estimated to be equivalent to that required to demonstrate competence in that task on the conventional FLY 6 check" (Browning, et al., 1977, p. 20). This standard focuses on the required terminal level of performance; i.e., the objective of training.

Actual grading of performance was accomplished using a dichotomous scale. Task performance that met or exceeded the standard was recorded as "P"; task performance that did not meet the standard was recorded as "1." The proficiency grading introduced by the TAEG had a further requirement. Performance was graded each time the task was performed and this series of graded trials was recorded and kept in the sequence of presentation. The procedure of grading each task trial as it was performed eliminated the requirement for the instructor to make a summary judgment of task proficiency based on pilot trainee performance of successive task trials during a flight.

The advantages of a proficiency grading system for increasing the precision of performance judgments have been incorporated in the CATES system. The performance standard used in the CATES system is defined as task performance estimated to be equivalent to that required to earn an adjective rating of "Qualified" and/or a numerical score of 4 on the Naval Air Training and Operating Procedures Standardization (NATOPS) Program flight evaluation. The CATES system uses the same proficiency grading procedure as discussed previously. Although the grading procedure increases the precision, it does not reduce several sources of variability in trainee performance; e.g., task difficulty and learning rates.

The proficiency grading procedure results in a task performance or training protocol for each task. Two hypothetical trainee records (protocols from the same trainee) are shown in table 1.

TABLE 1. HYPOTHETICAL TASK PERFORMANCE OF ONE TRAINEE FOR TWO DIFFERENT TASKS

Task	Training Protocol
Task A	1P1P1PPP1PP
Task B	1PPPPPPPPPPP

It could be inferred that "Task A" is more difficult than "Task B" or it could be inferred that the trainee is more proficient on "Task B" than "Task A."

Table 2 contains examples of trainee task performance protocols for two different kinds of tasks and hypothetical task protocols for a trained pilot. The pilot trainees exhibit different protocols initially (more "1's" than "P's") but the variability eventually will diminish. Learning rates differ among tasks as shown by comparing Task A with Task B. During later flights/sessions the protocols for the pilot trainee are not readily distinguishable from those of a trained pilot. A procedural problem remains in determining when task performance protocols for trainees matched the protocols of trained pilots.

TABLE 2. COMPARISON OF HYPOTHETICAL TASK PERFORMANCE PROTOCOLS FOR TWO DIFFERENT TASKS AND TWO LEVELS OF AVIATOR PROFICIENCY

Task/Aviator	Training Procotol During Flights/Sessions					
	One	Two	Three	Four	Five	Six
TASK A						
Pilot Trainee	111	P11	1P1	1PP	PPP	PP
Trained Pilot	PPP1	PPP	1PP	P	P1	PPP
TASK B						
Pilot Trainee	11	1P	P	PPP	PP	PP
Trained Pilot	PP	PPP	P1	PP	P	PP

The essence of the problem lies in assessing, with a specified degree of confidence, the point at which proficiency has been obtained.

Several ways to deal with the problem were explored. Two approaches were found in previous research concerned with proficiency assessment. The first approach was to arbitrarily define 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).

The second approach was used in the evaluation of the Initial Entry Rotary Wing Flight Training Program by the Army (USAAVNC Evaluation Team, 1979). The tasks were graded by daily performance rather than by individual trials; however, the approach used to determine proficiency could also be incorporated with graded trials.

> The point of principal concern was the training day on which the student achieved proficiency on each maneuver. Achievement of maneuver proficiency was defined as that training day on which the third successive (+) grade on the maneuver was given the student. That is, the student was required to perform a maneuver in accord with established USAAVNC standards on three successive occasions before he was judged to be proficient on that maneuver (USAAVNC Evaluation Team, 1979, p. 21).

While both of the above approaches are logical, objective, and expedient, they are faulty. Both require training protocols that include initial and final levels of proficiency to make accurate performance determinations. In other words, they are "after the fact" rather than predictive. Another flaw is that an arbitrary number of "P" trials is not realistic across all tasks due to differences in task difficulty. In addition, these approaches may not accommodate situations where only a small number of training trials are given or where there are wide differences in learning rates of trainees. Finally, the instructor's judgment may be biased if he has knowledge of an arbitrary decision rule.

SEQUENTIAL METHOD. Both of the above approaches require a sample of trials of trainee performance before the rule can be applied. An alternate approach would be to examine trials taken one at a time and accumulate the information for input into the decision model (Hoel, 1971). Using this approach, one would expect to be in a better position to make decisions than if no attempt were made to look at the data until a sample of fixed size had been taken.

There are methods available, using sequential sampling techniques and a statistical decision model, that operate on this accumulation of information basis and that require considerably less sampling on the average than the fixed-size sample methods. The statistical decision model is limited to two choices in decision making (three choices if one considers deferring a decision as a decision). This limitation is not troublesome when applied to proficiency determination. The decisions of primary concern are simply: Is the trainee proficient? or, alternatively, Is the trainee not proficient? Additional advantages are: (1) Decisions are reached based on a minimum number of trials and (2) Decisions are made with an established level of confidence.

SECTION III

CATES DECISION MODEL

One sequential method that may be used as a means for making statistical decisions with a minimum sample was introduced by Wald (1947). Probability ratio tests and corresponding sequential procedures were developed for several statistical distributions. One of the tests, the binomial probability ratio test, was formulated in the context of a sampling procedure 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. The sequential testing procedure also provides for a postponement of decisions concerning acceptance or rejection. This deferred decision is based on prescribed values of alpha (α) and beta (β). Alpha (α) limits errors of declaring something "True" when it is "False" (Type I error). Beta (β) limits errors of declaring something "False" when it is "True" (Type II error).

In an industrial quality control setting, the inspector needs a chart similar to figure 1 to perform a sequential test to determine if a manufacturing process has turned out a lot with too many defective items or whether the proportion of defects is acceptable. 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, the inspector will reject the production lot. If the plotted line crosses the lower parallel line, the lot will be accepted. If the plotted line remains between the two parallel lines of the sequential decision chart, another sample item will be drawn and observed/tested.

This sequential sampling procedure decision model has been previously used in educational and training settings. Ferguson (1969) used the sequential test to determine whether individual students should be advanced or given remedial assistance after they completed learning modules of instruction. Similarly, Kalisch (1980) employed the sequential test for an Air Force Weapons Mechanics Training Course (63ABR46320) conducted at Lowry Air Force Base, Colorado. Results from both applications of sequential testing indicate greater efficiency than for tests composed of fixed numbers of items. It appears sequential testing may substantially reduce testing time.

The CATES system decision model uses sequential testing similar to those applications previously cited. The decision model focuses on proportions of proficient trials (analogous to nondefectives or correct responses) whereas, in previous applications, proportions of defectives or incorrect responses were the items of interest. This approach does not alter the logic of the sequential sampling procedure or the decision model. It does enhance the "meaningfulness" of the procedure in decisions concerning proficiency because the ultimate goal is to determine "proficiency" rather than "nonproficiency." It should be noted that in the industrial quality control setting, sampling occurs after the manufacturing process. In the educational and training applications cited above (Ferguson, 1969 and Kalisch, 1980), sequential sampling occurred <u>after</u> the learning period. In the CATES system, the sequential sampling occurs <u>during</u> the learning period and eventually terminates it.

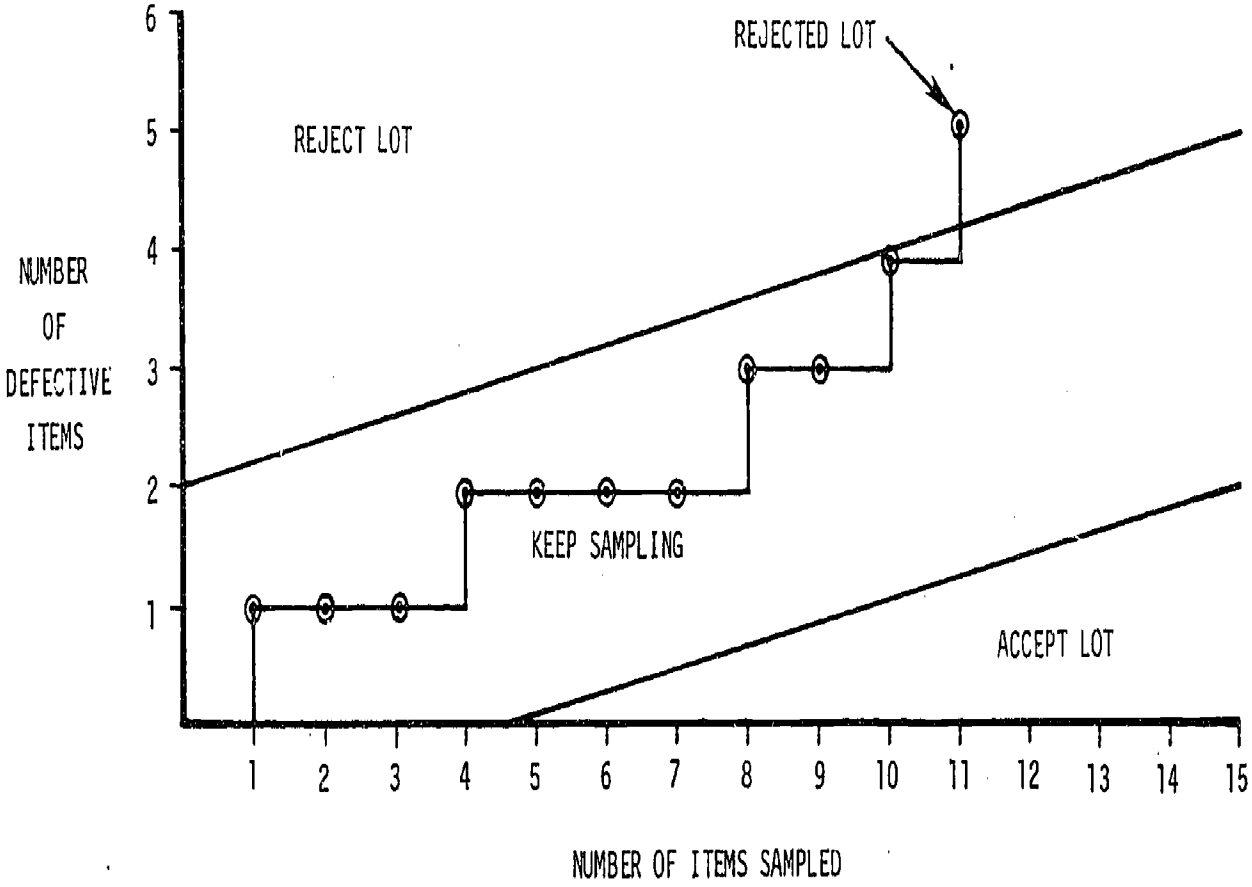


Figure 1. Hypothetical Sequential Sampling Chart

CATES SYSTEM MODEL PARAMETERS

The decision model can be described as consisting of decision boundaries. Referring to figure 1, the parallel lines represent those decision boundaries. Crossing the upper line, or boundary, results in a decision to "Reject Lot"; crossing the lower line, or boundary, results in a decision to "Accept Lot." In the CATES system, these decision boundaries translate to "Proficient" and "Not Proficient." Calculations of the decision boundaries require four parameters. These four parameters are:

P_1 Lowest acceptable proportion of proficient trials (P) required to pass the NATOPS flight evaluation with a grade of "Qualified." Passage of the NATOPS flight evaluation is required to be considered a trained aviator in an operational (fleet) squadron.

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

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

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

Parameter setting is a crucial element in the development of the sequential sampling decision model. Kalisch (1980) outlines three methods for selecting proficient/not proficient performance (q_0/q_1 values) as:

> Method 1--External Criterion. Individuals are classified as masters, non-masters, or unknown on the basis of performance on criteria directly related to the instructional objectives. These criteria can be in terms of demonstrated levels of proficiency either on the job or in a training environment. The mean proportion of items answered correctly by the masters on an objective would provide an estimate for q_0. Similarly, q_1 would be the proportion correct for the non-masters.
>
> Method 2--Rationalization. Experts in the subject area who understand the relation of the training objectives to the end result; e.g., on-the-job performance, select the q_0 and q_1 values to reflect their estimation of the necessary levels of performance. This method is probably the closest to that now used by the Air Force. The procedure may provide somewhat easier decision making since specifying two values creates an indecision zone--neither mastery nor non-mastery. This indecision zone indicates that performance

is at a level which may not be mastery but is not sufficiently poor to be considered at a non-mastery level.

Method 3--Representative Sample. The scores of prior trainees, who demonstrate the entire range from extremely poor to exemplary performance on objectives, are used to estimate q_0 and q_1. The proportion correct for the entire sample is used to obtain an initial cutting score C. Scores are separated into two categories: (a) those scores greater than or equal to C and (b) those less than C. For each category, the mean proportion correct score is computed. The mean for the first category equals q_0; the mean for the second category equals q_1.

Selection of values for P_1 and P_2 ($P_1 = q_1$ and $P_2 = q_0$ in Kalisch, 1980) for the CATES decision model incorporated Method 1 for setting of P_1 and Method 3 for setting of P_2.

The value selected for P_1 was based on the lowest proportion of P grades (numerical grade of 4.0 on the NATOPS flight evaluation) that may be given and still result in an overall rating of "Qualified." The NATOPS evaluation flight consists of a number of flight tasks grouped in areas and subareas. As the tasks or subareas are performed, the pilot's performance is graded using a numerical score. Three numerical scores may be awarded: Qualified performance is assigned a "4," Conditionally Qualified performance is assigned a "2," and Unqualified performance is assigned a "0." The numerical scores are averaged across all tasks and subareas to yield an overall numerical score. To receive an overall rating of "Qualified," the average of all tasks or subareas must fall within the range of 3.00 to 4.00. Thus, the criteria for passing the NATOPS flight evaluation with a "Qualified" rating require that at least 50 percent of the tasks be graded as "Qualified." Therefore, the lower limit of proficient performance was set at .50 for all tasks.

The value selected for P_2 was determined by examining performance scores of a sample of 49 Naval Aviators' NATOPS flight evaluations given at Helicopter Antisubmarine Squadron (HS-1), Naval Air Station (NAS) Jacksonville, Florida. The sample was restricted to only those aviators rated as "Qualified," thus representing exemplary performance. This examination revealed the proportion of "Qualified" scores for each subarea and/or flight task. This proportion is directly translated to P_2 values for each task in the training syllabus.

The selection of alpha (α) and beta (β) should be based on the criticality of accurate proficiency decisions. Small values of alpha (α) and beta (β) require additional task trials to make decisions with greater confidence. Factors that are important in selecting values for alpha (α) and beta (β) are outlined below:

1. Alpha (α) values

   a. Safety--potential harm to the trainee or to others due to the trainee's actual non-mastery of the task.

   b. Prerequisite in Instruction--potential problems in future instruction, especially if the task is prerequisite to other tasks.

   c. Time/Cost--potential loss or destruction of equipment either in training or upon fleet assignment.

   d. Trainee's View of the Training--potential negative view by trainee when classified as proficient although the trainee lacks confidence in that decision. Also, after fleet assignment if previous training has not prepared him sufficiently the trainee may also have a negative view of the training program.

2. Beta (β) values

   a. Instruction--requirement for additional training resources (personnel and materials) for unnecessary training in case of misclassification as not proficient.

   b. Trainee Attitudes--the attitude of trainees when tasks have been mastered yet training continues; trainee frustration; corresponding impact on performance in the remainder of the training program and fleet assignment.

   c. Cost/Time--the additional cost and time required for additional training that is not really needed.

Alpha (α) and beta (β) values used in the CATES decison model were arbitrarily selected as .10. A confidence level of 90 percent in decisions made by the model appears reasonable when the previously discussed factors are considered. As rigorous field testing of the model is conducted, these parameters may be modified as indicated by empirical evidence and command policy. At present, values of .10 appear quite reasonable.

After the model parameters have been selected, calculation of the decision boundaries may be accomplished using the Wald Binomial Probability Ratio Test. The appendix provides a formal mathematical discussion of this test.

To illustrate the differences in task difficulty, two tasks were selected from the HS-1 training syllabus, and the decision models for these tasks were calculated. To further show how the decision models serve to aid in making proficiency decisions, task protocols of a pilot trainee are imposed on the model.[1] Figure 2 shows the model for the task "Running Takeoff," and figure 3 shows the model for the task "Free Stream Recovery."

[1]Actual trial data for a pilot trainee undergoing training at HS-1, NAS Jacksonville, FL.

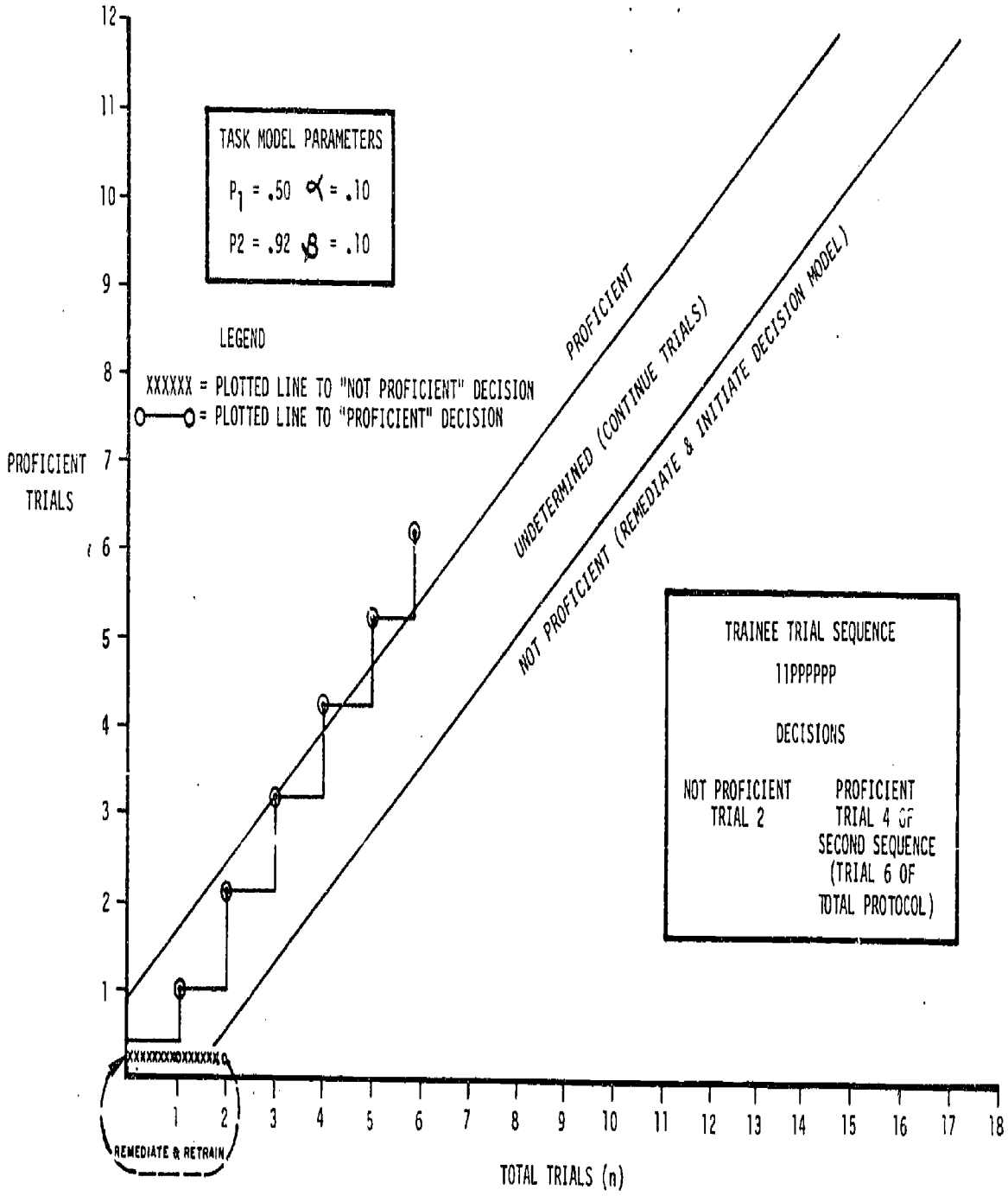


Figure 2. Sequential Sampling Decision Model for Running Takeoff Task

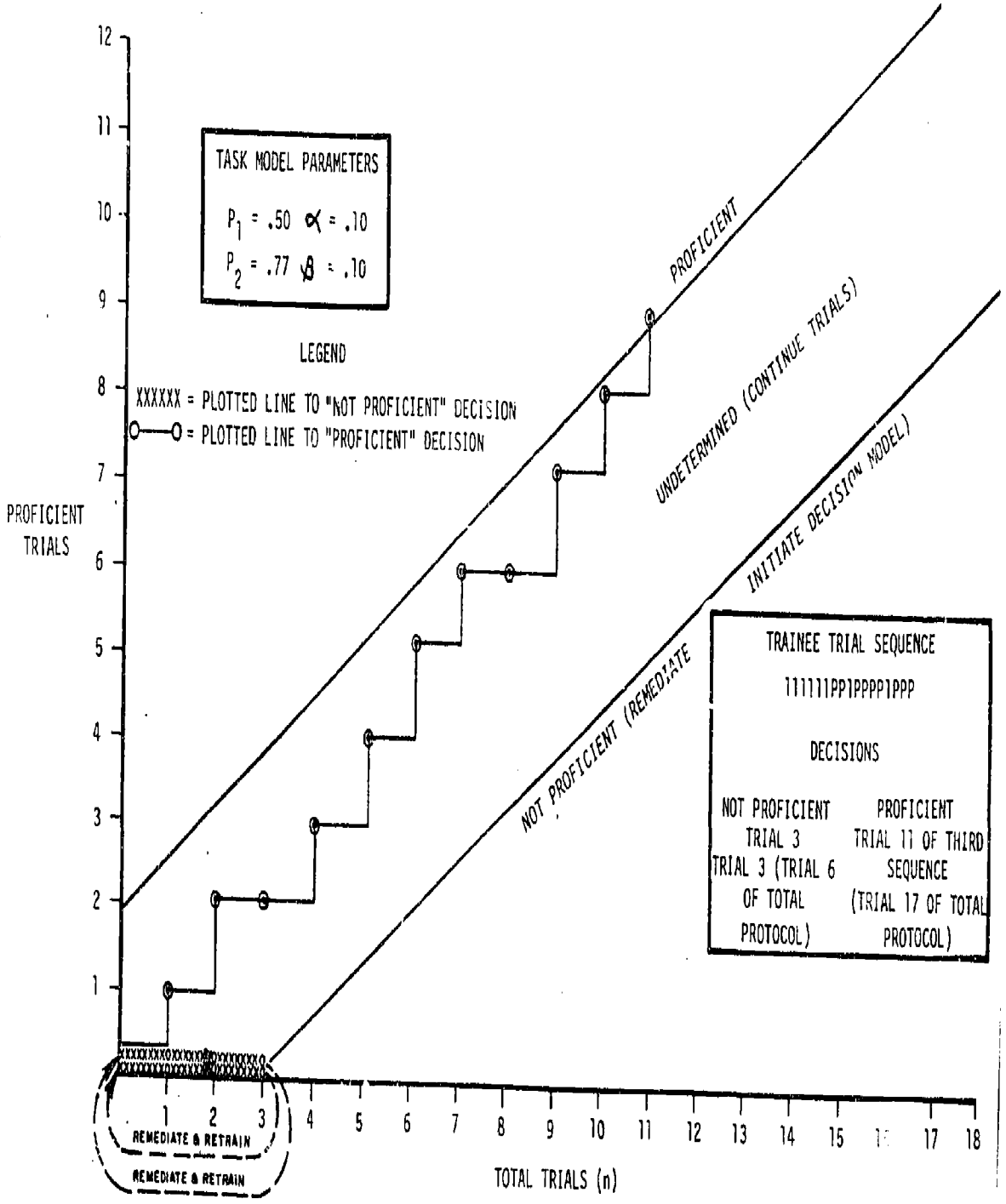


Figure 3. Sequential Sampling Decision Model for Freestream Recovery Task .

Empirical data reflect a relative difference in task difficulty. The sample of NATOPS evaluation scores indicates the proportion of "Qualified" scores on the Running Takeoff task was .92, while the proportion of "Qualified" scores on the Free Stream Recovery task was .77. This relative difference in task difficulty is represented in the model as differences between the slopes and the widths between the parallel lines of the two models. In the case of the Free Stream Recovery task (figure 3), the slopes are less steep (indicating more trials to reach proficiency) and the parallel lines are farther apart (indicating there will typically be more uncertainty about individual trials before a decision can be reached).

In these examples, the probability of making decision errors (both type I and type II) as indicated earlier was set at .10 for both tasks. If this level of confidence was increased (lower values of alpha (α) and beta (β)), the region of uncertainty would also increase. The overall result is that more trials are required to make a decision with increased confidence.

Both models, then, reflect rather well the true state of affairs between different tasks and their impact on a rational decision process. The differences in task difficulty relate directly to differences in the model parameters.

Figures 2 and 3 also show the decisions reached by the model on student performance. The student received a total of eight trials on the Running Takeoff task during the training program. The sequence of graded trials and the graphical plots of the sequence are shown in figure 2. The first two trials were judged to be below the standard of performance. On the second trial the decision model indicated the student was "Not Proficient" and logically should be given remedial or additional training. The sequence is initiated again on trial three, and on the fourth trial of that sequence (sixth trial given) the model decision was "Proficient."

Figure 3 shows the protocol for the Free Stream Recovery task. Perhaps because of slower acquisition of a more difficult task, two decisions were made declaring the student "Not Proficient" in the earlier sessions of task exposure. The model does show that more task trials were required before a decision could be made about proficiency. This can be attributed to increased task difficulty and variability of performance.

SECTION IV

PLANNING FOR IMPLEMENTATION

The role of sequential sampling decision models to determine aviation task proficiency must be operationally explored in terms of feasibility and subsequent validity. A study is currently underway to test the concept at the East Coast SH-3 FRS, HS-1, NAS, Jacksonville, Florida. The study is broadly planned as follows:

1. identify a syllabus of specific training tasks

2. establish proficiency decision model parameters from prior data collected at HS-1

3. train instructors to render performance judgments on task trials; i.e., was performance a "1" or a "P"?

4. collect data on each trainee's task performance by trial

   a. The current decision model (unique to each instructor) will determine when to terminate training the task.

   b. Instructors and training managers will have no knowledge of CATES system decisions regarding task proficiency.

5. compare analytically the models using final performance criterion (NATOPS flight evaluation performance).

6. make recommendations as to feasibility.

Assuming the results of the study are promising, it will be desirable to look toward incorporating or designing a CMI system for which these models are readily amenable. Semple, Cotton, and Sullivan (1980) have summarized the advantages of a CMI system for aircrew training devices applicable to all aspects of aircraft flight training. CMI systems compare a student's training history with a standard training syllabus made up of lists of clearly defined tasks. The "ideal" system assesses student performance on each task and compares this performance with criteria of acceptable performance. This comparison identifies tasks that the student can or still cannot perform. System software then composes an individualized set of instructional tasks that may be trained in subsequent training sessions or flights. Additional factors that may be considered in system design include training asset availability and prediction of training completion dates.

All the virtues of a well conceived CMI system are contingent upon an acceptable, workable performance assessment schema. Figure 4 is a functional flow diagram describing the CATES system to be operationally developed and tested for use by HS-1. It is premature to assert whether CATES will be a "stand alone" system or become an integral subsystem of the Aviation Training Support System (ATSS) (Naval Weapons Center, 1978). In either event, implementing the proficiency determination concept advanced in this report can only be done efficiently with on-line computer support. The work of Ferguson

COMPUTER AIDED TRAINING EVALUATION AND SCHEDULING SYSTEM
(CATES System)

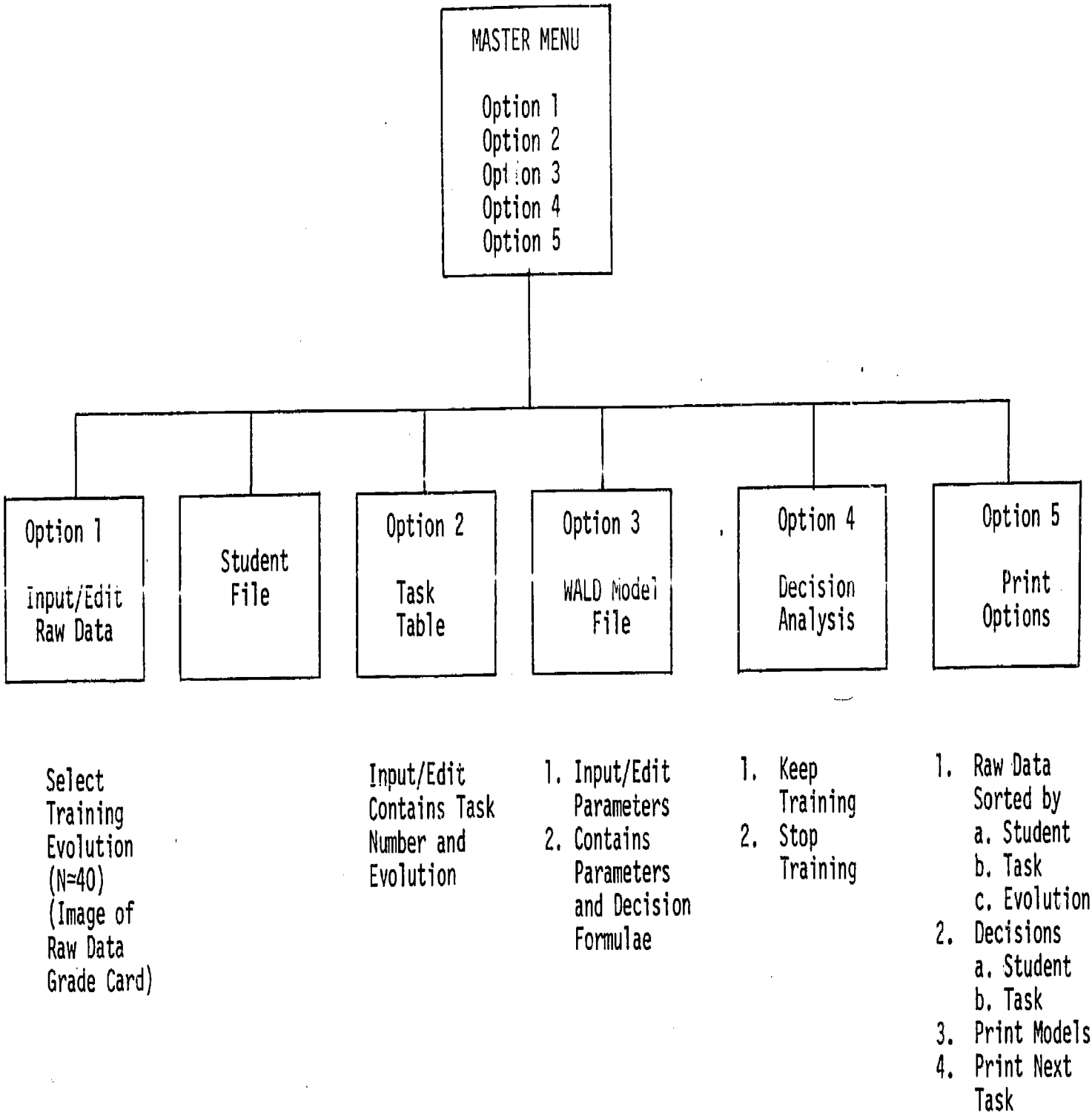


Figure 4. Functional Flow Diagram of CATES System

(1969) and Kalisch (1980) would have been virtually impossible without on-line computer support. Also planned are future efforts to determine the range of applicability to other FRS settings.

POST NOTE

In summary, this report has shown the variability of flight task performance and the difficulty encountered in making accurate proficiency determinations. The CATES system has been introduced as a method to formalize and quantify the parameters of the decision process used in making these determinations, thereby achieving a measure of control. Effort is underway to operationally test the CATES system concerning feasibility, validity, and range of applicability. This report is a prelude to that effort.

REFERENCES

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 No. 42. January 1977. Training Analysis and Evaluation Group, Orlando, FL 32813. (AD A035771)

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 No. 54. April 1978. Training Analysis and Evaluation Group, Orlando, FL 32813. (AD A053650)

Ferguson, R. The Development, Implementation, and Evaluation of a Computer-Assisted Branched Test for a Program of Individually Prescribed Instruction. Unpublished dissertation, University of Pittsburgh. 1969.

Ferguson, R. "A Model for Computer-Assisted Criterion-Referenced Measurement." Education. 1970. 91. pp. 25-31.

Fitts, P. M. and Posner, M. J. Human Performance. Belmont, CA: Brooks and Cole, 1968.

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

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.

Naval Weapons Center. Master Program Plan for the Aviation Training Support System (ATSS). Technical Report TM 3143-3347-77. September 1978. Naval Weapons Center, China Lake, CA.

Semple, C. A., Cotton, J. C., and Sullivan, D. J. Aircrew Training Device Instructional Support Features. AFHRL-TR-80-58. July 1980. Air Force Human Resources Laboratory, Brooks AFB, TX.

Sidman, M. Tactics of Scientific Research. New York: Basic Books. 1960.

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.

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

APPEN X

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).